Uni- and Multimodal and Structured Representations for Modeling Frame Semantics

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

Dissertation

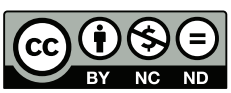
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

vorgelegt von
Teresa Isabel Botschen (geb. Martin), M.Sc.
geboren in Sigmaringen

Tag der Einreichung: 29. November 2018
Tag der Disputation: 24. Januar 2019

Referenten: Prof. Dr. Iryna Gurevych, Darmstadt
Prof. Dr. Stefan Roth, Darmstadt
Prof. Dr. Hinrich Schütze, München

Darmstadt 2018
D17



Abstract

Language is the most complex kind of shared knowledge evolved by humankind and it is the foundation of communication between humans. At the same time, one of the most challenging problems in Artificial Intelligence is to grasp the meaning conveyed by language.

Humans use language to communicate knowledge and information about the world and to exchange their thoughts. In order to understand the meaning of words in a sentence, single words are interpreted in the context of the sentence and of the situation together with a large background of commonsense knowledge and experience in the world. The research field of Natural Language Processing aims at automatically understanding language as humans do naturally.

In this thesis, the overall challenge of understanding meaning in language by capturing world knowledge is examined from the two branches of *(a) knowledge about situations and actions* as expressed in texts and *(b) structured relational knowledge* as stored in knowledge bases. Both branches can be studied with different kinds of vector representations, so-called embeddings, for operationalizing different aspects of knowledge: textual, structured, and visual or multimodal embeddings. This poses the challenge of determining the suitability of different embeddings for automatic language understanding with respect to the two branches.

To approach these challenges, we choose to closely rely upon the lexical-semantic knowledge base FrameNet. It addresses both branches of capturing world knowledge whilst taking into account the linguistic theory of frame semantics which orients on human language understanding. FrameNet provides frames, which are categories for knowledge of meaning, and frame-to-frame relations, which are structured meta-knowledge of interactions between frames. These frames and relations are central to the tasks of Frame Identification and Frame-to-Frame Relation Prediction.

Concerning branch *(a)*, the task of Frame Identification was introduced to advance the understanding of context knowledge about situations, actions and participants. The task is to label predicates with frames in order to identify the meaning of the predicate in the context of the sentence. We use textual embeddings to model the semantics of words in the sentential context and develop a state-of-the-art system for Frame Identification. Our Frame Identification system can be used to automatically annotate frames on English or German texts. Furthermore, in our multimodal approach to Frame Identification, we combine textual embeddings for words with visual embeddings for entities depicted on images. We find that visual information is especially useful in difficult settings with rare frames. To further advance the performance of the multimodal approach, we suggest to develop embeddings for verbs specifically that incorporate multimodal information.

Concerning branch *(b)*, we introduce the task of Frame-to-Frame Relation Prediction to advance the understanding of relational knowledge of interactions between frames. The task is to label connections between frames with relations in order to complete the meta-knowledge stored in FrameNet. We train textual and structured embeddings for frames and explore the limitations of textual frame embeddings with respect to recovering relations between frames. Moreover, we contrast textual frame embeddings versus structured frame embeddings and develop the first system for Frame-to-Frame Relation Prediction. We find that textual and structured frame

embeddings differ with respect to predicting relations; thus when applied as features in the context of further tasks, they can provide different kinds of frame knowledge. Our structured prediction system can be used to generate recommendations for annotations with relations. To further advance the performance of Frame-to-Frame Relation Prediction and also of the induction of new frames and relations, we suggest to develop approaches that incorporate visual information.

The two kinds of frame knowledge from both branches, our Frame Identification system and our pre-trained frame embeddings, are combined in an extrinsic evaluation in the context of higher-level applications. Across these applications, we see a trend that frame knowledge is particularly beneficial in ambiguous and short sentences.

Taken together, in this thesis, we approach semantic language understanding from the two branches of knowledge about situations and actions and structured relational knowledge and investigate different embeddings for textual, structured and multimodal language understanding.

Zusammenfassung

Sprache gilt als ein hochkomplexes Kulturgut der Menschheit und fungiert als Grundlage der Kommunikation zwischen Menschen. Gleichzeitig ist die Erfassung von Bedeutung in Sprache eine der größten Herausforderungen an die Forschung im Bereich der künstlichen Intelligenz.

Mittels Sprache tauschen Menschen Gedanken aus, vermitteln sich gegenseitig Wissen und teilen sich Informationen über die Welt mit. Die Bedeutung einzelner Wörter wird im Zusammenhang eines Satzes verstanden und wird weiterhin im Lichte des Allgemeinwissens und des Erfahrungsschatzes interpretiert. Die Forschung im Bereich der automatischen Sprachverarbeitung verfolgt das Ziel, Sprache automatisch so zu verstehen, wie es der Mensch auf natürliche Weise tut.

In dieser Dissertation nähern wir uns der übergeordenten Herausforderung der Erfassung von Bedeutung in Sprache vor einem Hintergrund an Weltwissen von zwei Seiten: *(a) Kenntnis über typische Situationen und Handlungen* wie sie zum Beispiel in Texten beschrieben werden und *(b) strukturiertes Wissen über Relationen* wie es in Wissensdatenbanken gespeichert wird. Beide Seiten können mit verschiedenartigen Vektordarstellungen (sogenannten verteilten Repräsentationen) untersucht werden, um unterschiedliche Aspekte von Hintergrundwissen abzudecken: textuelle, strukturierte und visuelle oder multimodale verteilte Repräsentationen. Daraus ergibt sich die konkrete Herausforderung, die Eignung der verschiedenen verteilten Repräsentationen in Bezug auf die Erfassung von Bedeutung in Sprache – entsprechend der beiden genannten Seiten – zu bestimmen.

Wir gehen die Herausforderungen der Erfassung von Bedeutung in Sprache mittels der lexikalisch-semantischen Wissensbasis FrameNet an. FrameNet widmet sich beiden Seiten der Erfassung von Weltwissen und beruht auf der linguistischen Theorie der Frame-Semantik, welche auf das Modellieren von menschlichem Sprachverstehen abzielt. FrameNet definiert Frames als Kategorien für Bedeutungseinheiten und weiterhin definiert es Beziehungen zwischen Frames als strukturiertes Metawissen über Zusammenhänge von Frames. Diese Frames und Beziehungen sind für die Aufgaben der Frame-Identifikation und der Frame-zu-Frame-Beziehungs-Vorhersage von zentraler Bedeutung.

Bezugnehmend auf Seite *(a)*, wurde die Aufgabe der Frame-Identifikation entwickelt, um so das Verstehen von Kontextwissen über typische Situationen, Handlungen und deren Akteure zu fördern. Bei dieser Aufgabe sollen Prädikate mit Frames annotiert werden, um so die Bedeutung des Prädikats im Satzkontext zu erfassen. Wir verwenden textuelle verteilte Repräsentationen, um die Bedeutung von Wörtern im Satzkontext zu modellieren und entwickeln ein System für Frame-Identifikation, das beste Leistungen im Vergleich zu Vorgängersystemen erzielt. Unser System für Frame-Identifikation kann zur automatischen Annotation von Frames in englischen oder in deutschen Texten genutzt werden. Darüber hinaus entwickeln wir einen multimodalen Ansatz zur Frame-Identifikation, in welchem wir textuelle verteilte Repräsentationen für Wörter mit visuellen verteilten Repräsentationen für auf Bildern dargestellte Entitäten kombinieren. Wir finden heraus, dass visuelle Informationen besonders in schwierigen Kontexten mit seltenen Frames hilfreich sind. Für künftige Arbeiten zur Weiterentwicklung des multimodalen Ansatzes schlagen wir vor, multimodale verteilte Repräsentationen gezielt für Verben zu entwickeln.

Bezugnehmend auf Seite *(b)*, führen wir die Aufgabe der Frame-zu-Frame-Beziehungs-Vorhersage ein, um so das Verstehen von strukturiertem Wissen über Zusammenhänge von Frames zu fördern. Bei dieser Aufgabe sollen Verbindungen zwischen Frames mit Beziehungsbeschreibungen annotiert werden, um so das strukturierte Metawissen über Frames in FrameNet zu erweitern. Wir trainieren textuelle und strukturierte verteilte Repräsentationen für Frames und erforschen die Grenzen der textuellen verteilten Repräsentationen beim Modellieren von Beziehungen. Darüber hinaus stellen wir textuelle und strukturierte verteilte Repräsentationen vergleichend gegenüber und entwickeln das erste System für die Frame-zu-Frame-Beziehungs-Vorhersage. Wir finden heraus, dass textuelle und strukturierte verteilte Repräsentationen bei der Vorhersage von Beziehungen Unterschiede aufweisen. Das bedeutet, dass diese beiden Repräsentationsarten unterschiedliches Frame-Wissen beisteuern können, wenn sie im Rahmen anderer Aufgaben angewendet werden. Weiterhin kann unser strukturiertes Vorhersagesystem genutzt werden, um Vorschläge für die Vervollständigung der Beziehungs-Annotation in FrameNet zu machen. Für künftige Arbeiten zur Weiterentwicklung des strukturierten Ansatzes für die Frame-zu-Frame-Beziehungs-Vorhersage schlagen wir vor, auch hier visuelle Informationen einzubinden. Zusätzlich kann ein solcher erweiterter Ansatz zur Einführung von neuen Frames und Beziehungen beitragen.

Die zwei Arten von Frame-Wissen der beiden Seiten – unser System für Frame-Identifikation und unsere verteilten Repräsentationen für Frames – werden für eine extrinsische Evaluierung im Rahmen anderer Aufgaben angewandt. Über die verschiedenen Anwendungen hinweg sehen wir einen Trend, dass Frame-Wissen besonders in mehrdeutigen und kurzen Sätzen hilfreich ist.

Zusammengefasst behandeln wir in dieser Dissertation zwei entgegengesetzte Seiten des Verstehens von Bedeutung in Sprache, nämlich das Verstehen von typischen Situationen und Handlungen sowie das Verstehen von strukturiertem Wissen über Relationen, und wir untersuchen beide Seiten mit unterschiedlichen verteilten Repräsentationen, wobei wir textuelles, strukturiertes und multimodales Hintergrundwissen abdecken.

Acknowledgments

I would like to express my warmest gratitude to all people who supported me as advisors or as friends and who, by this, made this thesis possible.

First, I would like to thank my supervisor Prof. Dr. Iryna Gurevych most sincerely for creating an excellent environment at the Ubiquitous Knowledge Processing Lab and at the graduate school on Adaptive Preparation of Information from Heterogeneous Sources. Thank you for contributing inspiring as well as challenging thoughts in many discussions, and for always having an open ear for me.

Second, I am very thankful to my co-supervisor, Prof. Dr. Stefan Roth, for providing valuable feedback ranging from the very broad perspectives to the very details. Thank you for contributing your expertise towards the integration of the visual modality into language processing.

Third, I would like to thank my external reviewer, Prof. Dr. Hinrich Schütze, and my committee members, Prof. Dr. Karsten Weihe, Prof. Dr. Johannes Fürnkranz, and Prof. Dr. Christian Reuter, for discussing my research with me.

Next, I am very thankful to Prof. Andrew McCallum and the Information Extraction and Synthesis Lab at UMass Amherst for offering the opportunity to do a research stay. Thank you all for supporting me in getting familiar with research in Deep Learning and Natural Language Processing in the early phase of my PhD.

Furthermore, I would like to express special thanks to my colleagues with whom I had the chance to collaborate with closely and who peer-reviewed parts of this thesis: Daniil Sorokin, Jan-Christoph Klie, Dr. Lisa Beinborn, Maxime Peyrard, Markus Zopf, Ilia Kuznetsov, Shweta Mahajan, Dr. Thomas Arnold, and Ines Zelch. Thank you all for being so supportive and, at the same time, for challenging me with critical questions. In addition, I thank all my other colleagues, with whom I had the chance to work together and to grow jointly by sharing thoughts and doubts: Aicha, Ana, Andreas, Avinesh, Benjamin, Claudia, Fabrizio, Gerold, Hatem, Silvana, Tobias, and Todor. Finally, I enjoyed the productive and welcoming atmosphere in all the UKP-lab and in the AIPHES graduate school. Thank you all for discussing research (and everything else) from so many perspectives, for being sometimes critical and sometimes philosophical. I am happy of having found some good friends in you.
Also, I appreciate the high motivation by Patricia, Ines, André, Anadi, and Jan-Christoph. It was inspiring to (co-)supervise your Bachelor or Master theses.

Moreover, I would like to thank Birgit and the Atelier for making Darmstadt colorful, and my close friend Eyleen for accompanying each other, being near or far.

After all, I am deeply grateful to my parents Carola and Karl, my brother Johannes and all my family and friends for making life so beautiful, for always being there and for growing up together.
With all my heart, I thank my husband Fiete for all and everything on our journey. Ich liebe Dich.

Contents

Chapter 1

Introduction

1.1 Motivation

Communication of meaning and knowledge is essential to humans and might even be the key to the development of humans as a species (Premack, 2004; Locke and Bogin, 2006). Humans use language to interact with other humans, to communicate information about the world they live in and to exchange their thoughts. In order to understand the meaning of words in a sentence, single words are interpreted in the context of the sentence and also in the context of the situation. Human language understanding relies on a large treasure of commonsense knowledge and experience in the world and links words to their referents in the real world (Barsalou, 1999).

The research field of Natural Language Processing (NLP) aims to model and to analyze language as used by humans as means of communication. The higher-order goal is to automatically understand language as humans do naturally (Jurafsky and Martin, 2017) given their shared background of commonsense knowledge. The NLP-perspective on commonsense knowledge branches out into two complementary directions: *(a) knowledge about situations and actions* as expressed in texts and *(b) structured relational knowledge* as stored in knowledge bases. An interdisciplinary perspective adds the grounding of language in different channels of the human sensomotoric inventory: *multimodal knowledge* such as visual experience.

Addressing aspects of automatic language understanding, current methods operate in 'embedding spaces' where human concepts, such as words in language, or artefacts from the world, such as objects depicted on images, are modeled as high-dimensional vectors. More broadly from the perspective of the research field of Artificial Intelligence (AI), representations from different kinds of sensors are accumulated to infer decisions or actions.

As the fundamental starting point of this thesis, we take position with respect to two open fields of discussion in Artificial Intelligence regarding human language understanding. On the one hand, the question is about whether Artificial Intelligence should aim to mimic or to inspire in humans. And on the other hand, the question is about the role of Natural Language Processing in the context of Artificial Intelligence.

Artificial Intelligence – to Mimic, to Inspire in, or to Ignore Humans? The field of Artificial Intelligence aims to build software and robots, which incorporate a range of abilities that is comparable to that of humans (Russell and Norvig, 1995). This aim is not limited to, but includes, human language understanding. Approaches to this aim range on a large continuum between two extremes. To the one end, the extreme is to aim at a detailed understanding of the human brain in order to exactly mimic human language processing. To the other end, the extreme is to ignore how humans or the human brain accomplishes certain abilities, as long as an automated system can deliver the desired output or action. In-between the two extremes but leaning towards the former, approaches tend to inspire in humans. According to Davis and Marcus (2015), Artificial Intelligence is not about directly mimicking human cognition, but about operating with representations of human common sense. Furthermore, Lake et al. (2017) express a need for algorithms to learn and think like people in terms of lifelong learning to generalize over tasks and to acquire meta-level skills.

In this thesis, we take the direction of inspiring in humans in order to pursue the goal of automatic language understanding in terms of human categories of meaning. Our work is not about finding biologically plausible models of the human mind, but about approximating human-like understanding of meaning by inspiring computational approaches in how humans process information and infer meaning.

Role of Natural Language Processing to Artificial Intelligence. The role of Natural Language Processing with respect to Artificial Intelligence can be discussed controversially.[1] Again, opinions on this topic range on a large continuum between two opposing extreme perspectives. The one extreme perspective regards Natural Language Processing as not being of major importance to Artificial Intelligence where more basic tasks should be solved first. The other extreme perspective regards Natural Language Processing as being key to Artificial Intelligence where language gives access to shared knowledge. In-between the two extremes but leaning towards the latter, recent trends in research on Artificial Intelligence suggest to incorporate human world knowledge into automatic approaches for improving automatic text processing (Marcus, 2018a), and also, the principle of *innateness* is identified as key for artificial intelligence (Marcus, 2018b). By this, 'Natural Language Processing is relevant to the goal of artificial intelligence in several ways'.[2]

In this thesis, we take a compromise stance and argue for combining Natural Language Processing with other disciplines (such as Computer Vision) in order to pursue the goal of automatic language understanding in terms of human categories of meaning. This is about leveraging different modalities in terms of multimodal embeddings in order to approximate a holistic incorporation of meaning.

Having clarified our starting position, next, we formulate our overall research question in the context of Natural Language Processing and elaborate on our proceeding.

[1] Controversy between Yann LeCun and Christopher Manning: `https://www.youtube.com/watch?v=fKk9KhGRBdI&feature=youtu.be`

[2] Jacob Eisenstein's 2018 draft on 'Natural Language Processing' (under contract with MIT Press, shared under CC-BY-NC-ND license): `https://github.com/jacobeisenstein/gt-nlp-class/tree/master/notes`

Research Question in the Context of Natural Language Processing. The higher-order goal of Natural Language Processing is to automatically understand language as humans do naturally (Jurafsky and Martin, 2017), which requires a holistic understanding of how humans express meaning in texts, including their large background of commonsense knowledge and experience in the world. As outlined in the beginning, the overall challenge of capturing meaning and world knowledge in language can be split into the two branches of *(a) knowledge about situations and actions* and *(b) structured relational knowledge*. These can be studied with different kinds of embeddings for operationalizing different aspects of knowledge, such as textual embeddings for modeling situations or actions expressed in texts, structured embeddings for relations stored in knowledge bases, or visual embeddings for objects depicted on images – where a combination of embeddings yields multimodal embeddings. With respect to modeling meaning in vector spaces, Liang (2016) points out the challenge of 'how to represent the semantics of natural language'. In this thesis, we pick up on this challenge and formulate our overall research question:

> **'What kind of vector representations are suitable for Natural Language Processing tasks involving semantic language understanding according to human categories of meaning?'**.

In Natural Language Processing, different tasks have been established in order to advance the computational understanding of meaning expressed in terms of language. Starting at syntax-level, typical NLP-tasks focus on the annotation of parts-of-speech (van Halteren et al., 2001), *e.g., nouns, verbs, adjectives*, or dependencies (Kübler et al., 2009), *e.g., nominal subject, verbal modifiers, clausal complements.* Further typical NLP-tasks continue at a semantic level where the focus is on identifying the underlying meaning of expressions or sentences. This includes the semantic analysis of predicate-argument structures (Gildea and Jurafsky, 2002), *e.g., finding out about 'who does what to whom'*, and also the disambiguation of words which can refer to several meanings, such as linking entities to knowledge bases (Erbs et al., 2011), *e.g., match the mention of 'Obama' to the Wikipedia entry of Barack Obama*, or disambiguating the sense of a word by linking it to a lexicon entry (Navigli, 2009; Mihalcea, 2007), *e.g., connect the mention of 'key' to the entry for 'keyboard'.* Above that, more abstract NLP-tasks pose the challenge of semantic text understanding as in sentence similarity (Agirre et al., 2012; Cer et al., 2017), *e.g., judge whether two sentences express similar meaning*, or in summarization (Nenkova and McKeown, 2011), *e.g., extract the most relevant expressions from a document.*

These tasks are designed to contribute as substeps to process natural language in order to, finally, model human language understanding and communication of meaning. Additionally, these tasks help to evaluate advances in vector representations that aim to incorporate meaning.

Moreover, an *interdisciplinary* perspective encourages the integration of multimodal information into NLP-tasks. According to the field of grounded cognition and embodied sentence comprehension (Barsalou, 2008), human language understanding incorporates different levels of sensomotoric experience in the world. To give an example, for understanding the meaning of an expression humans connect it to experiences of different modalities:

'She is running with a barking dog.'
→ Visual modality: we have seen instantiations of the entity dog
→ Auditory modality: we have heard the sound of barking
→ Motoric modality: we have performed the activity of running

Transferring the grounded perspective to NLP-tasks encourages the incorporation of multimodal embeddings, i.e. the combination of embeddings. Specifically, images are an orthogonal source of world knowledge to texts and recently, combinations of visual and textual information have been successfully applied to NLP-tasks (Beinborn et al., 2018). Furthermore, Schubert (2015) notes a list of desiderata and approaches for semantic representations which includes language-like expressivity as well as accord with semantic intuitions (amongst others). Even if in this list cognitive aspects of human language understanding are missing, Schubert (2015) still mentions to integrate specialized methods that orient on the human proficiency in spatial and imagistic reasoning.

In this thesis, we build upon these desiderata and take them on a cognitive level when extending our overall research question to the question of how multimodal representations could improve Natural Language Processing.

Next, we outline our approach to the overall research question, starting from the fundamental assumption about holistic language understanding requiring different aspects of world knowledge – which is inspired by human language understanding.

Approach to Our Overall Research Question. To approach our overall research question, we choose to closely rely upon the lexical-semantic knowledge base FrameNet as it addresses both branches of capturing world knowledge whilst taking into account the linguistic theory of frame semantics (Fillmore, 1976) which orients on human language understanding. FrameNet provides frames, which are categories for knowledge of meaning, and frame-to-frame relations, which are structured meta-knowledge of interactions between frames. These frames and relations are central to the tasks of Frame Identification and Frame-to-Frame Relation Prediction, respectively.

On the one hand, *(a) knowledge about situations and actions* refers to general commonsense knowledge of situations or actions (i.e., humans perceive a certain course of happenings as a situation or an action and agree on a set of expected participants) – and this is relevant to the task of Frame Identification in the context of Semantic Role Labeling (Das et al., 2014). The task of Frame Identification is to label predicates with frames in order to identify the meaning of the predicate in the context of the sentence. To give an example, an expression annotated with FrameNet frames (Baker et al., 1998) specifies the situation or action that is happening:

*'He **sat** down on a bench.'* *evokes* FrameNet *frame: Change_posture*

We use textual embeddings to model the semantics of words in the sentential context and develop a state-of-the-art system for Frame Identification. Our Frame Identification system can be used to automatically annotate frames on English or German texts. Furthermore, in our multimodal approach to Frame Identification, we combine textual embeddings for words with visual embeddings for entities depicted on

images. We find that visual information is especially useful in difficult settings with rare frames. To further advance the performance of the multimodal approach, we suggest to develop embeddings for verbs specifically that incorporate multimodal information.

On the other hand, *(b) structured relational knowledge* refers to concrete knowledge of relations between entities (i.e., documented relations about who did what, or what is located where) – and this is relevant to the task of Knowledge Base Completion (Wang et al., 2017), where world knowledge is formulated in terms of relational triples. The task of Knowledge Base Completion is to label relations between entities in order to complete knowledge bases such as Freebase (Bollacker et al., 2008). To give an example, a relation annotated between two entities forms a triple:

> *'Barack Obama' and 'Michelle Obama' form the* Freebase *triple:*
> *('Barack Obama', 'married to', 'Michelle Obama')*

FrameNet can also be regarded as a special case of a knowledge base storing relations as it provides meta-knowledge of interactions between frames (i.e., what action follows after another action) – and we propose the task of Frame-to-Frame Relation Prediction to complete the meta-knowledge stored in the FrameNet hierarchy. To give an example, a relation annotated between two frames forms a triple:

> *'Change_posture' and 'Posture' form the* FrameNet *triple:*
> *('Change_posture', 'causative_of', 'Posture')*

We train textual and structured embeddings for frames and explore the limitations of textual frame embeddings with respect to recovering relations between frames. Moreover, we contrast textual frame embeddings versus structured frame embeddings and develop the first system for Frame-to-Frame Relation Prediction. We find that textual and structured frame embeddings differ; thus when applied as features in the context of further tasks, they can provide different kinds of frame knowledge. Our prediction system leveraging the structure of the FrameNet hierarchy can be used to generate recommendations for annotations with relations. To further advance the performance of Frame-to-Frame Relation Prediction and also of the induction of new frames and relations (short frame-relation-induction), we suggest to develop approaches that incorporate visual information.

The two kinds of frame knowledge from both branches, our Frame Identification system and our pre-trained frame embeddings, are combined in an extrinsic evaluation in the context of higher-level applications: Summarization, Summary Evaluation, Motif Construction, Semantic Textual Similarity, and Argument Reasoning Comprehension. Across these applications, we see a trend that frame knowledge is particularly beneficial in ambiguous and short sentences. Thus, from a practical point of view, there are direct applications of our systems and embeddings in text processing.

Finally, we provide an outlook on the next challenges for multimodal language processing. Other than the development of multimodal verb embeddings and the integration of visual information for frame-relation-induction, we elaborate on the need to automatically learn how to combine complementary information and select relevant information from different modalities.

Taken together, in this thesis, we approach semantic language understanding from the two branches of *(a) knowledge about situations and actions* and *(b) structured relational knowledge*, and we investigate different embeddings for textual, structured and multimodal language understanding. In a broader sense, representations for meaning are the communication channel between human language (here in the form of text) and machines; thus, this thesis contributes towards improved automatic processing of meaning expressed by human language. Finally, when regarding language as the human way of incorporating shared knowledge aggregated from several modalities, then multimodal representations can be regarded as a computational way of modeling this shared knowledge.

1.2 Contributions and Findings

Here, we list our contributions and findings in order to provide a concise overview.
Contributions:

- Frame Identification **systems** that operate on FrameNets of different languages, namely English and German: `UniFrameId` (based on unimodal textual embeddings) and `MultiFrameId` (based on multimodal embeddings)
- Knowledge Base Completion **systems** for FrameNet's frame-to-frame relations with our `StruFFRel` approach (leveraging the structure of the FrameNet hierarchy to train the prediction)
- Different kinds of frame **embeddings**: textual and structured
- Extrinsic evaluation of the potential of frame knowledge in different **application** scenarios: Summarization, Summary Evaluation, Motif Construction, Semantic Textual Similarity, Argument Reasoning Comprehension
- Explorations for future work on **multimodality**: verb similarity, Knowledge Base Completion

Findings:

- Structured knowledge about frames complements textual knowledge about frames with respect to frame-to-frame relations.
- Visual commonsense knowledge about participants helps to identify the frames in a sentence.
- Semantic knowledge from FrameNet shows a trend to be helpful in applications such as Summarization, Motif Construction, Semantic Textual Similarity, and Argument Reasoning Comprehension.
- Multimodal approaches improve different tasks in the context of language understanding: tasks requiring knowledge about situations or actions as well as relational knowledge.
- Identification of next challenges: development of multimodal embeddings for verbs to improve Frame Identification, and integration of visual knowledge into Frame-to-Frame Relation Prediction and into induction of frames.

1.3 Publication Record

Several parts of this thesis have been published previously in international peer-reviewed conference and workshop proceedings from major events in natural language processing, e.g. ACL with NAACL and EACL, EMNLP, COLING. All the publications are listed below, together with indications for the chapters and sections of this thesis which build upon them, and with a notion of the author's contribution.

- **Teresa Botschen**, Iryna Gurevych, Jan-Christoph Klie, Hatem Mousselly-Sergieh and Stefan Roth: 'Multimodal Frame Identification with Multilingual Evaluation', in: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 1481–1491, New Orleans, USA, June 2018.
  My contributions in this paper are the following: `UniFrameId` system for FrameNet and SALSA, `MultiFrameId` system and analysis of experiments.
  (Chapters 2, 4 in Sections, 2.1.2 4.1, 4.2)

- **Teresa Botschen**, Hatem Mousselly-Sergieh and Iryna Gurevych: 'Prediction of Frame-to-Frame Relations in the FrameNet Hierarchy with Frame Embeddings', in: *Proceedings of the 2nd Workshop on Representation Learning for NLP (RepL4NLP, held in conjunction with ACL)*, pp. 146–156, Vancouver, Canada, August 2017.
  My contributions in this paper are the following: exploration of frame embeddings, `StruFFRel` approach and analysis of experiments.
  (Chapters 2, 5 in Sections 2.2.2, 5.1 and 5.2)

- Lisa Beinborn*, **Teresa Botschen*** and Iryna Gurevych: 'Multimodal Grounding for Language Processing', in: *Proceedings of the 27th International Conference on Computational Linguistics: Technical Papers (COLING)*, pp. 2325–2339, Santa Fe, USA, August 2018. (* equal contribution)
  My contribution in this paper is the following: exploration of verb embeddings.
  (Chapter 4 in Section 4.2.4.1)
  Further, joint contributions of myself together with my co-author Lisa Beinborn are: distinctions within models of multimodal information flow and within methods for learning multimodal embeddings, and a literature review on combining and selecting information from different modalities. For these, we refer to our survey in the background chapters and in the outlook.
  (Chapters 2, 3, 7 in Sections 2.3, 3.5, 7.1)

- **Teresa Botschen***, Daniil Sorokin* and Iryna Gurevych: 'Frame- and Entity-Based Knowledge for Common-Sense Argumentative Reasoning', in: *Proceedings of the 5th International Workshop on Argument Mining (ArgMin, held in conjunction with EMNLP)*, pp. 90–96, Brussels, Belgium, November 2018. (* equal contribution)
  My contributions in this paper are the following: annotation of texts with frames using `UniFrameId` system, input with frame embeddings, analysis with respect to frames.
  (Chapter 6 in Section 6.2.2)

- Hatem Mousselly-Sergieh, **Teresa Botschen**, Iryna Gurevych, and Stefan Roth: 'A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning', in: *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (StarSem, held in conjunction with NAACL)*, pp. 225–234, New Orleans, USA, June 2018.
  My contributions in this paper are the following: exploration of synset embeddings, extension of approach by Xie et al. (2017) for multimodal Knowledge Base Completion on WN9-IMG dataset.
  (Chapters 3, 5 in Sections 3.3, 5.2.3.1)

- Silvana Hartmann, Ilia Kuznetsov, **Teresa Martin** and Iryna Gurevych: 'Out-of-domain FrameNet Semantic Role Labeling', in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 471–482, Valencia, Spain, April 2017.
  My contribution in this paper is the following: `SimpleFrameId` system with `WSABIE` embeddings.
  (Chapters 3, 4 in Sections 3.2, 4.1.3)

- **Teresa Martin**, Fiete Botschen, Ajay Nagesh and Andrew McCallum: 'Call for Discussion: Building a New Standard Dataset for Relation Extraction Tasks', in: *Proceedings of the 5th Workshop on Automated Knowledge Base Construction (AKBC, held in conjunction with NAACL)*, pp. 92–96, San Diego, USA, June 2016.
  My contributions in this paper are the following: identification of weaknesses of datasets for Relation Extraction, roadmap for building a fully labeled dataset.
  (Chapter 2 in Section 2.2.1)

- Markus Zopf, **Teresa Botschen**, Tobias Falke, Benjamin Heinzerling, Ana Marasović, Todor Mihaylov, Avinesh P.V.S, Eneldo Loza Mencía, Johannes Fürnkranz, and Anette Frank: 'What's important in a text? An extensive evaluation of linguistic annotations for summarization.', in: *Proceedings of the 5th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Valencia, Spain, October 2018.
  My contribution in this paper is the following: annotation of texts with frames using `SimpleFrameId` system.
  (Chapter 6 in Section 6.1.1)

- Maxime Peyrard, **Teresa Botschen**, and Iryna Gurevych: 'Learning to Score System Summaries for Better Content Selection Evaluation', in: *Proceedings of the Workshop "New Frontiers in Summarization" (held in conjunction with EMNLP)*, pp. 74–84, Copenhagen, Denmark, September 2017.
  My contribution in this paper is the following: annotation of texts with frames using `SimpleFrameId` system.
  (Chapter 6 in Section 6.1.2)

1.4 Thesis Structure

In the following we present the structure of this thesis that is illustrated in Figure 1.1, and which will accompany us throughout the thesis.

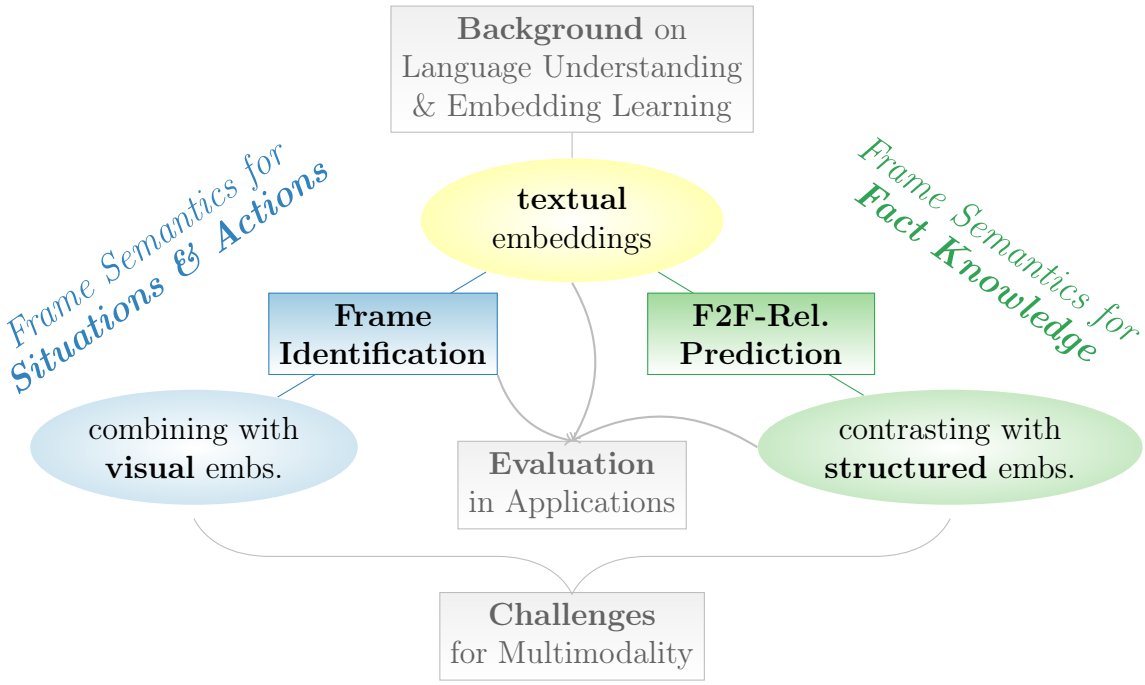


Figure 1.1: Overview of thesis structure. Upper gray box (Ch. 2, 3): theoretical and methodological background. Left blue branch (Ch. 4): knowledge about situations and actions with textual and visual word embeddings for Frame Identification. Right green branch (Ch. 5): knowledge about facts with textual versus structured frame embeddings for Frame-to-Frame Relation Prediction. Middle gray box (Ch. 6): evaluation of frame knowledge in applications. Lower gray box (Ch. 7): outlook on challenges for grounded language processing.

To start with, we provide the theoretical and methodological background on language understanding and on embedding learning (upper gray box). In Chapter 2 we review different facets of language understanding: textual semantics for situations and actions, structured relational knowledge and grounded language understanding. In Chapter 3 we review methods of representation learning which we apply to our data: textual, structured, visual and multimodal embedding approaches.

To study world knowledge as conceptualized by frame semantics and viable by embeddings, we branch out into two directions.

On the one hand (left **blue** branch), in Chapter 4, we model knowledge about situations and actions with textual word embeddings and in combination with visual ones for the task of Frame Identification.

On the other hand (right **green** branch), in Chapter 5, we contrast textual and structured frame embeddings to model knowledge about relational triples in the task of Frame-to-Frame Relation Prediction.

Subsequently, in Chapter 6, we extrinsically evaluate frame knowledge (from the two branches) in high-level tasks (middle gray box) by reporting about applications

of our unimodal Frame Identification system and of our textual and structured frame embeddings.

Finally, in Chapter 7, we resume with an outlook on the directly implied next challenges for grounded language processing as the combination of complementary information and also the selective grounding in different modalities and a comment on the trend for the role of Natural Language Processing (lower gray box).

Chapter 2

Understanding Meaning in Language

This chapter provides a theoretical overview by reviewing relevant literature of different facets of language understanding. We explain the conceptual foundation of this thesis by branching out into *(a) knowledge about situations and actions* and *(b) structured relational knowledge*, and by leveraging frame semantics.

On the one side *(a)*, we explore textual semantic language understanding (Section 2.1) where we review the corresponding part of the lexical-semantic knowledge base FrameNet and the task of Frame Identification. And on the other side *(b)*, we explore structured language understanding (Section 2.2) where we review the structured part of FrameNet and introduce the task of Frame-to-Frame Relation Prediction. Furthermore, both sides can be extended with the perspective of grounded language understanding (Section 2.3), which motivates the combination of different information channels for any complex task.

2.1 Textual Semantic Language Understanding

Knowledge of meaning enables humans to understand the semantics of language, texts or single words. Semantic knowledge of meaning can be incorporated either by the words themselves and, importantly, by the context around the words (Section 2.1.1) or by shared categories of meaning that several words can refer to depending on the context they appear in (Section 2.1.2).

2.1.1 Meaning via Context

Context words are crucial to understand the semantics of single words. As an example, after hearing a sequence of words in a statement, it is possible to guess which word will be said next. Also for ambiguous verbs like *'buy'* or *'play'*, the given context helps to further specify what aspect of word meaning is referred to, see for example:

$$buy \; < \; \begin{matrix} \textit{buy goods} \\ \textit{buy an excuse} \end{matrix} \quad ; \quad play \; < \; \begin{matrix} \textit{play a game} \\ \textit{play on an instrument} \end{matrix}$$

The importance of the context words for the meaning of a single word is expressed by linguistic philosophy.

Linguistic Philosophy. Linguistic philosophy, or ordinary language philosophy, analyzes natural language as it is used by humans in order to gain knowledge. With this approach and aim, linguistic philosophy is different from analytic philosophy of language formally analyzing language in order to improve natural language by the insights of formal logics ('Tractatus logico-philosophicus', Wittgenstein, 1922). As one representative work of linguistic philosophy, 'Philosophical investigations' states:

> *'The meaning of a word is its use in the language.'* Wittgenstein (1953)

This expresses that the meaning of a word is the context it appears in in language or in text and also that word meaning can be inferred from textual context. It motivates considering word meaning as observed contexts, basically as a distribution over context words. In the field of Natural Language Processing, this point of view is implemented in distributional approaches which will be described in Section 3.2.

2.1.2 Understanding Situations and Actions with Frames

Other than directly incorporating knowledge of meaning by words themselves (Section 2.1.1), the theory of frame semantics (Fillmore, 1976), organizes knowledge of meaning in categories – so called frames – and considers these as cognitive schemata:

> *'Frames are the cognitive schemata that underlie the meanings of the words associated with that Frame.'* Fillmore (2012)

Frame semantics uses frames to capture complex situations and states that the words in our language are understood with frames:

> *'The idea behind frame semantics is that speakers are aware of possibly quite complex situation types, packages of connected expectations, that go by various names – frames, schemas, scenarios, scripts, cultural narratives, memes – and the words in our language are understood with such frames as their presupposed background.'* Fillmore (2012)

The frame itself refers to a situation or action which, in turn, is further specified by linking to participants of the event – by this, a whole scenario is described. In parallel to Fillmore, several lines of research, not only linguistics but also cognitive science and artificial intelligence, were working on formalizations of human knowledge structures (for an overview see Minsky, 1988). Minsky describes earlier work by Fillmore (1967, foundation of later frame semantics) as a case-grammar sentence-analysis theory centered around verbs, and furthermore he describes work by Schank (1972, foundation of later script knowledge) as a collection of 'basic conceptualizations' and relations between them. Minsky himself initially proposes *frames* to structure human knowledge into 'stereotyped situations' – but these *frames* are not exactly the same as Fillmore's later frames in the context of frame semantics and as then incorporated by FrameNet. Still, Fillmore's frames can be seen as one outcome of the wide field of discussion and research on human knowledge structures that got implemented by FrameNet which is described in the next paragraph.

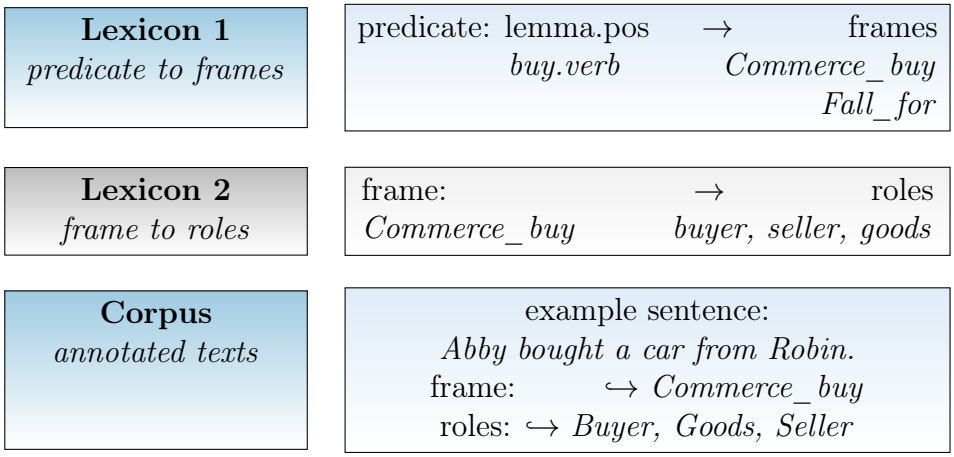


Figure 2.1: Sketch of the FrameNet resources providing semantic knowledge. Lexicon 1 (upper blue) provides a mapping from predicates to evokable frames. Lexicon 2 (gray, as not used in this thesis) provides a mapping from frames to roles. The corpus (lower blue) provides fully annotated sentences from news articles, where the annotations use frames and roles according to lexicon 1 and 2.

FrameNet incorporating Frame Semantics. The Berkeley FrameNet (Baker et al., 1998; Ruppenhofer et al., 2016, FN, common abbreviation) is an ongoing project for manually building a large lexical-semantic knowledge base (KB, common abbreviation) with expert annotations. FrameNet embodies the theory of frame semantics (Fillmore, 1976): frames capture units of meaning corresponding to prototypical situations. FrameNet provides two lexica with repertoires of situations and actions (frames for predicates) and participants (frame-specific roles for participants), and texts manually labeled with respect to these lexica. These knowledge resources are sketched in Figure 2.1. FrameNet differentiates between predicates together with frames and the participants of these frames. The first lexicon consists of a mapping from predicates to frames and the second lexicon consists of a mapping from frames to frame-specific roles. We extend the example of the verb *'buy'* (cf. Lexicon 1 in Figure 2.1) to showcase that ambiguous predicates can evoke different frames depending on the context:

$$buy \;<\; \begin{matrix} \textit{buy goods} & \rightarrow & \textit{evokes frame: Commerce\_buy} \\ \textit{buy an excuse} & \rightarrow & \textit{evokes frame: Fall\_for} \end{matrix}$$

Concerning the roles, the second lexicon gives access to frame-specific role-labels (e.g., *'Buyer', 'Goods'* or *'Deception', 'Victim'*) as applied in Semantic Role Labeling (SRL, common abbreviation).

As an overview of terms used in the context of FrameNet: a predicate can evoke several frames – thus, the predicate is also called frame evoking element. A predicate is captured in terms of a lexical unit (LU): the lemma of the predicate and its part-of-speech tag (POS tag). In FrameNet, importantly, predicates are not reduced to verbs only, but also nouns or adjectives can incorporate predicates. Each frame, in turn, provides a list of frame elements, also called roles, that can be assigned to the arguments of the predicate – then, these are called fillers of the frame elements. Together, the frames (for predicates) and the frame elements (for arguments) are the labels to assign to semantic predicate argument structures on the sentence level.

lexicon	**frames**	**LUs**	**avg**(fr/pred)	**%amb.pred.**
FrameNet	1020	11942	1.26	17.32
SALSA	1023	1827	2.82	57.56

Table 2.1: Lexicon statistics for FrameNet 1.5 and for SALSA 2.0: the total number of distinct **frames** and lexical units **LUs** (distinct predicate-frame combinations), the number of frames a predicate can evoke on average **avg**, and the % of **amb**iguous **pred**icates.

		sentences	**frames**
FrameNet	train	2819	15406
	dev	707	4593
	test	2420	4546
SALSA	train	16852	26081
	dev	3561	5533
	test	3605	5660

Table 2.2: Dataset statistics for FrameNet 1.5 corpus of fully annotated texts with split by Das et al. and for SALSA 2.0 with our split: number of **sentences** and **frames** (as used in our experiments).

In this work, we are working with the English FrameNet (Baker et al., 1998; Ruppenhofer et al., 2016) and the German counterpart, SALSA (Burchardt et al., 2006; Rehbein et al., 2012, short for Saarbrücken Lexical Semantics Annotation and Analysis). For a comparative overview of FrameNet versus SALSA, Table 2.1 contains the lexicon statistics and Table 2.2 the dataset statistics.

2.1.2.1 FrameNet Semantic Role Labeling

Semantic Role Labeling (SRL, common abbreviation) is a basic task in Natural Language Processing (NLP, common abbreviation), introduced by Gildea and Jurafsky (2002). Semantic Role Labeling aims at structuring the meaning of a sentence in order to answer the question of *'Who did what to whom?'*. To understand the meaning of a sentence it is important to identify and understand the situation or action that is happening and the participants that incorporate the roles involved in the event. Typically, a repertoire of situations, actions and participants is provided by a database for Semantic Role Labeling, such as FrameNet (Baker et al., 1998) or PropBank (Palmer et al., 2005).

FrameNet Semantic Role Labeling analyzes sentences with respect to frame-semantic structures based on FrameNet (Fillmore et al., 2003) and typically, this involves the following two steps. First, **Frame Identification** (FrameId, common abbreviation), which is to capture the context around a predicate and then to assign a frame to this predicate, i.e. a word sense label for a prototypical situation. For this step, lexicon 1 in Figure 2.1 is used. Second, **Role Labeling**, which is to identify the participants of the predicate and to connect them with predefined frame-specific role labels. For this step, lexicon 2 in Figure 2.1 is used.

For the full annotation, refer to the following example sentence from FrameNet's fully annotated corpus:

'Abby bought a car from Robin.'
⇒ Frame Identification: *'bought'* → *'Commerce_buy'*
⇒ Role Labeling: *'Abby'* → *'Buyer'*, *'a car'* → *'Goods'*, *'Robin'* → *'Seller'*

This example sentence describes the action of buying with the participants Abby, Robin and a car. The correct frame is *'Commerce_buy'* and the correct roles are *'Buyer'* for *Abby*, *'Seller'* for *Robin* and *'Goods'* for *a car*.

Annotating a sentence with situations, actions and participants is an abstraction of the sentence that structures the meaning. This, in turn, is used as input for higher-level tasks (Jurafsky and Martin, 2017) such as Question Answering (Surdeanu et al., 2011) or Machine Translation (Lo et al., 2013).

Importance of Frame Identification in Semantic Role Labeling. Frame Identification is crucial to the success of Semantic Role Labeling as errors in Frame Identification account for most wrong predictions in current systems (Hartmann et al., 2017). By definition, Frame Identification is more challenging than Role Labeling. This is because in Frame Identification a classification is done over more than 1000 frame-categories (cf. Table 2.1 for the number of frames in FrameNet and SALSA), whilst in Role Labeling not only the sentence but also the frame is known and so, the frame-specific roles cut down the number of categories for role classification to choose from: on average there are 9.7 frame elements per frame.

Hartmann et al. perform a comprehensive analysis of Semantic Role Labeling on several datasets, including out-of-domain datasets. There are two crucial observations: first, Frame Identification is more challenging on out-of-domain datasets and this effect is propagated to Role Labeling and full Semantic Role Labeling with system-predicted frames. Second, this effect is not found for Role Labeling and full Semantic Role Labeling with gold frames on out-of-domain datasets: the performance of full Semantic Role Labeling with gold frames is more than 70% F1 whereas that with system-predicted frames ranges from 29% F1 (out-of-domain) to 55% F1 (in-domain). This shows the dependence of FrameNet role labels on correct frame labels. Consequently, improving the step of Frame Identification (as the current bottleneck in Semantic Role Labeling) is of major interest.

2.1.2.2 Frame Identification

An essential step in FrameNet Semantic Role Labeling is the task of Frame Identification, which aims at disambiguating a situation around a predicate. The main challenge and source of prediction errors of Frame Identification systems are ambiguous predicates, which can evoke several frames. An ambiguous predicate evoking different frames was showcased above with the verb *'buy'* evoking the frames *'Commerce_buy'* or *'Fall_for'* – but there are also more fine-grained differences in the nuances of meaning as for example with the predicate *'sit'*:

$$sit \;<\; \begin{array}{lcl} \textit{a person is sitting back on a bench} & \rightarrow & \textit{evokes frame: Change\_posture} \\ \textit{a company is sitting in a city} & \rightarrow & \textit{evokes frame: Being\_located} \end{array}$$

In a context where a person is sitting somewhere, the verb *'sit'* evokes the frame *'Change_posture'*, while in a context where a company is sitting somewhere, it evokes *'Being_located'*. Understanding the context of the predicate, and thereby

the context of the situation (here, *'Who / what is sitting where?'*), is crucial to identifying the correct frame for ambiguous cases.

State-of-the-art systems for Frame Identification rely on pre-trained word embeddings as input (Hermann et al., 2014). This proved to be helpful: those systems consistently outperform the previously leading Frame Identification system `Semafor` (Das et al., 2014), which is based on a handcrafted set of features.

Definition of the Task. The task of Frame Identification is defined in the following. Given are a sentence S and a predicate $pred \in S$, plus optionally, a set of frames associated with this predicate via access to the FrameNet lexicon $F \in L$. The goal is to predict the correct frame $f_{correct}$ based on the context *cont* around the predicate in the sentence ($cont = words \in S$). See the following example:

> Given sentence: *'Abby bought a car from Robin.'* and predicate *'bought'*
> ⇒ predict *'Commerce_buy'*
> ⇒ or select *'Commerce_buy'* from the lexicon-list of all frames for *'buy'*.

Use of FrameNet Lexicon. For the evaluation of our systems, we consider two settings: with lexicon (standard procedure) and without lexicon (suggested in Hartmann et al. (2017)). In the with-lexicon setting, the lexicon is used to reduce the choice of frames for a predicate to only those listed in the lexicon. If the predicate is not in the lexicon, it corresponds to the without-lexicon setting, where the choice has to be done amongst all frames. During testing, a system for Frame Identification outputs weights for all the frames available in the lexicon, and the best-scoring frame is selected as frame prediction. From the machine learning perspective, the lexicon is an external resource of knowledge: after having the weights, additional filtering can be performed with the lexicon specifying the available frames for each lexical unit. By this, the prediction is made by selecting the highest weighted frame amongst only those available for the respective predicate. If the predicate is unknown to the lexicon, the overall best-scoring frame is chosen. If the predicate has only one entry in the lexicon, it is unambiguous and the frame is assigned directly. Thus, using the lexicon is increasing the performance of any system if the lexicon is reliable given the domain of the texts. However, the FrameNet lexicon has coverage problems when applied to new domains (Hartmann et al., 2017). On the one hand, rare and specific predicates can be missing in the lexicon and on the other hand, even if a certain predicate exists, it might not be linked to the correct frame for a specific context. Furthermore, frames for rare-domain contexts might be missing in the lexicon. For these different aspects of lexicon coverage issues, using the lexicon might obscure the differences between systems in the testing stage. To take this into account, a two-fold evaluation is the most comprehensive approach: one in the traditional `with-lexicon` setting, and one in the `no-lexicon` setting, where frames are assigned directly by the system and no lexicon-based filtering is performed.

Evaluation Metrics. Frame Identification systems are usually compared in terms of accuracy.

Accuracy. Accuracy (Equation 2.1) is defined as the fraction of the number of correct predictions divided by the number of samples – which is the total number of predictions:

$$accuracy := \frac{\text{number of correct predictions}}{\text{total number of predictions}} . \tag{2.1}$$

As a multiclass classification problem, Frame Identification has to cope with a strong variation in the annotation frequency of frame classes. Minority classes are frames that occur only rarely; majority classes occur frequently. Note that the accuracy is biased toward majority classes, explaining the success of majority baselines on imbalanced datasets such as FrameNet. Alternatively, the F1-score is sometimes reported as it takes a complementary perspective.

F1-score. The F-measure (Equation 2.2) is the harmonic mean of precision and recall, measuring exactness and completeness of a model, respectively:

$$F1 := 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} . \tag{2.2}$$

In previous work, micro-averaging is used to compute F1-scores. Yet, similar to the accuracy, micro-averaging introduces a bias towards majority classes. Furthermore, for a setup with multiclass classification, micro-averaging for F1 computes the same as accuracy. We compute F1-macro instead, for which precision and recall are computed for each class separately and averaged afterwards, giving equal weight to all classes.

Taken together, this yields scores that underestimate (F1-macro) and overestimate (average accuracy) on imbalanced datasets. Previous work just used the overestimate so that a comparison is possible in terms of accuracy in the with-lexicon setting. We suggest using F1-macro additionally to analyze rare, but interesting classes. Thus, a comparison within our work is possible for both aspects, giving a more detailed picture.

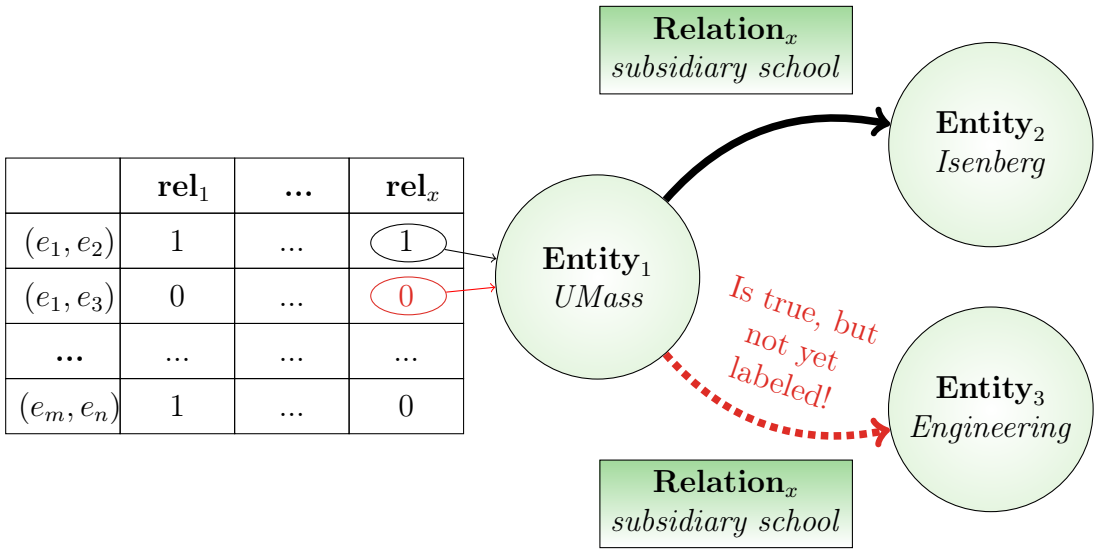
2.2 Structured Language Understanding

Knowledge bases organize knowledge of meaning: they structure relations about real-world entities or concepts in a graph structure. Typical knowledge bases with structured knowledge are organized in relational triples that form a graph (Section 2.2.1). Also the lexical semantic knowledge base FrameNet has a graph-structured component (Section 2.2.2).

2.2.1 Understanding Relations with Knowledge Bases

A knowledge graph (KG, common abbreviation) is a knowledge base that defines a graph structure, i.e. they store relational triples. A relational triple in a knowledge graph is structured as a triple of head and tail entities along with the relation that holds between them, for example:

(head-entity, relation, tail-entity)
('Michelle Obama', 'is married to', 'Barack Obama')



	rel$_1$	**...**	**rel**$_x$
(e_1, e_2)	1	...	1
(e_1, e_3)	0	...	0
...	...	...	...
(e_m, e_n)	1	...	0

Figure 2.2: Recall problem in knowledge bases. Left: Labelling structure of Freebase for triples of entity pairs (rows) and relations (columns). Cell labels: 1 for 'true, triple exists'; 0 for 'false, triple is not in Freebase' which means that it is either indeed false or actually true but not yet labeled. Right: The entities for University of Massachusetts and Isenberg college are connected via the relation *'subsidiary school'*. University of Massachusetts has several subsidiary colleges which do exist in Freebase, e.g., Engineering college, but the triple is not added to Freebase (red).

An example of a typical knowledge graph is the manually constructed Freebase (Bollacker et al., 2008, FB, common abbreviation), with the FB15k-dataset (Bordes et al., 2013) and its extension FB15k-237 (Toutanova et al., 2015) for Relation Extraction tasks. Knowledge graphs are crucial for various kinds of tasks, such as Question Answering and Information Retrieval. We denote the knowledge graph as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relations, and $\mathcal{T} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$ the set of triples in the knowledge graph.

Importance of Knowledge Base Completion. Relational knowledge is virtually infinite and is frequently subject to change. This raises the question of incompleteness of knowledge graphs. To address this problem, several methods have been proposed for automatic Knowledge Graph Completion (KGC, for a survey refer to Wang et al., 2017). In Martin et al. (2016) we illustrate one facet of incompleteness of knowledge graphs with the example of Freebase: the so-called *'recall problem'*. The recall problem produces misleading results concerning recall when evaluating on Relation Extraction tasks. As illustrated in Figure 2.2, the measure of recall is misleading when non-existent relations for entity pairs in the knowledge graph are assumed to be 'false' just because they do not appear so far. This assumption is dangerous as non-existent relations in the knowledge graph could indeed be 'true'. Knowledge Base Completion is to correctly label the triples for relations and pairs of entities in the graph in order to obtain more complete knowledge resources.

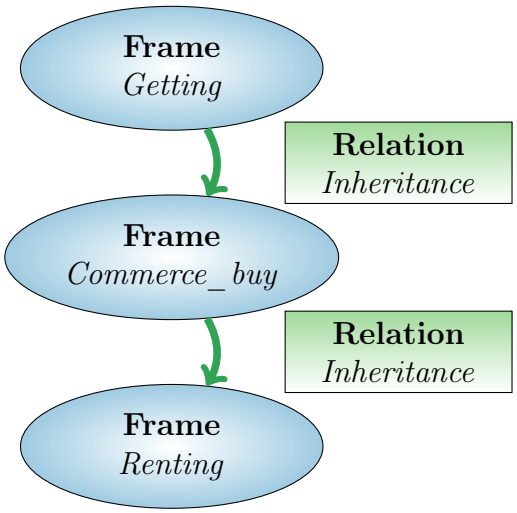


Figure 2.3: Sketch of the structure of FrameNet as a knowledge graph. Pairs of frames are connected via frame-to-frame relations.

2.2.2 Frame Semantics in a Knowledge Base

FrameNet as Knowledge Base. Most often, the definitions of frame-evoking elements and frame-specific roles are used for the task of Semantic Role Labeling. However, FrameNet also contains manual annotations for relations that connect pairs of frames. Figure 2.3 sketches the structure of FrameNet with respect to the knowledge graph of frames connected via frame-to-frame (F2F) relations: the frame *'Commerce_buy'* is in an *Inheritance*-relation with the frames *'Getting'* (*Inherits_from*) and *'Renting'* (*Is_Inherited_by*). The FrameNet hierarchy includes eight types of frame-to-frame relations of which five are inverse relations that exist in both directions between frames (e.g. *Inheritance*: *Inherits_from*, *Is_Inherited_by* or *Precedence*: *Precedes*, *Is_Preceded_by*), see Table 2.3. Table 2.3 also lists all frame-to-frame relation names with the number of frame pairs for each relation according to the FrameNet hierarchy, and also restricted counts. The restricted counts include only those frame pairs of which both frames have lexical units and thereby could potentially by evoked by predicates in texts (e.g., the frame *'Waking_up'* can be evoked by the verb *'awake'*). Thus, excluded are the 125 frames, which are used as meta-frames for abstraction purposes and do not have lexical units .

The FrameNet hierarchy lists the frame-to-frame relations to other frames for each of the overall 1,019 frames. We denote with G the collection of triples (f_1, r, f_2), where the notation stands for frame 'f_1 is in relation r to frame f_2'. The frame pair $f_1, f_2 \in F_h$ is part of the set of frames in the FrameNet hierarchy and the relation $r \in R$ is part of the set of frame-to-frame relations. As listed in Table 2.4, there are 2,912 triples in the FrameNet hierarchy with 1,913 triples remaining if considering only those where both frames have lexical units, and with 1,447 triples remaining if considering only those where both frames occur in the textual data. We split the obtained triples into a training and a test set so that the training set contains the first 70% of all the triples for each relation. Table 2.4 summarizes frame counts per data source together with counts of frame-to-frame relations where both frames occur in the underlying source.

F2F Relation Name	Total	Restricted
Inherits_from	617	383
Is_inherited_by	617	383
Uses	491	430
Is_used_by	490	430
Subframe_of	119	29
Has_subframes	117	29
Perspective_on	99	15
Is_perspectivized_in	99	15
Precedes	79	48
Is_preceded_by	79	48
Causation	48	47
See_also	41	40
Inchoative	16	16
Sum	2,912	1,913

Table 2.3: Frame-to-frame relation pair counts and restricted pair counts of frames with lexical units.

Corpus	Frames	F2F Relations
FN Hierarchy	1,019	2,912
FN Hierarchy restricted to frames with LU	894	1,913
Textual data: FN 1.5 sentences	700	1,447

Table 2.4: Counts for frames and frame-to-frame relations.

Figure 2.4 visualizes a more complex interplay of frames with several frame-to-frame relations and also points out missing annotations for relations between frames. The frame *'Waking_up'* is in a *Precedence*-relation to the frame *'Being_awake'* and both frames are subframes of *'Sleep_wake_cycle'*. Between the two frames, also as a subframe of the *'Sleep_wake_cycle'*, a frame such as *'Biological_urge'* could fit in – which can be evoked by adjectives like *'tired', 'sleepy', 'exhausted'*. The frame *'Sleep_wake_cycle'* has no lexical unit, i.e. it cannot be evoked within a text. The FrameNet hierarchy does not provide lexical units for 125 frames. In fact, such frames are used as meta-frames for abstraction purposes, thus, they exist only to participate in frame-to-frame relations with other frames (Ruppenhofer et al., 2016). In general, each frame pair is connected via only one frame-to-frame relation with occasional exceptions (Ruppenhofer et al., 2016).

2.2.2.1 Frame-to-Frame Relation Prediction

Automatic completion of frame-to-frame relations in the FrameNet hierarchy has received little attention although they incorporate meta-level commonsense knowledge and are used in downstream approaches. We address the problem of sparsely annotated frame-to-frame relations.

The task of Relation Prediction originates from automatic Knowledge Graph Completion and is known as 'Link Prediction' (Bordes et al., 2011, 2012, 2013). We will transfer this task to Frame-to-Frame Relation Prediction for frame pairs.

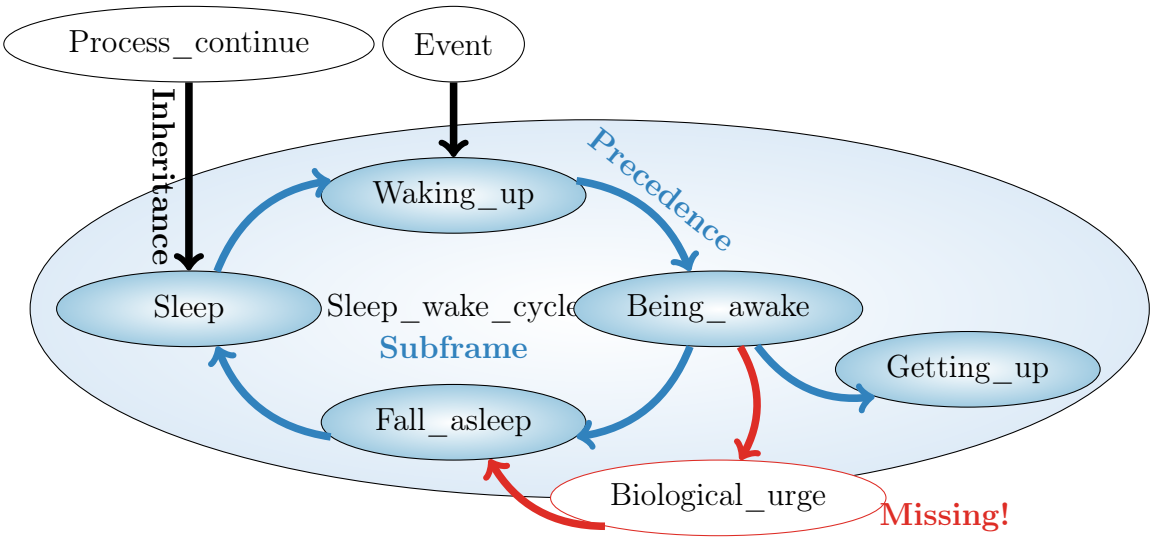


Figure 2.4: Frame-to-frame relations example. Ellipses contain frames. Frame-to-frame relations from FrameNet hierarchy: Inheritance (black arrows), Precedence (blue arrows), Subframe (in largest ellipse). Red arrows: missing annotation with Precedence relationship.

Definition of the Task. Given two FrameNet frames (f_1, f_2) and the set of FrameNet relations $r \in R$, predict the correct relation r for the given pair of frames. The task is to train and test on the existing FrameNet hierarchy, G the collection of triples (f_1, r, f_2), and to apply the best system on pairs of frames (f_1, f_2) which are not yet connected.

Importance of Frame-to-Frame Relation Prediction. Frame-to-frame relations are used in the context of other tasks, such as text understanding (Fillmore and Baker, 2001), paraphrase rule generation with the *'Perspective_on'*-relation for the system LexPar (Coyne and Rambow, 2009) or with the *'Using'*-relation (Sikos and Padó, 2018a) and recognition of textual entailment (Aharon et al., 2010). Furthermore, frame-to-frame relations can be used as a form of commonsense knowledge as they connect frames on a higher abstraction level; Rastogi and Van Durme (2014) remark that meta-level knowledge incorporated into frame-to-frame relations is of interest for intelligent systems. Rastogi and Van Durme (2014) give the example of the frames *Experience_bodily_harm* and *Hostile_encounter* which are not yet connected with the relation *Is_Causative_Of*, even this causation would be 'reasonable to expect'. This is the point where frame-to-frame relations are relevant to intelligent systems: they can help to train or to evaluate expectations or chains of reasoning about possible causations or interactions of situations or actions.

The incompleteness of the FrameNet hierarchy is a known issue not only at the frame level (Rastogi and Van Durme, 2014; Pavlick et al., 2015; Hartmann and Gurevych, 2013) but also at the F2F relation level. Figure 2.4 exemplifies a missing `precedence` relation: 'Fall_asleep' is preceded by 'Being_awake' but in-between yet another frame could be added, e.g. 'Biological_urge' (evoked by adjectives like *'tired', 'sleepy', 'exhausted'*). Rastogi and Van Durme (2014) note a lack of research on automatically enriching the frame-to-frame relations, which would be

beneficial given the large number of possible frame pairs for a relation and their use in other tasks. The automatic annotation of frame-to-frame relations involves three difficulties accounted for by the nature of FrameNet. First, frame-to-frame relation annotations occur sparsely and for the majority of pairs in each relation there are few instances (see Table 2.3). Second, the relations themselves have no direct lexical correspondences in text and hence inferring them from text is not trivial. Third, if a relation involves a frame that does not have any lexical unit (see restricted counts in Table 2.3), this frame does not occur in text and hence inferring this relation from text is even more difficult.

2.3 Grounded Language Understanding

Humans are in constant interaction with their environment and constantly enrich their experience. From this perspective, knowledge of meaning can be intertwined with the experience of certain aspects of meaning. This kind of knowledge of meaning is grounded in experience from multiple modalities (Section 2.3.1), and computational approaches to model the information flow between multiple modalities traditionally orient towards theories in Cognitive Science (Section 2.3.2).

2.3.1 Meaning via Experience

The importance of context for determining the meaning of a word or even longer expressions was explained in Section 2.1.1. Whilst Section 2.1.1 refers to 'context' as 'context words', here, the notion of 'context' will be extended to 'context in the world of experience'. Whilst Section 2.1.1 argues for 'the meaning of a word being determined by the context the word it appears in', here, it is extended to 'the meaning of a word being determined by the context of impressions it was experienced in'. This perspective originates from psycholinguistics.

Psycholinguistics. For humans, language understanding or text understanding incorporates different levels of experience and therefore involves many modalities when interpreting a text within its situational context. In psycholinguistics, or psychology of language, there is evidence for humans understanding scenes and also texts via mental simulations (see Barsalou, 1999; Zwaan and Madden, 2005). From the perspective of psycholinguistics, Fillmore's frames (Fillmore, 2012, as introduced in Section 2.1.2) are regarded as an approach to 'capture the structure of situations' (Barsalou, 2008) in the context of amodal symbol systems (Barsalou, 1999). Referring back to the explanatory example in Chapter 1, humans understand words when knowing what they refer to in their world of experiences (e.g., understanding of the entity 'dog' is grounded in the visual, acoustic and haptic modality) and they understand descriptions of whole situations when knowing again what they refer to in their world of experiences (e.g., understanding of the expression 'running after a barking dog' is grounded in multiple modalities).

> *'As people comprehend a text, they construct simulations to represent its perceptual, motor, and affective content.'* (Barsalou, 2008)

This states that the meaning of a word can be inferred from the situational context it appears in in the world of experiences. It motivates regarding word meaning as experienced impressions from different sensory modalities. In the field of Natural Language Processing, this perspective is implemented in multimodal approaches for embedding learning that integrate information from multiple modalities, which will be described in Section 3.5. The term 'multimodal' has been used in a broad range of different interpretations. In the common interpretation, modalities refer to sensory input in humans, such as audio, vision, touch, smell, and taste. Further definitions expand to different communicative channels such as language and gesture, or simply different 'modes' of the same modality (e.g., day and night pictures). Grounding in (human) modalities has different foci in Natural Language Processing so far; Beinborn et al. (2018) partition the foci into concepts, phrases or whole sentences – which are reviewed in the following based on the survey by Beinborn et al. (2018).

Grounding Concepts. Modeling semantic relations between concepts is foundational to process language and to generalize from known concepts to new ones. Beinborn et al. (2018) review literature about the grounding of concepts in the field of multimodal Natural Language Processing, of which the relevant parts to this thesis are summarized in the following.

The quality of concept representations, multimodal as well as unimodal, is commonly evaluated by their ability to model semantic properties as for example relations between concepts. The similarity-relation is a basic but still challenging semantic property to be modeled: there are several similarity datasets to compare the performance of uni- and multimodal approaches to learning concept representations, e.g., *WordSim353* (Finkelstein et al., 2002), *SimLex-999* (Hill et al., 2015), *MEN* (Bruni et al., 2012), *SemSim* and *VisSim* (Silberer and Lapata, 2014)). These datasets contain pairs of words that have been annotated with similarity scores for the two concepts, e.g., journey and voyage are rated by humans as highly similar, whereas professor and cucumber are rated as highly different – according to *WordSim353*. The similarity is easy to judge by humans, however when only using words to describe the difference it would take longer than simply looking at the two images.

Grounding in perception motivates and requires multimodal concept representations. So far, research and corpus creation has mostly focused on combining the textual and the visual modality to ground concept representations. Still, for dedicated tasks, perceptual information from further modalities have also been explored, e.g., the auditory (Kiela and Clark, 2015, 2017) and olfactory channel (Kiela et al., 2015).

Multimodal (textual plus visual) concept representations are found to outperform unimodal ones in modeling semantic similarity by evaluation studies of semantic models (Feng and Lapata, 2010; Silberer and Lapata, 2012; Bruni et al., 2014; Kiela et al., 2014) and by comparative studies of image sources and architectures (Kiela et al., 2016).

However, it remains an open question whether multimodal concept representations contribute to an approximation of human conceptual grounding that is cognitively more plausible. Contradictory results by Bulat et al. (2017a) and Anderson et al. (2017) demonstrate the openness and difficulty of this question: both experiment with human brain activity scans of the perception of concepts and compare

different distributional models for concept similarity but on the one hand, Bulat et al. (2017a) observe visual information as beneficial for modeling concrete concepts, whereas on the other hand, Anderson et al. (2017) conclude that textual models sufficiently integrate visual properties.

Over a broad range of concept-related tasks, if there are multimodal studies, multimodal approaches seem to be advantageous: multimodal information was successfully integrated for the disambiguation of concepts (Xie et al., 2017) and named entities (Moon et al., 2018).

Imaginability of Abstract Concepts. Conceptual grounding of concrete words is straight-forward as they have direct reference in sensory experience (e.g., 'cup' has an obvious visual correspondent). Building multimodal representations for abstract concepts is more challenging due to the lack of perceptual patterns associated with abstract words (Hill et al., 2014). In the same line, Bruni et al. (2014) and Hill and Korhonen (2014) find that multimodal representations are helpful for modeling concrete words, but have little to no impact when evaluating abstract words.

Unseen concepts can be modeled in multimodal space when projected into the representation space based on their textual relations to seen concepts. However, it is questionable whether the information about the textual relation is sufficient to infer relations between abstract concepts in multimodal space. Lazaridou et al. (2015) analyze projected abstract concepts and confirm that concrete objects are more likely to be captured adequately by multimodal representations. Still, they also find illustrating examples of situations or objects which represent abstract words surprisingly well (e.g., *freedom* can be associated with an image of a revolution-scene or *theory* can be associated with an image of a bookshelf full of books, lexica and papers).

Grounding Phrases: Abstract versus Concrete. In order to ground phrases, adding the meaning of abstract concepts to that of concrete ones is essential. The most straight-forward approach to compose phrases is the extension of concepts (nouns) by adjectives (adjective plus nouns). In the following, we summarize the relevant parts to this thesis of the literature review by Beinborn et al. (2018) with respect to the grounding of phrases.

With respect to the compositional meaning of adjective-noun combinations in terms of color adjectives, Bruni et al. (2012) find multimodal representations to be superior compared to unimodal ones, but difficulties remain regarding literal versus non-literal usage of color adjectives (e.g., *green/black cup* versus *green/black future*). Furthermore, Winn and Muresan (2018) propose to ground comparative adjectives describing colors in RGB space where their approach is able to model unseen colors and comparatives.

Concerning figurative language, Lakoff and Johnson (1980) argue that abstract concepts can be grounded metaphorically in embodied and situated knowledge. Lakoff and Johnson's theory of metaphor assumes metaphors to be a mapping from a concrete source domain to a more abstract target domain (e.g., *future* can be viewed *as a place in front of us, which we are approaching or which is flowing towards us*).

Turney et al. (2011) implement the theory of metaphor by leveraging the discrepancy in concreteness of source and target term to identifying metaphoric phrases. This approach is in turn applied to adjective-noun combinations: Shutova et al. (2016) and Bulat et al. (2017b) use multimodal models for identifying metaphoric word usage in combinations of adjective plus noun. Their models show that adjectives and nouns used in a metaphorical sense (*dry wit*) are less similar than words in literal phrases (*dry skin*).

Taken together, this indicates that multimodal compositional grounding is crucial for a more holistic understanding of figurative language processing.

Grounding Sentences. Finally, for the grounding of sentences, we summarize the relevant parts to this thesis of the literature review by Beinborn et al. (2018).

Multimodal representations of sequences or sentences are crucial when grounding descriptions of actions; following this, Regneri et al. (2013) ground descriptions of actions in videos showing these actions. Studies that require sentence representations go even further in terms of sequence length. Shutova et al. (2017) find promising tendencies regarding the use of multimodal information for the disambiguation of sentences. Still, the underlying compositional principles of combining multiple modalities for sentence comprehension are yet to be understood. Furthermore, interdisciplinary research is required to obtain a deeper understanding of cognitively plausible language processing (Embick and Poeppel, 2015).

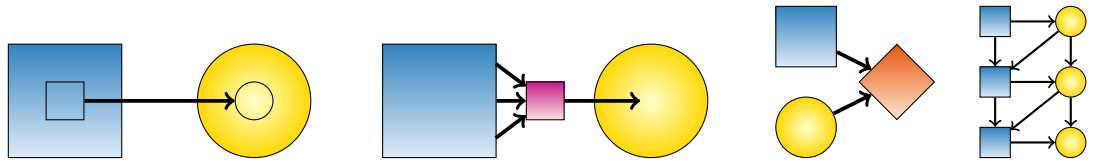
2.3.2 Multimodal Information Flow

A foundational design question in multimodal tasks is about how to exchange information between modalities, this is also called information flow. In the following, we explain the same distinctions within models of multimodal information flow as published in our survey (Beinborn et al., 2018).[1] In this survey, we propose a classification of multimodal tasks with respect to the information flow between modalities into cross-modal transfer, cross-modal interpretation, and joint multimodal processing – which are explained along Figure 2.5, together with typical tasks requiring this kind of information flow in Figure 2.6. From a historical perspective, progress in multimodal processing can be aligned with cognitive theories of multimodal organization in the human brain. We exemplify different approaches for multimodal information flow with the two modalities of text and images, whilst a broad overview of the challenges and machine learning methods for multimodal information processing can be found in Baltrušaitis et al. (2018).

2.3.2.1 Cross-modal Transfer

The information flow in the category of cross-modal transfer is inspired by the theory of the modularity of mind (Fodor, 1988) and information from different modalities is processed independently.

[1] The distinction within models of multimodal information flow is a joint contribution of myself together with my co-author Lisa Beinborn, for which we refer to our survey (Beinborn et al., 2018).

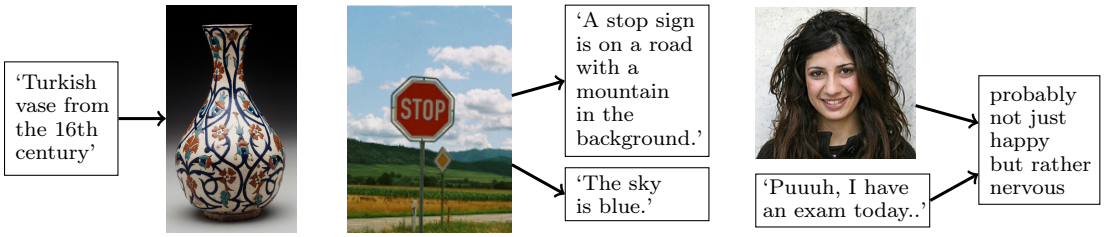


(a) Cross-modal transfer. Information from modality A is aligned to comparable information in B.

(b) Cross-modal interpretation. Relevant information in modality A is summarized and interpreted in modality B.

(c) Joint multimodal processing. Left: Modality A and B both contribute to a joint prediction. Right: Interactive exchange of information between modalities.

Figure 2.5: Information flow in multimodal tasks. Blue and yellow shapes refer to modality A and B.



(a) Cross-modal transfer: text-based image retrieval.

(b) Cross-modal interpretation: image captioning.

(c) Joint multimodal processing: emotion recognition.

Figure 2.6: Examples for typical tasks requiring different kinds of information flow. The examples use textual and visual modalities, however the ideas are not restricted to these two specific modalities.

Modularity of Mind – Independent Modalities. The theory of the modularity of mind about how the human brain processes information (Fodor, 1988) assumes domain-specific encapsulated modules that do not interact with each other. Regarding language processing, this view corresponds to mental models of a language hub in the brain that does not directly incorporate perceptual information (Chomsky, 1986).

The modular perspective is adopted in earlier computational approaches to multimodal processing: information is processed in each modality independently and the final outcome is obtained via transfer or alignment to another modality. Information from the output modality is mainly used for the reranking of the input hypotheses. The category of cross-modal transfer is to group tasks in which one modality serves as the interface to query or to represent the content from another modality, see Figure 2.5a.

Examples of Cross-modal Transfer. Typical examples for cross-modal transfer are search and retrieval tasks. Figure 2.6a shows an example for a typical task requiring cross-modal transfer: text-based image retrieval, where given the textual input query of a specific vase, the output shall be returned in the visual modality as an image of exactly this specific vase. This means, the human user provides a natural language description to query an artefact from a database, i.e. an image,

video, or audio file (Atrey et al., 2010). Socher et al. (2014) introduce a neural network model which uses dependency trees of sentences sentences to build embeddings that facilitate to retrieve corresponding images. The query and the output are from different modalities, which requires cross-modal alignment. Besides text-based image retrieval tasks, there are further transfer tasks: speech-related transfer tasks that map texts to audio for users who cannot read (Zen et al., 2009) or also, the other way round, the transcription of audio and video content in speech recognition (Juang and Rabiner, 2005), in subtitle generation (Daelemans et al., 2004) and in lipreading tasks (Ngiam et al., 2011).

Taken together, in cross-modal transfer tasks, information is processed synchronously and in parallel in several modalities without directly influencing each other. The main challenge lies in finding appropriate translations or alignments from one modality to the other.

2.3.2.2 Cross-modal Interpretation

The information flow in the category of cross-modal interpretation is inspired by the concept of attention to mediate between modalities.

Attention for Selection – Mediator between Modalities. The human brain perceives much information from different channels simultaneously but seems to have a successful strategy to prevent information overload. The concept of attention is well recognized for explaining how humans can select relevant information from perceptual input: Bridewell and Bello (2016) argue that attention serves as 'a bottleneck for information flow in a cognitive system' that redirects mental resources.

In computational processing in the sense of cross-modal interpretation, the concept of attention as a mediator between modalities is important to extract the relevant information, see Figure 2.5b. In order to identify relevant information, attention mechanisms (Bahdanau et al., 2014) are used. This means, first, a compressed and structured intermediate representation is created out of the relevant parts of the input in order to, then, generate an interpretation in the target modality out of the intermediate step.

Examples of Cross-modal Interpretation. Figure 2.6b shows an example for a typical task requiring cross-modal interpretation: image captioning (Xu et al., 2015), where given an image of a stop sign, different textual 'summaries' or 'interpretations' can be produced depending on the decision of what is relevant. Thus, this is not a one-to-one mapping but rather an interpretation. The same holds for sketch recognition (Li et al., 2015). Relevant elements need to be identified, individual elements need to be grouped to semantic concepts with relations between concepts, and finally, all the relations between elements need to be output in natural language while paying attention to different salient areas in the image. Besides image captioning tasks, there are also reverse 'summarization' or 'interpretation' tasks where visual representations need to be generated to summarize documents and present the most relevant information in an intuitive way (Kucher and Kerren, 2015). Other approaches include semantic relations between words for a more conceptual-driven

interpretation (Xu et al., 2016). Concept maps highlight structural relations between concepts in a graph-based visualization (Zubrinic et al., 2012).

Taken together, in cross-modal interpretation tasks, information from one modality is selectively interpreted in another modality. One key challenge lies in the evaluation of the output: for interpretations – which are by definition subjective – several divergent solutions can be equally valid. Accumulations over various human ratings are currently considered to be better quality approximations than any automatic metrics (Vedantam et al., 2015).

2.3.2.3 Joint Multimodal Processing

The information flow in the category of joint multimodal processing is inspired by the cognitive theory of embodied processing (see Section 2.3.1) and information from different modalities is combined during processing.

Embodiment – Combining Modalities. A broad range of experiments in psycholinguistics provides evidence for the cognitive theory of embodied processing and, consequently, the separation of different modalities has become blurred (Pulvermüller et al., 2005). A similar development can be observed in multimodal machine learning and in line with this, the category of joint multimodal processing is to group tasks which explicitly require the combination of knowledge from different modalities, see Figure 2.5c.

Examples of Joint Multimodal Processing. Figure 2.6c shows an example for a typical task requiring joint multimodal processing: emotion recognition, where the input is given in two different modalities with an image of a smiling girl and a text snippet of what she is saying, and the output shall be an estimate about her emotions. This means, a complex interplay between facial expression and sentiment of text (and more) has to be solved – an ironic tone of voice might influence and reverse the perception and interpretation of the language content. Emotion recognition (Morency et al., 2011) as well as persuasiveness prediction (Santos et al., 2016) require to evaluate jointly the textual and paraverbal cues (e.g., pitch, facial expression). Going even further, visual question answering is designed to require an interactive flow of information as a human user can ask questions about an image that the system should answer (Malinowski et al., 2015). This not only requires understanding the question and determining the relevant elements in the image, but also interpreting the image with respect to the question and generating a coherent natural language answer that matches the question. Hence, exchange of information between the modalities is crucial. In an overview, Wu et al. (2017b) compare 29 approaches to visual question answering where 23 of these use a joint representation of textual and visual information. (The remaining 6 approaches organize the exchange either through a coordinated network architecture or through shared external knowledge bases.) Novel interactive approaches enable direct modulation of the information flow in one modality by input from another modality (de Vries et al., 2017) or by human feedback (Ling and Fidler, 2017).

Taken together, in joint multimodal processing tasks, information is processed

while enabling an exchange between modalities on the fly. The multimodal approaches in this thesis, follow the category of joint multimodal processing.

In Section 3, we discuss different methods for obtaining joint representations computationally.

2.4 Summary of the Chapter

This chapter consolidated three different facets of language understanding – semantic, structured and grounded (Sections 2.1, 2.2 and 2.3, respectively) – in the literature of their respective research fields. Regarding NLP-tasks, the semantic and structured viewpoints traditionally apply for different kinds of tasks, whilst the grounded aspect motivates the combination of different information-channels for any complex task. We reviewed how success in grounded language processing in terms of grounding concepts, phrases and sequences. We analyzed how multimodal processing has developed from transfer between encapsulated modalities to interactive processing over joint multimodal representations – and showcased typical multimodal tasks.

We provide a summary of this chapter in terms of bullet points in the following box (see next page):

FOUNDATIONS IN A NUTSHELL

In **textual** semantic language understanding, the meaning of words is determined by the textual context around the word.

- Frame semantics focuses on situations and actions, and interactions between them
- FrameNet implements frame semantics
- Frame Identification aims at labeling situations and actions:
  Given the sentence *'Abby bought a car.'* and the predicate *'bought'*,
  the frame *'Commerce_buy'* is evoked.

Structured language understanding represents the meaning of words in terms of relation to other words in knowledge bases.

- Knowledge bases organize information in terms of relational triples
- The task of knowledge base Completion infers new relational triples
- FrameNet is a knowledge base: relations between frames

→ We introduce the task of Frame-to-Frame Relation Prediction:
Given the frame pair *(Commerce_buy, Getting)*
the correct frame-to-frame relation is *'Inherits_from'*.

Grounded language understanding assumes that, for humans, the meaning of words is interlinked to the experience of these words.

- Experience refers to a multimodal end-product of sensory perception
- Multimodal strategies ground concepts, phrases and sequences

→ We categorize the multimodal information flow:
Cross-modal transfer *(e.g., text-based image retrieval)*
Cross-modal interpretation *(e.g., image captioning)*
Joint multimodal processing *(e.g., emotion recognition)*

Chapter 3

Methods for Learning Meaning Representations

This chapter provides a methodological overview of representation learning. We are interested in representations for categories of meaning such as words, frames, or images; and in using such representations across different semantic tasks in Natural Language Processing. Meaning representations model different aspects of meaning in a coherent vector space, respectively. After we define a vector space model and introduce our notation, we give some background on neural networks (Section 3.1). Following, we review methods for textual (Section 3.2), structured (Section 3.3), visual (Section 3.4), and multimodal (Section 3.5) embedding learning. Different viewpoints orient towards different tasks for which in turn different methods for learning embeddings have evolved. Thus, this chapter provides the methodological foundation of this thesis with respect to representation learning.

Definition of Vector Space Models. We denote the vector space model *vsm* (Equation 3.1) that defines a mapping from each concept c (for example a word) to an m-dimensional vector $\overrightarrow{v_{(c)}}$ in the following way:

$$vsm(c) = \overrightarrow{v_{(c)}} \text{ , with } \overrightarrow{v_{(c)}} = \left[v_{(c)1}\ v_{(c)2}\ \ldots\ v_{(c)m}\right]. \tag{3.1}$$

Lowe (2001) defines the characteristics of a semantic vector space model as a quadruple $\langle B, A, S, M\rangle$, where:

B is the basis of the semantic model in form of a collection of source documents to learn the vectors from, e.g., for every word the co-occurrence with every other word can be counted.

A defines a lexical association function that converts co-occurrence counts to association weights, e.g., normalization.

S defines a similarity metric to measure the distance between pairs of vectors, e.g., cosine similarity as explained below.

M is the model that actually transforms the vector space, e.g., by reducing the dimensionality.

The most common measure for similarity in vector space models is the cosine distance d_{cos} (Equation 3.2), which is the complement of cosine similarity s_{cos} (Equation 3.3):

$$d_{cos}(\overrightarrow{v_{(c_A)}}, \overrightarrow{v_{(c_B)}}) := 1 - s_{cos}(\overrightarrow{v_{(c_A)}}, \overrightarrow{v_{(c_B)}}) \; . \tag{3.2}$$

Cosine similarity calculates the cosine of the angle between two vectors $\overrightarrow{v_{(c_A)}}$ and $\overrightarrow{v_{(c_B)}}$, i.e. vectors pointing in similar directions have a high (near 1) similarity and a low (near 0) distance:

$$s_{cos}(\overrightarrow{v_{(c_A)}}, \overrightarrow{v_{(c_B)}}) := \frac{\sum_{i=1}^{n} \overrightarrow{v_{(c_A)i}} \; \overrightarrow{v_{(c_B)i}}}{||\overrightarrow{v_{(c_A)}}|| \cdot ||\overrightarrow{v_{(c_B)}}||} \text{ , with } ||\overrightarrow{v_{(c)}}|| = \sqrt{\sum_{i=1}^{n} v_{(c)i}^2} \; . \tag{3.3}$$

Meaning representations are known in the literature with different names, even if they all refer to the vectors $\overrightarrow{v_{(c)}}$ for concepts c obtained by the vector space model *vsm*:

- 'vector representations' – a certain aspect of meaning corresponds to a certain vector in the vector space.
- 'dense vector representations' – dense representations are continuous vectors of reduced dimensionality. In contrast, for 'one-hot' representations as well as for co-occurrence counts, the dimensionality is the size of the vocabulary: every word is assigned to a unique combination of a single 1 amongst 0s, or, for every word, each dimension notes the co-occurrence with one other word. Dense vector representations, however, reduce the dimensionality by learning continuous vectors – this refers to M in the quadruple $\langle B, A, S, M \rangle$ by Lowe (2001).
- 'distributed representations' – the distribution of different aspects of meaning from the source documents (B in $\langle B, A, S, M \rangle$ by Lowe (2001)) is modeled in the vector space. In case of 'distributional representations', this goes back to the 'distributional hypothesis' (Harris, 1954) about similar meaning being manifested by occurrence in similar textual contexts. However, 'distributed representations' are not restricted to modeling the 'distributional hypothesis', but could also model yet other aspects of meaning, e.g., for source documents B not being texts but knowledge bases.
- 'embeddings' – the representations 'embed' certain aspects of meaning from the source documents in the vector space.

We adopt the formulation 'embeddings' and 'embedding space'; textual embeddings are also known as 'word embeddings', similar to visual embeddings that are also known as 'image embeddings'.

In NLP-applications, the embedding space model *vsm* is used as a simple word lookup function to get the embedding for a word and then process it for a downstream application. Before neural networks were broadly used, embeddings for words were computed by counting co-occurrences of words (see survey for the pre-neural era of word embeddings, Erk, 2012). Current research on NLP-applications uses pretrained embedding space models that have learned to optimize a training objective,

often implemented with neural network architectures. Common embedding learning approaches implement two underlying ideas in their training objectives (related to S in $\langle B, A, S, M \rangle$ by Lowe (2001)):

distance in the embedding space: the vectors of similar concepts (e.g., words or images) are close to each other in the vector space whilst dissimilar concepts are far apart from each other.

direction in the embedding space: the vectors of two concepts that fulfill a certain relation (e.g., singular and plural form of a word, usual and upside-down version of an image) point into a certain direction.

In the following sections, several training objectives will be explained with the respective embedding learning approaches, within the tasks they optimize training.

3.1 Foundation – Background on Neural Networks

Neural networks (NN, common abbreviation) are a powerful approach to representation learning. When optimizing a training objective, they learn internal representations, which can be extracted afterwards in order to be applied in the context of further tasks. Thus, with such extracted internal representations, some independence of hand-crafted features is gained.

For the typical case of supervised learning, example instances of input data and output labels are necessary to optimize a training objective and to learn the relationship from input to output. After training, the neural network should be able to predict correct labels for novel input instances. In this section, we review the background on neural networks and their training procedures.

Architecture of Neural Networks. Intuitively, neural networks build on computational models for information processing of biological neurons (McCulloch and Pitts, 1943; Rosenblatt, 1958) and also the term 'neural networks' originally stems from information processing of neurons in the brain. McCulloch and Pitts (1943) present a model of a neuron: the activation of a biological neuron is computed from the input signals it receives from predecessor neurons then yielding an internal state, which is confronted with a threshold, which finally results in either an activation or non-activation as output. The computation of the internal state follows a weighted sum over the input signals.

The perceptron learning algorithm (Rosenblatt, 1958) fits a binary logistic regression model to estimate the weights which are crucial to determine the internal state. In the notation of MacKay (2003) (cf. Equations 3.4 and 3.5), the outputs y are obtained by applying a non-linear activation function f over the neurons' activations a:

$$t \approx y = f(a(x, w, b)) \; . \tag{3.4}$$

The aim is to approximate the intended targets t by the outputs y. The inputs x are confronted with learnable weights w and biases b to compute the activation a of every neuron:

$$a(x, w, b) = \sum_i (w_i * x_i + b) \; . \tag{3.5}$$

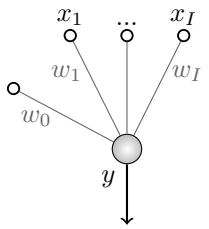


Figure 3.1: Model for a neuron. The neuron (central circle) receives several input signals x_1 to x_I, which are weighted by the weights w_1 to w_I and also a bias term w_0 is added. Finally, the neuron outputs the activation y.

The model in Figure 3.1 visualizes a single neuron computing the output activation out of input signals to which weights and a bias term are applied, corresponding to Equations 3.4 and 3.5.

Crucially, the non-linearity in the activation function f is the key to the success of neural networks as it allows for learning complex feature combinations that linear classifiers could not solve. Typical choices for non-linear activation functions are the sigmoid function (Equation 3.6):

$$sig(x) = \frac{1}{1 + \exp(-x)} \,, \tag{3.6}$$

the rectified linear unit (Nair and Hinton, 2010, ReLU, common abbreviation) (Equation 3.7):

$$ReLU(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \,, \tag{3.7}$$

or the hyperbolic tangent (Equation 3.8):

$$tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \,. \tag{3.8}$$

Several layers of neurons can be connected: the output of the predecessor layer is the input to the successor layer and the connections carry the weights. With respect to the depth of the stack, the neural net is called a single layer perceptron, multilayer perceptron or deep neural network.

The sketch in Figure 3.2 visualizes a hidden layer neural network (with one hidden layer) where the neurons of predecessor and successor layer are fully connected, orienting on the notation of Hastie et al. (2009). As can be seen in Figure 3.2, the last layer with the final activations contains several neurons, thus several final activations. In classification setups, where the final prediction should be a decision for one class, yet another layer can be stacked on top. This final layer is to apply a softmax-classifier (Equation 3.9) as in the following:

$$softmax(y)_j = \frac{\exp(y_j)}{\sum_k (\exp(y_k))} \,, \tag{3.9}$$

in order to map the activations to values in a range from 0 to 1 in a way so that they sum up to 1. By this, the final outputs express probabilities for specific classes and the class with the highest probability can be selected for prediction.

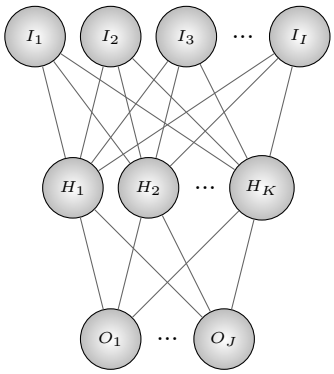


Figure 3.2: Model for a hidden layer neural network. The first row of neurons is the input layer (I), followed by one hidden layer (H), which represents the learned non-linear combination of the input data. More hidden layers can be added. The last layer is the output layer (O) (last row), which contains the final activations. All layers are fully connected.

Backpropagation. A 'forward pass' in a neural network means to traverse it once from input to output, where in the beginning the weights are initialized randomly. Thus, the first output, after the first forward pass, rarely succeeds to closely approximate the intended targets.

In order to improve the prediction, the internal weights and biases need to be adapted towards the training targets (also called 'learning'), which is implemented with stochastic gradient descent by the back propagation algorithm (Rumelhart et al., 1986). Given the actual outputs y and the intended targets t, an error function (or 'cost function') captures the difference. The mean squared error (Equation 3.10) can be used as error function:

$$E(x, W, b) = 1/2 \sum_i (y_i - t_i)^2 \ . \tag{3.10}$$

In the context of backpropagation, the parameter updates (Equations 3.11 and 3.12) for the weights and the biases are computed:

$$w_{(t+1)} = w_{(t)} - \gamma \frac{\partial}{\partial w_{(t)}} E(x, W, b) \ , \tag{3.11}$$

$$b_{(t+1)} = b_{(t)} - \gamma \frac{\partial}{\partial b_{(t)}} E(x, W, b) \ ; \tag{3.12}$$

in order to minimize the error with better values for the parameters. The error function is partially derived with respect to every learnable parameter. Figuratively, the derivative of the error function is used to point towards a minimum in the error space, where the weights and biases get updated following the negative gradient. Updating the weights and biases from output to input is a 'backward pass'.

In one training 'epoch', the computation of stochastic gradient descent passed over the entire data once, thus the error function decreases in every epoch and finally converges to a local minimum. The learning rate γ is to control that the process does not end in a poor local minimum but eventually a better local optimum is found. This rate adapts the size of the update towards the negative gradient, e.g.,

a larger rate in the beginning transforming into a smaller rate once a promising are is reached.

Regularization. Approaches on regularization are to prevent 'overfitting'. A scenario where a neural network fits the training examples perfectly but is not able at all to generalize to the unseen test instances is called 'overfitting'. In such a scenario the learnable parameters are not found in the intended way. To overcome this, regularization approaches penalize for overfitting too much on a development set.

LSTMs. When applying neural methods to textual data, multi-layer perceptrons (MLP, common abbreviation) and long-short term memories (Hochreiter and Schmidhuber, 1997, LSTM, common abbreviation) are most commonly used. Importantly, in LSTMs, special input-, output- and forget-gates are integrated into each neuron to enable the memorization of important information until the end of an input sequence. By this, long-short term memory networks are recommended for dealing with long-distance relationships in textual data.

CNNs. Neural methods applied to images differ from those applied to texts: for images convolutional neural networks (CNN, common abbreviation) are standard.

Convolutional neural networks are designed for image processing specifically: other than the basic neural network architecture, they use a filter mask, also called *kernel*, that is slid over groups of pixels of regions of an input image. The kernel computes a local representation of the corresponding input area and, furthermore, shares the weights across the layer, which reduces the number of parameters. Importantly, by this, the aim of translation invariance is followed to take into account that patterns should be recognized independent of their position in the image.

Convolutional neural networks have a multi-layer structure with several layers of non-linear feature extractors for recognizing visual patterns directly from pixel images. The large number of parameters within these layers is learned from data, for example on the task of image classification or for digit recognition (see special CNN-architecture 'LeNet-5' for hand-written digits, LeCun et al., 1998).

3.2 Textual Embeddings

Textual embeddings for words can be learned by two different lines of methods: neural networks and matrix factorization. These methods were developed independently of each other and they formulate their training objective via different tasks. For their respective training objectives, neural methods for textual embedding learning focus on the prediction of context words, whereas the matrix factorization methods that we will use in our work focus on the prediction of coarser-grained categories (e.g., frame labels).

Neural Methods

We will review some common neural approaches such as `Word2Vec` (Mikolov et al., 2013a) and `GloVe` (Pennington et al., 2014) for textual embedding learning, which we will apply throughout the thesis. We start with explaining the underlying task which is optimized during training.

Task: Prediction of Word or Context. Textual embedding learning methods are usually regarded as 'unsupervised' learning methods as no dataset needs to be labeled for training. The setup for 'unsupervised' textual embedding learning is special to Natural Language Processing, whereas in the field of Computer Vision, almost all embeddings arise from supervised learning (see Section 3.4 for details on visual embedding learning). However, also in textual embedding learning, there is still an underlying task – which indeed does not require the creation of a labeled dataset for training – but which still provides supervision in the sense of providing a training objective. The training data are freely available texts, such as Wikipedia articles or news paper articles. The task is either predicting a target word given its context words:

> Given context: *'dogs sometimes _____ too loudly'*
> ⇒ predict word *'bark'*,

or predicting context words given their target word:

> Given word: *'_____ _____ bark _____ _____'*
> ⇒ predict context *'dogs', 'sometimes', 'too', 'loudly'*.

This means, out of the texts, a 'corpus' is created to map every word to its context words as training instances. After training embeddings for words, the embeddings mirror the similarities of words in the source texts: embeddings of synonym words will be similar to each other (e.g., close in embeddings space) as synonyms appear in similar contexts, for example:

> Given context: *'dogs sometimes _____ too loudly'*
> ⇒ not only *'bark'* can appear in text, but also *'bay'*, *'snarl'* or *'yelp'*.

There are different methods for optimizing the training objective. In the following, we will review the approaches that we will apply throughout the thesis.

Common Approaches. In the following, we review several common approaches to learning textual embeddings with neural methods, all pursuing the task of predicting a word given the context (or predicting the context given a word). The presented approaches are: `Word2Vec` for words and sequences, `dependency`-based embeddings and `GloVe`.

Word2Vec. The neural network (NN, common abbreviation) architecture of the `Word2Vec` approach (Mikolov et al., 2013a) learns word embeddings by either predicting a target word given c context words before and after the word (`CBOW`) or by predicting the c context words given their target word (`skip-gram`). The acronyms

are common abbreviations for the following: `Word2Vec` for word-to-vector, `CBOW` for continuous Bag-of-Words model, `skip-gram` for continuous Skip-gram model. Both the continuous Bag-of-Words model and the continuous Skip-gram model are log-linear classifiers with the training objective to optimize the prediction of either the current word or the context words.

Both models implement the following neural network architecture: after the word-level input layer and the projection layer, the network further comprises a hidden layer and an output layer. At the input layer, each word of the c context words (`CBOW`) or just the current word (`skip-gram`) is encoded with a 'one-hot' representation where the dimensionality is the size of the vocabulary v and the position of the single 1 amongst 0s encodes the word. The hidden layer H maps the input layer of both models to same reduced dimensionality d using a projection matrix of size $v \times d$. The continuous Bag-of-Words model averages all context words and projects them into the same position, which is a continuous distributed dense representation of the context. With the continuous Skip-gram model, the current word is directly processed by the continuous projection layer to obtain a dense representation. Finally, the output layer consists of v neurons where it incorporates the log-linear classifier by using the softmax function to compute a probability distribution over all the words in the vocabulary (cf. Section 3.1); here the weight matrix is of size $d \times v$. By adjusting the size d of the hidden layer H (the one which is before the softmax layer) different sizes of embeddings can be obtained.

Mikolov et al. (2013a) provide[1] different sizes of embeddings, amongst which the most commonly used are 50-, 100-, and 300-dimensional, that have been pre-trained on a part of the Google News dataset (about 100 billion tokens), yielding a vocabulary size of 3,000,000 words. Reimers et al. (2014) provide[2] 100-dimensional embeddings for German. For training, they applied the `Word2Vec` approach to the German Wikipedia, and to additional German newswire text to cover more domains, yielding a vocabulary size of 3,363,088 words. However, the `Word2Vec` approach can be applied to texts of any domain and language.

Dependency-based Embeddings. Levy and Goldberg (2014a) extend the Skip-gram model in `Word2Vec` to not only word contexts but specifically dependency-based contexts. In comparison to the original `skip-gram` embeddings, these `dependency`-based embeddings incorporate more of the functional similarity and less of the topical similarity. Thus, by using `dependency`-based embeddings, additional part-of-speech features as extension to word embeddings can be avoided.

Levy and Goldberg (2014a) provide[3] 300-dimensional `dependency`-based embeddings, pre-trained on English Wikipedia, yielding a vocabulary size of 174,015 words.

GloVe. Pennington et al. (2014) introduce the method `GloVe` (as abbreviation for global vectors) to obtain global vectors for word representation. The global log-bilinear regression model is a mixture of global matrix factorization models and of

[1] pre-trained `Word2Vec` embeddings: `https://code.google.com/archive/p/word2vec/`

[2] pre-trained `Word2Vec` embeddings for German, `Reimers` embeddings: `https://www.informatik.tu-darmstadt.de/ukp/research_6/ukp_in_challenges/germeval_2014/`

[3] pre-trained `dependency`-based embeddings: `https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/`

local context window models (as the Skip-gram model in `Word2Vec`).

Pennington et al. (2014) provide[4] 300-dimensional `GloVe` embeddings that have been pre-trained on the English Wikipedia and on additional English newswire text (about 6 billion tokens), yielding a vocabulary size of 400,000 words. In addition to this most commonly used set of `GloVe` embeddings, they also provide a larger set that has been pre-trained on Common Crawl texts (about 840 billion tokens), yielding a vocabulary size of 2,200,000 words.

Tasks for Evaluation of Textual Word Embeddings. There are further tasks specifically designed for the evaluation of word embeddings (Mikolov et al., 2013b). These tasks shall check whether the initial training objective is actually met in the embedding space. They are formulated as analogy tasks about syntax or semantics, addressing the underlying ideas of distance and direction in embedding spaces:

> Analogy task: *'a is to b as c is to _ _ _ '*
> Example: *'man is to woman as king is to _ _ _ '*
> ⇒ predict the correct d , e.g., *'queen'*.

Mikolov et al. (2013b) provide word pairs incorporating such syntactic or semantic relations as a corpus to perform the analogy task on:

> Corpus of word pairs for analogy task: *(a, b)* and *(c, d)* ⇒ relation :
> e.g., *(man, woman)* and *(king, queen)* ⇒ *male-female*-relation
> e.g., *(Germany, Berlin)* and *(France, Paris)* ⇒ *country-capital*-relation.

Mikolov et al. (2013b) suggest a vector offset method to solve analogy tasks. As in Equation 3.13, it calculates the offset $\overrightarrow{o_{(c_A,c_B)}}$ between two vectors $\overrightarrow{v_{c_A}}$ and $\overrightarrow{v_{c_B}}$:

$$\mathit{offset}(\overrightarrow{v_{c_A}}, \overrightarrow{v_{c_B}}) = \overrightarrow{o_{(c_A,c_B)}} = \overrightarrow{v_{c_B}} - \overrightarrow{v_{c_A}} \ . \tag{3.13}$$

The vector offset method (Equation 3.14) assumes that relationships are expressed by vector offsets: for the word pair *(man, woman)*, the offset expresses the *male-female*-relation. Given the next word pair to check, *(king, queen)*, the offset $\overrightarrow{o_{(c_A,c_B)}}$ between the embeddings for man ($\overrightarrow{v_{c_A}}$) and woman ($\overrightarrow{v_{c_B}}$) is added to the embedding for king ($\overrightarrow{v_{c_C}}$). The sum should end up in the close neighborhood of the embedding for queen ($\overrightarrow{v_{c_D}}$):

$$\overrightarrow{v_{c_C}} + \overrightarrow{o_{(c_A,c_B)}} = \overrightarrow{v_{c'_D}} \approx \overrightarrow{v_{c_D}} \ . \tag{3.14}$$

The 'close neighborhood' is judged by cosine distance d_{cos} (cf. Equation 3.2) of the obtained vector $\overrightarrow{v_{c'_D}}$ and the actual embedding for queen ($\overrightarrow{v_{c_D}}$). Thus, given two word pairs *(a, b)* and *(c, d)*, the analogy task checks to what extent the relations within the pairs are similar. Figure 3.3 illustrates the intuition of the vector offset method, where *'man'* is to *'woman'* as *'king'* is to *'queen'* as both pairs are in a *male-female*-relation (represented by the green arrow in the figure).

[4] pre-trained `GloVe` embeddings: `https://nlp.stanford.edu/projects/glove/`

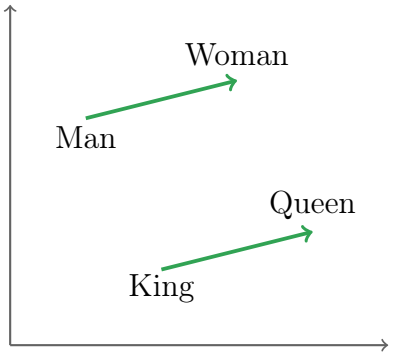


Figure 3.3: Intuition of the vector offset method for evaluating `Word2Vec`-embeddings. The embeddings should incorporate semantic regularities such as *'man'* is to *'woman'* as *'king'* is to *'queen'* where both pairs are in a *male-female*-relation (green arrow).

Word2Vec Beyond Words. Traditionally, embedding learning methods are used on free text in order to model linguistic or thematic relations between words. There is an interest in abstracting from word embeddings towards embeddings for more coarse grained units: `Word2Vec` is used to learn embeddings for senses (Iacobacci et al., 2015) or for supersenses (Flekova and Gurevych, 2016). Iacobacci et al. (2015) use the continuous Bag-of-Words model on texts annotated with BabelNet senses (Navigli and Ponzetto, 2012a). Flekova and Gurevych (2016) use the Skip-gram model on texts with mapped WordNet supersenses (Miller, 1990; Fellbaum, 1990). For evaluation, both works are oriented towards the analogy tasks by Mikolov et al. (2013b) and perform qualitative analyses for the top k most similar embeddings for (super)senses or visualize the embeddings in a vector space.

In our case, we not only use pre-trained textual embeddings for words (cf. Chapter 4), but we also apply the `Word2Vec` approach to frame-annotated texts in order to obtain textual embeddings for frames (cf. Chapter 5). When applying the `Word2Vec` approach to learn frame embeddings, we orient ourselves to the related work on (super)sense embeddings (Iacobacci et al., 2015; Flekova and Gurevych, 2016) as reviewed above.

The embedding methods for words are extended to longer sequences such as embeddings for multi-word expressions, sentences (e.g., `Sent2Vec` or `InferSent`, Pagliardini et al., 2018; Conneau et al., 2017) or paragraphs (e.g., `Paragraph Vector` also known as `Doc2Vec`, Le and Mikolov, 2014).

Matrix Factorization Methods

Matrix factorization denotes a factorization of a matrix into a product of matrices. One possible form is the following: an initial vector $(1 \times m)$ is multiplied by a matrix $(m \times n)$ to yield a goal vector $(n \times 1)$. Matrix factorization methods learn latent feature vectors, which in turn can also be regarded as embeddings. To our knowledge, the only approach so far actually learning frame embeddings is a matrix factorization approach (`WSABIE`, Weston et al., 2011, applied by Hermann et al. (2014) and explained in the following paragraphs). In the context of the task of Frame Identification with matrix factorization, it learns frame embeddings as a by-

product. As we will work with frame embeddings (cf. Chapter 5), we review this matrix factorization approach to embedding learning.

Task: Prediction of Category. The task is Frame Identification as explained in Section 2.1.2.2. The setup is to learn latent representations for frames and to learn a matrix-based mapping from sentences to the latent space so that the overall task of predicting the frame for the sentence is successfully performed. The state-of-the-art system (Hermann et al., 2014) for Frame Identification projects frames and predicates with their context words into the same latent space by using the algorithm for Web Scale Annotation by Image Embedding (`WSABIE`, Weston et al., 2011). As the focus of such systems is on the task of Frame Identification, the latent representations of the frames are rather a substep contributing to Frame Identification but not studied further or applied to other tasks. We will extract these frame embeddings and explore them with respect to frame-to-frame relations (cf. Section 5.1).

Approach: WSABIE. The `WSABIE` algorithm (Weston et al., 2011) originates from research in user-item recommendation, where a user (say a person interested in watching a movie) is associated with certain suggestions for items (say a title of a movie). The recommendation shall be made based on the person's pattern of interests, the user-item interactions; and the recommendation for a user is the first item in the ranked list of items. The `WSABIE` algorithm uses a Weighted Approximate-Rank Pairwise loss (WARP, common abbreviation) and gradient-based updates to minimize the distance between the latent representations of user and the correct item label, while maximizing the distance to all the other irrelevant item labels. Weston et al. (2011) transfer the setting of user-item interactions to an image-annotation setup where the 'recommendation' for an image is the first item in the ranked list of annotations.

Hermann et al. (2014) suggest to use the `WSABIE` algorithm for Frame Identification to map sentences and frame representations to a common latent space. In the context of Frame Identification, 'users' are the predicates within the sentences and 'items' are the frame-labels. For training with `WSABIE`, a user-item-interaction matrix is created (of the size: number of users $\times$ number of items). To allow for estimation of user-item interactions for a new unseen user (the test data), 'user features' are added: these are the initial context representations.These user features are shared between training set and test set. For these user features, the training procedure yields estimated latent vectors in the lower-dimensional space, which are called 'user embeddings'. Analogous to this, it yields 'item embeddings' for the items. At test time, the new user gets projected into the lower-dimensional space by multiplying the new user's features with the learned user embeddings. In the projection, the closest item embedding (corresponding to the most likely frame-label) can be found. Two projection matrices (one for frames and one for predicates) are learned using WARP loss and gradient-based updates so that the distance between the predicate's latent representation and that of the correct frame are minimized. Consequently, latent representations of frames will end up close to each other if they are evoked by similar predicates and context words. During testing, the cosine distance d_{cos} (cf. Equation 3.2) is used to find the closest frame given the input.

One advantage of this approach is that similar frames are positioned close to each other in the latent space, which allows information to be shared between similar predicates and similar frames.

We follow their approach in Section 4.1.3 and further experiment with the obtained embeddings in Section 5.1.

Comment on Neural versus Matrix Factorization Approaches. Goldberg (2016) points out that neural embedding approaches are connected to matrix factorization approaches and thus, embedding approaches should not be hyped too much (Levy et al., 2015a). In fact, (Levy and Goldberg, 2014b) remark that neural embedding approaches are implicit matrix factorizations.

3.3 Structured Embeddings

Structured embedding approaches leverage the information of a knowledge base to learn embeddings for the entities and the relations that are organized in triples of *(entity_1, relation, entity_2)*, or in short (e_1, r, e_2). In the following, we provide background knowledge about a widely-used method for learning structured embeddings, `TransE` (Bordes et al., 2013, as a common abbreviation for Translating Embeddings), which we will apply in our work.

Tasks: Link Prediction and Triple Classification. The two tasks of Link Prediction and Triple Classification are intertwined with each other as they are both subgoals of Knowledge Base Completion: Link Prediction is to predict new triples, and Triple Classification is to check whether a triple is correct or not. However, both tasks work with the rich structure of the knowledge graph and do not aim at creating triples of unknown and new entities or relations.

Link Prediction. The task of Link Prediction is illustrated in Figure 3.4. Given a pair of a head/tail and a relation, the goal of link prediction is to identify the missing tail/head. As suggested by Bordes et al. (2013), Link Prediction is usually evaluated in terms of *(1)* the mean rank (MR, common abbreviation) of the correctly predicted entities and *(2)* the proportion of correct entities in the top-10 ranked ones (Hits@10, common abbreviation).

Triple Classification. The task of Triple Classification is illustrated in Figure 3.5. Triple classification is a binary classification task, in which the knowledge graph triples are classified as correct or not according to a given dissimilarity measure (Socher et al., 2013). For this purpose, a threshold for each relation δ_r is learned. Accordingly, a triple (h, r, t) is considered correct if its energy is less than δ_r, and incorrect otherwise.

Approach: TransE. The basic assumption leading to the training objective is the translation within a triple: one entity in the triple can be obtained by combining the other entity with the relation so that they translate into the first entity (see illustration in Figure 3.6). As the entities and relations are modeled as embeddings,

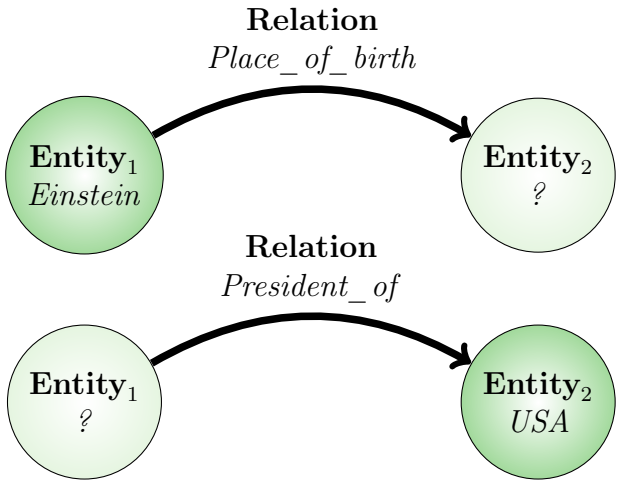


Figure 3.4: Link Prediction. Given one entity and a relation, predict the missing other entity.

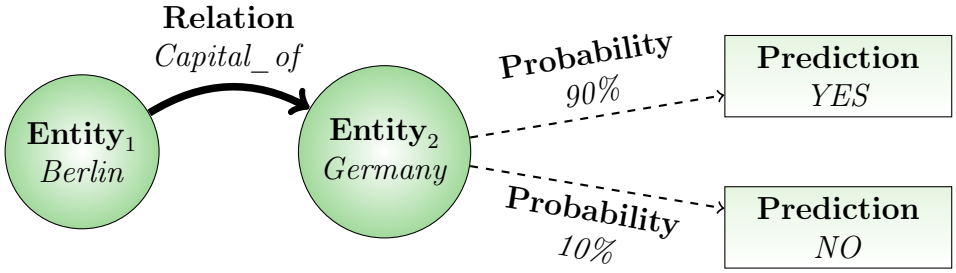


Figure 3.5: Triple Classification. Given a triple, classify whether it is correct or not.

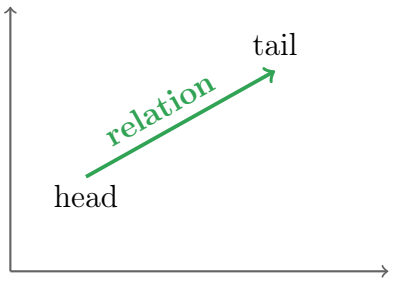


Figure 3.6: Translation assumption in vector space. The addition of head entity and relation shall end up in (translate to) the tail entity.

these, in turn, can be used to perform different kinds of inferences on the knowledge graph. These include identifying new relations or validating existing ones. However, translation-based methods rely on the rich structure of the knowledge graph and generally ignore any type of external information about the included entities.

`TransE` (Bordes et al., 2013) is the foundational and straight-forward translation-based approach for knowledge graph representation learning. `TransE` represents entities and relations as vectors in the same space, where the relation is considered a translation operation from the representation of the head to that of the tail entity. Following the translational assumption (Equation 3.15), given a triple (h, r, t) where $\boldsymbol{h}$, $\boldsymbol{r}$, $\boldsymbol{t}$ are the vector representations of the head, relation, and tail, respectively, we have:

$$\boldsymbol{h_s} + \boldsymbol{r_s} \approx \boldsymbol{t_s} \ . \tag{3.15}$$

Additionally, `TransE` uses a dissimilarity (or distance) measure d to define the energy of a given triple as $d(\boldsymbol{h} + \boldsymbol{r}, \boldsymbol{t})$. Finally, the representations of knowledge graph entities and relations are learned by minimizing a margin-based ranking objective that aims to score positive triples higher than negative triples based on their energies and a predefined margin.

In general, previous works such as Bordes et al. (2013) start from Equation 3.15 and build models for minimizing a margin-based ranking criterion as a loss function, with m being the margin ($m > 0$). Conventionally, negative triples are sampled from the knowledge graph by corrupting the head, the tail, or the relation of correct triples, e.g., $(\boldsymbol{h'} + \boldsymbol{r}, \boldsymbol{t'})$. Then, the following ranking loss (Equation 3.16) between positive and negative triples is minimized:

$$loss = [m + d(h + r, t) - d(h' + r, t')]_+ \ , \tag{3.16}$$

where $[x]_+$ denotes the positive part of x.

`TransE` is a simple and effective method, however, the simple translational assumption constrains the performance when dealing with complex relations, such as one-to-many (e.g., one actor can play the main character in several movies) or many-to-one (e.g., many people can have the same surname, 'Smith' as an example). To address this limitation, some extensions of `TransE` have been proposed. Wang et al. (2014) introduced `TransH`, which uses translations on relation-specific hyperplanes and applies advanced methods for sampling negative triples. Lin et al. (2015b) proposed `TransR`, which uses separate spaces for modeling entities and relations. Entities are projected from their space to the corresponding relation space by relation-specific matrices. Moreover, they propose an extension, `CTransR`, in which instances of pairs of head and tail for a specific relation are clustered in a way that the members of the clusters exhibit similar meanings of this relation. Lin et al. (2015a) proposed another extension of `TransE`, `PTransE` that leverages multi-step relation path information in the process of representation learning.

The above models all rely only on the structure of the knowledge graph, and learning better knowledge graph representations is dependent on the complexity of the model.

3.4 Visual Embeddings

Visual embedding approaches learn embeddings for objects that are present in images. The task giving the training objective is the other way round: given an image, predict the object it shows. Neural methods applied to images differ from those applied to texts: Whilst for text, multi-layer perceptrons (MLP, common abbreviation) and long-short term memories (Hochreiter and Schmidhuber, 1997, LSTM, common abbreviation) are most commonly used, for images convolutional neural networks (CNN, common abbreviation) are standard (cf. Section 3.1).

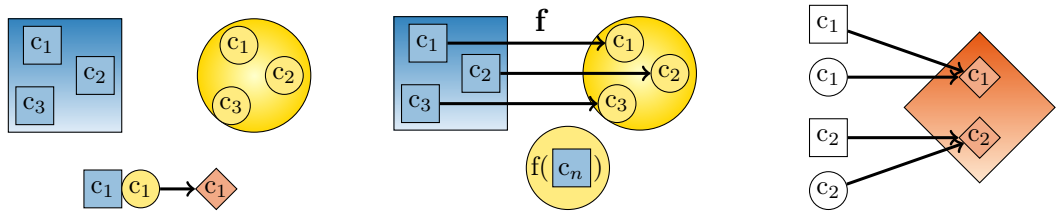
In the following, we provide background knowledge about a widely-used model for learning visual embeddings, `VGG` model (Chatfield et al., 2014, as a common abbreviation for Visual Geometry Group), which we will apply in our work.

Convolutional neural networks, as for example the `VGG` model, have a multi-layer structure with several layers of non-linear feature extractors for recognizing visual patterns directly from pixel images. The large number of parameters within these layers is learned from data, for example on the task of image classification.

Task: Image Classification. Image classification is the task of assigning a label to an image. The label is assigned based on the information in the image, as for example instantiations of cats or bikes. There are different datasets providing collections of images labeled with the object(s) shown, for example ImageNet (Deng et al., 2009) and the WN9-IMG dataset (Xie et al., 2017). The WN9-IMG dataset links WordNet synsets to a collection of ten ImageNet images and all synsets in WN9-IMG are part of triples of the form entity-relation-entity, i.e. synset-relation-synset. Further datasets to evaluate on image understanding are MS COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014).

Approach: VGG. In our work, we use the `VGG-m-128` Convolutional Neural Network (Chatfield et al., 2014). It is a pre-trained `VGG` model (as a common abbreviation for Visual Geometry Group) that can process an input image of size 224 x 224 in order to extract the 128-dimensional activation of the last fully-connected layer (the pre-final layer which is before the softmax layer for final classification). This 128-dimensional vector is then regarded as a visual embedding for the input image. Usually, L_2-normalization is applied for consistency. The pre-trained `VGG-m-128` model consists of eight learnable layers (five convolutional layers) where the last three layers are fully-connected. It is pre-trained on the images from ImageNet (Deng et al., 2009) when classifying images into more than 1,000 object categories. During training, the parameters of the convolutional neural network, i.e. the weights and biases of the layers, are determined. Finally, this optimized pre-trained model can be used to obtain a visual embedding for a new image.

Alternatives to the `VGG-m-128` model are: `VGG-16` (Simonyan and Zisserman, 2014) and `VGG-19` which are both extensions of `VGG-m-128`, and furthermore also `AlexNet` (Krizhevsky et al., 2012), `GoogLeNet` (Szegedy et al., 2015), (Szegedy et al., 2016), `ResNet` (He et al., 2016), and `Inception` (Szegedy et al., 2017). These models differ in depth (number of layers) and in the size of the filters (the smallest building block) in the layers. We decide for the `VGG-m-128` model as this one is most often used by prior work in Natural Language Processing that includes visual embeddings.



(a) Multimodal fusion. Concatenate known representations from modality A and B and apply dimensionality reduction.

(b) Mapping. Learn a mapping function f from modality A to B that can be applied on unknown concepts c_n.

(c) Joint learning. Optimize two objectives simultaneously: quality of unimodal representations and cross-modal alignment.

Figure 3.7: Methods for learning multimodal representations. Blue and yellow shapes indicate the representation space of modality A and B.

3.5 Multimodal Embeddings

Multimodal representations combine information from separate modalities. Specifically, there are different approaches on how to combine representations from separate channels in order to obtain a joint representation. In the following, we explain the same distinctions within methods for learning multimodal embeddings as published in our survey (Beinborn et al., 2018).[5] In this survey, we group existing approaches into three categories: multimodal fusion, mapping, and joint learning. Whereas fusion is a straight-forward combination of two modalities, mapping and joint learning address the problem that the concept representation is only available in one modality: they map into the second modality or project into a new shared modality. Furthermore, joint learning enables an exchange between both source modalities. In the following sections, these categories are explained along Figure 3.7.

Tasks for Evaluation of Multimodal Embeddings. Analogous to the evaluation of textual word embeddings in terms of modeling word pair similarities (Section 3.2), multimodal embeddings are also evaluated in terms of modeling pair-wise similarities on the textual level or visual level.

Word pair similarity ratings on the textual level are provided by Mikolov et al. (2013b, see Section 3.2), but also in datasets such as *WordSim353* (Finkelstein et al., 2002), *SimLex-999* (Hill et al., 2015), *MEN* (Bruni et al., 2012), *SemSim* and *VisSim* (Silberer and Lapata, 2014), where *VisSim* extends the notion of similarity from the textual level to the visual level (Section 2.3.1). In order to measure quantitatively the extent to which the cosine similarity s_{cos} (cf. Equation 3.3) of two embeddings agrees with the human similarity rating for the two concepts, respectively, the Pearson product moment correlation coefficient (Equation 3.17)

$$r_p(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \,, \tag{3.17}$$

[5] The distinction within methods for learning multimodal embeddings is a joint contribution of myself together with my co-author Lisa Beinborn, for which we refer to our survey (Beinborn et al., 2018).

or Spearman rank correlation coefficient (Equation 3.18)

$$r_s(X,Y) = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (\text{rank}(x_i) - \text{rank}(y_i))^2 \tag{3.18}$$

are used. The correlation coefficients measure the statistical correlations of two variables X and Y (Pestman, 1998). More specifically, the Pearson correlation score measures a linear relationship between the similarity ratings, whereas the Spearman correlation score is a rank correlation metric that considers the ordering of the similarity ratings. In the described evaluation, $x_i \in X$ are the cosine similarities of the embedding pairs and $y_i \in Y$ are the human similarity ratings of the corresponding two concepts. High correlations are obtained for the following scenario: the higher the similarity rating, the higher the cosine similarity and, at the same time, the lower the similarity rating, the lower the cosine similarity.

Multimodal fusion

Multimodal fusion is the most straight-forward approach to combine representations from separate channels (Figure 3.7a). The multimodal representation $\overrightarrow{v_{mm(c)}}$ for a concept c is built by concatenating $\frown$ the two unimodal embeddings $\overrightarrow{v_{m_1(c)}}$ and $\overrightarrow{v_{m_2(c)}}$ (Equation 3.19) from the separate modalities m_1 and m_2:

$$\overrightarrow{v_{mm(c)}} = \overrightarrow{v_{m_1(c)}} \frown \overrightarrow{v_{m_2(c)}} \,. \tag{3.19}$$

Furthermore, the concatenation can be subject to a weighing (Equation (3.20)) in terms of a tunable parameter α:

$$\overrightarrow{v_{mm(c)}} = \alpha \cdot \overrightarrow{v_{m_1(c)}} \frown (1-\alpha) \cdot \overrightarrow{v_{m_2(c)}} \,. \tag{3.20}$$

When simply concatenating embeddings, there is no adaptation in the separate unimodal spaces. To give an example, the concatenated representation for *cat* could give us the information that, on the one hand, *cat* is visually similar to *panther* (similar cues in appearance), and on the other hand, *cat* is textually similar to *dog* (occurring in similar textual contexts), but it is not possible to determine cross-modal similarity.

Concatenation occurs directly on the concept representation level (e.g., concatenation of textual and visual embeddings for the concept *cat*) and is thus also known as feature-level fusion or early fusion (Leong and Mihalcea, 2011; Bruni et al., 2011). Early fusion is to be understood in contrast to late fusion, where late fusion means obtaining two separate unimodal predictions and only afterwards combining the two predictions to one final multimodal prediction. In both cases, every single modality contributes to the final prediction.

By design, concatenation increases the dimensionality of the resulting multimodal embedding. In order to reduce the dimensionality and to smooth the concatenated representations while maintaining multimodal correlations, dimensionality reduction techniques such as singular value decomposition (Bruni et al., 2014, SVD, common abbreviation) or canonical correlation analysis (Silberer and Lapata, 2012, CCA, common abbreviation) have been applied.

Mapping

For very common, mostly concrete concepts such as *cat*, for example, representations are available in several modalities, textual and visual for example. However, for rare or abstract concepts such as *God*, for example, there might be a textual representation but it might be hard to form a visual representation. In such cases, the multimodal fusion strategy in terms of concatenation cannot produce a multimodal embedding. Mapping approaches make it possible to still obtain a representation, in this example a visual representation even if there was no image available: the textual representation is mapped to visual space (Figure 3.7b). Crucially, this approach does not require a corresponding image to exist. First, a mapping function f from m_1 to m_2 is learned so that the similarity between all known representations of c in m_2 and their projections from the representation in m_1 is maximized: $c_{m_2} \sim f(c_{m_1})$. Concerning the loss measure, a margin-based ranking criterion minimizing the distance between true pairs of concept representations (c_{m_1}, c_{m_2}) and maximizing the distance for pairs with random target representations $(c_{m_1}, random_{m2})$ has been shown to be a good choice for image labelling (Frome et al., 2013). In image labelling, the mapping is applied in the opposite direction as in the above example in order to classify unknown objects in images based on their semantic similarity to known objects (Socher et al., 2013).

There is a growing interest in Natural Language Processing for enriching traditional approaches with knowledge from the visual domain, as images capture qualitatively different information compared to text and words benefit from grounding in the visual world (Lazaridou et al., 2014; Collell et al., 2017). Multimodal approaches based on pre-trained embeddings are reported to be superior to unimodal approaches. Textual embeddings have been enriched with information from the visual domain, e.g., for Metaphor Identification (Shutova et al., 2016), Question Answering (Wu et al., 2017b), and Word Pair Similarity (Collell et al., 2017). The latter presents a simple, but effective way of extending textual embeddings with so-called multimodal Imagined embeddings by a learned mapping from language to vision. We apply the Imagined method in the context of multimodal Frame Identification in Section 4.2.1.

Imagined Method for Language and Vision. The Imagined method (Collell et al., 2017) learns a mapping function: it maps from the word embedding space to the visual embedding space given those words that occur in both pre-trained embedding spaces. The Imagined method is promising for cases where one embedding space (here, the textual one) has many instances without correspondence in the other embedding space (here, the visual one), but the user still aims at obtaining instances of the first in the second space. The mapping is a nonlinear transformation using a simple neural network. The objective is to minimize the cosine distance d_{cos} (cf. Equation 3.2) between each mapped representation of a word and the corresponding visual representation. Finally, a multimodal representation for any word can be obtained by applying this mapping to the word embedding. Interestingly, in a later study Collell and Moens (2018) remark, that the initial formulation of obtaining 'imagined' visual representations after applying the learned mapping can in fact be interpreted in a misleading way. They find, that the mapped vectors remain more

similar to the input vectors (word embeddings) than to the target vectors (image embeddings). Knowing this, the Imagined method is still appealing to incorporate some visual information into the mapped embeddings whilst keeping information from the initial embeddings, and specifically, to obtain 'imagined' embeddings for words that do not have a visual embedding.

Joint learning

Apart from tolerating unseen concepts in one modality, joint learning (Figure 3.7c) also implements an exchange between both source modalities. This is crucially different to Fusion and Mapping, as these assume pre-trained unimodal embeddings which are transformed in a directed way. Joint learning specifically projects the unimodal embeddings into a new, joint modality and by this, both source modalities together have the potential to shape the shared representations in the new modality.

Joint learning of multimodal representations is successfully applied to problems in image understanding such as image captioning, image retrieval, or feature learning from the textual, visual and auditory modalities (Karpathy et al., 2014; Srivastava and Salakhutdinov, 2012; Ngiam et al., 2011) and there is a need of methods for aligning multimodal embeddings.

Concerning textual embeddings learned by the Skip-gram model in `Word2Vec`, Lazaridou et al. (2015) enrich the learning approach with visual features. Their model optimizes two constraints: the embeddings for concepts c with respect to their textual contexts (this is the objective in m_1, known from Skip-gram) and the similarity between textual embeddings and their visual counterparts (supervised objective with margin-based ranking criterion for (c_{m_1}, c_{m_2})). In their approach, the visual embeddings remain fixed, but the textual representations are learned from scratch in order to fulfill both constraints. Going further, Silberer and Lapata (2014) use stacked multimodal autoencoders to simultaneously learn embeddings for each modality (unsupervised reconstruction objective for both, m_1 and m_2) and their optimal multimodal combination (supervised classification objective for (c_{m_1}, c_{m_2})). Both approaches implicitly also learn a mapping between the two modalities.

In this thesis, we apply multimodal fusion and mapping strategies to obtain multimodal representations.

3.6 Summary of the Chapter

This chapter introduced different approaches for learning textual, structured, visual and multimodal embeddings (Sections 3.2, 3.3, 3.4 and 3.5, respectively) – together with the tasks they have evolved with.

We provide a summary of this chapter in terms of bullet points in the box below:

FOUNDATIONS IN A NUTSHELL

Textual embedding learning uses neural nets or matrix factorization.

- Neural methods (e.g., `Word2Vec`, `GloVe`) train on predicting words:
  Given *'dogs sometimes ___ too loudly'*, predict *'bark'*
- Matrix factorization methods (e.g., `WSABIE`) train on predicting categories:
  Given *'Abby bought a car.'*, predict *'Commerce_buy'*
- Evaluation of embeddings: vector offset method to check whether relationships are expressed by vector offsets
  For the *male-female*-relation: *man is to woman as king is to queen*

Structured embeddings are learned on knowledge bases.

- Translation-based methods (e.g., `TransE`) train on classifying triples:
  Given *(Berlin, Capital_of, Germany)*, predict *correct*

Visual embeddings are learned on images and their descriptions.

- Neural methods (e.g., `VGG`) train on recognizing objects:
  Given *an image of a house*, predict *house or building*

Multimodal embedding learning combines information from separate modalities (e.g., textual and visual).

→ We categorize the methods for learning multimodal representations:
Multimodal fusion *(e.g., concatenation)*
Mapping *(e.g.,* `Imagined` *method)*
Joint learning *(e.g.,* `Word2Vec` *for words and images)*

Chapter 4

Frame Semantics for Situations and Actions

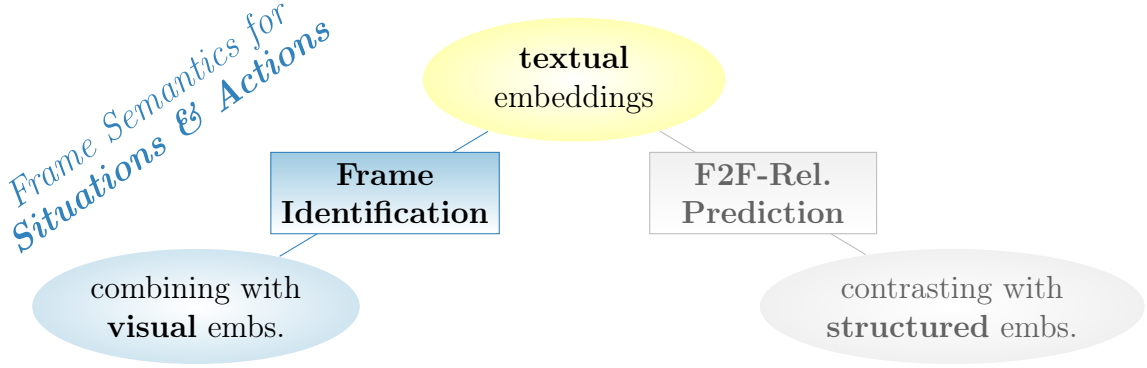


Figure 4.1: Structure of Chapter 4. Left blue branch: knowledge about situations and actions with textual and visual word embeddings for Frame Identification. (Right branch is focus of Chapter 5.)

In this chapter, we assume context knowledge to be crucial for the abstraction from single words to categories of meaning. We present and discuss our contributions and findings in the context of language understanding with frame semantics for modeling knowledge of situations and actions as outlined in Figure 4.1 (left blue branch). We model knowledge about situations and actions with textual word embeddings and in combination with visual ones for the task of Frame Identification. The immediate background for textual and grounded language understanding and for embedding learning based on the textual, the visual, and multiple modalities was given in Sections 2.1 and 2.3, and 3.2, 3.4 and 3.5, respectively.

In the first part, Section 4.1, we focus on Frame Identification with textual embeddings as context representations. We develop a state-of-the-art Frame Identification `UniFrameId` system that operates on FrameNets of two languages, namely English and German. The underlying assumption is about context knowledge being necessary for the abstraction from single words to categories of meaning in terms of frames. We find that taking the context words into account in terms of textual embeddings in a straight-forward neural network architecture yields state-of-the-art results for English as well as for German data. This part is to lay the foundations

of Frame Identification with textual embeddings which we will build upon in the succeeding part.

In the second part, Section 4.2, we focus on Frame Identification with multimodal embeddings to enrich the context representations. We extend our unimodal Frame Identification system to a use-case with multimodal embeddings, `MultiFrameId`, which improves the performance on English data. The underlying hypothesis is that language understanding requires implicit commonsense knowledge which is rarely expressed textually but can be extracted from images. We find that additional information from images is beneficial to Frame Identification. In an outlook, we explore the textual and visual grounding of highly embodied verbs (Section 4.2.4.1) and suggest to develop multimodal embeddings for verbs specifically that incorporate sensomotoric information.

4.1 Frame Identification with Textual Embeddings

In this section, we present and discuss our contributions and findings in the context of textual language understanding with frame semantics, where we model knowledge about situations and actions with textual word embeddings for the task of Frame Identification. We introduce our new state-of-the-art system for Frame Identification that uses textual embeddings, `UniFrameId`, and we approach the development of the system with the underlying assumption about context knowledge being necessary for abstracting from single words to categories of meaning in terms of frames. Our research question asks about which approach with textual embeddings to use for the task of Frame Identification: the previous state-of-the-art approach by Hermann et al. (2014) using a matrix factorization architecture to learn `WSABIE` embeddings (cf. Sections 3.2), or alternatively, a neural network architecture (cf. Section 3.2).

> RQ: *Which architecture to recommend for Frame Identification, according to experiments?*

First, we review previous systems (cf. Section 4.1.1) in chronological order. Next, we re-implement the matrix factorization architecture of the previous state-of-the-art approach by Hermann et al. (2014) and contrast this with a straight-forward neural network approach, which is our prototype system for Frame Identification `SimpleFrameId` (cf. Section 4.1.2). As we find an advantage of the neural network approach, we introduce our optimized system `UniFrameId`, which we develop out of the prototype `SimpleFrameId`. In particular, we analyze the performance of `UniFrameId` in detail, especially for difficult cases (cf. Section 4.1.3). Finally, we expound a multilingual evaluation and contrast Frame Identification on English versus German data (cf. Section 4.1.4).

Taken together, this section is to lay the foundations of Frame Identification with textual embeddings which we will build upon in Section 4.2. Our papers (Hartmann et al., 2017)[1] and (Botschen et al., 2018a)[2] are foundational to this chapter.

[1] My contribution in this paper is the following: `SimpleFrameId` system with `WSABIE` embeddings.
[2] My contributions in this paper are the following: `UniFrameId` system for FrameNet and SALSA, `MultiFrameId` system and analysis of experiments.

model	FrameId (accuracy)	full SRL (F1-score)
`Semafor` - Das et al.	83.60	64.54
`mateplus (Framat)` - Roth and Lapata	`Semafor`	67.88
`PathLSTM` - Roth and Lapata	`Semafor`	70.0_
`Hermann-14`* - Hermann et al. (own full lexicon)	88.41	69.91
`Hermann-14`* - Hermann et al. (`Semafor` lexicon)	86.49	68.69
`NNs for SRL` - FitzGerald et al.	`Hermann-14`	70.9_
`Open-SESAME` - Swayamdipta et al.	`Hermann-14`	70.9_

Table 4.1: Scores of previous models. Results are listed for Frame Identification (first column) and full Semantic Role Labeling (second column). The dashed line separates the approaches into `Semafor` and those that build on top (upper part) and into `Hermann-14` and those that build on top. The underscores in the last column mark scores for which the second digit after the comma is not reported.
* `Hermann-14` is not publically available.

4.1.1 Previous Systems

The task of Frame Identification aims at disambiguating a situation around a predicate as introduced in Section 2.1. For a long time, traditional feature-based models such as `Semafor` (Das et al., 2014, relying on manually defined features) were state-of-the-art approaches to Frame Identification until these were outperformed by latent feature models as proposed by Hermann et al. (2014, further referred to as `Hermann-14`).

In the following, we review these two approaches to Frame Identification, `Semafor` and `Hermann-14`, and we additionally give an overview of further approaches that are build on top of them for full Semantic Role Labeling. The respective scores of the systems in Frame Identification and in full Semantic Role Labeling are listed in Table 4.1.

Semafor. The system `Semafor` (Das et al., 2014, short for SEMantic Analyzer Of Frame Representations) implements a pipeline approach for the full Semantic Role Labeling, including a model for Frame Identification. `Semafor` reaches an accuracy of 83.60% for Frame Identification and an F1-score of 64.54% for full Semantic Role Labeling (cf. Table 4.1).

The Frame Identification part of the system `Semafor` relies on an elaborate set of syntactic and lexical features to represent the context of the predicate. To give an example, the WordNet hierarchy is used as a source of lexical information, and a label propagation-based approach is implemented to account for unknown predicates. The feature-based classifier is a conditional log-linear model for supervised classification, where the features are similar to previous work.

Hermann-14. Hermann et al. (2014) present a new state-of-the-art system for Frame Identification (referred to as `Hermann-14`). They outperform `Semafor` with an accuracy of 88.41% for Frame Identification and an F1-score of 69.91% for full Semantic Role Labeling on the same dataset split as published by Das et al. (2014) but using their own full lexicon. 'Full' lexicon means, that they prepare a lexicon

that renders every predicate (lexical unit) in the test set as seen, i.e. at test time, they can look up the possible set of frames for every instance. However, with the `Semafor` lexicon, there occur some rare predicates at test time that are unknown to the lexicon and thus, the prediction has to be done over all frames in the lexicon – which is a more difficult setup. For a direct comparison of the Frame Identification systems `Semafor` and `Hermann-14`, Hermann et al. (2014) also report their performance when using the `Semafor` lexicon: in this setting `Hermann-14` reaches 86.49% accuracy for Frame Identification and 68.69% F1 for full Semantic Role Labeling (cf. Table 4.1).

The approach by Hermann et al. (2014) uses word embeddings augmented by syntactic information to represent the context of the predicate. In the syntax-augmented context representations, a certain region of the vector, a *container*, is reserved for each possible dependents that a syntactic parser can find to a predicate in the training data. Regarding a sentence with a predicate to identify the frame for, any container is filled with the word embedding of the corresponding syntactic dependent, if this syntactic path exists in the sentence. All the remaining containers, for which the syntactic path does not exist in the sentence, stay empty (implemented by filling it with zeros instead of word embeddings).

The system `Hermann-14` uses the `WSABIE` algorithm (Weston et al., 2011, cf. Section 3.2) to map context representations and frame representations to a common latent space. During testing, cosine distance d_{cos} (cf. Equation 3.2) is used to find the closest frame given the input. One advantage of this approach is that similar frames are positioned close to each other in the latent space which allows information to be shared between similar predicates and similar frames.

A disadvantage of `Hermann-14` is that the context representations build out of the syntactic containers is sparse as only a few syntactic paths exist in every sentence and thus, many reserved containers are not filled with the word embedding of the corresponding syntactic dependent. Furthermore, the system as well as their own pre-trained word embeddings are not publicly available and the exact scores are difficult to replicate.

Full Semantic Role Labeling. Further contributions to Role Labeling use the Frame Identification systems `Semafor` and `Hermann-14` to automatically identify frames in their full Semantic Role Labeling approaches. Neural network approaches for role labeling on top of the existing Frame Identification systems can boost the performance of full Semantic Role Labeling up to an F1-score of 70% (cf. Table 4.1), see `mateplus-Framat` (Roth and Lapata, 2015) and `PathLSTM` (Roth and Lapata, 2016) on top of `Semafor`, and `NNs for SRL` (FitzGerald et al., 2015) and `Open-SESAME` (Swayamdipta et al., 2017, short for SEmi-markov Softmax-margin ArguMEnt parser) on top of `Hermann-14`.

4.1.2 Frame Identification System `SimpleFrameId`

In the previous Section 4.1.1, we have seen that state-of-the-art systems for Frame Identification encode the situational context of the predicate using pre-trained textual embeddings for words (see Hermann et al., 2014). Hence, it is assumed that the context of the situation is explicitly expressed in words. Two aspects are important. First, the textual embedding of the predicate itself is promising as this embedding

contains information about contexts the predicate appeared in during training. Second, the textual embeddings of the context words in the sentence are promising as they reveal the actual context, the actual meaning of the predicate in question. This is, Frame Identification systems using textual embeddings in these two ways assume and implement the idea of context carrying meaning of single words (distributional hypothesis, Harris, 1954).

We follow this assumption with our Frame Identification systems that are based on textual embeddings (Hartmann et al., 2017; Botschen et al., 2018a). In this section, we discuss our prototype system `SimpleFrameId` for Frame Identification, out of which we then develop our optimized system `UniFrameId` for English and German FrameNets (see next Section 4.1.3).

4.1.2.1 Architecture: Matrix Factorization versus Neural Approach

We explain the development of the Frame Identification classifier in the context of `SimpleFrameId` (Hartmann et al., 2017). First, we re-implement the matrix factorization architecture of the previous state-of-the-art approach by Hermann et al. (2014), we aim at replicating the results of the state-of-the-art system `Hermann-14`. Then, we explore an alternative approach with a neural network architecture, which is our prototype system for Frame Identification `SimpleFrameId`.

Textual Input Embeddings for both Approaches. The input representation (Equation 4.1) for both approaches is a simple concatenation $\frown$ (cf. Equation 3.19) of the predicate's pre-trained embedding $\overrightarrow{v_{(pred)}}$, and an embedding of the predicate context $\overrightarrow{v_{(cont)}}$:

$$\begin{aligned} \overrightarrow{v_{(in)}} = \ & \overrightarrow{v_{(cont)}} \frown \overrightarrow{v_{(pred)}} \ ; \\ & \text{with } \overrightarrow{v_{(cont)}} = \frac{\sum_{w \in cont} vsm(w)}{|cont|} \ , \\ & \text{and } \overrightarrow{v_{(pred)}} = vsm(pred) \ . \end{aligned} \tag{4.1}$$

Regarding the predicate context *cont*, we experiment with two kinds of contexts to build a dimension-wise mean of the pre-trained embeddings of a set of selected words w in the sentence. First, we orient ourselves to `Hermann-14` by considering the dependency parse of the sentence: we include only words, which are direct dependents of the predicate, to build an average of the respective word embeddings (we will refer to this as dependency-based bag-of-words approach `DepBOW`). Second, we include all the words in the sentence to build an average of the respective word embeddings (we will refer to this as sentence-based bag-of-words approach `SentBOW`). Thus, in both cases, we consider the average embedding of the pre-trained embeddings of the predicate's dependents or of all words in the sentence.

As the pre-trained word embeddings by Hermann et al. (2014) are not publicly available, we choose other pre-trained word embeddings that are public. Hermann et al. (2014) incorporate the notion of context in terms of syntactic dependents into their approach, thus we decide to choose `dependency`-based embeddings (Levy and Goldberg, 2014a, cf. Section 3.2). By this choice, syntactic knowledge of co-occurrence of syntactic dependents is integrated into the word embeddings directly.

We experiment with two different classification methods to process the input representations: one is a matrix factorization approach following the line of `Hermann-14`, the other one is a straight-forward two-layer neural network.

Matrix Factorization Approach. With the matrix factorization approach we follow the line of the current state-of-the-art system `Hermann-14` (Hermann et al., 2014, cf. Section 4.1.1) and learn representations for frames and predicates in the same latent space using the `WSABIE` algorithm (Weston et al., 2011, cf. Section 3.2).[3] We will refer to this approach as `WSB`. Note that a by-product of the approach oriented on `Hermann-14` are the `WSABIE` embeddings for frames that will be further examined in Section 5.1.

The outputs are scores for each frame known to the system by the lexicon, such that the frame with the highest score is selected as prediction.

Neural Network Approach. With the neural approach we follow the recent success of neural methods, which improved the performance of role labeling (cf. `PathLSTM`, `NNs for SRL`, and `Open-SESAME` in Table 4.1), but was not yet implemented for Frame Identification. We decide for a conceptually simple prototype to explore the potential of neural methods for Frame Identification. We will refer to this approach as `NN`.

Our neural network-based system is a two-layer feed-forward neural network, implemented with 'adagrad' optimizer. The first hidden layer comprises 256 neurons, followed by 100 neurons in the second hidden layer. Each node in the output layer corresponds to one frame-label class known from the lexicon. We use rectified linear units (Nair and Hinton, 2010, ReLU, common abbreviation) as activation function for the hidden layers, and a softmax activation function for the output layer yielding a multinomial distribution over frames. We take the highest activated neuron (arg max) to obtain the most likely frame label according to the classifier as the final prediction at test time. Optionally, filtering based on the lexicon can be performed on the predicted probabilities for each frame label. As this is a prototype, no hyperparameters have been optimized yet – this is done with `UniFrameId` (cf. Section 4.1.3).

Note that the classifier itself is agnostic to the predicate's part-of-speech and exact lemma and only relies on word representations from the *vsm*.

4.1.2.2 Experimental Setup and Data

We contrast the performance of four systems with respect to Frame Identification: dependency- versus sentence-based bag-of-words for input embeddings in the matrix factorization approach, `WSB+DepBOW` and `WSB+SentBOW`, and also in the neural network approach, `NN+DepBOW` and `NN+SentBOW`. Regarding the approach, `WSB+DepBOW` is most similar to `Hermann-14` (Hermann et al., 2014). We compare the performances of our systems to the state-of-the-art system `Hermann-14`.

[3] In our implementation, we use the LightFM package (Kula, 2015) for matrix factorization with the `WARP` option for a Weighted Approximate-Rank Pairwise loss.

	model	acc	acc amb
	Hermann-14	**88.41**	73.10
	WSB+DepBOW	85.69	69.93
	WSB+SentBOW	84.46	67.56
	NN+DepBOW	87.53	73.58
SimpleFrameId:	NN+SentBOW	87.63	**73.80**

Table 4.2: FrameId results (in %) on FrameNet test data. Reported are overall **acc**uracy and **acc**uracy for **amb**iguous predicates. Best results highlighted in bold. Models: *(a)* State of the art Hermann-14, *(b)* WSB+DepBOW, *(c)* WSB+SentBOW, *(d)* NN+DepBOW, *(e)* SimpleFrameId: NN+SentBOW

Data and Data Splits: Berkeley FrameNet. The Berkeley FrameNet (Baker et al., 1998; Ruppenhofer et al., 2016), as presented in Section 2.1.2, is a lexical resource for English with annotations based on frame semantics (Fillmore, 1976). The fully annotated texts provide the sentences with frame-labels for the predicates for training and evaluation. The lexicon, mapping predicates to the frames they can evoke, can be used to facilitate the identification of the frame for a predicate. (Table 2.1 contains the lexicon statistics, Table 2.2 the dataset statistics.)

In this work, we use FrameNet 1.5 to ensure comparability with the previous state of the art. Also, we use the common evaluation split for Frame Identification systems introduced by Das and Smith (2011) together with the development split of Hermann et al. (2014). Due to having one single annotation as consent of experts, it is impossible to determine the performance of a single human based on the experts' agreement.

4.1.2.3 Results and Discussion

We present the results of our four systems in Table 4.3.

Interestingly, we find that our straight-forward neural approach NN+SentBOW using sentence-based bag-of-words embeddings achieves results (accuracy of 87.63%) comparable to the state-of-the-art system, Hermann-14 (accuracy of 88.41%). From now on, we refer to our best system NN+SentBOW as SimpleFrameId.

However, the performance of WSB+DepBOW (accuracy of 85.69%) is worse than that of Hermann-14 even if the approach with WSB+DepBOW is the most similar one to Hermann-14. This gap in performance is relativized, but not nullified, by taking into account the slightly worse performance of Hermann-14 (accuracy of 86.49%) when using the Semafor lexicon to be directly comparable.

Our initial attempts to replicate Hermann-14, which is not publicly available, revealed that the container-based input feature space is very sparse: there exist many syntactic paths that can connect a predicate to its arguments, but a predicate instance rarely has more than five arguments in the sentence. So, by design, the input representation bears no information in most of its path containers. Moreover, Hermann-14 makes heavy use of automatically created dependency parsers, which might decline in quality when applied to a new domain or another language.

With respect to the input representation, we find an interesting tendency. On the one hand, for the matrix factorization approach WSB, the dependency-based

input representation DepBOW is the better choice compared to the sentence-based input representation SentBOW. This mirrors the strength of the dependency-based input representation in the by matrix factorization approach proposed by Hermann et al. (2014). On the other hand, for the neural approach NN, using all words of the sentence as context representation SentBOW performs slightly better than the complex context representation that uses dependency parses DepBOW. This could be an effect of the network leveraging the dependency information incorporated into the dependency-based word embeddings by Levy and Goldberg (2014a).

We demonstrate that the straight-forward neural approach SimpleFrameId, which is a simpler system compared to Hermann-14, achieves competitive performance on the FrameNet test data. Importantly, the performance on ambiguous predicates is slightly higher with SimpleFrameId (accuracy of 73.80%) than with Hermann-14 (accuracy of 73.10%). Furthermore, its performance is competitive even in out-of-domain performance – for details on this, the interested reader is referred to Hartmann et al. (2017).

As we find an advantage of the neural approach (accuracy of 87.63%) over the matrix factorization approach (accuracy of 85.69%) in terms of performance and in terms of simplicity, we decide to further explore the potential of the neural approach.

4.1.3 Frame Identification System UniFrameId

In this section, we build upon the prototype system SimpleFrameId for Frame Identification (see previous Section 4.1.2) to develop our optimized system UniFrameId for English and also for German FrameNets. The name UniFrameId indicates that we are using uni-modal embeddings, namely textual embeddings. This is extended to the multimodal case with MultiFrameId in Section 4.2.

4.1.3.1 Architecture and Textual Input Embeddings

Starting from the system SimpleFrameId, where we find an advantage of the neural approach over the matrix factorization approach in terms of performance and in terms of simplicity, we now further explore the potential of the neural approach with the optimized system UniFrameId.

Textual Input Embeddings. As input embeddings, we use the 300-dimensional GloVe embeddings (Pennington et al., 2014) as presented in Section 3.2. This is different to SimpleFrameId which used dependency-based word embeddings by Levy and Goldberg (2014a). The decision for GloVe embeddings is justified by hyperparameter studies by Klie (2017). Using our architecture, Klie (2017) experiments with different embeddings (amongst others Word2Vec and dependency-based embeddings). His experiments point out GloVe embeddings as reaching the best performance on the development set. The largest set of GloVe embeddings (2,200,000 words) did perform slightly better, but we decided for the standard set of GloVe embeddings (400,000 words) as described in Section 3.2, as this is most consistent with other research and the best trade-off with respect to vocabulary size and out-of-vocabulary tokens in the FrameNet corpus.

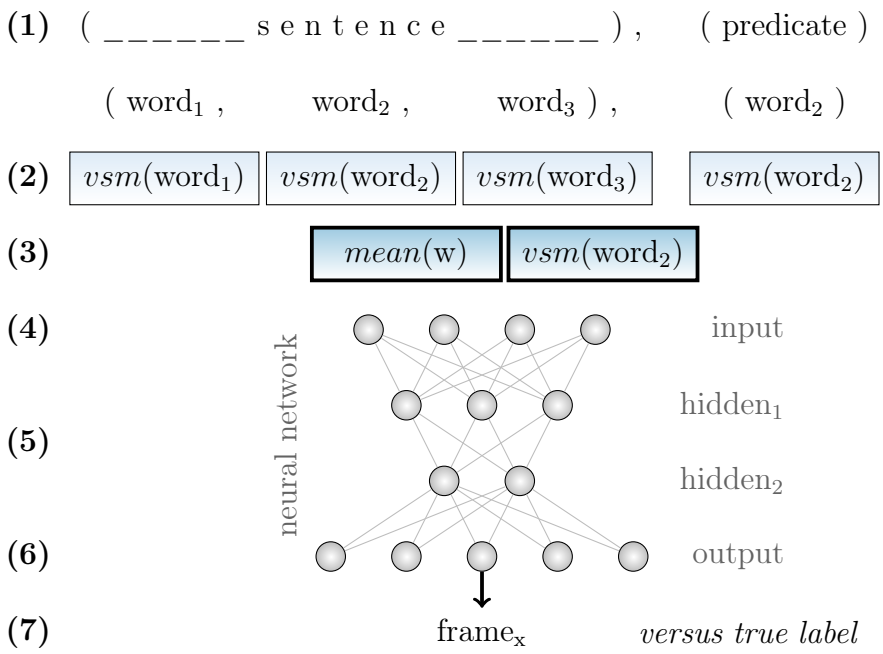


Figure 4.2: Sketch of the pipeline. (1) Input data: sentence with predicate, (2) Mapping: words to embeddings, (3) Input representation: concatenation of embeddings of the sentence mean and predicate, (4-6) Classifier: neural network, (4) Input layer, (5) Two hidden layers with ReLu activation function, (6) Output layer with SoftMax activation function, (7) Prediction of frame (plus comparison with true label).

Neural Architecture for `UniFrameId` versus `SimpleFrameId`. The system `UniFrameId` builds upon the `SimpleFrameId` (Hartmann et al., 2017) system for English Frame Identification based on textual word embeddings.

We explain the system pipeline along with Figure 4.2 that shows a sketch of the `UniFrameId` pipeline that arouse out of `SimpleFrameId`. Same as `SimpleFrameId`, the `UniFrameId` system is based on pre-trained word embeddings to build the input representation out of the predicate context and the predicate itself. The input representation is a concatenation (Equation 4.1, cf. step (3) in Figure 4.2) of an embedding of the predicate context $\overrightarrow{v_{(cont)}}$ and the predicate's pre-trained embedding $\overrightarrow{v_{(pred)}}$. We adopt the input representation from `SimpleFrameId` which we found to be most promising: as context *cont*, we consider the dimension-wise mean of the pre-trained embeddings of all words in the sentence.

The input representation is processed by a two-layer Multilayer Perceptron (MLP, Rosenblatt, 1958) – see step (5) in Figure 4.2, implemented with 'nadam' optimizer. Different to `SimpleFrameId`, in `UniFrameId` the first hidden layer comprises 512 neurons, followed by 256 neurons in the second hidden layer. Other than `SimpleFrameId`, we do apply dropout to all hidden layers to prevent overfitting (Srivastava et al., 2014). Each node in the output layer corresponds to one frame-label class known from the lexicon – see step (6) in Figure 4.2. Same as `SimpleFrameId`, we use rectified linear units (Nair and Hinton, 2010, ReLU, common abbreviation) as activation function for the hidden layers, and a softmax activation function for the output layer yielding a multinomial distribution over frames. We take the highest activated neuron (arg max) to obtain the most likely frame label according to the

classifier as the final prediction at test time – see step (7) in Figure 4.2). Optionally, filtering based on the lexicon can be performed on the predicted probabilities for each frame label. The differences in hyperparameters of `UniFrameId` to `SimpleFrameId` are listed in Section 4.1.3.2.

4.1.3.2 Experimental Setup and Majority Baselines

We compare the performance of the `UniFrameId` system to the baselines and to previous work. We run the prediction ten times to reduce noise in the evaluation (cf. Reimers and Gurevych, 2017) and report the mean for each metric.

Hyperparameters. Using our experimental setup, Klie (2017) identifies the best hyperparameters based on the development set. He contrasts different architectures and finds the Multilayer Perceptron architecture to perform consistently better than a more complex Gated Recurrent Unit model (Cho et al., 2014). For this reason, we continue with the Multilayer Perceptron architecture and find that adding more than two hidden layers does not yield any improvement over two layers. Furthermore, using dropout on the hidden layers helps to increase the accuracy. Among the various input representations, a concatenation of the representations of context and predicate is the best amongst others, including dependencies, lexicon indicators, and part-of-speech tags. Training is done using Nesterov-accelerated Adam (nadam, Dozat, 2016) with default parameters. A batch size of 128 is used. Learning stops if the development accuracy has not improved for four epochs, and the learning rate is reduced by factor of two if there has not been any improvement for two epochs. The differences in hyperparameters of `UniFrameId` to `SimpleFrameId` are the following: here we use 'nadam' as optimizer instead of 'adagrad', furthermore we apply dropout on hidden layers and use early stopping to regularize training. Finally, the number of hidden units is different, as now it is optimized by grid search.

Majority baselines. We propose a new strong baseline combining two existing ones. These are: first, the most-frequent-sense baseline using the data majority (Data Baseline) to determine the most frequent frame for a predicate; second, the baseline introduced by Hartmann et al. (2017) using a lexicon (Lexicon Baseline) to consider the data counts of the Data Baseline only for those frames available for a predicate. We propose to combine them into a Data-Lexicon Baseline, which uses the lexicon for unambiguous predicates, and for ambiguous ones it uses the data majority. This way, we trust the lexicon for unambiguous predicates but not for ambiguous ones, there we rather consider the data majority. Comparing a system to these baselines helps to determine whether it just memorizes the data majority or the lexicon, or actually captures more.

4.1.3.3 Results and Discussion

We present the results of `UniFrameId` on English data (see Table 4.3). Here, the comparison with the baselines is of special interest in order to find out whether the trained system can actually contribute more than what can be achieved by counting majority occurrences as the baseline does.

model	with lexicon				without lexicon			
	acc	**acc amb**	**F1-m**	**F1-m amb**	**acc**	**acc amb**	**F1-m**	**F1-m amb**
Data Bsl	79.06	69.73	33.00	37.42	79.06	69.73	33.00	37.42
Lexicon Bsl	79.89	55.52	65.61	30.95	–	–	–	–
Data-Lexicon	86.32	69.73	64.54	37.42	–	–	–	–
`Hermann-14`	88.41	73.10	–	–	–	–	–	–
`SimpleFrameId`	87.63	73.80	–	–	77.49	–	–	–
`UniFrameId`	**88.66**	**74.92**	**76.65**	**53.86**	**79.96**	**71.70**	**57.07**	**47.40**
*(`UniFrameId`)	*89.35*	*76.45*	*77.76*	*55.24*	*80.36*	*73.38*	*58.21*	*49.05*

Table 4.3: FrameId results (in %) on English data with and without using the lexicon. Reported are **acc**uracy and **F1-m**acro, both also for **amb**iguous predicates (mean scores over ten runs). Best average results highlighted in bold. Models: *(a)* Data, Lexicon, and Data-Lexicon Baselines, *(b)* Previous models for English, *(c)* Ours: unimodal `UniFrameId`, *(d)* *(`UniFrameId`): maximum performance of best run, scores printed in italics.

Baseline Results. The new Data-Lexicon Baseline reaches a considerable accuracy of 86.32%, which is hard to beat by trained models. Even the most recent state-of-the-art system `Hermann-14` only beats it by about two points: 88.41% (Hermann et al., 2014) and `SimpleFrameId` can only slightly outperform it by about one point 87.63%. However, the accuracy of the baseline drops for ambiguous predicates (69.73%) and the F1-macro score reveals its weakness toward minority classes (drop from 64.54% to 37.42%). Furthermore, the new strong Data-Lexicon Baseline, by design, depends on the lexicon. Thus, in the setting without the lexicon, only the weaker Data Baseline can be considered, which shows a drastic drop in performance when evaluating with the F1-macro measurement.

Insights from Baseline. Many indicators point to our approach not just learning the data majority: the trained models have better F1-macro and especially much higher ambiguous F1-macro scores with the lexicon. This clearly suggests that the system `UniFrameId` is capable of acquiring more expressiveness than the baselines do by counting majorities. This advantage can be attributed to the textual context representation in `UniFrameId` that helps to disambiguate and prevents from only reproducing majority counts.

Unimodal Results. The unimodal system `UniFrameId` trained and evaluated on English data slightly exceeds the accuracy of the previous state of the art (88.66% on average versus 88.41% for Hermann et al., 2014); the best run's accuracy is 89.35%. Especially on ambiguous predicates, i.e. the difficult and therefore interesting cases, the average accuracy surpasses that of previous work by more than one point (the best run by almost three points). Considering the proposed F1-macro score (cf. Section 2.1.2.2) for an assessment of the performance on minority classes and ambiguous predicates reveals the main improvement: the system `UniFrameId` substantially outperforms the strong Data-Lexicon Baseline, demonstrating that `UniFrameId` differs from memorizing majorities and actually improves minority cases.

Significance Tests. We conduct a single sample t-test to judge the difference between the previous state-of-the-art accuracy (Hermann et al., 2014) and the unimodal approach with `UniFrameId`. The null hypothesis (expected value of the sample of ten accuracy scores equals the previous state-of-the-art accuracy) is rejected at a significance level of $\alpha = 0.05$ ($p = 0.0318$). In conclusion, the unimodal approach `UniFrameId` outperforms prior state of the art (`Hermann-14`) in terms of accuracy (and also its predecessor system `SimpleFrameId`).

Impact of Lexicon. We report results achieved without the lexicon to evaluate the system's performance independent of the lexicon quality (Hartmann et al., 2017). `UniFrameId` outperforms `SimpleFrameId` by more than two points in accuracy and achieves a large improvement over the Data Baseline. Comparing the F1-macro score with and without lexicon, it can be seen that the additional information stored in the lexicon strongly increases the score by about 20 points for English data.

4.1.4 Multilingual Evaluation – the case of German

After evaluating `UniFrameId` on English FrameNet data, we extend our system to the multilingual use case and aim at assessing the applicability to the German language. Note that there is a general lack of Frame Identification systems for languages other than English. This is problematic as different languages yield different challenges; German, for example, due to long distance dependencies. Furthermore, word embeddings trained on different languages have different strengths in ambiguous words due to differences in ambiguities between languages. We elaborate on insights from using different datasets by language.

Frame identification in German. `Shalmaneser` (Erk and Pado, 2006, short for SHALlow seMANtic parSER) is a toolbox for semantic role assignment on FrameNet schemata of English and German, with the German part integrated into the SALSA project. `Shalmaneser` uses a Naive Bayes classifier to identify frames, together with features for a bag-of-word context with a window over sentences, bigrams, and trigrams of the target word and dependency annotations. They report an F1-score of 75.1% on FrameNet 1.2 and of 60% on SALSA 1.0. These scores are difficult to compare against more recent work as the evaluation uses older versions of datasets and different splits. `Shalmaneser` requires software dependencies that are not available anymore, hindering evaluation on new data. To the best of our knowledge, there is no Frame Identification system evaluated on SALSA 2.0.

Johannsen et al. (2015) present a simple, but weak translation baseline for cross-lingual Frame Identification. A system based on `Semafor` is trained on English FrameNet and tested on German Wikipedia sentences, translated word-by-word to English. This translation baseline reaches an F1-score of 8.5% on the German sentences when translated to English. The performance of this weak translation baseline is worse than that of another simple baseline: a 'most frequent sense baseline' (basically the Data Majority Baseline) – computing majority votes for German (and many other languages) – reaches an F1-score of 53.0% on the German sentences. This shows that pure translation does not help with Frame Identification and, furthermore, indicates a large potential for improvement on Frame Identification in

languages other than English. Furthermore, all of the majority baselines we use (as presented in Section 4.1.3.2) strongly outperform the translation baseline of Johannsen et al. (2015) when training the system on English data and evaluating it on German data. This supports the development of systems trained on the specific target language that do not need to rely on translations.

4.1.4.1 Experimental Setup

We re-use the experimental setup of the English system with German data and German embeddings and compare the performance on English against German data.

Data and Data Splits: SALSA. The SALSA project (Burchardt et al., 2006; Rehbein et al., 2012) is a completed annotation project, which serves as the German counterpart to FrameNet. Its annotations are based on FrameNet up to version 1.2. SALSA adds proto-frames which are predicate-specific frames to properly annotate senses that have not yet been covered by the English FrameNet. For a more detailed description of differences between FrameNet and SALSA, see Ellsworth et al. (2004); Burchardt et al. (2009). SALSA also provides a lexicon (see Table 2.1 for statistics) and fully annotated texts. There are two releases of SALSA: 1.0 (Burchardt et al., 2006) as used for `Shalmaneser` (Erk and Pado, 2006, see Section 4.1.4), and the final release 2.0 (Rehbein et al., 2012), which contains more annotations and adds nouns as predicates. We use the final release.

SALSA has no standard evaluation split; Erk and Pado (2006) used an undocumented random split. Neither is it possible to transfer the splitting method of Das and Smith (2011), as the SALSA project distributions do not map to single documents. We suggest splitting based on sentences, i.e. all annotations of a sentence are in the same set to avoid mixing training and test sets. We assign sentences to 100 buckets based on their identifier-numbers and create a 70/15/15 split for training, development, and test sets based on the bucket order. This procedure allows future work to be evaluated on the same data[4]. Table 2.2 shows the dataset statistics.

Textual Input Embeddings. We use the 100-dimensional embeddings of Reimers et al. (2014) for German. Similar to `GloVe` embeddings, `Reimers` embeddings have been trained on Wikipedia (now German version) and on additional news text to cover more domains, resulting in similarly low out-of-vocabulary scores.

4.1.4.2 Results and Discussion – German versus English

As we re-use the experimental setup of `UniFrameId` with English data, we directly report the results on German data and then compare the performance on German data against English. The comparison of results obtained on German data and on English data manifests crucial differences in the two underlying lexica and datasets (or data splits), which we will discuss here. For the comparison, we repeat the results of `UniFrameId` for English data (see Table 4.4, top) and now we add those for German data (see Table 4.4, bottom).

[4] Our split is publicly available in *salsa_splits.txt* at: `https://public.ukp.informatik.tu-darmstadt.de/naacl18-multimodal-frame-identification/`

	model	with lexicon				without lexicon			
		acc	**acc amb**	**F1-m**	**F1-m amb**	**acc**	**acc amb**	**F1-m**	**F1-m amb**
FrameNet	Data Bsl	79.06	69.73	33.00	37.42	79.06	69.73	33.00	37.42
	Lexicon Bsl	79.89	55.52	65.61	30.95	–	–	–	–
	Data-Lexicon	86.32	69.73	64.54	37.42	–	–	–	–
	`Hermann-14`	88.41	73.10	–	–	–	–	–	–
	`SimpleFrameId`	87.63	73.80	–	–	77.49	–	–	–
	`UniFrameId`	**88.66**	**74.92**	**76.65**	**53.86**	**79.96**	**71.70**	**57.07**	**47.40**
	*(`UniFrameId`)	*89.35*	*76.45*	*77.76*	*55.24*	*80.36*	*73.38*	*58.21*	*49.05*
SALSA	Data Bsl	77.00	70.51	37.40	28.87	77.00	70.51	37.40	28.87
	Lexicon Bsl	61.57	52.5	19.36	15.68	–	–	–	–
	Data-Lexicon	77.16	70.51	38.48	28.87	–	–	–	–
	`UniFrameId`	**80.76**	**75.59**	**48.42**	**41.38**	**80.59**	**75.52**	**47.64**	**41.17**
	*(`UniFrameId`)	*80.99*	*76.00*	*49.40*	*42.55*	*80.80*	*75.90*	*48.60*	*42.23*

Table 4.4: FrameId results (in %) on English (upper) and German (lower) with and without using the lexicon. Reported are **acc**uracy and **F1-m**acro, both also for **amb**iguous predicates (mean scores over ten runs). Best average results highlighted in bold. Models: *(a)* Data, Lexicon, and Data-Lexicon Baselines. *(b)* Previous models for English. *(c)* Ours: unimodal `UniFrameId`, *(d)* *(`UniFrameId`): maximum performance of best run, scores printed in italics.

Unimodal Results versus Baseline Results for German Data. The system `UniFrameId` sets a new state of the art on the German corpus with 80.76% accuracy (the best run's accuracy is 80.99%), outperforming the baselines (77.16%; no other system evaluated on this dataset). The difference in the F1-macro score between the majority baselines and `UniFrameId` is smaller than for the English FrameNet. This indicates that the majorities learned from data are more powerful in the German case with SALSA than in the English case, when comparing against `UniFrameId`.

Impact of Lexicon: English versus German. For German data, the increase of the F1-macro score with lexicon versus without is small (one point). This indicates that, once having seen the training data, the lexicon is not needed to obtain the correct predictions on the test data. Even if both lexica approximately define the same number of frames (see Table 2.1), the number of defined lexical units (distinct predicate-frame combinations) in SALSA is smaller. This leads to a lexicon that is a magnitude smaller than the FrameNet lexicon. Thus, the initial situation for the German case is more difficult. The impact of the lexicon for SALSA is smaller than for FrameNet (best visible in the increase in the F1-macro score with using the lexicon compared to without), which can be explained by the larger percentage of ambiguous predicates (especially evoking proto-frames) and the smaller size of the lexicon. The evaluation on two different languages highlights the impact of an elaborate, manually created lexicon: it boosts the performance on frame classes that are less present in the training data. English Frame Identification benefits from the large high-quality lexicon, whereas German Frame Identification currently lacks a high-quality lexicon that is large enough to benefit the Frame Identification task.

	model	with lexicon				without lexicon			
		corr	**err uns**	**err unsLab**	**err normal**	**corr**	**err uns**	**err unsLab**	**err normal**
FN	`UniFrameId`	89.35	0.40	3.04	7.22	80.36	1.32	7.68	10.65
S	`UniFrameId`	80.99	0.49	0.97	17.54	80.80	0.49	1.10	17.61

Table 4.5: Error analysis of best unimodal systems on English (upper) and German (lower). Reported is the percentage of predictions in each category. Categories for predictions are: **corr**ect predictions versus **e**rroneous predictions. Prediction errors can occur for predicates which were **uns**een during training, for predicates which were **uns**een with the target **lab**el, or they can be a **n**ormal classification error.

Dataset Properties: English versus German. To better understand the influence of the dataset on the prediction errors, we further analyze the errors of our approach (see Table 4.5) following Palmer and Sporleder (2010). A wrong prediction can either be a normal classification error, or it can be the result of an instance that was unseen at training time, which means that the error is due to the training set. The instance can either be completely unseen or unseen with the target label. We observe that FrameNet has more problems with unseen data compared to SALSA, especially data that was unseen with one specific label but seen with another label. This is due to the uneven split of the documents in FrameNet, leading to data from different source documents and domains in the training and test split. SALSA does not suffer from this problem as much since the split was performed differently. It would be worth considering the same splitting method for FrameNet.

Comments on Full Semantic Role Labeling. Kabbach et al. (2018) pose the question of the true state of the art in full Semantic Role Labeling when ensuring equal preprocessing steps. According to their analysis, `Semafor` is still stronger compared to `Open-SESAME` in full Semantic Role Labeling when observed under fixed experimental settings. For this, Kabbach et al. (2018) use our previous Frame Identification system (`SimpleFrameId`, which we later extended to our current state-of-the-art system `UniFrameId`), reproduce our previous results on Frame Identification and use it in their pipeline for full Semantic Role Labeling.

We do not focus on full Semantic Role Labeling in this thesis, however, we report on some explorations for determining the difficulties in full Semantic Role Labeling for English and German. Building up on identified frames, Markard (2018) finds that after Frame Identification, there is a further bottleneck in full Semantic Role Labeling, which is not the actual role label assignment, but the identification of the correct span for roles. Interestingly, this seems to be more straight-forward for German compared to English. In both languages, full Semantic Role Labeling highly profits from frame knowledge compared to role labeling without knowing the frames.

4.2 Grounded Frame Identification: Combining Textual with Visual Embeddings

In this section, we present and discuss our contributions and findings in the context of multimodal language understanding with frame semantics, where we complement textual knowledge about situations and actions with visual word embeddings for the task of Frame Identification. We extend our unimodal Frame Identification system to a use-case with multimodal embeddings by grounding in images of entities (cf. Section 4.2.1). The basic hypothesis of this section is the following:

> HYPO: *Frame Identification requires implicit commonsense knowledge which is rarely expressed textually but can be extracted from images.*

We connect our hypothesis to an example of sentences in the FrameNet dataset to showcase how implicit commonsense knowledge is obvious enough to be rarely expressed in sentences, but is more likely to be present in images. Figure 4.3 takes the ambiguous predicate *sit* to illustrate how images can provide access to implicit commonsense knowledge crucial to Frame Identification: *'people can sit back on a bench, but companies cannot', 'companies are built in cities'* (also see Bruni et al. (2014)). In particular, we analyze the performance of our Frame Identification system for difficult cases where we find visual information to be beneficial. Furthermore, we study whether the improvement is due to the visual modality, we alternatively extend the textual embeddings with random embeddings and with structure-enhanced embeddings (cf. Section 4.2.2). Finally, we draw a comparison between performance on English and on German data (cf. Section 4.2.3).

Taken together, this chapter implements a broader view that includes the visual modality in addition to textual and structured approaches. In an outlook, we explore the textual and visual grounding of highly embodied verbs (cf. Section 4.2.4.1) and suggest to develop multimodal embeddings for verbs specifically that incorporate sensomotoric information. Our papers (Botschen et al., 2018a)[5] and (Beinborn et al., 2018)[6] are foundational to this chapter.

Participant Knowledge Complements Knowledge about Situations and Actions. Situational background knowledge can be described in terms of frames (Fillmore, 1985) and scripts (Schank and Abelson, 2013). Whilst we focus on frames with the lexical-semantic knowledge base FrameNet, scripts are also an event-centered form of structured world knowledge to model human knowledge structures. To give an example, Schank and Abelson (2013) expound the *Restaurant Script* where roles (e.g., *customer, waiter, cook, owner*) interact (with e.g., *menu, food, money*) and thereby change the entry conditions (e.g., *customer is hungry and has money*) into the results (e.g., *customer is less hungry and has less money*, *owner has more money*).

[5] My contributions in this paper are the following: `UniFrameId` system for FrameNet and SALSA, `MultiFrameId` system and analysis of experiments.

[6] My contribution in this paper is the following: exploration of verb embeddings

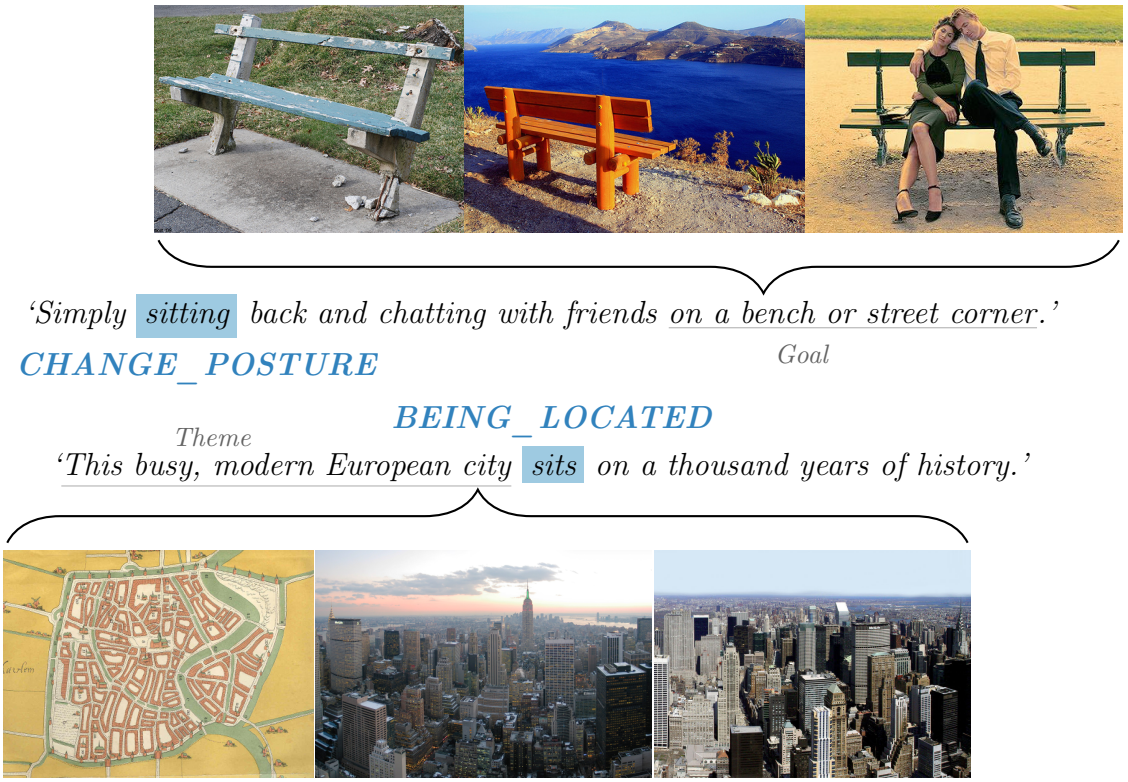


Figure 4.3: Example sentences demonstrating the potential benefit of visual knowledge inferred from images in order to disambiguate ambiguous predicates (here *'sit'*). As an example, the pictures of a bench or a city might help to infer that the frame *'Change_posture'* is more likely to occur in the context of furniture that looks like a bench or chair, whilst the frame *'Being_located'* is more likely to occur in the context of cities or locations that look similar.

When looking at the semantics of situations and actions, Frame Identification has commonalities with event prediction tasks which aim at linking events and their participants to script knowledge and at predicting events (or situations and actions) in narrative chains. Ahrendt and Demberg (2016) report that knowing about a script's participants (e.g., *customer, waiter, cook*) aids in predicting events (e.g., *Restaurant Script*) linked to script knowledge. This finding suggests the need for implicit context knowledge about participants also for Frame Identification for ambiguous predicates. Addressing ambiguous predicates where participants have different properties depending on the context, Feizabadi and Padó (2012) give some examples where location plays a discriminating role as participant: motion verbs that have both a concrete motion sense and a more abstract sense in the cognitive domain, e.g., *struggle*, *lean*, *follow*.

Transferring these insights to Frame Identification, we assume that a rich context representation including information about participants helps to identify the sense of ambiguous predicates. This specifically applies to images, which can reflect properties of the participants of a situation in an inherently different way, see Figure 4.3. There are several established corpora containing images of entities (e.g., objects) from web crawls. The more concrete an entity is, the more obvious and similar to each other are the images.

(a) KEY.N.01, definition: 'metal device shaped in such a way that when it is inserted into the appropriate lock the lock's mechanism can be rotated'.

(b) KEY.N.15, definition: 'a lever (as in a keyboard) that actuates a mechanism when depressed'.

Figure 4.4: ImageNet images for different senses of WordNet noun synset 'key'.

However, same as with predicates, there are also ambiguous entities, e.g., *'key'* – the tool to open and lock a door and also the pushbuttons on a keyboard. Thus, it is not trivial to assign images to nouns. This is why we will use images for noun synsets. To continue the above example, the noun *'key'* can occur in several synsets, see Figure 4.4: For each noun synset, a visual embedding is learned out of a collection of images showing exactly this sense of the noun – where different approaches will be explained in Section 4.2.1.1.

4.2.1 Multimodal Frame Identification System `MultiFrameId`

Whilst current Frame Identification methods rely only on textual representations (including our state-of-the-art system `UniFrameId` as in Section 4.1.3), we hypothesize that Frame Identification can profit from a richer understanding of the situational context. Such contextual information can be obtained from commonsense knowledge, which is richly represented in images. We examine whether multimodal representations grounded in images can encode commonsense knowledge to improve Frame Identification. Thus, we extend our unimodal Frame Identification system `UniFrameId` in order to effectively leverage multimodal representations and develop the multimodal Frame Identification system `MultiFrameId`. Regarding Frame Identification, to the best of our knowledge, multimodal approaches have not yet been investigated. We aim to uncover whether representations that are grounded in images can help to improve the accuracy of Frame Identification.

We extend the representation of the predicate context $\overrightarrow{v_{(in)}}$ (so far as given in Equation 4.1) with multimodal embeddings. Furthermore, we assess the applicability to another language, namely German – following our cross-lingual evaluation setup as introduced in Section 4.1.4 with `UniFrameId`.

4.2.1.1 Architecture and Multimodal Input Embeddings

The multimodal architecture for the system `MultiFrameId` is analogous to the unimodal architecture for `UniFrameId` as described in Section 4.1.3.

Multimodal Pipeline. Different from `UniFrameId`, in `MultiFrameId` the representation of the predicate context now is multimodal (Equation 4.2). We extend Equation 4.1 beyond textual embeddings ($\overrightarrow{v_{m_1}}$) by also using visual ($\overrightarrow{v_{m_2}}$) and IMAGINED embeddings ($\overrightarrow{v_{m_3}}$), which are explained in the next paragraph:

$$\overrightarrow{v_{mm(in)}} = \overrightarrow{v_{m_1(cont)}} \frown \overrightarrow{v_{m_2(cont)}} \frown \overrightarrow{v_{m_3(cont)}} \frown \overrightarrow{v_{m_1(pred)}} . \tag{4.2}$$

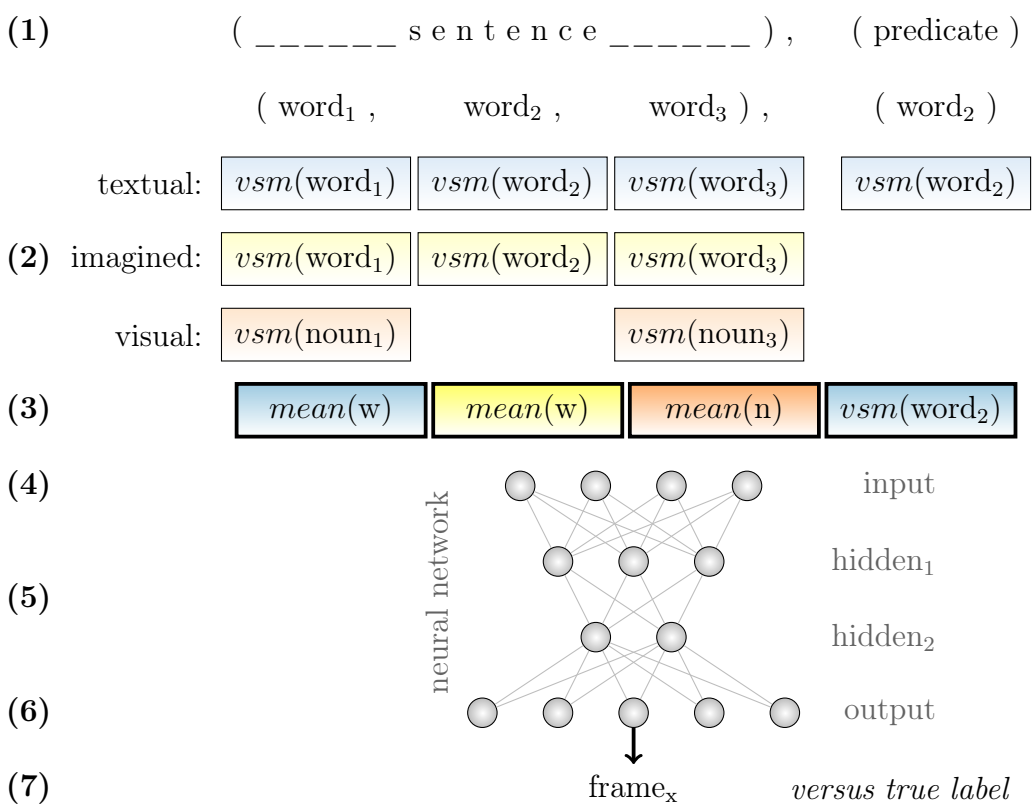


Figure 4.5: Sketch of the pipeline. (1) Input data: sentence with predicate, (2) Mapping: words to embeddings, (3) Input representation: concatenation of modality-specific means, (4-6) Classifier: neural network, (7) Prediction of frame.

More precisely, we concatenate all single modalities' representations of the predicate context, which in turn are the single modalities' mean embeddings of all words in the sentence. We use concatenation $\frown$ (cf. Equation 3.19) for fusing the different embeddings as it is the simplest yet successful fusion approach (Bruni et al., 2014; Kiela and Bottou, 2014); see Section 3.5 for the description of multimodal fusion. Figure 4.5 provides a sketch of the `MultiFrameId` system pipeline extending the `UniFrameId` system pipeline (Figure 4.2). Now, given the multimodal input representation, we use the same Multilayer Perceptron architecture as for the unimodal input representation in Section 4.1.3 where we adapt the number of nodes to the increased input size.

Visual Input Embeddings. The preparation of textual embeddings for words was explained in Section 4.1.3 with `UniFrameId`. Now, we explain the preparation of visual embeddings for synsets and of Imagined embeddings for words so that, in combination, we can represent multiple modalities.

Visual Embeddings for Synsets. We obtain visual embeddings for WordNet synsets (Miller, 1995; Fellbaum, 2000): we apply the pre-trained `VGG` Convolutional Neural Network model (Chatfield et al., 2014, cf. Section 3.4) to images for synsets from ImageNet (Deng et al., 2009), we extract the 128-dimensional activation of

the second last layer and then we L_2-normalize it. We use the images of the WN9-IMG dataset (Xie et al., 2017), which links WordNet synsets to a collection of ten ImageNet images. Examples for ImageNet images for different synsets of the same noun are depicted in Figure 4.4. We average the embeddings of all images corresponding to a synset, leading to a vocabulary size of 6,555 synsets. All synsets in WN9-IMG are part of triples of the form *entity-relation-entity*, i.e. *synset-relation-synset*. Such synset entities that are participants of relations with other synset entities are candidates for incorporating the role fillers for predicates and, therefore, may help to find the correct frame for a predicate (see Section 4.2.1.2 for details about sense-disambiguation for synsets.)

Imagined Embeddings for Words. We use the Imagined method (Collell et al., 2017) for learning a mapping function (which we categorize as a typical 'Mapping' approach in Section 3.5 for learning multimodal representations): it maps from the word embedding space to the visual embedding space given those words that occur in both pre-trained embedding spaces (7,220 for English and 7,739 for German). To obtain the English synset lemmas, we extract all lemmas of a synset and keep those that are nouns. We automatically translate English nouns to German nouns using the Google Translate API (Application Programming Interface) to obtain the corresponding German synset lemmas. The Imagined method is promising for cases where one embedding space (here, the textual one) has many instances without correspondence in the other embedding space (here, the visual one), but the user still aims at obtaining instances of the first in the second space. We aim to obtain visual correspondences for the textual embeddings in order to incorporate regularities from images into the system `MultiFrameId`. The mapping is a nonlinear transformation using a simple neural network. The objective is to minimize the cosine distance d_{cos} (cf. Equation 3.2) between each mapped representation of a word and the corresponding visual representation. Finally, a multimodal representation for any word can be obtained by applying this mapping to the word embedding.

Our application case of Frame Identification is more complex than the original setting of Collell et al. (2017), which is a comparison on the word-pair level, as we consider a whole sentence in order to identify the predicate's frame. However, we see some potential for multimodal Imagined embeddings to help: their mapping from text to multimodal representations is learned from images for nouns. Such nouns, in turn, are candidates for role fillers of predicates. In order to identify the correct sense of an ambiguous predicate, it could help to enrich the representation of the context situation with multimodal embeddings for the entities that are linked by the predicate.

4.2.1.2 Experimental Setup

The multimodal setup for the `MultiFrameId` system is analogous to the unimodal setup for the `UniFrameId` system as described in Section 4.1.3. Now, we contrast the performance of `UniFrameId` for context representations based on unimodal (textual) against the performance of `MultiFrameId` for context representations based on multimodal (combinations of textual and visual) embeddings.

		sentences	frames	reduced sentences synsVis
FrameNet	train	2819	15406	1310
	dev	707	4593	320
	test	2420	4546	913
SALSA	train	16852	26081	4707
	dev	3561	5533	1063
	test	3605	5660	1032

Table 4.6: Dataset statistics for FrameNet 1.5 corpus of fully annotated texts with split by Das et al. and for SALSA 2.0 with our split: number of **sentences** and **frames** (as used in our experiments). Rightmost column: number of sentences when reduced to only those having synsets in the visual embeddings.

Preprocessing: Synsets in FrameNet and SALSA. To prepare the datasets for working with the synset embeddings, we sense-disambiguate all sentences using the API (Application Programming Interface) of BabelNet (Navigli and Ponzetto, 2012b), which returns multilingual synsets. BabelNet is a multilingual semantic network covering lexicographic and encyclopedic knowledge from WordNet (Miller, 1990; Fellbaum, 1990) and Wikipedia and its performance on standard word sense disambiguation is state of the art in three different SemEval evaluation tasks (Navigli and Ponzetto, 2012a, 2010). We thus depend on the performance of BabelNet when using synset embeddings on sense-disambiguated sentences. However, this dependence does not hold when applying Imagined embeddings to sentences, as the mapping from words to Imagined embeddings does not need any synset labeled in the sentences. After sense-disambiguation some sentences do not contain any synset available in our synset embeddings. The statistics of those sentences that have at least one visual synset embedding is given in Table 4.6 (extending Table 2.2).

4.2.1.3 Results and Discussion

We report the multimodal results in contrast to unimodal results of Section 4.1.3.2 by extending Table 4.4 to the full Table 4.7.

The multimodal system `MultiFrameId` slightly exceeds the overall accuracy of the unimodal state of the art: 88.82% on average versus 88.66% for `UniFrameId`; the best run's accuracy is 89.09%. Interestingly, we observe the advantage of the multimodal approach to be more pronounced in difficult setups: rare and ambiguous cases and without lexicon. We elaborate on this in the following.

Multimodal Results. The most helpful context representations is the concatenation of Imagined embeddings and visual synset embeddings. This is determined experimentally by contrasting a range of multimodal context representations as extensions to `MultiFrameId`. The concatenation of Imagined embeddings and visual synset embeddings outperforms the unimodal approach slightly in all measurements. We observe that the improvements are more pronounced for difficult cases, such as for rare and ambiguous cases (one point improvement in F1-macro), as well as in the absence of a lexicon (up to two points improvement).

	model	with lexicon				without lexicon			
		acc	**acc amb**	**F1-m**	**F1-m amb**	**acc**	**acc amb**	**F1-m**	**F1-m amb**
FrameNet	Data Bsl	79.06	69.73	33.00	37.42	79.06	69.73	33.00	37.42
	Lexicon Bsl	79.89	55.52	65.61	30.95	–	–	–	–
	Data-Lexicon	86.32	69.73	64.54	37.42	–	–	–	–
	Hermann-14	88.41	73.10	–	–	–	–	–	–
	SimpleFrameId	87.63	73.80	–	–	77.49	–	–	–
	UniFrameId	88.66	74.92	76.65	53.86	79.96	71.70	57.07	47.40
	MultiFrameId	**88.82**	**75.28**	**76.77**	**54.80**	**81.21**	**72.51**	**57.81**	**49.38**
	*(UniFrameId)	*89.35*	*76.45*	*77.76*	*55.24*	*80.36*	*73.38*	*58.21*	*49.05*
	*(MultiFrameId)	*89.09*	*75.86*	*78.17*	*57.48*	*81.57*	*73.29*	*58.63*	*51.29*
SALSA	Data Bsl	77.00	70.51	37.40	28.87	77.00	70.51	37.40	28.87
	Lexicon Bsl	61.57	52.5	19.36	15.68	–	–	–	–
	Data-Lexicon	77.16	70.51	38.48	28.87	–	–	–	–
	UniFrameId	**80.76**	**75.59**	**48.42**	**41.38**	**80.59**	**75.52**	**47.64**	**41.17**
	MultiFrameId	80.71	75.58	48.29	41.19	80.51	75.51	47.36	40.93
	*(UniFrameId)	*80.99*	*76.00*	*49.40*	*42.55*	*80.80*	*75.90*	*48.60*	*42.23*
	*(MultiFrameId)	*80.99*	*75.95*	*49.91*	*43.34*	*80.78*	*75.85*	*49.11*	*42.88*

Table 4.7: FrameId results (in %) on English (upper) and German (lower) with and without using the lexicon. Reported are **acc**uracy and **F1-m**acro, both also for **amb**iguous predicates (mean scores over ten runs). Best average results highlighted in bold. Models: *(a)* Data, Lexicon, and Data-Lexicon Baselines. *(b)* Previous models for English. *(c)* Ours: unimodal `UniFrameId`, multimodal on top of `UniFrameId` – `MultiFrameId` – with Imagined embeddings (and synset visual embeddings for English). *(d)* *(`MultiFrameId`): maximum performance of best run, scores printed in italics, with best marked in red.

Furthermore, when comparing average performance and best performance of the unimodal and the multimodal approach, `MultiFrameId` is strongest with respect to the average whilst `UniFrameId`, with respect to the best run, reaches highest accuracy in some cases. This is an effect of `UniFrameId` having larger variance in the results, whilst `MultiFrameId` is more robust. Interestingly, `MultiFrameId` consistently reaches best performance with respect to F1-macro, even when looking at the best run's results. This shows the strength of `MultiFrameId` with respect to rare frame classes, in particular.

Significance tests. To judge the difference between the unimodal and the multimodal approach, we conduct a t-test for the means of the two independent samples. The null hypothesis states identical expected values for our two samples of ten accuracy scores. Regarding the setting with lexicon, the null hypothesis cannot be rejected at a significance level of $\alpha = 0.05$ ($p = 0.2181$). However, concerning accuracy scores without using the lexicon, the null hypothesis is rejected at a significance level of $\alpha = 0.05$ ($p < 0.0001$). In conclusion, the multimodal approach has a slight overall advantage and, interestingly, has a considerable advantage over the unimodal one when confronted with a more difficult setting of not using the lexicon.

Impact of Multimodal Representations. Multimodal context representations improve the results compared to unimodal ones. It helps to incorporate visual commonsense knowledge about the situation's participants. Referring back to the example of the ambiguous predicate *sit*, the multimodal approach is able to transfer the knowledge to the test sentence *'Al-Anbar in general, and Ramadi in particular, are set with the Americans in Jordan.'* by correctly identifying the frame *Being_located* whilst the unimodal approach fails with predicting *Change_posture*. The increase in performance when adding information from visual synset embeddings is not simply due to higher dimensionality of the embedding space. To verify this, we further investigate extending the unimodal system (`UniFrameId`) with random word embeddings. This leads to a drop in performance compared to using just the unimodal representations or using these in combination with the proposed multimodal embeddings, especially in the setting without lexicon.

Future Work. As stated previously, Frame Identification has commonalities with event prediction. Since identifying frames is only one way of capturing events or situations and actions, the approach is transferable to other schemes of event prediction and visual knowledge about participants of situations should be beneficial there, too. It would be interesting to evaluate the multimodal architecture on other predicate-argument frameworks, e.g., script knowledge or VerbNet style Semantic Role Labeling. In particular the exploration of our findings on visual contributions to Frame Identification in the context of further event prediction tasks may form an interesting next step.

More precisely, future work should consider using implicit knowledge not only from images of the participants of the situation, but also from the entire scene in order to directly capture relations between the participants. This could provide access to a more holistic understanding of the scene. The following visual tasks with accompanying datasets could serve as a starting point: *(a)* visual Verb Sense Disambiguation with the VerSe dataset (Gella et al., 2016) and *(b)* visual SRL with several datasets, e.g., imSitu (Yatskar et al., 2016, linked to FrameNet), V-COCO (Gupta and Malik, 2015, verbs linked to COCO), VVN (Ronchi and Perona, 2015, visual VerbNet) or even SRL grounded in video clips for the cooking-domain (Yang et al., 2016) and visual Situation Recognition (Mallya and Lazebnik, 2017). Such datasets could be used for extracting visual embeddings for verbs or even complex situations in order to improve the visual component in the embeddings for `MultiFrameId`. Vice versa: visual tasks could profit from multimodal approaches (Baltrušaitis et al., 2018) in a similar sense as the textual task of Frame Identification profits from additional information encoded in further modalities. Moreover, visual SRL might profit from the multimodal system `MultiFrameId` to a similar extent as any FrameNet SRL task profits from correctly identified frames (Hartmann et al., 2017).

Regarding the combination of embeddings from different modalities, we suggest to experiment with different fusion strategies complementing the middle fusion (concatenation) and the mapping (Imagined method). This could be a late fusion at decision level operating like an ensemble.

Conclusion. We investigated multimodal representations for Frame Identification by incorporating implicit knowledge, which is better represented in the visual domain. We presented `MultiFrameId`, a flexible Frame Identification system that is independent of modality and language in its architecture. With this flexibility, it is possible to include textual and visual knowledge and to evaluate on gold data in different languages. We created multimodal representations from textual and visual domains and showed that for English FrameNet data, enriching the textual representations with multimodal ones improves the accuracy toward a new state of the art. For German SALSA data, we set a new state of the art with textual representations only and discuss why incorporating multimodal information is more difficult. For both datasets, `MultiFrameId` is particularly strong with respect to ambiguous and rare classes, considerably outperforming the new Data-Lexicon Baseline and thus addressing a key challenge in Frame Identification.

4.2.2 Alternatives to Visual Embeddings

In a strict sense, 'multimodal' approaches refer to the combination of different human sensory input channels. In a broader sense, 'multimodality' can also include any combination of different input channels, independent of the human sensoric repertory. Thus, structured knowledge can also be regarded as a 'modality' that is to combine or to contrast with other modalities. We contrast the visual modalitiy with the structured modality with respect to the performance in the Frame Identification task, when added to the textual embeddings.

After confirming that simply increasing the dimensionality by adding random embeddings instead of visual embeddings does not yield any improvement, it is still to be found out whether *any* information in the added dimensions leads to improvement or especially *visual* information is helpful. We contrast visual information with structured information.

Structure-enhanced Input Embeddings for Synsets. We obtain 300-dimensional linguistic structure-enhanced synset embeddings in the following way. We apply the `AutoExtend` approach (Rothe and Schütze, 2015) to `GloVe` embeddings and produce synset embeddings for all synsets having at least one synset lemma in the `GloVe` embeddings. This leads to a synset vocabulary size of 79,14. `AutoExtend` builds linguistic synset embeddings based on textual word embeddings by enriching them with the synset information contained in the knowledge base WordNet (Miller, 1990; Fellbaum, 1990). These structure-enhanced embeddings for synsets are an alternative to the visual synset embeddings and incorporate rather linguistic information the visual commonsense knowledge. The statistics of those sentences that have at least one synset embedding with the `AutoExtend` approach is given in Table 4.8 (extending Table 4.6).

Impact of Multimodal Representations: Visual versus Structure-enhanced. Interestingly, replacing visual synset embeddings with linguistic synset embeddings (`AutoExtend` by Rothe and Schütze (2015), see above) in further investigations also showed that visual embeddings yield better performance. This points out the potential for incorporating even more image evidence to extend our approach.

				reduced sentences	
		sentences	**frames**	**synsVis**	**synsAutoExt**
FrameNet	train	2819	15406	1310	2714
	dev	707	4593	320	701
	test	2420	4546	913	2318
SALSA	train	16852	26081	4707	16736
	dev	3561	5533	1063	3540
	test	3605	5660	1032	3570

Table 4.8: Dataset statistics for FrameNet 1.5 corpus of fully annotated texts with split by Das et al. and for SALSA 2.0 with our split: number of **sentences** and **frames** (as used in our experiments). Rightmost: number of sentences when reduced to only those having synsets in the visual and in the linguistic structure-enhanced AutoExtend embeddings.

	model	**with lexicon**				**without lexicon**			
		corr	**err uns**	**err unsLab**	**err normal**	**corr**	**err uns**	**err unsLab**	**err normal**
FN	`UniFrameId`	89.35	0.40	3.04	7.22	80.36	1.32	7.68	10.65
	`MultiFrameId`	89.79	0.58	3.55	6.08	80.63	1.91	8.50	8.96
SALSA	`UniFrameId`	80.99	0.49	0.97	17.54	80.80	0.49	1.10	17.61
	`MultiFrameId`	81.24	1.94	1.88	14.94	80.96	1.94	2.05	15.05

Table 4.9: Error analysis of best uni- and multimodal systems on English (upper) and German (lower). Reported is the percentage of predictions in each category. Categories for predictions are: **correct** predictions versus **e**rroneous predictions. Prediction errors can occur for predicates which were **uns**een during training, for predicates which were **uns**een with the target **lab**el, or they can be a **n**ormal classification error.

From this we conclude that for language understanding in terms of situations and actions, visual information is helpful and that the approach we tried for including structured information cannot compete.

4.2.3 Multilingual Evaluation – the case of German

Analogous to the evaluation of `UniFrameId` on English FrameNet data and on German SALSA data (cf. Section 4.1.4), here we evaluate `MultiFrameId` in the same multilingual setup. However, for the German case with the SALSA dataset, the multimodal context representations cannot show an improvement over the unimodal ones.

Dataset Properties: English versus German. We extend Table 4.5 of Section 4.1.3 with the error analysis of the `MultiFrameId` system in Table 4.9. Our observations with `UniFrameId` are confirmed with `MultiFrameId`: again, we observe that FrameNet has larger issues with unseen data compared to SALSA.

Difficulties for German Data. The impact of multimodal context representations is more difficult to interpret for the German dataset. The fact that they have not helped here may be due to mismatches when translating the English nouns of a synset to German in order to train the Imagined embeddings. Here, we see room for future work to improve on simple translation by sense-based translations. In SALSA, a smaller portion of sentences has at least one synset embedding, see Table 2.2. For further investigations, we reduced the dataset to sentences actually containing at least one synset embedding. Then, minor improvements of the multimodal approach were visible for SALSA. This points out that a dataset containing more words linking to implicit knowledge in images (visual synset embeddings) can profit more from visual and Imagined embeddings.

The analysis shows that for the German data, textual representations are still competitive with multimodal ones. However, concerning the English data, the multimodal Frame Identification approach outperforms its unimodal counterpart, setting a new state of the art. Its benefits are particularly apparent in dealing with ambiguous and rare instances, the main source of errors of current systems.

4.2.4 Recommendation for Grounded Frame Identification

In the previous Section 4.2 we have seen that visual information in terms of embeddings for noun synsets is useful to Frame Identification especially in difficult settings with rare frames. Thus, visual embeddings complement textual ones for concrete nouns, but it is questionable whether the visual domain beneficially adds information to abstract nouns, verbs, or stop words.

For the interpretation of sequences, it is fundamental to include verbs and their arguments into methods for multimodal representations and into evaluation (Beinborn et al., 2018). The *imSitu* dataset (Yatskar et al., 2016) addresses this need by collecting images of verbs, i.e. situations where actions are performed, and also by providing annotations which link the verb arguments to visual reference in the image. This dataset is used for the task of multimodal situation recognition (Mallya and Lazebnik, 2017; Zellers and Choi, 2017). Grounding verbs is particularly challenging because of the variety of their possible visual instantiations (e.g., an image of an adult drinking beer has very little in common with a zebra drinking water – even if in both the action of drinking is taking place).

Next, we explore the textual and visual grounding of highly embodied verbs and suggest to develop multimodal embeddings for action- and motion-verbs specifically that incorporate sensomotoric information. Our paper (Beinborn et al., 2018)[7] is foundational to this exploration.

4.2.4.1 Excursion – Multimodal Grounding of Verbs

In order to build a broader understanding of the situation or the action that is described in a sentence, it might not be enough to depict the participants (cf. Section 4.2.1 for noun synsets), but it might be of help to depict the situation or action itself. Verbs play a fundamental role for expressing relations between concepts and their situational functionality (Hartshorne et al., 2014). The dynamic nature of verbs

[7] My contribution in this paper is the following: exploration of verb embeddings

poses a challenge for multimodal grounding. To the best of our knowledge, only Hill et al. (2014) and Collell et al. (2017) consider verbs when evaluating multimodal embeddings. However, they only report that results for verbs are significantly worse than for nouns, but they do not further elaborate on this finding. Most multimodal research to date focuses on the representation of individual concepts (nouns) and their properties (adjectives). The benefit of multimodal embeddings for language tasks going beyond concept similarity needs to be examined in more detail from both, engineering and theoretical perspectives.

Here, we analyze the potential of using images of verbs when learning multimodal embeddings and we present first steps towards an investigation of verb grounding. For this, images of verbs, as the unit that carries the situation or action, are crucial. However, verbs are more difficult to grasp in pictures than entities are, as they involve several participants; thus they depict the interaction of several objects.

Embodiment of Verbs. From a cognitive perspective, verbs can be categorized according to their degree of embodiment. While some verbs directly refer to body movements (e.g., dance), others are connected to physical activity. A measurement of embodiment for verbs indicates to which extent verb meanings involve bodily experience (Sidhu et al., 2014). Sidhu et al. (2014) collect embodiment ratings for 687 English verbs and infer that bodily experience plays an important role to the meanings of some verbs (e.g., *dance, breathe*) when compared to other verbs (e.g., *evaporate, expect*).

Setup for Verb Similarity with Embodied Verbs. We present first steps towards an investigation of verb grounding in terms of multimodal embeddings for verbs. We hypothesize that considering multimodal verb embeddings and human similarity ratings for verb pairs, highly embodied verbs (e.g., dance) yield a higher agreement when compared to all verbs. This means, we measure to what extent the multimodal verb embeddings mirror the human verb similarity ratings. Ideally, the more (less) a verb pair is judged as similar by humans, the closer (farther apart) it should be in embeddings space.

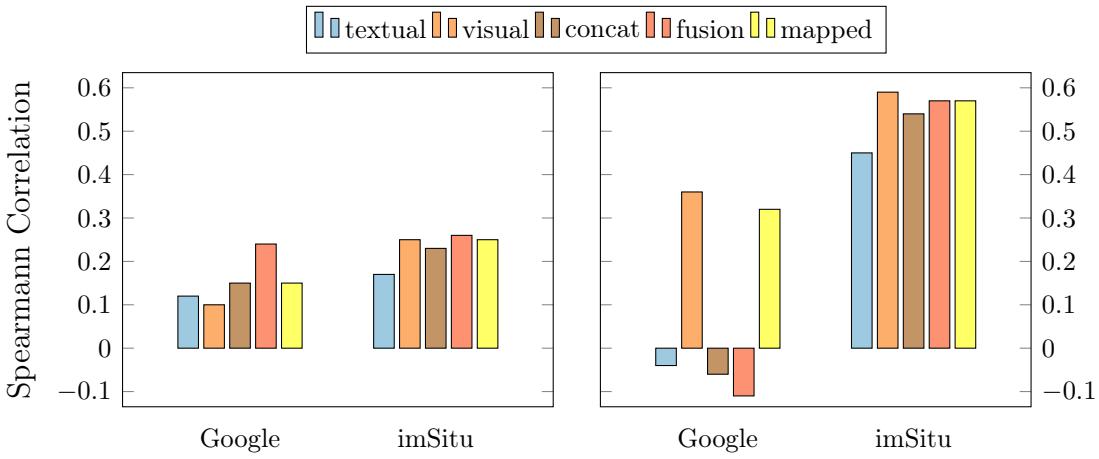
In line with previous work, the quality of the representations is evaluated as the Spearman rank correlation coefficient $r_s(X, Y)$ (cf. Equation 3.18) between the cosine similarity s_{cos} (cf. Equation 3.2) of two verb embeddings and their corresponding human similarity rating in the *SimVerb* dataset (Gerz et al., 2016). We obtain embodiment ratings for 1163 verb pairs.[8] The class *high embodiment* contains pairs like *fall-dive* in which the embodiment of both verbs can be found in the highest quartile (135 pairs), *low embodiment* contains pairs with embodiment ratings in the lowest quartile (81 pairs) like *know-decide*.[9]

Figure 4.6 illustrates the quality of verb representations in the most common publicly available approaches for multimodal representations.[10] The quality of the

[8] `https://psychology.ucalgary.ca/languageprocessing/node/22`. We only include a pair, if an embodiment rating is available for both verbs.

[9] It should be noted that not all instances of the two classes are covered by the visual representations. The small number of instances might have an impact on the correlation values.

[10] The pre-trained embeddings and the script to reproduce our results are available for research purposes: https://github.com/UKPLab/coling18-multimodalSurvey.



(a) All verb pairs.

(b) Highly embodied verb pairs.

Figure 4.6: Illustration for the quality of verb representations indicated as Spearman correlation between the cosine similarity of verbs and their corresponding similarity rating in the *SimVerb* dataset. Images for verbs are contrasted for the image resource of the Google and the imSitu datasets.

representations is evaluated in terms of the correlation as described above.

We compare the quality of 3498 verb pairs[11] in textual `GloVe` representations (Pennington et al., 2014) and two visual datasets: the Google dataset that performed best in Kiela et al. (2016) and has the highest coverage for the verb pairs (493 pairs, 14%)[12] and the *imSitu* dataset which has been intentionally designed for verb identification (354 pairs, 10%).

The results show that models which include visual information outperform purely textual representations for known concepts. However, the general quality of the verb representations is much lower than the quality reported for nouns. As a consequence, the mapping to unseen verb pairs yields unsatisfactory results for the full *SimVerb* dataset. Our encouraging results for the *imSitu* dataset indicate that it is recommended to directly obtain visual representations for verbs instead of projecting the meaning. Building larger multimodal datasets with a focus on verbs seems to be a promising strand of research for future work.

Coherent with previous work on concrete and abstract nouns (Hill et al., 2014), it can be seen that visual representations better capture the similarity of verbs with a high level of embodiment. The mapped representations maintain this sense of embodiment, whereas the concatenated and fused representations better capture the similarity for verbs referring to more conceptual actions. This finding indicates that multimodal information is not equally beneficial for all words.

[11] Two pairs had to be excluded because *misspend* was not covered in the textual representations.
[12] The coverage in *WN9-IMG* (Xie et al., 2017) and the dataset used by Collell et al. (2017) is lower.

4.3 Summary of the Chapter

We provide a summary of this chapter in terms of bullet points in the box below:

INSIGHTS IN A NUTSHELL

For language understanding with frame semantics for situations and actions, we develop the **Frame Identification** system, setting a new state of the art.

- We apply the notion of context to represent meaning in Frame Identification by using textual embeddings for all words in the sentence.
- `SimpleFrameId`: strongest performance with neural approach, frame embeddings as a by-product of the matrix factorization approach
- `UniFrameId`: optimized architecture achieving state of the art
- multilingual evaluation for English versus German
  - FrameNet lexicon: benefits English more than German
  - FrameNet test data: more difficult for English than for German
- strongest baseline has difficulties with rare ambiguous predicates
- We apply visual commonsense knowledge to represent meaning in Frame Identification by using visual embeddings for noun synsets.
- `MultiFrameId` benefits from knowledge about participants
- For incorporating commonsense knowledge about participants, visual embeddings are superior to structured embeddings
- Highly embodied verbs profit from multimodal embeddings

→ We recommend to ground verbs multimodally.

Chapter 5

Frame Semantics for Relational Knowledge

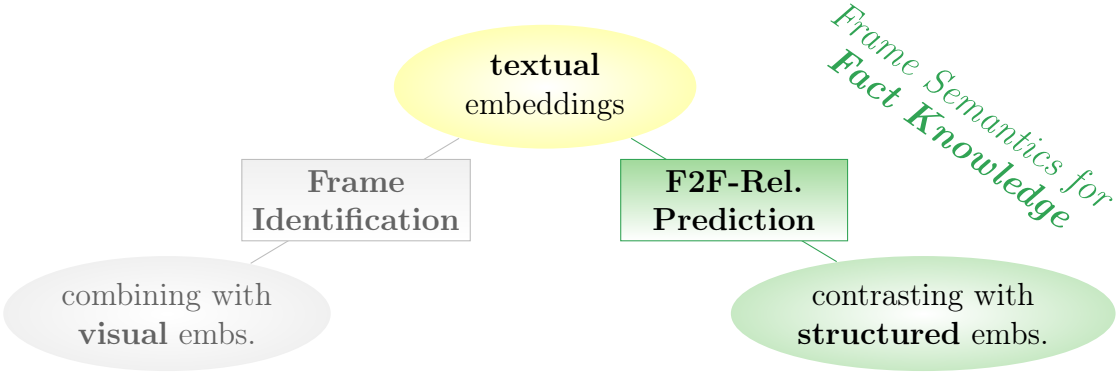


Figure 5.1: Structure of Chapter 5. Right green branch: knowledge about facts with textual versus structured frame embeddings for Frame-to-Frame Relation Prediction. (Left branch was focus of Chapter 4.)

In this chapter, we focus on relations between single categories of meaning in order to model meta-knowledge of interactions, procedures or associations. Novel knowledge about relations could either be inferred from textual data, or from structured data in knowledge bases. For modeling relational knowledge, we contrast textual versus structured embeddings for predicting relations between frames. We present and discuss our contributions and findings in the context of language understanding with frame semantics for relational knowledge in structured triples as outlined in Figure 5.1 (right green branch). The immediate background for structured language understanding and for embedding learning based on knowledge bases was given in Sections 2.2 and 3.3, respectively.

In the first part, Section 5.1, we examine textual frame embeddings with respect to recovering frame-to-frame relations. The underlying research question is whether frame-to-frame relations can be directly inferred from text. We point out the limitations of textual embeddings in mirroring frame-to-frame relations. This also hints at textual frame embeddings incorporating other semantic information than the one contained in the relations.

In the second part, Section 5.2, we introduce the new task of Frame-to-Frame Relation Prediction as a Knowledge Base Completion task for FrameNet. On this

task, we contrast the performance of textual versus structured frame embeddings and point out the advantage of structured embeddings in correctly predicting relations between frame pairs.

On the one hand, we address this task with a supervision-less approach in order to explore the predictive power despite the small number of the triples available for training in the FrameNet hierarchy. Thus, we experiment with textual frame embeddings (pre-trained without triples, but on annotated texts) as the basis of a 'supervision-less' prediction that does not involve training of weights and biases in a prediction system that would be tuned to fit the triples in the hierarchy.

On the other hand, we approach the task in a supervised way by making use of the subset of triples available for training. We introduce the `StruFFRel` approach using the structure of the FrameNet hierarchy and we contrast a collection of systems that process different input embeddings. More precisely, the `StruFFRel` approach leverages the structure of the FrameNet hierarchy to train a prediction system on the training triples. A prediction system processes pre-trained embeddings for frames and the best performance is achieved when using the structured frame embeddings. Our best structured prediction system can be used to generate recommendations for annotations with relations.

In an outlook, we explore the potential of multimodal approaches to Knowledge Base Completion (Section 5.2.3.1) and suggest to develop approaches that incorporate visual information about frames to benefit Frame-to-Frame Relation Prediction and also frame induction (short frame-relation-induction).

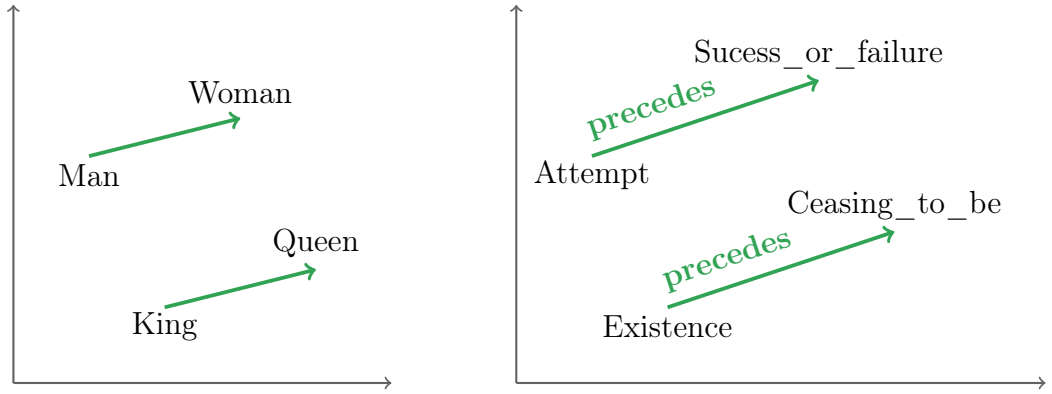
In this chapter we finalize the applicability of Frame Identification in higher-level tasks by complementing the Frame Identification system (cf. Chapter 4) with different sets of frame embeddings.

5.1 Frame-to-Frame Relations in Textual Embeddings for Frames

In this section, we present and discuss our contributions and findings in the context of structured language understanding with frame semantics, where we initially examine textual frame embeddings with respect to recovering frame-to-frame relations. The underlying research question of this section is the following:

RQ: *Can frame-to-frame relations be directly inferred from text?*

We aim at empirically analyzing whether frame-to-frame relations are mirrored in textual frame embeddings, which were learned on frame-labeled texts in the context of other language understanding tasks. We inspire in textual word embeddings being evaluated on syntactic or semantic analogy tasks with the vector offset method (cf. Equation 3.13), where it is known that these embeddings implicitly learn syntactic or semantic relations from texts (Mikolov et al., 2013b). However, for textual frame embeddings it is yet to investigate whether they implicitly learn frame-to-frame relations from texts. Thus, we want to find out whether a statistical analysis of textual frame embeddings naturally yields the relations of the FrameNet hierarchy. Indeed, the frame-to-frame relations are manually annotated by expert linguists but there is no guarantee that frame-to-frame relations directly emerge from text.



(a) Word2Vec embedding space. (b) FrameNet embedding space.

Figure 5.2: Intuition for frame embeddings incorporating frame-to-frame relations in vector space, following the idea of `Word2Vec`.

If these relations could emerge from raw text it would be reassuring for the definitions of the frame-to-frame relations that led to annotations of frame pairs and furthermore the annotations could be generated automatically. We hypothesize that distances and directions between frame embeddings learned on textual data can correspond to frame-to-frame relations. Figure 5.2 illustrates the intuition by following the findings within word embeddings by Mikolov et al. (2013b). In a textual embedding space, *'man'* is to *'woman'* as *'king'* is to *'queen'* as both word pairs are in a *male-female*-relation; and transferred to frames the question is whether for the two frame pairs *'Attempt'* is to *'Success_or_failure'* as *'Existence'* is to *'Ceasing_to_be'* as both pairs are in a *precedence*-relation. Our analysis of the textual frame embeddings on the training set of the triples reveals insights about the difficulty of reconstructing frame-to-frame relations purely from text.

Taken together, this hints at textual frame embeddings incorporating other semantic information than the one contained in the relations. Our paper (Botschen et al., 2017)[1] is foundational to this chapter.

5.1.1 Experimental Setup for Exploration of Textual Embeddings

To learn textual frame embeddings, we make use of embedding learning methods (cf. Section 3.2) applied on frame-annotated texts provided by FrameNet. FrameNet additionally provides frame-to-frame relations that link frames to other frames in the hierarchy. Frame-labeled text can only serve to directly learn textual frame embeddings, but not textual embeddings for frame-to-frame relations. Thus, in a first step, we learn textual embeddings for frames. Then, in a second step, we combine the frame embeddings of those frames forming a pair in a relation in order to approximate embeddings for frame-to-frame relations, which we call 'prototypical' embeddings for frame-to-frame relations. We use two different approaches to learn textual frame embeddings: on the one hand, we apply a matrix factorization ap-

[1] My contributions in this paper are the following: exploration of frame embeddings, `StruFFRel` approach and analysis of experiments.

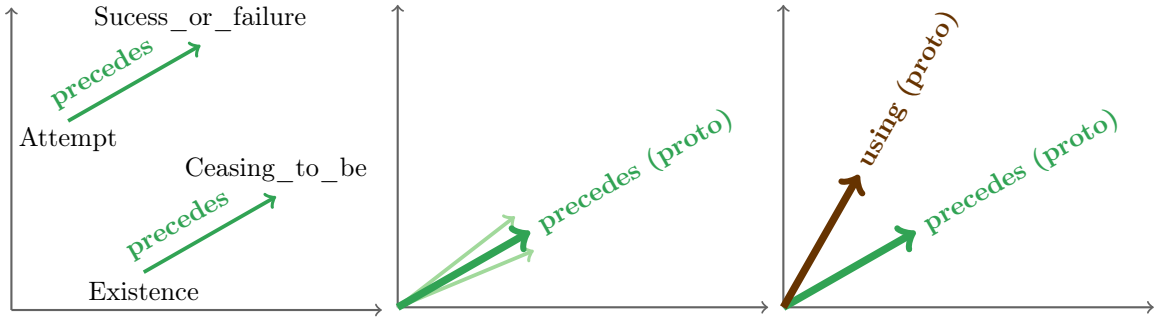
proach for learning WSABIE embeddings for frames (as explained in Section 3.2) on the task of Frame Identification. On the other hand, we apply a neural network approach for learning Word2Vec embeddings for frames (as explained in Section 3.2) on the task of predicting context words given the target frame.

WSABIE Embeddings for Frames. Referring to the matrix factorization approach for learning textual frame embeddings, we reuse our own publicly available code from the SimpleFrameId system (Hartmann et al., 2017, as introduced in Section 4.1.3). To make an example of how the WSABIE embeddings for frames are learned, let us look at the sequence *'Officials claim that Iran has produced bombs'* for which the annotation with frames labels the predicate 'claim' with the frame *'Statement'*. The latent representation for the frame *'Statement'* is learned in a way that it is close to the concatenation of the embedding for the predicate *'claim'* and of the context embedding. The implementation for learning WSABIE embeddings for frames is based on the state-of-the-art system Hermann-14 (Hermann et al., 2014) and achieves comparable results on Frame Identification, though not exactly reproducing their results. Our hyperparameter choices are oriented towards our system SimpleFrameId (Hartmann et al., 2017): embedding dimension 100, maximum number of negative samples: 100, epochs: 1000, and an initial representation of predicate and context: concatenation of pre-trained dependency-based word embeddings (Levy and Goldberg, 2014a).

Word2Vec Embeddings for Frames. Concerning the neural network approach for learning textual frame embeddings, we use the Word2Vec implementation in the python library gensim (Řehůřek and Sojka, 2010). To obtain frame embeddings we follow the same steps as if we learned word embeddings on FrameNet sentences. Above that, we replace all predicates with their frames in the FrameNet sentences. For instance, in the sequence *'Officials claim that Iran has produced bombs'* the predicates *'claim'* and *'bombs'* are replaced by *'Statement'* and *'Weapon'*, respectively. This procedure of replacing words with their respective higher-level labels corresponds to Flekova's setup for learning supersense embeddings (Flekova and Gurevych, 2016) and our hyperparameter choices are oriented towards their best performing ones: training algorithm: Skip-gram model, embedding dimension: 300, minimal word frequency: 10, negative sampling of noise words: 5, window size: 2, initial learning rate: 0.025 and iterations: 10. Referring to the example sentence, the Skip-gram model learns the embeddings so that given *'Statement'*, the context words can be predicted.

Prototypical Embeddings for Frame-to-Frame Relations. We denote learned embeddings with $\overrightarrow{v_{f_1}}$ (for frame f_1). We use the frame embeddings to infer prototypical frame-to-frame relation embeddings $\overrightarrow{v_r}$ with the vector offset method (cf. Equation 3.13) in the following way: we denote with I_r the relation-specific subset of G with all the instances (f_1, r, f_2) for this relation (see frame pair counts in Table 2.3). The vector offset $\overrightarrow{o_{(f_1,f_2)}}$ for two frames (f_1, f_2) is the difference of their embeddings (Equation 5.1), which transfers Equation 3.13 to frame embeddings:

$$offset(\overrightarrow{v_{f_1}}, \overrightarrow{v_{f_2}}) = \overrightarrow{o_{(f_1,f_2)}} = \overrightarrow{v_{f_2}} - \overrightarrow{v_{f_1}} \; . \tag{5.1}$$



(a) FrameNet embedding space with relations as offset vectors.

(b) Prototypical relation embedding as the mean of single relation instances.

(c) Different prototypical relation embeddings as the mean of their respective single relation instances.

Figure 5.3: By averaging, we obtain the prototypical relation embeddings out of single relation examples.

We denote with O_r the relation-specific set of vector offsets of all $(f_1, f_2) \in I_r$. We define the prototypical embedding $\overrightarrow{v_r}$ for a relation r as the dimension-wise mean over all $\overrightarrow{o_{(f_1,f_2)}} \in O_r$. For visualizations in vector space, we use t-SNE-plots (t-distributed Stochastic Neighbor Embedding algorithm, van der Maaten and Hinton, 2008). Figure 5.3 explains step by step how we obtain the prototypical relation embeddings out of single relation examples.

Difficulty of Associating Frame Pairs with Prototypical Relations. The association of the embedding of a frame pair $\overrightarrow{o_{(f_1,f_2)}} \in O_r$ with the correct prototypical relation embedding $\overrightarrow{v_r}$ is easier if the intra-relation variation (i.e. the deviation of frame pair embeddings from their prototypical embedding) is smaller than the inter-relation variation (i.e. the distances between prototypical embeddings). This means, the association is easier if two frame pairs which are members of the same frame-to-frame relation, on average, differ less from each other than they would differ from a member of another relation. As a way to capture this difficulty of association we compare the mean cosine distance d_{cos} (cf. Equation 3.2) between all prototypical relation embeddings $\overrightarrow{v_r}$ of all $r \in R$ to the relation-specific mean cosine distance between the frame pair embeddings in O_r and the prototypical embedding $\overrightarrow{v_r}$.

5.1.2 Results and Discussion

We ask the question whether frame-to-frame relations are learned implicitly from the text when learning textual embeddings for frames on the frame-labeled text. To illustrate the intuition with an example referring to Figure 5.3: Given a sequence like *'After all the effort, they aced it (or failed it)'* with 'effort' evoking the frame 'Attempt' and 'aced' evoking the frame 'Success_or_failure', a reader understands that 'Attempt' precedes 'Success_or_failure' – and after reading many similar sequences the reader could infer that the 'precedence'-relation holds true here. With the experiments we explore whether frame-to-frame relations are incorporated in the textual frame embedding spaces.

	Top 10 most similar frames	
frame	`WSABIE`	`Word2Vec`
Weapon	Substance, **Shoot_projectiles**, Manufacturing, **Bearing_arms**, **Toxic_substance**, **Hostile_encounter**, Ingredients, Information, **Smuggling**, Active_substance	**Military**, Substance, Operational_testing, Store, Electricity, Process_completed_state, Active_substance, Range, Estimated_value, Cause_to_make_progress
Statement	**Evidence**, Causation, **Topic**, **Chatting**, **Point_of_dispute**, **Request**, **Text_creation**, Cognitive_connection, **Make_agreement_on_action**, **Communication**	**Reveal_secret**, **Telling**, **Complaining**, **Reasoning**, **Communication_response**, Awareness, **Reassuring**, **Bragging**, **Questioning**, Cogitation

Table 5.1: Top 10 most similar frames to two exemplary most frequent frames (Weapon, Statement) for frame embeddings learned with *WSABIE* and *Word2Vec*. Marked in bold are frames which are obviously semantically related to the exemplary frame.

Frame Embeddings. Once the frame embeddings are learned, we perform a sanity check for frames. For this, we orient ourselves to Iacobacci et al. (2015) and Flekova and Gurevych (2016) who qualitatively check their embeddings for (super-)senses by looking at the most similar (super-)sense embeddings (cosine similarity). We also qualitatively check the frame embeddings in terms of most similar frames in the embedding space. Checking the top 10 most similar frame embeddings confirms that known properties from word or sense embeddings also apply to frame embeddings: their top 10 most similar frames are semantically related, both for frame embeddings learned with `WSABIE` and with `Word2Vec`. This is exemplified in Table 5.1 for the two most frequently occurring frames in the text data evoked by nouns, e.g., *'Weapon'* and by verbs, e.g., *'Statement'*. For both `WSABIE` and `Word2Vec`, in many cases the most similar frames are obviously semantically related (which we marked in bold), with some exceptions where it is hard to judge or where the relation works via an association chain. For the frame 'Weapon', the most similar frames with embeddings learned by `Word2Vec` are weaker compared to those with embeddings learned by `WSABIE` — this is an example for the qualitative differences between `WSABIE` and `Word2Vec`, this however does not allow a general conclusion over all frames learned with `Word2Vec` or `WSABIE`.

Frame-to-frame Relations. To check whether the textual frame embeddings directly mirror frame-to-frame relations, we measure the difficulty of associating frame pairs with the correct prototypical relation embedding.

In a first step, we visually inspect some examples of single relation embeddings (obtained from single frame pairs) in the training set and we also visualize the inferred prototypical relation embeddings in a vector space with t-SNE-plots.

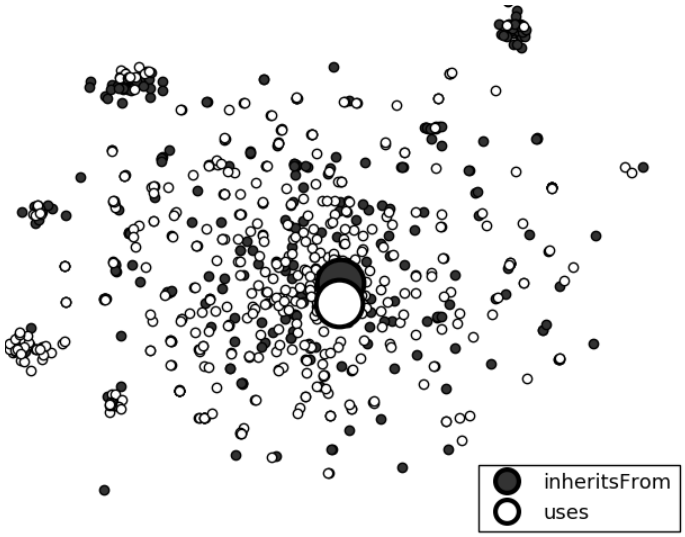


Figure 5.4: t-SNE plot of WSABIE-embeddings for the two most frequent frame-to-frame relations. Small: frame pair embeddings (offset). Large: prototypical embeddings (mean).

Mean distances between	WSABIE	Word2Vec
inter-relation variation (between prototypes)	0.73 ± 0.28	0.76 ± 0.28
intra-relation variation (between frame pairs and their prototypes)	0.75 ± 0.04	0.78 ± 0.05

Table 5.2: Cosine distances between the frame-to-frame relations in embedding space.

Figure 5.4 depicts examples of WSABIE embeddings for the two most frequently occurring frame-to-frame relations `inherits_from` and `uses`. It demonstrates that the prototypical embeddings are very close to each other, whilst there are no separate relation-specific clusters for frame pairs. Vector space visualizations of embeddings trained with Word2Vec and WSABIE hint that the embeddings have difficulties in mirroring the frame-to-frame relations.

In a second step, we quantify the insights from the plots by comparing the distances between all prototypical embeddings to the mean over all mean distances between frame pair embeddings and their prototypical embeddings. Table 5.2 lists these cosine distances. It shows that the distance between the prototypical embeddings (inter-relation) is smaller than that between frame pair embeddings and corresponding prototypical embeddings (intra-relation). In other words, two frame pairs which are members of the same relation, on average, differ as much from each other as they would differ from a member of another relation.

To sum up, our analysis of the textual frame embeddings on the training set of the triples reveals insights about the limitations of textual frame embeddings in reconstructing frame-to-frame relations. We find that embeddings of frame pairs that are in the same relation do not have a similar vector offset which corresponds to the frame-to-frame relation. The FrameNet hierarchy could not be reconstructed by

the statistical analysis of textual embeddings because there is as much intra-relation variation as inter-relation variation. We conclude that, in terms of the methods we explored, the frame embeddings learned with `WSABIE` and `Word2Vec` have difficulties in showing structures in vector space corresponding to frame-to-frame relations and that frame-to-frame relations might not emerge purely from textual data.

In the next section, we address the prediction of frame-to-frame relations with algorithms involving learning from the knowledge contained in the FrameNet hierarchy; and we propose a new task, namely Frame-to-Frame Relation Prediction on the FrameNet hierarchy.

5.2 Frame-to-Frame Relation Prediction: Contrasting Textual versus Structured Embeddings

In this section, we present and discuss our contributions and findings in the context of structured language understanding with frame semantics, where we model knowledge about relations with textual versus structured frame embeddings for the task of predicting relations between frames.

First, in Section 5.2.1, we approach our newly proposed task, namely Frame-to-Frame Relation Prediction on the FrameNet hierarchy (Botschen et al., 2017), which is novel to research on FrameNet (cf. Section 2.2). This task is about predicting the correct relation between two frames. A well-performing prediction system can then be used to complete the sparse coverage of relation annotations in FrameNet. We elaborate on the differences between Frame-to-Frame Relation Prediction and traditional Knowledge Base Completion, one of which is the small number of the triples provided by the FrameNet hierarchy for training (cf. Table 2.3 for the relation-specific frame-to-frame pair counts). Thus, we first explore the task of Frame-to-Frame Relation Prediction with a supervision-less approach in order to assess the predictive power despite the small number of the triples available for training. To this end, we experiment with textual frame embeddings (pre-trained without triples) as the basis of a 'supervision-less' prediction that does not involve training of weights and biases in a prediction system which would be tuned to fit the triples in the hierarchy. In a straight-forward way, we apply the vector offset method (cf. Equation 3.13) to the pre-trained textual frame embeddings on the test triples.

Next, in Section 5.2.2, we experiment with a supervised setup for the task of Frame-to-Frame Relation Prediction, following related work on Knowledge Base Completion (cf. Section 2.2) – or Knowledge Base Question Answering – as closely as possible. This means that, on the one hand, we use the training triples to learn structured frame embeddings (cf. Section 3.3), which are designed for the task of relation prediction. And on the other hand, we use the training triples to train relation prediction systems on top of the pre-trained frame embeddings (`StruFFRel` approach). We contrast the performance of textual versus structured frame embeddings. According to related work on Knowledge Base Completion, we expect:

> EXPECTATION: *For the task of Frame-to-Frame Relation Prediction structured embeddings are more informative than textual embeddings.*

A comparison of systems and embeddings exposes the crucial influence of structured frame embeddings on a system's performance in predicting frame-to-frame relations. We propose our best-performing system of our `StruFFRel` approach for automatically generating recommendations for new annotations of relations between frames.

Taken together, we contrast the potential of textual versus structured frame embeddings as input representations to different approaches to Frame-to-Frame Relation Prediction. In an outlook, we explore the potential of multimodal approaches to Knowledge Base Completion (Section 5.2.3.1) and suggest to develop approaches that incorporate visual information about frames to benefit Frame-to-Frame Relation Prediction and also frame induction (short frame-relation-induction). Our papers (Botschen et al., 2017)[2] and (Mousselly-Sergieh et al., 2018)[3] are foundational to this chapter.

5.2.1 Supervision-less Frame-to-Frame Relation Prediction

We introduce *Frame-to-Frame Relation Prediction* as a new task for finding the correct frame-to-frame relation given two frames, which can potentially be used for automatic completion of the frame-to-frame relation annotations in the FrameNet hierarchy. Whilst the definition of the task was presented in Section 2.2.2.1, here, we approach the task of Frame-to-Frame Relation Prediction with using the triples in the FrameNet hierarchy as little as possible in order to assess the predictive power despite the small number of the triples available for training.

Frame-to-Frame Relation Prediction versus Link Prediction in Knowledge Bases. This task transfers the principles of Link Prediction from Knowledge Base Completion (KBC, common abbreviation) to the case of FrameNet (for Link Prediction see Section 3.3). Frame-to-Frame Relation Prediction, however, is different to traditional Knowledge Base Completion in several ways. First, FrameNet operates with less relations compared to traditional knowledge bases, whilst FrameNet's relations are considered to be more abstract than those of a traditional knowledge bases. Second, traditional Knowledge Base Completion is often formulated as Link Prediction which is the prediction of an entity as the missing link (as explained in Section 3.3, Figure 3.4), whereas Frame-to-Frame Relation Prediction refers to the explicit prediction of a relation. Third, Frame-to-Frame Relation Prediction can be considered more challenging with respect to the training data as there are much less training triples compared to traditional knowledge bases.

The first two points show that the setup of the tasks differs in details and the third points makes clear that approaches which are promising on traditional Knowledge Base Completion can not be expected to yield the same top performance on Frame-to-Frame Relation Prediction.

Taking into account the small amount of training triples compared to traditional knowledge bases, we first explore the minimal setup of textual frame embeddings

[2] My contributions in this paper are the following: exploration of frame embeddings, `StruFFRel` approach and analysis of experiments.

[3] My contributions in this paper are the following: exploration of synset embeddings, extension of approach by Xie et al. (2017) for multimodal Knowledge Base Completion on WN9-IMG dataset.

which does not require further training on the triples. 'Minimal setup' means that the only point where the training triples are used is when building the prototypical mean relation embeddings $\overrightarrow{v_r}$ of the training set, which we will explain in the following section.

5.2.1.1 Experimental Setup and Baselines

We test the performance of the learned textual frame embeddings on the task of Frame-to-Frame Relation Prediction. In a straight-forward way, we apply the vector offset method (cf. Equation 3.13) to the pre-trained textual frame embeddings on the test set of the triples.

Given a triple (f_3, r, f_4) from the test set, we want to predict the correct relation r for (f_3, f_4). As described in Section 2.2.2, 30% of the triples in the FrameNet hierarchy are used for testing.

Baselines. The baselines are listed in the following, and from now on we will refer to them with their numbers and names, e.g., *system 0a ('random baseline')*.

System 0a: Random Baseline. A random guessing baseline that chooses a frame-to-frame relation randomly out of the set of all possible relations R.

System 0b: Majority Baseline. Informed majority baseline that leverages the skewed distribution in the training set and predicts the most frequent relation.

Vector Offset Method for Frames and Relations. We extend the list of systems by adding a system leveraging pretrained frame embeddings, and from now on we will refer to it with its number and name: *system 1 ('vector offset')*.

System 1: Vector Offset. A test of the pre-trained frame embeddings (`WSABIE` and `Word2Vec`) as introduced in Section 5.1. It computes the vector offset $\overrightarrow{o_{(f_3,f_4)}}$ between the test frame embeddings, measures the similarity with the prototypical mean relation embeddings $\overrightarrow{v_r}$ of the training set and ranks the relations in terms of cosine distance to output the closest one. No further training with respect to the FrameNet hierarchy takes place.

Evaluation Measurements. To evaluate the predictions of the systems for the Frame-to-Frame Relation Prediction task, we compare the measurements of accuracy, mean rank of the true relation and hits amongst the 5 first predictions, see Table 5.4.

Most straight-forward, *accuracy* measures the proportion of correctly predicted relations amongst all predictions. For the next two measures, not only the one predicted relation is of interest, but the ranked list of all relations with the predicted relation at rank 1. *Mean rank* measures the mean of the rank of the true relation label over all predictions, aiming at a low mean rank (best is $mr = 1$). *Hits@5* measures the proportion of true relation labels ranked in the top 5.

System	Embeddings	acc ↑	mr ↓	hits@5 ↑
0: random baseline	-	7.69	6.5	38.46
0: majority baseline	-	22.48	3.27	87.51
1: vector offset	WSABIE	25.22	4.50	68.52
1: vector offset	Word2Vec	30.61	4.53	66.96

Table 5.3: Supervision-less performances on Frame-to-Frame Relation Prediction.

5.2.1.2 Results and Discussion

The results are listed in Table 5.3. In the following, we discuss the performance of the baseline in comparison to that of the supervision-less vector offset approach with textual frame embeddings.

Baseline Results. The random guessing baseline, system 0a ('random baseline'), is a weak baseline that is outperformed by all approaches. The informed majority baseline, system 0b ('majority baseline'), however, is a strong baseline given the skewed distribution of frame-to-frame relations in the FrameNet hierarchy.

Baseline versus Vector Offset Approach. A comparison of the strong baseline with system 1 ('vector offset'), using the textual frame embeddings (WSABIE and Word2Vec) and the similarity with prototypical relation embeddings, emphasize the difficulties of these embeddings for reconstructing the frame-to-frame relations. In terms of accuracy scores, system 1 ('vector offset') performs slightly better than the strong baseline but concerning the other two measures, mean rank and hits at 5, it is the other way round. Another point made by system 1 ('vector offset') is the fact that it does not involve further training on the triples but is still competitive with the strong baseline that leverages the underlying distribution from the triples. This indicates that, to some extent, the textual frame embeddings still capture useful information for the Frame-to-Frame Relation Prediction Task.

However, these textual frame embeddings cannot be used as such to reliably infer the correct relation for a frame pair but might need some advanced learning. In the next section, we address the prediction of frame-to-frame relations with algorithms involving learning from the knowledge contained in the FrameNet hierarchy.

5.2.2 Trained Frame-to-Frame Relation Prediction System StruFFRel

As the examination of textual frame embeddings with respect to emergence of frame-to-frame relations suggest that textual frame embeddings do not mirror the frame-to-frame relations (cf. Section 5.1), we experiment with a supervised setup, following related work on Knowledge Base Completion (cf. Section 2.2) as closely as possible.

Related work in Knowledge Graph Completion demonstrates the strengths of representations trained directly on the knowledge graph for this task (cf. Bordes et al. (2011, 2012, 2013)). For this, we apply the TransE algorithm (as explained in Section 3.3) to the knowledge base incorporated by the FrameNet hierarchy.

This means that, on the one hand, we apply the `TransE` algorithm (as explained in Section 3.3) to FrameNet's training triples (f_1, r, f_2) to learn structured frame embeddings (cf. Section 3.3). And on the other hand, we use the training triples to train relation prediction systems on top of the pre-trained frame embeddings. We contrast the performance of textual versus structured frame embeddings and hypothesize that structured embeddings are more informative than textual embeddings for the task of Frame-to-Frame Relation Prediction.

For the structured frame embeddings learned with `TransE`, we do not need to calculate prototypical relation embeddings as `TransE` provides embeddings for frames and relations. Thus, system 1 ('vector offset') uses the `TransE` embeddings directly to calculate the similarity of the frame embeddings' vector offset and the relation embeddings.

We present the best-performing system of our `StruFFRel` approach for predicting frame-to-frame relations, which leverages the structure of the FrameNet hierarchy to (a) pre-train frame embeddings and to (b) further train a prediction model. Thus, it involves learning from the knowledge contained in the FrameNet hierarchy on top of pre-trained frame embeddings.

We quantify which input representations together with which approach for learning from the FrameNet hierarchy is most promising for the Frame-to-Frame Relation Prediction task. Regarding the approach for learning, we contrast a straight-forward regression model with a neural network model. Regarding the input representations to the system, in addition to the textual frame embeddings, we also learn structured embeddings for frames and for frame-to-frame relations.

We find a large advantage of structured frame embeddings compared to textual frame embeddings for predicting the correct relation, and also, we find a large advantage of the learning-based approaches compared to directly predicting the relation given the pre-trained frame embeddings. With a slight advantage of the neural network approach compared to the regression model, the best system we determine experimentally for predicting frame-to-frame relations is a neural approach together with structured frame embeddings. We introduce our new system for Frame-to-Frame Relation Prediction which follows the `StruFFRel` approach of leveraging the structure of the FrameNet hierarchy.

5.2.2.1 Architecture and Structured Input Embeddings

We learn structured embeddings for frames and for frame-to-frame relations and then use them in a trainable network for relation selection.

TransE Embeddings for Frames. We learn embeddings for frames as well as for frame-to-frame relations by applying the translation model `TransE` (as in Section 3.3) to the structure of the knowledge base incorporated by the FrameNet hierarchy. The structure defined by FrameNet refers to the collection of the *(frame, relation, frame)*-triples, which is then split into a training and a test set such that the training set contains the first 70% of all the triples for each relation (cf. Section 2.2.2). `TransE` learns low dimensional vector representations for frames and for frame-to-frame relations in the same space. By that, these embeddings will have the property of being learned explicitly for incorporating the annotations from the

FrameNet hierarchy. Concerning this knowledge-based approach for learning frame and frame-to-frame relation embeddings, we use an implementation of `TransE` provided by Lin et al. (2015b) yielding embeddings of dimension 50.

Learning-to-Rank Model. When relating Frame-to-Frame Relation Prediction to Knowledge Base Completion (see Section 5.2 for commonalities and differences), our task can be approached as a Link Prediction task from Knowledge Base Completion (cf. Section 3.3). Link Prediction is methodologically related to the key-task of Answer Selection from Question Answering (QA, common abbreviation). The task is to rank a set of possible answer candidates with respect to a given question (Tan et al., 2015), this is why this line of work is also called 'Learning-to-Rank'. State-of-the-art Question Answering models are presented by Feng et al. (2015) and by Tan et al. (2015). They jointly learn vector representations for both the questions and the answers in the training set. Representations of the same dimensionality in the same space allow to compute the cosine distances between these vectors.

We decide to build upon neural network models for Answer Selection and adapt the ideas to frame-to-frame Relation Prediction for the following reasons: We aim at having an alternative to the traditional representation learning paradigm where either, in a supervision-less way, pre-trained embeddings are compared with cosine similarity, or, in a supervised way, pretrained embeddings are taken as input representations to a classifier. We decide for the alternative to be a neural network approach with a ranking loss in order to follow related work on Knowledge Base Completion and Answer Selection as closely as possible.

In our case, a question corresponds to a frame pair and an answer corresponds to a frame-to-frame relation. Optionally, pre-trained frame embeddings can be used as initialization, which allows us to contrast different pre-trained frame embeddings, such as learned with `Word2Vec`, `WSABIE` and `TransE`.

Neural Network for Relation Selection. We propose a nonlinear model based on neural networks to identify the best frame-to-frame relation r between a frame pair (f_1, f_2). Figure 5.5 demonstrates the proposed neural network architecture. Given a training instance, i.e. a triple (f_1, r, f_2), we feed a pre-trained embedding for each element into the neural network. The pre-trained frame embeddings are learned with `Word2Vec`, `WSABIE` and `TransE`; the frame-to-frame relation embeddings come from pre-training for `TransE` only, whereas for `Word2Vec` and `WSABIE` we use the prototypical relation embeddings, respectively. Within the neural network, the initial vector representations of the two frames are combined into an internal dense layer c, followed by the calculation of the distance measure d between this combination and the representation for the frame-to-frame relation r. Here, the distance measure d is the cosine distance d_{cos} (cf. Equation 3.2). Meanwhile, a negative relation r' is sampled randomly (by selecting a frame-to-frame relation which does not hold between the two frames) and its vector representation is also fed into the neural network. This strategy is known as 'negative sampling'. The negative relation is processed in the same way as the correct one, yielding a second cosine distance (same distance measure d). Finally, the internal representations are trained to maximize the similarity between frame pair and correct relation and to minimize it for the negative relation.

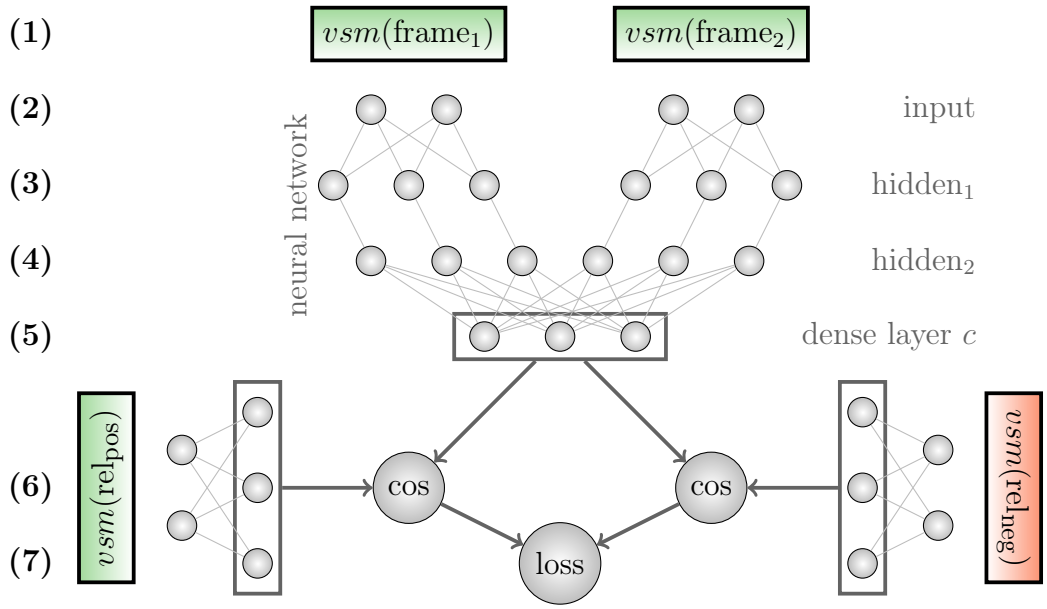


Figure 5.5: Architecture for training relation selection with a learning-to-rank model. (1) Pair of pre-trained frame embeddings. (2) Input as embeddings. (3) First dense hidden layer with tanh activation function. (4) Second hidden layer: concatenation. (5) Dense layer c with tanh activation: approximate of correct relation. (6) Cosine distance of c and of correct vs. wrong relation. (7) Loss function based on distances.

The neural network minimizes a margin-based ranking criterion as the loss function (Equation 5.2), with m being the margin ($m > 0$):

$$loss = [m + d(c, r) - d(c, r')]_+ \,, \tag{5.2}$$

where $[x]_+$ denotes the positive part of x. We exemplify three extreme cases (best case, worst case, just-enough case) with the margin of $m = 0.1$:

- best case:
  minimal distance for correct triples: $d(c, r) = 0$,
  maximal distance for incorrect triples: $d(c, r') = 1$
  $\Rightarrow$ $loss = [0.1 + 0 - 1]_+ = [-0.9]_+$
  $\Rightarrow$ not positive, therefore no further minimization.

- worst case:
  maximal distance for correct triples: $d(c, r) = 1$,
  minimal distance for incorrect triples: $d(c, r') = 0$
  $\Rightarrow$ $loss = [0.1 + 1 - 0]_+ = [1.1]_+ = 1.1$
  $\Rightarrow$ positive, therefore further minimization needed.

- just-enough case:
  distance for correct triples exactly by margin smaller than distance for incorrect triples: $d(c, r) + 0.1 = d(c, r')$,
  e.g., $d(c, r) = 0.4$ and $d(c, r') = 0.5$
  $\Rightarrow$ $loss = [0.1 + 0.4 - 0.5]_+ = [0]_+$
  $\Rightarrow$ exactly zero, not positive, therefore no further minimization.

We found the following hyperparameter choices to yield the best results: number of epochs: 550, size of dense layers: 128, dropout: 0.2, margin: 0.1, activation function: hyperbolic tangent (tanh, common abbreviation), batch size: 2, learning rate: 0.001.

5.2.2.2 Experimental Setup

Here, we extend the experimental setup of Section 5.2.1.1 with the supervised approach and the structured embeddings. Again, given a triple (f_3, r, f_4) from the test set, we want to predict the correct relation r for (f_3, f_4). As described in Section 2.2.2, 70% of the triples in the FrameNet hierarchy are used for training and the remaining 30% for testing.

Systems with Supervised `StruFFRel` Approach. The supervised systems are listed in the following, and from now on we will refer to them with their numbers and names, e.g., *system 3 ('NN')*.

System 2: Regression. A test of the pre-trained frame embeddings (`WSABIE` and `Word2Vec`) as introduced in Section 5.1 involving training with respect to the FrameNet hierarchy. It is a multinomial logistic regression model that trains the weights and the biases on the training triples. It takes the test frame embeddings $\overrightarrow{v_{f_3}}, \overrightarrow{v_{f_4}}$ as input and ranks the prediction for a relation via the softmax function.

System 3: NN. Neural network architecture for training with respect to the FrameNet hierarchy in the training triples (as described in the previous paragraph and illustrated in Figure 5.5). By default, it uses randomly initialized input representations, but it can also take pre-trained representations as input: *(a)* the pre-trained frame embeddings (`WSABIE` and `Word2Vec`) and inferred prototypical mean relation embeddings as introduced in Section 5.1. and *(b)* the `TransE` frame and relation embeddings trained on the training triples from the FrameNet hierarchy as introduced in Section 5.2.2.

The evaluation measurements are described in Section 5.2.1.1.

5.2.2.3 Results and Discussion

The results are listed in Table 5.4. In the following, we discuss the performance with the textual frame embeddings in comparison to that with the structured frame embeddings and we discuss the performance of the systems following the supervised `StruFFRel` approach by leveraging the triples in the FrameNet hierarchy for training.

Textual versus Structured Frame Embeddings. A comparison of systems and embeddings exposes the crucial influence of structured frame embeddings on a system's performance in predicting frame-to-frame relations.

On the one hand, the large improvement in all performance measures shows the strength of structured embeddings over textual embeddings and confirms the difficulty of textual embeddings in reconstructing the frame-to-frame relations.

	System	Embeddings	acc ↑	mr ↓	hits@5 ↑
	0: random baseline	-	7.69	6.5	38.46
	0: majority baseline	-	22.48	3.27	87.51
	1: vector offset	WSABIE	25.22	4.50	68.52
	1: vector offset	Word2Vec	30.61	4.53	66.96
	1: vector offset	TransE	51.13	2.99	83.30
StruFFRel approach	2: regression	WSABIE	35.65	3.14	84.00
	2: regression	Word2Vec	41.91	2.81	88.00
	2: regression	TransE	**66.61**	**1.93**	**93.22**
	3: NN	random	26.89	3.67	77.00
	3: NN	WSABIE	27.46	3.59	79.98
	3: NN	Word2Vec	30.55	3.27	82.61
	3: NN	TransE	**67.73**	**1.83**	**94.39**

Table 5.4: Supervised performances on Frame-to-Frame Relation Prediction with systems of the StruFFRel approach.

Thus, the final system incorporates structured frame embeddings for Frame-to-Frame Relation Prediction: StruFFRel.

On the other hand, structured embeddings are expected to perform better than textual ones on the task of Frame-to-Frame Relation Prediction: the training objective of TransE directly optimizes for predicting the relations and it does so by the direct association of frame embeddings and embeddings for frame-to-frame relations. WSABIE and Word2Vec, however, are trained on different objectives and on different data – they do not explicitly optimize for relation prediction. Taking this into account, it is impressive that WSABIE and Word2Vec still deliver meaningful results for the task of frame-to-frame relation prediction. This could encourage future work on unsupervised relation prediction or frame-to-frame relation induction, i.e. the annotation of new frame-to-frame relations.

Systems with Supervised StruFFRel Approach. Performance increases with system 2 ('regression'), the softmax regression model involving learning. This shows the effect of training on the FrameNet hierarchy with respect to frame-to-frame relations. The performance increase in textual frame embeddings from the supervisionless vector offset method to the StruFFRel approach with system 2 ('regression') indicates that training should be involved for leveraging the textual frame embeddings in the Frame-to-Frame Relation Prediction Task. Using embeddings pre-trained on the frame-to-frame relations of the FrameNet hierarchy (TransE) instead, again leads to a large improvement in all performance measures. This confirms that embeddings designed to incorporate the knowledge from the FrameNet hierarchy are better suited for the frame-to-frame relation prediction task and it emphasizes the large improvement over the textual embeddings.

Overall, we achieve best results in all performance measures with system 3 ('NN'), the neural network approach, in combination with the structured TransE embeddings as input representations. Interestingly, the difference between the neural network and the regression model is only marginal when using the TransE embeddings, indicating the crucial influence of the structured embeddings and not

necessarily of the approach to learn weights for the prediction. Moreover, when using the textual `WSABIE` and `Word2Vec` the system 2 ('regression') implementing a softmax regression model is stronger than the neural network, which might be due to little training data. Furthermore, the randomly initialized embeddings for system 3 ('NN') could be seen as another baseline which is not only beaten by the structured `TransE` embeddings but also by the textual `WSABIE` and `Word2Vec` embeddings in system 2 ('regression') and system 3 ('NN'). This again indicates the capability of the textual frame embeddings for capturing useful information for the frame-to-frame Prediction Task to at least some extent.

Future Work. The systems could reach higher scores if the split of the data into training and test triples was done randomly per relation so that the train and test set had some (random) relation-specific overlap in frames on the position f_1 in the triple. But in this case, it would not be clear whether the systems would just perform 'lexical memorization' as pointed out by Levy et al. (2015b) in cases where the test set contains partial instances that were in the training set. We leave it for future work to contrast and explore different splits, e.g., random split, zero-overlap by relation or by all relations.

The large gain in performance obtained by the approaches involving learning compared to directly predicting the relation given the frame embeddings shows a considerable amount of knowledge contained in the FrameNet hierarchy which is not yet incorporated in the frame embeddings but learned by the models. Interestingly, `TransE` embeddings and the approaches involving learning are trained on the same set of training triples. For `TransE` embeddings this means that they could not learn all relevant information that is contained in these triples. We leave it for future work to include this knowledge gain from training the regression or neural model directly into the frame embeddings.

Furthermore, when looking at textual frame embeddings, we notice an advantage of system 2 ('regression') implementing a softmax regression model compared to system 3 ('NN'). As discussed, this might be an effect of the small number of training triples – but it could also be an effect of implementing a classification versus a ranking approach, which should be explored in future work. System 3 ('NN') implements a learning-to-rank approach with a ranking loss inspired in related work on answer selection in Question Answering, whereas the implementation of a classification approach (e.g., with a softmax) is left for future exploration.

To sum up, on the one hand, the results confirm the conclusions from the exploration in Section 5.1: the frame embeddings learned on frame-labeled text in the context of other tasks are not able to reliably mirror the frame-to-frame relations, not even when used as input representations to a classifier. On the other hand, the results clearly emphasize the influence of the structured embeddings on the performance of the best system. Thus, we propose our best-performing system of our `StruFFRel` approach for automatically generating recommendations for new annotations of relations between frames. This is the first system for an automatic frame-to-frame relation annotation for frame pairs in the FrameNet hierarchy.

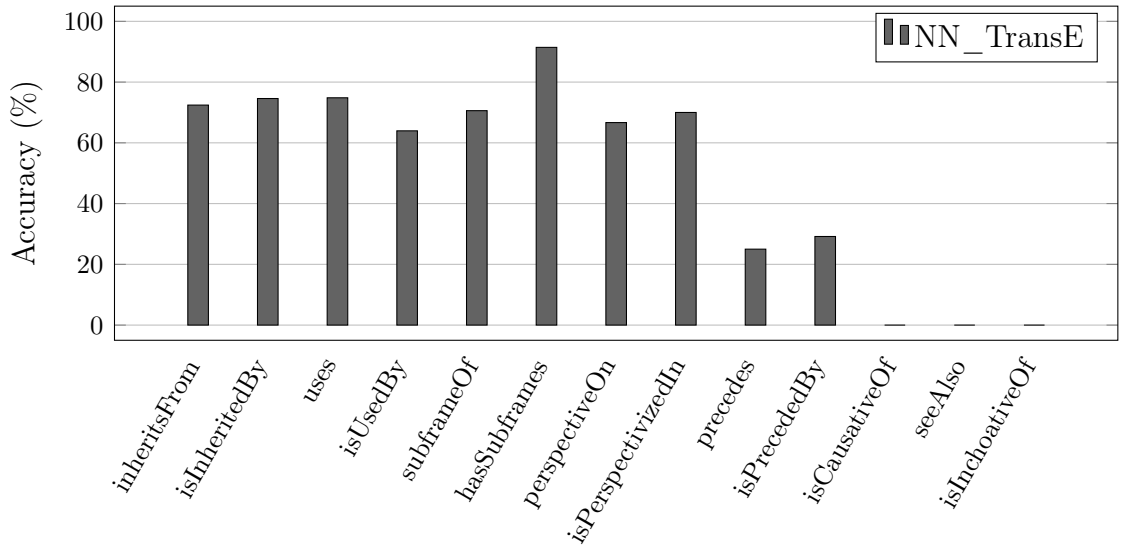


Figure 5.6: Relation-specific analysis of the best-performing model with respect to accuracy. The relation types on the x-axis are sorted by frequency.

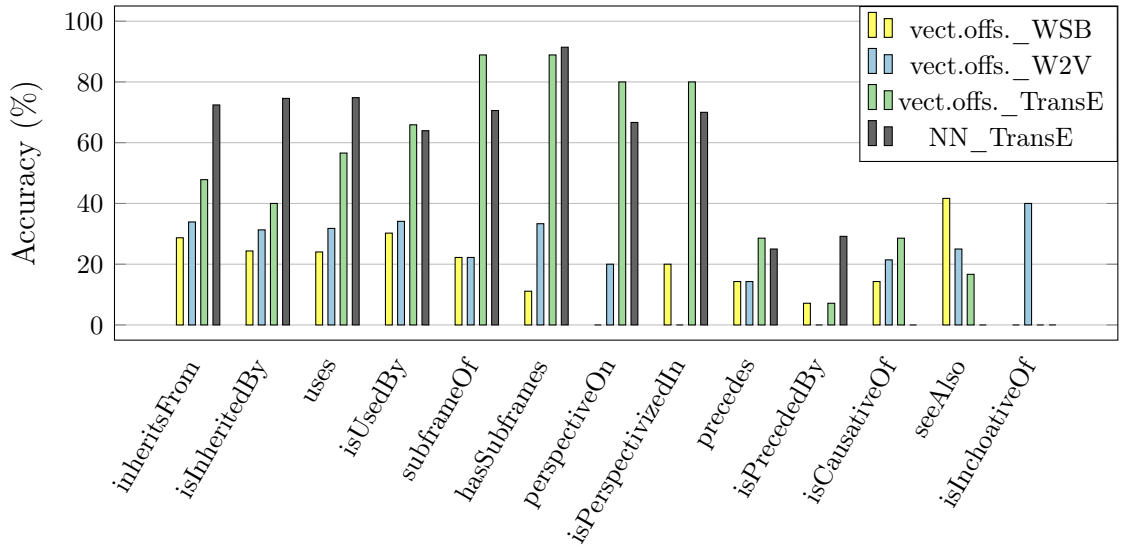


Figure 5.7: Relation-specific model comparison with respect to accuracy. The relation types on the x-axis are sorted by frequency. Again, the best model (`StruFFRel` approach with neural network and TransE embeddings) is depicted in black.

5.2.2.4 Relation-specific Analysis

We analyze the performance of the best model, system 3 ('NN') with `TransE` embeddings, by looking at all single relations. Figure 5.6 depicts a relation-specific performance of the best-performing model showing good performances (above 60% accuracy) for frequent relations, a drop for the less frequent *precedence*-relations and no capability at all in predicting infrequent relations, such as *is_Causative_of*, *see_Also* and *is_Inchoative_of*. This suggest that the predictive power for a relation in a supervised setup depends on the number of training instances available for

this respective relation (in relation to the number of training instances for the other relations). In our setup, frequent relations are trained on about 70 instances (or even more instances for the most frequent relations). For future work, we suggest to explore the performance with respect to reduced training instances in order to find out which relations suffer from loosing training data and which ones are obvious enough in embedding space to be learned from even less training data.

To compare and to refer back to our focus in Section 5.1, Figure 5.7 depicts the same relation-specific analysis for system 1 ('vector offset') with not only `TransE`- but also `WSABIE`- and `Word2Vec` embeddings.

An interesting effect is revealed by contrasting `TransE` embeddings in system 3 ('NN') that involves training and those in system 1 ('vector offset') that does not involve further training other than averaging over all relations inferred from the training frame pairs to obtain prototypical relation embeddings. The strength of the vector offset approach does not lie in the most frequent relations (as for the neural network approach) but in medium-frequency relations. Moreover, the drop for rare relations is less drastic than for the neural approach. This indicates that the neural approach loses some performance in rare relations when fitting more to the frequent relations in order to improve overall accuracy.

When comparing `TransE` embeddings to `WSABIE`- and `Word2Vec` embeddings in the system 1 ('vector offset'), it becomes apparent that `TransE` embeddings are the most informative ones for almost all relations (but *see_Also* and *is_Inchoative_of*), which supports our findings from Section 5.1. The difference in performance of structured embeddings and textual embeddings in the supervision-less vector offset approach is smaller for the most frequent relations (*inheritance*- and *using*-relations) than for the medium frequent relations (*subframe*- and *perspective*-relations). Interestingly, these medium frequency relations are very difficult ones for the textual embeddings, as only few frames with lexical units are involved in these relations (cf. Table 2.3). Thus, with respect to the most frequent relations (*inheritance*- and *using*-relations) the frame embeddings trained on text show the highest potential – but still fall behind the frame embeddings trained on the structure of the FrameNet hierarchy.

5.2.2.5 Demonstration of Predictions

We generically demonstrate the predictions of the best-performing system of our `StruFFRel` approach for unannotated frame pairs and suggest its application for automatic FrameNet completion on the relation level.

Looking back at the motivational example from the beginning, Figure 2.4 illustrated the incompleteness of the FrameNet hierarchy at the frame-to-frame relation level with the example of a possibly missing `precedence` relation from 'Being_awake' to 'Biological_urge' (evoked by the predicate 'tired'). Table 5.5 displays the top 3 frame-to-frame relation predictions for the frame pairs around 'Biological_urge' in the figure. The expected frame-to-frame relation is indeed amongst the top 3 predictions of the best performing system for this example, even for the `precedence` relation, which is rather underrepresented in the data. If this system was used to make suggestions to human expert annotators, they should be informed about the system being biased against the infrequent relations.

frame pair (f_1, f_2)	**top 3 F2F relation predictions**	**expected F2F relation**
(Biological_urge, Sleep_wake_cycle)	Subframe_of, Inherits_from, Uses	Subframe_of
(Biological_urge, Being_awake)	Is_Inherited_by, Precedes, Is_Preceded_by	Is_Preceded_by
(Biological_urge, Fall_asleep)	Is_Inherited_by, Precedes, Is_Preceded_by	Precedes

Table 5.5: Demonstration of Frame-to-Frame Relation Predictions of the best system. Given a yet unlinked frame pair (first column), our system ranks the relations (middle column) and we compare the system predictions to the expected prediction (last column).

However, it is hard to do a proper manual evaluation as judging the suggested relations requires expert knowledge of the definitions and annotation best-practices for the frame-to-frame relations. We propose using the best-performing system for semi-automatic FrameNet completion on the relation level in cooperation with FrameNet annotation experts. The system can be used to make suggestions of relations for frame pairs and the final decision could be made by experienced FrameNet annotators. This would be the first step towards improving the incompleteness of frame-to-frame relation annotations in FrameNet, which in turn could improve the performance in other tasks that take these frame-to-frame relations as input.

Following the example in our demonstration, the annotation of new frame-to-frame relations could also inspire future work on frame induction in the following was. In a strict sense, the frame *Biological_urge*, is not a full subframe of *Sleep_wake_cycle* as the majority, but not all, lexical units that evoke *Biological_urge* are also part of *Sleep_wake_cycle*. As an example, take the adjective *'hungry'*. It would be intersting to study how to split the lexical units of *Biological_urge* in order to form a new specific frame *Biological_urge_sleepy* for those lexical units that fit into the *Sleep_wake_cycle* (such as adjectives like *'tired', 'sleepy', 'exhausted'*).

5.2.2.6 Reflection

As frame-to-frame relations of the FrameNet hierarchy did not emerge from frame embeddings learned on frame-labeled text, the frame-to-frame relations should be seen as metastructures not having direct evidence in text. On the one hand, more advanced approaches might be needed to distill frame-to-frame relations for frames occurring in raw text, by learning about commonsense knowledge involving frames, and then inferring the implicit relations. Here, it could also be helpful to exploit inter-sentential clues e.g., event chains, to enrich the frame embeddings which so far are built on sentence level. On the other hand, the automatic completion of frame-to-frame relations can rely on structured embeddings trained on the hierarchy. To this end, an expert evaluation of the best-performing system's predictions for frame pairs could give clues for further system improvements. It could also yield an expert upper bound and may pave the way for developing advanced systems using frame embeddings for the prediction of frame-to-frame relations. Finally, future work could

investigate the case of FrameNet for embeddings learned on both, frame-labeled texts and frame-to-frame relation annotations. By having such a combination, the limitation of the textual embeddings on frames that have lexical units (and hence occur in text) can be overcome as the structured embeddings also have access to frames without lexical units.

Advantage of Textual Frame Embeddings not Mirroring the Hierarchy: WordNet versus FrameNet embeddings. Last but not least, for different tasks, different representations of frames and relations might be better suited: embeddings purely learned on text, or embeddings purely learned on the FrameNet hierarchy, or a combination of both. Textual frame embeddings not mirroring the FrameNet hierarchy turns into an advantage when comparing embeddings learned with WordNet versus embeddings learned with FrameNet. It indicates that the FrameNet hierarchy indeed offers additional information in the frame-to-frame relations that is not incoporated into textual frame embeddings. Concerning WordNet, however, previous work argues that information in WordNet overlaps with word embeddings (Zhai et al., 2016). This suggests that frame semantic information might contribute additional knowledge to semantic tasks which is not accessible via standard word embeddings or via other external lexical knowledge bases such as WordNet. The actual potential of frame knowledge when applied in higher-level tasks is studied in the next chapter 6.

5.2.3 Recommendation for Visual Frame-Relation-Induction

In the previous Sections 5.1 and 5.2 we have seen that textual and structured frame embeddings differ in modeling and prediction frame-to-frame relations. It would be interesting to explore to what extent structured frame embeddings can benefit from addition information in embeddings from further modalities. To this respect, we see a large potential of leveraging the visual modality. In the previous Chapter 4 we have seen that the frame-semantic task of Frame Identification can benefit from additional visual information. This leads to the question whether the task of predicting frame-to-frame relations could also benefit from additional visual information. However, for Frame-to-Frame Relation Prediction, visual embeddings for frames are needed, but these are not trivial to provide. As a first approach, the *imSitu* dataset (intentionally designed for verb identification) and its links to FrameNet could be used to provide initial images for some frames. It is subject to discussion and research how to assign images to frames. Furthermore, it is yet to be determined whether frames are best to capture by the visual modality or rather by further modalities.

In lack of images for frames with a broad coverage, we explore the potential of multimodal approaches to traditional Knowledge Base Completion (Section 5.2.3.1). Finally, we suggest to develop approaches that incorporate visual information about frames to benefit Frame-to-Frame Relation Prediction and also frame induction (short frame-relation-induction).

Embedding Space	Top Similar Synsets
Linguistic	n02472987_world, n02473307_Homo_erectus, n02474777_Homo_sapiens, 02472293_homo, n00004475_organism, n10289039_man
Visual	n10788852_woman, n09765278_actor, n10495167_pursuer, n10362319_nonsmoker, n10502046_quitter, n09636339_Black
Structure (TransE)	_hypernym, n00004475_organism, n03183080_device, n07942152_people, n13104059_tree, n00015388_animal, n12205694_herb, n07707451_vegetable

Table 5.6: Closest synsets to the person synset (n00007846) according to different embedding spaces.

5.2.3.1 Excursion – Multimodal Knowledge Base Completion

In addition to our findings with respect to multimodal Frame Identification, which is a semantic task, we also explore the effect of information form multiple modalities with respect to the structure-focused task of Knowledge Base Completion (as explained in Section 2.2.1), see (Mousselly-Sergieh et al., 2018). We explore whether Knowledge Base Completion, with the established baseline approach `TransE` using structured embeddings (as in Section 3.3), can profit from additional modalities, namely textual and visual embeddings.

To gain initial insights into the potential benefits of external information for the task of Knowledge Base Completion, we consider the embeddings produced by the translation-based `TransE` method (Bordes et al., 2013) on the WN9-IMG dataset (Xie et al., 2017), which we used in Section 4.2.1 to obtain image embeddings for the synsets. This dataset contains a subset of WordNet synsets, which are linked according to a predefined set of linguistic relations, e.g. *hypernym*. We observe that `TransE` fails to create suitable representations for entities that appear frequently as the head/tail entity of *one-to-many/many-to-one*-relations. For example, the entity *person* appears frequently in the dataset as a head/tail entity of the *hyponym/hypernym* relation; the same holds for entities like *animal* or *tree*. `TransE` represents such entities as points that are very close to each other in the embedding space (cf. Table 5.6). Furthermore, the entity embeddings tend to be very similar to the embeddings of relations in which they frequently participate. Consequently, such a representation suffers from limited discriminativeness and can be considered a main source of errors for different knowledge graph inference tasks.

To understand how multimodal representations may help to overcome this issue, we perform the same analysis by considering two types of external information: textual and visual embeddings. The textual embeddings are created by implementing the `Word2Vec` approach (Mikolov et al., 2013a), and the visual ones are obtained from the feature layers of the `VGG` Convolutional Neural Network model (Chatfield et al., 2014, cf. Section 3.4) on images that correspond to the entities of the dataset. For the same category of entities discussed above, we observe that both the visual and the textual embeddings are much more robust than the structure-based embeddings of `TransE`. For instance, *person* is closer to other semantically related

concepts, such as *Homo_ erectus* in the textual embedding space, and to concepts with common visual characteristics (e.g., *woman*, *actor*) in the visual embedding space (cf. Table 5.6). Furthermore, the textual and the visual embeddings seem to complement each other and hence are expected to enhance knowledge graph representations if they can be leveraged during the representation learning process.

We analyze an extension of Xie's approach for knowledge graph representation learning by leveraging two different types of external, multimodal representations: not only *visual* embeddings obtained from images corresponding to the knowledge graph entities, but also *textual* embeddings created by analyzing the usage patterns of knowledge graph entities in text corpora.

We test our multimodal representations on the tasks of link prediction and triple classification as explained in Section 3.3. The results indicate that for these structure-focused tasks, a combination of all 'modalities' (structured, textual and visual) is of benefit compared to other multimodal approaches that enrich the traditional approach of structured embeddings with visual embeddings.

From this, we conclude that for language organization in terms of relations and knowledge bases, in addition to structured information, the combination of visual information and textual information is helpful. Also, we conclude that for each task it is worth investigating which combination mechanism is the most successful one – for Knowledge Base Completion, simple concatenation $\frown$ (cf. Equation 3.19) outperforms the Imagined method for mapping into the same space, whereas for Frame Identification, the Imagined method brings the advantage of obtaining embeddings in the visual space for every word.

5.3 Summary of the Chapter

Taken together, in this chapter we finalize the applicability of Frame Identification in higher-level tasks by complementing the Frame Identification system (cf. Chapter 4) with different sets of frame embeddings: textual versus structured frame embeddings. These contributions will be applied to higher-level tasks in the following (cf. Chapter 6), where, in particular, we showcase the application pipeline and the potential of frame knowledge with the task of Argumentative Reasoning Comprehension (cf. Section 6.2.2).

We provide a summary of this chapter in terms of bullet points in the box below:

INSIGHTS IN A NUTSHELL

For language understanding linked to structured knowledge bases, we focus on FrameNet's **frame-to-frame relations**.

- Textual frame embeddings do not mirror frame-to-frame relations
- Structured frame embeddings in a supervised approach are best at predicting frame-to-frame relations
- We introduce the approach `StruFFRel` leveraging the structure of the FrameNet hierarchy for Frame-to-Frame Relation Prediction
- Frequent relations have an advantage over rare relations
- Multimodal Knowledge Base Completion profits from combining textual, structured and visual embeddings

→ We recommend to develop visual embeddings for the induction of frame-to-frame relations and of frames.

Chapter 6

Extrinsic Evaluation: Applications of Unimodal Frame Knowledge

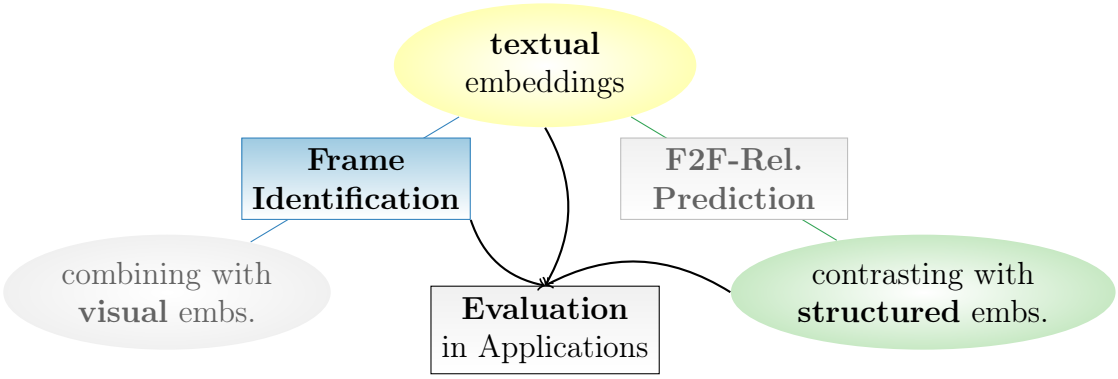


Figure 6.1: Chapter 6. Middle gray box: evaluation of frame knowledge (Frame Identification system, textual and structured frame embeddings) in high-level applications.

After having introduced a new state-of-the-art system for Frame Identification (cf. Chapter 4) and also different kinds of frame embeddings (cf. Chapter 5), we now address the following question, as indicated in Figure 6.1:

> RQ: *What is the potential of FrameNet's knowledge in terms of identified frames and pre-trained frame embeddings for higher-level language understanding tasks?*

Regarding the higher-level language understanding tasks, we consider the tasks of Summarization, Evaluation of Summaries, Motif Construction, Semantic Textual Similarity, and Argument Reasoning Comprehension. These tasks all require a semantic understanding of a source text to then extracting information, detecting patterns, or judging similarity on top of the acquired understanding. Frame knowledge could crucially enhance semantic understanding. Consequently, the Frame Identification system provides frame annotations which can serve as abstractions from the actual words used in texts, and the frame embeddings provide semantic knowledge

in terms of frames and frame-to-frame relations which provide meta-knowledge of how different frames relate to each other.

Next, we motivate the use of frame knowledge as an external source of knowledge for application in semantic tasks and then we give an overview of the organization of this chapter.

Motivation: Semantic Tasks Profit from External Knowledge. Language understanding requires more complex knowledge than that contained in current systems and word embeddings. For the task of semantic plausibility, Wang et al. (2018) reveal the failure of models only relying on distributional data, whilst the injection of world knowledge helps. Glockner et al. (2018) point out the deficiency of state-of-the-art approaches for understanding entailment on the large-scale SNLI corpus (Bowman et al., 2015, short for Stanford Natural Language Inference). In their study, the model incorporating external lexical information from WordNet, KIM (Chen et al., 2018, short for Knowledge-based Inference Model), does not yield the awaited improvements — where the crucial point might be WordNet (Miller, 1990; Fellbaum, 1990) which does not contain explicit world knowledge in the form of knowledge about situations and actions and about structured relations.

Previous work argues that information in WordNet overlaps with word embeddings (Zhai et al., 2016), therefore we focus on other types of knowledge in our work, namely frame semantic information. In Section 5.1 we have seen evidence that the frame semantic information contained in the frame-to-frame relations of the FrameNet hierarchy is not covered by word embeddings, but better incorporated by structured embeddings. Taken together, this suggests that frame semantic information might contribute additional knowledge to semantic tasks which is not accessible via standard word embeddings or via other external lexical knowledge bases such as WordNet.

Overview: Evaluation of Frame Knowledge in Applications. In this chapter, we aim at examining the hypothesis of FrameNet's knowledge contributing to a semantic understanding that is beneficial for higher-level tasks.

We elaborate on several applications of frame semantic information, namely in the context of the tasks of Summarization (Section 6.1.1), Evaluation of Summaries (Section 6.1.2), Motif Construction (Section 6.1.3), Semantic Textual Similarity (Section 6.2.1) and Argument Reasoning Comprehension (Section 6.2.2).

These tasks all have in common that first, they require a semantic understanding of a text. And then, they perform subsequent steps on top of that, such as extracting important information, detecting relevant patterns which repeat, judging semantic similarity, or choosing a plausible chain of reasoning.

We sort the applications by the amount of knowledge they use from FrameNet in our experimental setups. First, only using frame labels (Summarization, Evaluation of Summaries and Motif Construction in Section 6.1), and second, additionally leveraging information from frame embeddings and frame-to-frame relations (Semantic Textual Similarity and Argument Reasoning Comprehension in Section 6.2).

On the one hand, we find a tendency of frame information to be beneficial for these higher-level tasks in specific cases where they can bridge a semantic gap. On the other hand, when looking at the whole datasets, the impact of frame-enhanced

approaches is rather small. Thus, we finally conclude the chapter with reflecting on the potential of including frame information versus using end-to-end approaches that do not require any linguistic pre-processing (Section 6.3).

The publications stem from collaborations where we provide annotations with frames or frame embeddings. Thus, we summarize the approaches whilst referring the interested reader to the original papers for more details. Importantly, we elaborate on the insights with respect to frame knowledge. Especially with respect to the application case of Argument Reasoning Comprehension (Section 6.2.2), we discuss it in more detail, as in this application, we involve all our work on frame knowledge.

Our papers (Zopf et al., 2018a)[1] and (Peyrard et al., 2017)[2] and (Botschen et al., 2018b)[3] are foundational to this chapter.

6.1 Applications of Unimodal Frame Identification

Our Frame Identification system has been applied to different higher-level tasks: `SimpleFrameId` for Summarization (Section 6.1.1) and the evaluation of summarization systems (Section 6.1.2) (both English only), and the tuned new state-of-the-art Frame Identification system `UniFrameId` for Motif Construction (Section 6.1.3) (English and German). We report about the applications to showcase the potential of Frame Identification towards language processing.

Applying Frame Labels in Summarization, Evaluation of Summaries, and Motif Construction. In these three tasks, the extraction of important information and the detection of repetitive patterns is central (the tasks we will explained in detail in the respective sections).

We hypothesize that frame information can benefit these tasks in terms of *(a)* providing abstractions from word-level to frame-level, *(b)* disambiguating predicates with respect to the context, *(c)* offering meta-knowledge about the frame labels (e.g., frame-to-frame relations), and *(d)* eventually giving access to frame-specific role labels. Applying Frame Identification systems to obtain frame labels yields linguistically motivated abstractions from the text, which, in turn, can be used as input features. Whilst WordNet also implements the benefit of word sense disambiguation, FrameNet additionally offers access to information about frame-to-frame relations for broader context or world knowledge and frame-specific roles for syntactic information.

We observe positive tendencies obtained by the abstraction with frame labels.

6.1.1 Summarization – Estimating Importance with Frames

The effect of frame labels by `SimpleFrameId` (Hartmann et al., 2017) to estimate importance in the context of summarization is explored by Zopf et al. (2018a).

[1] My contribution in this paper is the following: annotation of texts with frames using `SimpleFrameId` system.

[2] My contribution in this paper is the following: annotation of texts with frames using `SimpleFrameId` system.

[3] My contributions in this paper are the following: annotation of texts with frames using `UniFrameId` system, input with frame embeddings, analysis with respect to frames.

Summarization is a complex high-level task that requires the detection of important information. For learning a notion of importance, automatic systems can be trained to regard patterns (e.g., keywords) as important if they appear in source texts as well as in the corresponding reference summaries. With respect to frame labels, this means to learn a distinction between important and less important frames – such that sentences containing *important* frames are selected for the system summary.

Task of Summarization. The research field of text summarization aims at automatically extracting the gist of source documents such as news articles. Current approaches can be grouped into extractive and abstractive summarization. In extractive summarization, the task is to extract the most important sentences of the collection of sentences in the source text. Abstractive summarization is not restricted to an exact repetition of source sentences but allows for reformulations in order to create new sentences that contain the condensed information which is most important. In both cases, the basic goal is to judge the importance of single information nuggets to then decide which sentences are important enough to be reported in a summary. To do so, automatic systems can be trained to learn a notion of importance on a training set of source texts corresponding reference summaries.

Below, we give an example of a source text with its reference summary. The source text is shortened from originally 28 sentences to the first three sentences. It stems from the DeepMind Q&A dataset (Hermann et al., 2015) which contains documents from news articles of CNN (common abbreviation for Cable News Network) together with bullet points (here 3 sentences) to summarize the article.

> **Source text:**
> *Abu Dhabi, United Arab Emirates (CNN) – Thirty-three people have now died from the MERS coronavirus, the World Health Organization said Friday. Three new cases, including a fatality, were recorded in Saudi Arabia, the kingdom announced this week. Another previously reported Saudi Arabian patient also died from the disease, which lacks a cure or vaccine. ...*
> **Reference summary:**
> *The WHO says 58 cases of MERS have been reported; 33 have died. Researchers have mapped the genetic characteristics of the virus. While cases are limited, MERS has killed more than half of its victims.*
> **System summary:**
> The system summary should be as similar as possible to the reference summary, ideally it would be exactly the same. The length is restricted to the number of sentences in the reference summary.

Evaluation Metric ROUGE. ROUGE (Lin, 2004, short for Recall-Oriented Understudy for Gisting Evaluation) is the standard evaluation measurement in summarization to score the similarity between system summaries and reference summaries with respect to the detection of importance and the avoidance of redundancy. ROUGE-N measures the n-gram overlap of system summaries and reference summaries, i.e. ROUGE-2 considers the bigram overlap.

ROUGE recall captures how many text elements (e.g., bigrams) of the reference summary are contained in the system summary. However, high recall can be

obtained even if the system summary includes irrelevant text elements, as long as many text elements are overlapping in reference and system summary.

ROUGE precision captures how many text elements of the system summary were in fact relevant or needed. However, high precision can be obtained by including only those text elements into the system summary for which the system is extremely confident, and thus, not risking to include those for which the system is unsure.

Potential of Frames in Summarization. We annotate frames for verbs in the above example of a (shortened) source text with its reference summary to showcase the potential of frames for detecting *important* text elements in the source text when knowing the reference summary.

> **Source text:**
> *Abu Dhabi, United Arab Emirates (CNN) – Thirty-three people have now died* DEATH *from the MERS coronavirus, the World Health Organization said* STATEMENT *Friday. Three new cases, including a fatality, were recorded* RECORDING *in Saudi Arabia, the kingdom announced* STATEMENT *this week. Another previously reported* STATEMENT *Saudi Arabian patient also died* DEATH *from the disease, which lacks* POSSESSION *a cure or vaccine. ...*
> **Reference summary:**
> *The WHO says* STATEMENT *58 cases of MERS have been reported* STATEMENT *; 33 have died* DEATH *. Researchers have mapped* SCRUTINY *the genetic characteristics of the virus. While cases are limited, MERS has killed* KILLING *more than half of its victims.*

The example demonstrates that the reference summary incorporates the statements about death from the source text in terms of the frames *'Statement'* and *'Death'*. As the system by Zopf et al. (2016a) learns to estimate the importance of text elements, here it could learn that the frames *'Statement'* and *'Death'* might be important since they appear both in source text and in reference summary, whilst *'Recording'* and *'Possession'* might be less important since they do not appear in the summary. Finally, this notion of importance for frames (learned on samples of sources texts and corresponding reference summaries) could then help the system to make a frame-semantically informed decision about which sentences from a source text to include into the system summary.

Setup. Initially, (Zopf et al., 2016a) introduce `CPSum` (Zopf et al., 2016a, short for contextual preferences in summarization), a systems for extractive summarization that learns to estimate the importance of text elements from pairwise preferences. Then, Zopf et al. (2018a) point out that `CPSum` judges the importance only by considering bigrams of all words in the source texts – thus, they propose to investigate the potential of other text annotations such as frames, named entities or sentiment annotations by extending `CPSum`. To obtain a summary for a source text, sentence utilities are estimated on the basis of the utilities of smaller text elements such as bigrams or frames, resulting in a ranked list of sentences. Finally, this ranked list of sentences from the source text (that could result in a system summary by selecting the x highest ranked sentences) should be as similar as possible to the ranked list

of sentences with respect to the reference summary. For more details, the interested reader is referred to the original papers (Zopf et al., 2016a, 2018a).

The sentence scores are either optimized for ROUGE recall or ROUGE precision when evaluation system summaries versus reference summaries. (For different properties of recall and precision with respect to summarization, see (Zopf et al., 2018b).) For ROUGE recall, Zopf et al. (2018a) simply add the learned utility scores of all text elements (e.g., frames) that are annotated in a sentence, whilst for ROUGE precision they compute the mean with the sentence length.

As the output of the system is not a system summary but a ranking over all sentences (from which the first ones could be selected for a summary), Zopf et al. (2018a) do not evaluate in terms of ROUGE, but with rank correlation metrics such as Kendall's tau (Kendall, 1938) for capturing the quality of the predicted ranking.

The experiments are conducted with data from three well-known datasets for multi-document summarization datasets: the DUC 2004 (common abbreviation for Document Understanding Conferences), TAC 2008 (common abbreviation for Text Analysis Conference), and TAC 2009 corpora.[4]

The impact of different annotations was tested in a collaborative effort. We contribute annotations with frames identified by `SimpleFrameId` for source documents and for reference summaries. Here, we report the insights concerning frames. We contrast frame annotations for nouns only and for verbs only in order to separately evaluate the effect of different parts-of-speech.

Results and Take-away. Zopf et al. (2018a) report that annotations other than bigrams have shown potential in specific situations, even though uni- and bigrams are hard to outperform. More specifically, n-grams are hard to beat if sentences from the input documents have to be ranked. On the other hand, n-grams are surprisingly weak if source and reference sentences have to be distinguished.

Amongst the annotations used to extend the ranking-based approach, frame annotations could increase the performance in some cases. We can identify three major findings with respect to frame annotations, which we will report in the following: nouns versus verbs, recall versus precision, and ablation study. Finally, we report additional insights from an analysis of frequent frames versus frames with high utility scores according to the system.

Frame Annotations for Nouns versus Verbs. Interestingly, Zopf et al. (2018a) report that the results obtained with frame annotations for nouns only strongly outperform those for verbs only (difference of .27 in the correlation metric of Kendall's tau on unseen test data). This suggests that, for this dataset, the importance is better captured by entities (nouns) than by actions (verbs). As the DUC and TAC corpora contain newswire texts from the general domain, the crucial information nuggets are indeed persons, locations, international players that tend to be expressed by nouns. Zopf et al. (2018a) did not only experiment with frame annotations, but also with 20 other annotations; and we can observe the effect of nouns being more suitable than verbs for estimating importance also in other kinds of annotations: This effect is also mirrored by the rather strong performance of

[4] DUC 2004: `https://duc.nist.gov/` and TAC 2008/9: `https://tac.nist.gov/`

entity annotations (amongst the top 5 annotations) and at the same the rather weak performance of verb-stem annotations. Despite some verbs being important and likely to be selected for the summary (e.g., 'kill' or 'betray'), in news articles verbs are often used to report about statements (e.g., a person 'said' something or 'expressed' an opinion) where the verb does not indicate general importance per-se.

Scoring with respect to ROUGE Recall versus Precision. Zopf et al. (2018a) report that for noun frames, the correlation when scoring the sentences in the source text with respect to ROUGE recall has been improved (difference of .36 in the correlation metric of Kendall's tau on unseen test data). As can be expected, the advantage in correlation for ROUGE recall versus precision is visible in all annotations (average difference of .21 for Kendall's tau on unseen test data), not only for frames. This can be attributed to the following interplay of longer sentences, recall, and importance of text elements. On the one hand, longer sentences are not punished in the measurement of ROUGE recall when comparing system summaries to reference summaries. On the other hand, they are more likely to contain relevant text elements by design. Thus, the recall-focused system scores and ranks longer sentences higher than shorter ones – and therefore gets better correlation performance than the precision-focused system.

Ablation Study of all Annotations in Ensemble. Zopf et al. (2018a) perform an ablation study with an aggregated ranking of all annotation elements with always excluding one annotation (e.g., frames) in order to find out how much performance is lost if one annotation element is removed from the ensemble. With respect to Kendall's tau on unseen test data, frame-based annotations show reasonably good performance: they are 3rd in causing a performance drop when being removed. Interestingly, now the frames are 3rd when scoring the sentences with respect to ROUGE recall as well as precision – even though they did not perform well with respect to precision in the first experiment (see above). Thus, frame annotations seem to contribute to the ensemble.

Frequent Frames versus Frames with High Utility Scores. In a further analysis, Zopf et al. (2018a) compare the ten most frequent frames annotated in the source documents with the ten most likely frames to appear in the reference summary if they appear in a source document. Precisely, the ten most likely frames are those with the highest utility scores according to the system. On the one hand, the frames *'Cardinal_numbers'* or *'Measure_duration'*, which are evoked by numbers or durations, are amongst the ten most frequent frames, but not very likely to be indicative with respect to relevance for the reference summary. And on the other hand, the frames *'Killing'* or *'Kidnapping'* are amongst the ten frames with highest utility scores, even if not being frequent. This illustrates that pure frequency is not equal to importance with respect to summaries in terms of frames.

Finally, from Zopf et al. (2018a), it can be taken away that FrameNet frames could help to learn a reasonable notion of importance for summarization systems. Especially when annotating frames for nouns and also when scoring sentences with respect to ROUGE recall, frames are suitable to convey importance. Most interestingly, an ablation study indicates the positive effect of frames in an ensemble.

6.1.2 Summary Evaluation – Judging Quality with Frames

The effect of frame annotations with `SimpleFrameId` (Hartmann et al., 2017) to judge the quality of system summaries in the context of the evaluation of summaries is explored by Peyrard et al. (2017). In the research field of text summarization, the evaluation of automatically produced summaries is an open question, which is approached by Peyrard et al. (2017). They suggest to not only rely on the metric ROUGE (as introduced in Section 6.1.1) but to learn an automatic scoring metric based on human judgments. In their setup, any already existing automatic metric (such as ROUGE) can be incorporated as a feature – and also new metrics (e.g., leveraging frame information) can be included. Then, the model learns the best combination of features for matching human judgments.

Task of Summary Evaluation. In general, evaluation metrics for judging the quality of system summaries are compared based on their ability to correlate with scores provided by humans for the quality of these system summaries. Thus, Peyrard et al. (2017) suggest to automatically learn how to score system summaries based on human judgments for the quality of system summaries with respect to reference summaries. They design a learning setup where an automatic scoring function is learned based on a training session with reference summaries, the corresponding candidate summary, and a human judgment of how well they fit. The automatic scoring function shall approximate the human score in terms of correlation, i.e. the function shall rank the system summaries similar to how humans ranked it. The scoring function is learned based on features which can be any already existing automatic metric (such as ROUGE) – or, it can also be new metrics (e.g., leveraging frame information). The respective metric, i.e. a feature, expresses a judgment of the quality of each system summary with respect to the corresponding reference summaries. On this basis, the scoring function learns to approximate the human judgments of quality by considering different automatic measurements for similarity of reference and system summary.

Below, we give an example of a reference summary with its system summary from the TAC 2009 corpus.[5]

> **Reference summary:**
> *The first US offshore wind turbine plan was in Nantucket Sound. Opponents of offshore wind power sought a more comprehensive strategy for wind energy and argued it would spoil vacation area aesthetics, disrupt commercial fishing and harm birds and marine life. One study found the proposal would have little effect on the surrounding air, sea, and animal life or fishing conditions. Another study concluded the turbines could alter local weather. Fishermen opposed clusters of windmills on Spain's southern shore because they would disturb tuna migration and necessitate dangerous detours.*
> **System summary:**
> *Wind power is widely used in Europe, both on land and offshore. Opponents emphasize the aesthetic impact of the project, while a draft environmental report said that it would not hinder commercial or sport activities nor would it kill birds or fish or affect currents, water quality, or noise levels. Now the first*

[5] TAC 2008/9: `https://tac.nist.gov/`

U.S. off-shore wind farm has been proposed for Nantucket Sound. Opponents say that the turbines are ugly, disrupt fishing, harm birds and marine wildlife, and affect local weather. Proponents of wind power assert that it is a safe, non-polluting, renewable alternative to fossil fuel.

Potential of Frames in Summary Evaluation. Frames provide semantic information by abstracting from word-level to frame-level and by disambiguating predicates with respect to the context. When judging the quality of a system summary with respect to the similarity it shows to the reference summary, frame labels can serve as a word sense disambiguation. Ambiguous expressions in system and reference summary can be disambiguated to either referring to the same sense or to referring to different senses.

In order to showcase the intention, we annotate frames for some selected nouns and adjectives in the above example of a reference summary and system summary:

Reference summary:
The first US offshore wind turbine ELECTRICITY *plan was in Nantucket Sound. Opponents of offshore wind power* ELECTRICITY *sought a more comprehensive strategy for wind energy* ELECTRICITY *and argued it would spoil vacation area aesthetics, disrupt commercial fishing and harm birds and marine life. One study found the proposal would have little effect on the surrounding air, sea, and animal life or fishing conditions. Another study concluded the turbines* ELECTRICITY *could alter local weather. Fishermen opposed clusters of windmills on Spain's southern shore because they would disturb tuna migration and necessitate dangerous* RISKY_SITUATION *detours.*
System summary:
Wind power ELECTRICITY *is widely used in Europe, both on land and offshore. Opponents emphasize the aesthetic impact of the project, while a draft environmental report said that it would not hinder commercial or sport activities nor would it kill birds or fish or affect currents* ELECTRICITY *, water quality, or noise levels. Now the first U.S. off-shore wind farm has been proposed for Nantucket Sound. Opponents say that the turbines* ELECTRICITY *are ugly, disrupt fishing, harm birds and marine wildlife, and affect local weather. Proponents of wind power* ELECTRICITY *assert that it is a safe* RISKY_SITUATION *, non-polluting, renewable alternative to fossil fuel.*

The example showcases that the system summary has an overlap with the reference summary in terms of frames (here dominant frames of *'Electricity'* and a mention of the frame *'Risky_situation'*). Here, two different sets of words ((turbine, power, energy, turbines, dangerous), (power, currents, turbines, power, safe)) are correctly mapped to the same set of frames (*'Electricity'*, *'Risky_situation'*). This illustrates one advantage of the abstraction with frames for ambiguous words: texts can use different words to express the same meaning and these words should be mapped to the same frame.

Now, the scoring function learns to approximate the human judgment of quality given the similarity of reference summary and system summary in terms of frames.

Approach. Peyrard et al. (2017) learn an automatic scoring metric for system summaries based on human judgments for the quality of these summaries in com-

parison to reference summaries. The training of the function is performed with reference summaries, the corresponding candidate summary, and a human judgment of how well they fit – so that the function approximates the human score in terms of correlation. More specifically, the function computes a score for each system summary based on its similarity with the reference summary. This score should rank system summaries in the same order as humans – and by this, achieve a high correlation.

The correlation metrics are Pearson product moment correlation coefficient (cf. Equation 3.17) or Spearman rank correlation coefficient (cf. Equation 3.18). Other than measuring *system-level correlation* where ROUGE performs well, Peyrard et al. (2017) suggest to measure *summary-level correlation* where performance of ROUGE drops and where, crucially, the correlation between human judgments and automatic scores is calculated for each topic and then averaged across topics. Thus, *summary-level correlation* measures how well evaluation metrics correlate with human judgments for summaries and not only for systems. A metric with a high summary-level correlation will be more robust, which is particularly important when this metric is used for training summarization systems.

Peyrard et al. (2017) point out ROUGE-N (Lin, 2004) to be the most straightforward feature: it computes the n-gram overlap between the candidate summary and the pool of reference summaries. Also, they include as features the variants identified by Owczarzak et al. (2012) as strongly correlating with humans: ROUGE-2 recall with stemming and stopwords not removed (giving the best agreement with human evaluation), and ROUGE-1 recall (the measure with the highest ability to identify the better summary in a pair of system summaries). Furthermore, Peyrard et al. (2017) experiment with frames as features for evaluating summaries.

Frames are more abstract than words, thus different but related words might be associated with the same frames depending on the meaning of the words in the respective context. All nouns and verbs of the reference and candidate summaries are replaced with their frames. This frame annotation is done with the best-performing system configuration from `SimpleFrameId` pre-trained on all FrameNet data. It assigns a frame to a word based on the word itself and the surrounding context in the sentence. ROUGE-N enriched with frame annotations can now match related words through their frames. Analogous to the variants of ROUGE-N, ROUGE-1 and ROUGE-2, with respect to frames the frame-enriched metric is called Frame-N, and the unigram and bigram variants are called Frame-1 and Frame-2.

Results and Take-away. With respect to the analysis of the frame features Peyrard et al. (2017) report the following. The simple and straight-forward metric ROUGE-N (for uni- and bigrams) is hard to outperform by Frame-N and also by other semantically motivated metrics like ROUGE-N-WE using word embeddings. ROUGE-N achieves slightly better correlation scores than Frame-N and considerably better scores than ROUGE-N-WE. Interestingly, when comparing both semantically motivated metrics, Frame-N and ROUGE-N-WE, the frames perform consistently better.

We manually explore the data and the results with respect to frame annotations in oder to find hints on why the frame information could not further improve the correlation scores. We observe the two major properties of the data with respect to

frame annotations which we will report in the following and with respect to which we make suggestions for future work.

Texts are rarely ambiguous. The system and reference summaries in TAC 2008/9 are rarely ambiguous; they are newswire texts that try to convey their messages to the point and therefor use a precise writing style with little ambiguity. More specifically, there are *(a)* not many cases where two different words refer to the same frame and there are *(b)* almost no cases where the same words refer to different frames. Due to little ambiguity, in many cases the most frequent frame is the correct label.

Noise of incorrect labels. Obviously, the most frequent frame label is of help when mapping two different words to the same most frequent frame. However, this benefit of correct frame annotations is outweighed by the noise introduced with incorrect frame annotations – which can happen if a frame label other than the most frequent frame is chosen, for example.

Moreover, it is hard to identify single examples where the scoring with and without frame enrichment works well or badly as a correlation of a ranking is measured (not single scores, but the correlation of two ranked lists).

Taken together, the benefit by frames is small due to low ambiguity of data and this benefit is hidden by some noise from wrong frame predictions. Consequently, simple lexical comparison still seems to be better for evaluation of summaries.

Suggestions for future work. Based on our observations of properties of the data with respect to frame annotations, we suggest the following for future work in this specific approach.

Going beyond TAC data about news in a precise writing style with little ambiguity, there are broader use-cases for summarization and evaluation thereof. To give an example, the *h*MDS corpus (Zopf et al., 2016b), a heterogeneous, multi-genre corpus for Multi-Document Summarization contains more heterogeneous data which, for this reason, could be more ambiguous. In more ambiguous settings, we expect frames to be more beneficial.

Even if it is hard to pinpoint single examples of success or failure, we would suggest the following approach to improve for future work. With a correlation plot of ROUGE-N, the most extreme outliers could be identified. Then, the same correlation plot could be done for frame-N in order to see where these outliers have moved, i.e. to see whether the frame enrichment helps to improve upon outliers in ROUGE-N.

Finally, form Peyrard et al. (2017), it can be concluded, that improving summary evaluation with frames as features does not work straight away on any texts of any domain. We suggest to explore the approach with more heterogeneous texts.

6.1.3 Motif Construction – Identifying Patterns with Frames

The effect of frame annotations with `UniFrameId` (Botschen et al., 2018a) to identify patterns in the context of motif construction is explored by Arnold (2018). He uses

motifs to assess properties of texts such as text quality or changes of topics – where a motif is a recurring pattern in a graph built out of the sequence of words in textual data. In order to assess the change of topics over time, he introduces *metamotifs* – which are motifs of motifs. Interestingly, traditional motifs are contrasted with metamotifs, where he uses frame labels to first generalize from the surface text level and to then build metamotifs out of these abstractions.

Arnold (2018) builds metamotifs out of frames that were identified by the system `UniFrameId` (English and German variants) on US presidency and German Bundestag data in order to predict parties from speeches. He finds that the proposed metamotifs, using the abstraction layer of frame annotations, show a significant increase in discriminatory power compared to simpler methods – which, concerning the frames, re-assures their abstractive power for higher-level tasks.

Motifs and Metamotifs. Arnold (2018) describes *motifs* as small, connected subgraphs of a larger network. Whilst the size of motifs can vary, they often consist of three or four nodes. Applied to textual data, motif analysis in the approach by Arnold (2018) uses a sentence level graph representation based on shared nouns. This means, starting with a selected Wikipedia article, every sentence is represented by a node and two nodes are connected if and only if *(a)* they share at least one noun token, and *(b)* they are separated by at most two other sentences.

Taken to a next level, Arnold (2018) builds motifs out of frame labels annotated for textual data by first creating graph structures from the annotations and then extracting *frame-motifs*.

Furthermore, he builds *metamotifs* as motifs of frame-motifs in order to classify texts on the basis of these high-level abstractions. He defines a metamotif as a connected graph, that is, in contrast to a motif, a specific combination of its included motifs.

Potential of Frames for Metamotifs. In order to build higher-level abstractions such as metamotifs, frames can be leveraged as a first abstraction from the text and also to capture universal language and content patterns independent of the respective speaker. The intention by Arnold (2018) for detecting recurrent patterns and obtaining metamotifs from texts via frame annotations is illustrated in the example we give below. It is an excerpt of a political debate text (United States presidential election 2016: Trump versus Clinton, extracted from the website of the American Presidency Project[6]) for which Arnold (2018) builds metamotifs out of the frame annotations with `UniFrameId`:

> *A Trump Administration will secure and defend our borders.*
> *And yes, we will build a wall. ...*

With frames annotations for nouns and verbs, the above example of a Trump-statement in the presidential election is a foundation to build metamotifs on:

> *A Trump Administration will* *secure* PROTECTING *and* *defend* DEFEND *our* *borders* BOUNDARY *.*
> *And yes, we will* *build* BUILDING *a* *wall* ARCHITECTURAL_PART *. ...*

[6] American Presidency Project: `http://www.presidency.ucsb.edu`

The example illustrates that here, protection and defence are connected to the action of building a wall. The annotation of many political debate texts can reveal whether this is a recurring pattern and also which parties or people make use of this pattern. Frames as (meta-)motifs could give insights into the patterns used by specific parties.

Setup. Arnold (2018) investigates whether metamotif analysis is useful as features for classification tasks, such as predicting political parties or politicians from political speeches. He contrasts the following annotations or abstractions as features in several machine learning classifiers for the prediction: part-of-speech tags, simple motifs, frame-motifs and metamotifs built on top of frames. He aims at finding out whether the metamotifs of frames help to generalize better than the motifs or frame-motifs alone in order to identify a party from a speech.

The data to experiment with contains transcripts of official political speeches (US) and political party programs and debates (Germany). Concerning the US, the speeches range form 1789 to 2016, and concerning Germany, the programs range from 1949 to 2017 and the debates range from 2013 to 2017. Amongst others, the data contains speeches of the United States presidential election 2016: Trump versus Clinton (see above for the link to the American Presidency Project) and also German political debate texts (deutsche Bundestagswahl, 2017: Merkel versus Schulz, extracted from the Manifesto Project website[7] and from the German Bundestag: [8]). For the German data, Arnold (2018) uses the German annotations of `UniFrameId` in order to obtain the frame annotations.

He creates an embedding for every party or politician, a so-called *'motif signature'*. The embeddings have the dimensionality of the amount of all motifs that occur at least three times in the speeches of any party or politician. Every single dimensionality notes the occurrence count of a specific motif in the texts of the respective party or politician and finally, the embedding is normalized to sum up to one. This way, a party-specific embedding stores the relative occurrences of all motifs (occurring at least three times) and can therefore be regarded as a signature for this party that can be compared to that of other parties.

To obtain frame-motifs, he creates a graph structure out of the frame annotations for texts in the following way. Starting with a frame annotated text, every frame is represented by a node and two nodes are connected if and only if *(a)* the respective source tokens come from the same sentence, and *(b)* no other predicted token lies in between. Any path consisting of two to four nodes is considered as a frame-motif. By this, he obtains chains of nodes from single sentences that might mirror local patterns in the political language. The metamotifs are motifs of frame-motifs across two consecutive sentences: the last node of a sentence is connected to the first node of the next sentence by using a special type of edge. Thus, each metamotif is a unique combination of two frame-motifs, where each frame-motif can consist of up to four semantic frames. Similar to *'motif signatures'*, Arnold (2018) also creates *'frame-motif signatures'* and *'metamotif signatures'* that rely on the frame annotations. Such a signature of a single politician is a training example from which a system has to predict the respective party.

[7] Manifesto Project: `https://manifesto-project.wzb.eu/`
[8] German Bundestag: `https://www.bundestag.de/dokumente/protokolle/plenarprotokolle`

Results and Take-away. Arnold (2018) finds the metamotifs to reveal higher predictive power than traditional motifs; thus, in this setup frames seem to be useful when contributing to higher-level abstractions such as metamotifs.

Specifically, he finds a tendency for the metamotifs to outperform a majority-baseline, and also the features of parts-of-speech, motifs and frame-motifs with almost any classifier. However, this tendency is statistically significant only on the German datasets, not on the English dataset. With some classifiers, the frame-motifs slightly outperform the metamotifs as features.

In a qualitative analysis, he comments on the most prominent motifs of individual politicians and parties.

Politician-specific analysis. Referring to the example of a Trump-statement above, Arnold (2018) finds that the frame-motif of *'Building - Architectural_part'* is a typical example for an individual motif for Trump. Even if metamotifs of frames showed to be the most promising features, he finds them to be difficult to interpret in a qualitative analysis.

Party-specific analysis. Arnold (2018) finds that Republicans (including Donald Trump) use more frames like *'Law'*, *'Opinion'*, *'Motion'*, whereas Democrats (including Hillary Clinton) use more frames like *'People'*, *'Education'*, *'Intentionally_create'*.

Concerning the German political debate texts, he finds the most prominent party-specific frame-motifs to capture characteristic attitudes and core topics of the parties. As an example, he reports the following frame-motifs to be dominant for specific parties:

- the frame-motif *'Collaboration - Collaboration - Collaboration - Collaboration'* for parties like CDU/CSU and SPD,
- the frame-motif *'People - People - People - People'* for parties like B90/Die Grünen, Die LINKE and SPD,
- the frame-motif *'Commerce - Commerce'* for parties like FDP.

After the qualitative analysis, Arnold (2018) concludes that, one the one hand, frame-motifs are suitable to capture the attitudes of political parties (and also people if there is enough data of their speeches), and on the other hand, single frames capturing topical aspects but not sentiment can be evoked by speeches from opposed camps. This demonstrates the need to leverage motifs or metamotifs out of frames instead of using isolated frames.

Taken together, he finds that combining motif analysis with frame annotations helps to identify typical patterns of political parties. However, he points out to finding even more abstract annotations in order to capture general strategies for persuasion, where in his approach, the frames do not generalize to this high level.

Finally, from Arnold (2018), it can be taken away, that using frames as a level of abstraction from which to build further features (e.g., frame-motifs or metamotifs) is a promising approach, although he remains with the open question of the adaptability and scalability of his method to different tasks and data types.

Conclusion of Applying Frame Labels in Summarization, Evaluation of Summaries, and Motif Construction. In retrospection of applying frame labels in higher-level tasks, we find evidence for the hypothesis of FrameNet's knowledge contributing to a semantic understanding in the context of summarization and motif construction.

To conclude, on the one hand, we observe positive tendencies obtained by the abstraction with frame labels, and on the other hand, we identify the need for expressing similarities between frames. We address this need in the next Section 6.2 by experimenting not only with frame labels but also with frame embeddings that incorporate similarities between frames.

6.2 Applications of Frame Embeddings

Whilst in the previous Section 6.1, the applications used frame information in terms of frame labels, now we also provide frame knowledge in terms of frame embeddings. Crucially, frame labels only allow to decide whether two labels are equal or not. But with frame embeddings, it can be judged to which extent two labels are similar.

To this end, textual frame embeddings encode information whether two frames appear in similar textual context, whereas structured frame embeddings encode information whether two frames form similar triples of (frame_1, relation, frame_2). In Chapter 5 we have seen that textual and structured frame embeddings differ in modeling and prediction frame-to-frame relations. We examine the impact of frame embeddings in two different high-level applications: first, we report on frame information being used for semantic sentence similarity (cf. Section 6.2.1) and second, we study the impact of frame information for commonsense reasoning (cf. Section 6.2.2).

We discuss these applications to showcase and elaborate on the potential of semantic knowledge in terms of frames and frame-to-frame relations towards language processing. By this, we also collect evidence for the formulation of the higher-level task crucially determining the gain by knowledge about situations and actions in terms of frames.

Applying Frame Information in Semantic Similarity and Commonsense Reasoning. These two tasks both rely on semantic text understanding in order to then judge semantic similarity, or identify a plausible chain of reasoning.

We hypothesize that frame information can benefit semantic tasks such as sentence similarity and we go even further in challenging this hypothesis with high-level (semantic) reasoning. For both tasks, frame knowledge can be beneficial to achieve the foundational text understanding in terms of *(a)* providing abstractions from word-level to frame-level, *(b)* disambiguating predicates with respect to the context, *(c)* offering meta-knowledge about the frame labels (e.g., frame-to-frame relations), *(d)* informing about the degree of relatedness of frames by their frame embeddings, *(e)* offering complementary measures for relatedness (textual versus structured frame embeddings), and *(f)* eventually giving access to frame-specific role labels. Whilst WordNet also implements the benefit of word sense disambiguation, FrameNet additionally offers access to information about frame-to-frame relations for broader context or world knowledge and frame-specific roles for syntactic information.

score	***explanation: The two sentences are ...*** **two example sentences**
5	*... completely equivalent, as they mean the same thing.* The bird is bathing in the sink. Birdie is washing itself in the water basin.
4	*... mostly equivalent, but some unimportant details differ.* Two boys on a couch are playing video games. Two boys are playing a video game.
3	*... roughly equivalent, but some important information differs.* John said he is considered a witness but not a suspect. 'He is not a suspect anymore.' John said.
2	*... not equivalent, but share some details.* They flew out of the nest in groups. They flew into the nest together.
1	*... not equivalent, but are on the same topic.* The woman is playing the violin. The young lady enjoys listening to the guitar.
0	*... completely dissimilar.* The black dog is running through the snow. A race car driver is driving his car through the mud.

Table 6.1: Explanation of similarity scores with examples [Agirre et al., 2013].

We combine Frame Identification with the use of our frame embeddings and find evidence for knowledge about frames in terms of embeddings to be helpful in both applications. However, our analyses hint that frame information is of more help in tasks where a semantic gap needs to be bridged and of less help in tasks where logical reasoning is the key.

6.2.1 Semantic Textual Similarity – Judging Similarity

The impact of frame knowledge in the context of the task of semantic textual similarity is explored by Zelch (2018). Judging the semantic similarity of sentences is a challenging task and an important foundation to higher-level NLP-tasks, such as question answering or text summarization. Semantic textual similarity requires an understanding of the context of the described situations where background knowledge about the frames and their interactions is hypothesized to play a key role. Building up on frame labels and using our frame embeddings (cf. Chapter 5), Zelch (2018) extend a current approach on semantic textual similarity with frame knowledge. She finds that an enrichment with frame knowledge contributes to improvement over the non-enriched approach for sentences lacking semantic information.

Task of Semantic Textual Similarity. Following the demand for judging sentence similarity in higher-level NLP-tasks, Semantic Textual Similarity (STS, common abbreviation) was conducted annually as a shared task at the SemEval Workshop (Agirre et al., 2012; Cer et al., 2017). The task of STS measures the degree of semantic equivalence of two sentences on a scale from 0 (completely different) and 5 (exactly equivalent), see Table 6.1 for examples corresponding to the different scores. An example for a sentence pair labeled with 5 is:

'The bird is bathing in the sink. - Birdie is washing itself in the water basin.'.

The most recent dataset provides 625 sentence pairs in the development set and 250 pairs in the test set (STS-2017).

Baseline. The system `InferSent` (Conneau et al., 2017) is the best performing sentence-level baseline in STS-2017 (Cer et al., 2017). It operates with pre-trained sentence embeddings (`InferSent` embeddings) for the two sentences to compare and then judges the similarity using the cosine distance d_{cos} (cf. Equation 3.2). Thus, for the STS task, `InferSent` is regarded as an unsupervised system: the STS training data is not used as the sentence embedding method is pre-trained on the SNLI corpus (Bowman et al., 2015).

In contrast, the current winners (`ECNU` (Tian et al., 2017, short for East China Normal University) and `BIT` (Wu et al., 2017a, short for Beijing Institute of Technology)) of the shared task STS-2017 are complex ensemble models using a supervised setup.

Potential of Frame Information in STS. An analysis of the currently most powerful systems, including `ECNU`, `BIT` and `InferSent`, identifies the weaknesses in word sense disambiguation (WSD, common abbreviation) as a major source of error (Cer et al., 2017). As an example, they all fail to judge the two sentences displayed in the following example, which are labeled as exactly equivalent:

Sentence pair (exactly equivalent):
There is a cook preparing COOKING_CREATION *food.*
A cook is making COOKING_CREATION *food.*

This example showcases the intuition for involving frames: they could help with proper word sense disambiguation. The difficulty are different words (e.g., *'prepare'* and *'make'*) that can be synonyms in the context of the two sentences. In these cases, word sense disambiguation seems promising. This can be done with the help of FrameNet frames, but also other approaches to word sense disambiguation are possible, such as WordNet's schema for disambiguating senses. We are interested in investigating into the potential of FrameNet, which, on the one hand seems to be promising for word sense disambiguation, but also offers meta-knowledge in terms of frame-to-frame relations (incorporated into our structured frame embeddings) or frame-specific roles.

The benefit of FrameNet for the task of semantic textual similarity was initially reported by Wang et al. (2013). They combine similarity models based on words, WordNet and FrameNet (using `Semafor` (Das et al., 2014)). Interestingly, they find that the FrameNet-model is particularly strong on short texts with similar structure. However, they are not using current embedding techniques.

Taken together, *(a)* FrameNet provides several kinds of helpful information to identify word meaning and to add context between words, *(b)* a previous FrameNet-model showed to be strong in specific cases in semantic textual similarity, also when contrasted with WordNet (Wang et al., 2013), but embedding techniques were not implemented yet, and *(c)* whilst WordNet is reported to overlap with word embeddings (Zhai et al., 2016), we developed frame embeddings which either overlap with word embeddings (textual frame embeddings, learned with `Word2Vec` on textual

data) or which do not overlap with word embeddings (structured frame embeddings, learned with `TransE` on hierarchical data).

Zelch (2018) investigates the potential of FrameNet for the challenge of word sense disambiguation semantic textual similarity, integrates this feature into the system `InferSent` and analyzes the effects with respect to two categories of failure of current systems where frames could help: 'underpredicted' and 'overpredicted' – which they study on the development set with respect to the potential for frame-based improvements in a semantic gap.

Finally, frame information can only have an impact on a small subset of the sentence pairs in the task of semantic textual similarity, which Zelch (2018) categorizes into *underpredictions* and *overpredictions* (counts are given in the categories' descriptions below). However, Zelch (2018) is interested in exactly these cases the more so as current systems fail in them.

Underpredictions. For some instances `InferSent` predicts lower than the gold annotations are, e.g., the instance *'A double decker red bus driving down the road. - A red double decker bus driving down a street.'* has a gold score of 5, but `InferSent` predicts 3.83. For some of these instances, Zelch (2018) sees the potential for frames to be helpful: two sentences meaning the same but expressing it with different words (*'road'* and *'street'*) – which map to the same frame ('Roadways'). In her manual analysis on the development set, Zelch (2018) could identify 9 sentence pairs in the category of underpredictions, out of which they could identify 8 where frame information could help in terms of mapping different words to the same frame.

Overpredictions. For some instances `InferSent` predicts higher than the gold annotations are, e.g., the instance 'A man is playing soccer. - A man is playing flute.' has a gold score of 1, but `InferSent` predicts 1.73. Zelch (2018) also sees the potential for frames to be helpful in some of these instances: two sentences with different meaning but expressing it with the same words ('playing' and 'playing') – which map to the different frames ('Competition' and 'Cause_to_make_noise'). In her manual analysis on the development set, Zelch (2018) could identify 3 sentence pairs in the category of overpredictions, out of which they could identify 2 where frame information could help in terms of mapping different words to the same frame. Thus, overpredictions are less frequent than underpredictions in this dataset.

Approach. Zelch (2018) extends the baseline model `InferSent` by Conneau et al. (2017) with the information from FrameNet: frame annotations are obtained by `Open-SESAME` (Swayamdipta et al., 2017), then our pre-trained frame embeddings (trained with Word2Vec and TransE, cf. Chapter 5) are retrieved. `InferSent` provides sentence embeddings to which Zelch (2018) concatenates the frame information in terms of embeddings. This corresponds to a *late fusion* strategy: obtain frame annotations independent of the `InferSent` embedding $\overrightarrow{v_{(sentence)}}$ and combine them by simple concatenation $\frown$ (cf. Equation 3.19) on the sentence level in the form of $\overrightarrow{v_{(sentence)}} \frown \overrightarrow{v_{(frames)}}$. The vector $\overrightarrow{v_{(frames)}}$ is the mean vector of the embeddings of all frames that were identified in the sentence. Then, the cosine distance d_{cos} (cf. Equation 3.2) is used as the final prediction.

Model	Pearson	Spearman
original `InferSent`	80.999	79.664
`InferSent`, frame embeddings (`TransE`)	81.325	80.032
`InferSent`, frame embeddings (`Word2Vec`)	**81.404**	**80.095**
`InferSent`, parents' frame embeddings (`Word2Vec`)	**81.407**	**80.210**

Table 6.2: Results for Semantic Textual Similarity (best highlighted in bold) on the test set. Pearson (middle column) and Spearman (right column) correlation are measured in percent.

Results and Take-away. To evaluate the performance of the frame-enhanced approach, the Pearson product moment correlation $r_p(X, Y)$ (cf. Equation 3.17) and the Spearman rank correlation $r_s(X, Y)$ (cf. Equation 3.18) are used. Zelch (2018) reports the most successful settings of frame information improving the `InferSent` system, see Table 6.2, where the nine most frequent frames are excluded. She reports a tendency for frame information leading to small improvements compared to the original `InferSent` approach, even if not statistically significant. The results indicate a slight advantage of the `Word2Vec` frame embeddings which are learned on a broader text collection compared to `TransE` frame embeddings which are learned on FrameNet's hierarchy. Initially, it was hypothesized that `TransE` frame embeddings could add most additional information as they do not overlap with textual word embeddings since they were learned with `TransE` on hierarchical data. However, this theoretical advantage might be outweighed by the practical situation of more textual data and less structured data to learn embeddings from.

The best setting involves frame information in terms of the `Word2Vec` frame embeddings of the parent frame for every evoked frame. This way, the two advantages of (practically) more textual data to learn embeddings and (theoretically) more additional information in the structure of the FrameNet hierarchy when incorporating the *inheritance*-relation.

Zelch (2018) finds a tendency of frame information being beneficial for improving prediction towards higher similarity of sentence pairs, i.e. where current systems fail due to underpredictions. This fits to the observation by Wang et al. (2013) who also used FrameNet for the task of semantic textual similarity and found the strength of their FrameNet-model being in short texts with similar structure.

In order to improve predictive power towards lower similarity, i.e. where current systems fail due to overpredictions, the nature of the dataset has to be taken into account. The current dataset does not contain a considerable number of sentence pairs of low similarity which use similar words to express different meaning. But exactly this situation would profit from frame annotations. Thus, to test the potential of frames with respect to low similarity, one could think of augmenting the data with more instances corresponding to exactly this scenario: sentence pairs of low similarity which use similar words to express different meaning and which evoke at least one frame per sentence. As on the current dataset, the frame annotations seem not to be indicative for low similarity, one could also think of including new features to capture these aspects.

Furthermore, Zelch (2018) finds that the bottleneck is not the noise introduced by wrongly predicted frames, but rather missing lexical units in FrameNet or missing

links between lexical units and certain frames in FrameNet. On the one hand, the predictions depend on the quality of the Frame Identification system. On the other hand, however, a small manual analysis using manually annotated frames for a subset of the dataset reveals that correcting wrong predictions does not improve the final prediction. This tendency hints at using frame predictions that are mostly correct being sufficient. Thus, in order to further improve the quality of applications of frame information in higher-level tasks, we recommend enlarging the coverage of FrameNet.

To sum up, from Zelch (2018), it can be taken away that frame information could slightly improve the similarity prediction of sentence pairs compared to the original embedding-based baseline system `InferSent`. The extension of the current best baseline system `InferSent` with frame embeddings is promising where current systems underpredict.

6.2.2 Commonsense Reasoning – Judging Plausibility in Arguments

Commonsense argumentative reasoning is a challenging task that requires holistic understanding of the argumentation where external knowledge about the world is hypothesized to play a key role. We explore the idea of using knowledge about prototypical situations and actions from FrameNet and relational knowledge about concrete entities from Wikidata to solve the task (Botschen et al., 2018b). We find that both resources can contribute to an improvement over the non-enriched approach and point out two persisting challenges: first, integration of many annotations of the same type and second, fusion of complementary annotations. After our explorations, we question the key role of external world knowledge with respect to the argumentative reasoning task and rather point towards a logic-based analysis of the chain of reasoning.

Recently, Habernal et al. (2018) introduced a challenging dataset for Argument Reasoning Comprehension (ARC, common abbreviation) used in the SemEval-2018 shared task. After reviewing the participating systems, they hypothesize that

> *external world knowledge may be essential for Argument Reasoning Comprehension.*[9]

We explore enriching models with knowledge about situations and actions and about structured relations on Argument Reasoning Comprehension to investigate this hypothesis. After elaborating on the potential of combining frame and entity information to represent world knowledge, in Sections 6.2.2.3 and 6.2.2.4, we *(1)* present a proof of concept for semantic enrichment for the ARC task, *(2)* identify the importance of advanced combinations of complementary semantic annotations and finally *(3)* question the key role of external world knowledge with respect to Argument Reasoning Comprehension.

[9] SemEval-2018 Task 12: `https://competitions.codalab.org/competitions/17327`

6.2.2.1 Potential of Combining Frame and Entity Information to Represent World Knowledge

For different high-level tasks it was shown that multiple levels of knowledge processing are beneficial. Combining several kinds of annotations benefits question answering (Khashabi et al., 2018), external knowledge about synonyms enhances inference (Chen et al., 2018), and jointly modeling several tasks (e.g., frame-semantic parsing and dependency parsing) is fruitful (Peng et al., 2018). In particular, the idea of connecting semantics about situations and actions with relational knowledge was confirmed by Guo et al. (2016): they jointly formalize semantic role labeling and relation classification and thereby improve upon PropBank semantic role labeling. In the following, we will explain to what extend the knowledge resources FrameNet and Wikidata complement each other and how they could represent a holistic world knowledge.

Complementary Sources of External Knowledge. We experiment using the lexical-semantic resource FrameNet (FN) and the knowledge base Wikidata (WD). These resources provide information beyond the lexical relations encoded in WordNet and thus have a potential to enhance the underlying model with other kinds of external world knowledge. On the one hand, FrameNet provides qualitative knowledge about prototypical situations and actions. Thus, identifying frames unveils the situation or action that is happening. On the other hand, Wikidata provides relational knowledge about concrete entities. So, linking entities to a knowledge base disambiguates the participants. Furthermore, both resources provide meta-knowledge about how their frames or entries relate to each other.

Complementarity of Annotations. Work on event semantics hints at two annotation types complementing each other: additional information about participants benefits event prediction (Ahrendt and Demberg, 2016; Botschen et al., 2018a) and context information about events benefits the prediction of implicit arguments and entities (Cheng and Erk, 2018). Complementarity is further affirmed by efforts on aligning Wikidata and the FrameNet lexicon: the best alignment approach only maps 37% of the total Wikidata properties to frames (Mousselly-Sergieh and Gurevych, 2016). The complementarity of FrameNet and Wikidata annotations is the reason for also testing a model with joint annotation '+FN/WD'.

Wikidata is a collaboratively constructed knowledge base of high quality (Färber et al., 2015) that encodes world knowledge in the form of binary relations. It contains more than 40 million entities and 350 million relation instances.[10]

Whilst FrameNet formalizes knowledge about situations and actions (e. g., depending on the context, the verb *buy* evokes either *Commerce_buy* (buying goods) or *Fall_for* (buying a lie)), Wikidata implements relational knowledge (e. g., CAPITAL(Hawaii, Honolulu); INSTANCE OF(Hawaii, location)). FrameNet specifies high-level relations (e.g., *inherit, precede*) between frames, forming a hierarchy with a collection of *(frame, relation, frame)*-triples. Wikidata focuses on relational triples in the form of *(entity, relation, entity)*-triples.

[10] `www.wikidata.org/wiki/Special:Statistics`

6.2.2.2 Task of Argument Reasoning Comprehension

We explain the task of Argument Reasoning Comprehension together with the baseline that we will build upon.

The Argument Reasoning Comprehension task (Habernal et al., 2018) is formulated as follows: given a debate title (a), claim (b) and reason (c), a system chooses the correct warrant (i, green) over the other (ii), see the following examples for an instance of the ARC corpus:

Example 1 - companies and trust:

(a) title: *Can companies be trusted?*
(b) claim: *Companies can't be trusted.*
(c) reason: *Corporations have only one goal: to make a profit.*
(i) warrant: *they do not have to satisfy customers to make a profit*
(ii) warrant: *they have to satisfy customers to make a profit*

Example 2 - police and use of force:

(a) title: *Do Police use deadly force too often?*
(b) claim: *Police is too willing to use force.*
(c) reason: *Police use the excuse of fear for life to abuse use of force.*
(i) warrant: *the excuse is rarely warranted*
(ii) warrant: *the excuse is sometimes warranted*

The warrants vary only slightly, e.g., by a single negation. The argumentation chain is sophisticated and uses logical reasoning and language understanding. In order to automatically draw the correct decision, a holistic understanding of the context of both, claim and reason, is crucial, for which Habernal et al. (2018) recommend the inclusion of external knowledge.

Baseline. The baseline provided by Habernal et al. (2018) is an intra-warrant attention model that reads in `Word2Vec` vectors (Mikolov et al., 2013a) of all words in (a-c) and adapts attention weights for the decision between (i) and (ii).

In contrast, the shared task winner, GIST (Choi and Lee, 2018), transfers inference knowledge (SNLI, Bowman et al., 2015) to the task of Argument Reasoning Comprehension and benefits from similar information in both datasets.

6.2.2.3 Approach

We investigate whether external information in terms of frames (FN) and entities (WD) can contribute to holistic understanding of the argumentation in the Argument Reasoning Comprehension task. First, we examine the effect of both annotations separately and second, we explore whether a joint annotation benefits from the inherent complementarity of the schemata in FN and WD and eventually leads to a better annotation coverage. We enhance the baseline model provided with the Argument Reasoning Comprehension task in order to contrast three system configurations: '+FN', '+WD' and '+FN/WD'.

Our approach extends the baseline model with two external knowledge schemata, FN and WD, to explore their effects. The intuition can be explained with the instance in example 1 - 'companies and trust': FN could be helpful by disambiguating

Figure 6.2: Different embeddings from layers of annotations for an example sentence of the Argument Reasoning Comprehension task: words, frames, entities.

'companies' and 'corporations' to the same frame with meta-knowledge how it relates to other frames and WD could be of additional help by mapping them to entities with detailed information and examples for such institutions. We focus on utilizing the two knowledge schemata of FN and WD and thus, our interest is orthogonal to GIST. The advantage of our approach is independence of domain and task, which becomes especially relevant in scenarios lacking large-scale support data.

Preprocessing. We use two freely available systems to obtain semantic annotations for claim (b), reason (c) and alternative warrants (i, ii): the frame identifier by Botschen et al. (2018a) for frame annotations and the entity linker by Sorokin and Gurevych (2018a). We employ pre-trained vector representations to encode information from FN and WD. We use the pre-trained frame embeddings (50-dim.) that are learned with `TransE` (Bordes et al., 2013) on the structure of the FN hierarchy with the collection of *(frame, relation, frame)*-triples (Botschen et al., 2017). We also use `TransE` to pre-train entity embeddings (100-dim.) on the WD graph. The annotation of the Argument Reasoning Comprehension data leads to more frames per sentence (6.6 on avg.) than entities per sentence (0.7 on avg.).

Model. We extend the baseline model by Habernal et al. (2018) with embeddings for frames and entities (cf. previous paragraph for frame embeddings and entity embeddings). The baseline model is an intra-warrant attention model that only uses conventional pre-trained word embeddings as an input. We apply a *late fusion* strategy: obtain the annotations separately and combine them afterwards by appending the frame and entity embeddings to the word vectors on the token level. Each input sentence is processed by a bidirectional long short-term memory (LSTM) network that reads not only word embeddings, but also frame embeddings for all mentions of situations or actions and entity embeddings for all entity mentions (Figure 6.2).[11] Within our extension, the attention weights for the decision between the two warrants are adapted based on the semantically enriched representation of claim (b) and reason (c).

We optimize hyperparameters on the development set with random search. All models are trained using the Adam optimizer (Kingma and Ba, 2014) with a batch size of 16. For evaluation, we perform ten runs and report the mean and max accuracy together with the standard deviation.

Model	Accuracy mean Dev.	(±)	Test	(±)	max Test
Habernal et al. (2018) (reimpl.)	67.12	1.55	55.70	1.84	58.78
+WD	66.23	0.71	**56.80**	2.35	60.36
+FN	**67.41**	1.19	**56.76**	2.57	**61.04**
+FN/WD	66.30	0.88	55.92	1.64	59.46

Table 6.3: Mean and max accuracy over ten runs on the ARC development and test sets (best results highlighted in bold).

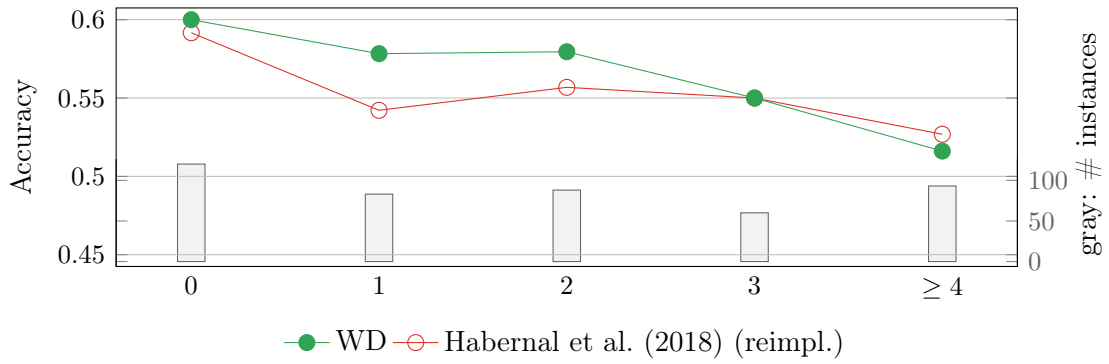


Figure 6.3: Performance for the '+WD' approach by the number of Wikidata entities in a test set instance.

6.2.2.4 Results and Analysis

In Table 6.3 we report the results on the Argument Reasoning Comprehension task. The extended approaches '+FN' and '+WD' for semantic enrichment with information about frames and entities increase the averaged performance by more than one percentage point against the baseline. For the best run, the advantage of '+FN' and '+WD' becomes even clearer (+2.2 pp.). On the other hand, the straight-forward combination of the two external knowledge sources, '+FN/WD', does not lead to further improvements. This points out the need for advanced models that are able to fuse annotations of different types. Albeit positive, the results do not seem to be a strong support for the hypothesis of Habernal et al. (2018) about external knowledge being beneficial for the defined task, as we observe only moderate improvements. Overall, the enriched models ('+WD', '+FN' and '+FN/WD') make mostly the same predictions as the baseline system. For instance, for '+WD' there is 79,5% overlap of the predictions with the baseline, and for '+FN', it is 76.6%. In the following section, we try to identify the reasons why structured knowledge from FrameNet and WikiData does not further improve the results.

Error Analysis. The amount of semantic information that the model can utilize depends on the number of annotations for an instance[12]. We analyze the performance of the enriched models by the number of annotations for '+WD' and for '+FN'.

[11] We refer to Habernal et al. (2018) for more details.
[12] Each instance is four sentences: a claim, a reason, a debate title and a candidate warrant.

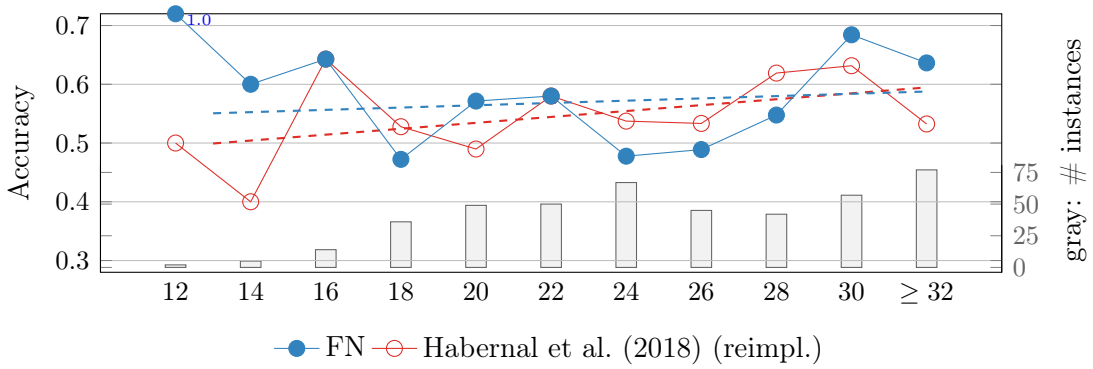


Figure 6.4: Performance for the '+FN' approach by the number of frames in a test set instance.

Figure 6.4 shows the performance of '+WD' in comparison to the baseline against the number of WD entities per test instance. As expected, there is no difference in performance for the instances without WD annotations. We can see a clear improvement for the instances with one or two entities, which indicates that some semantic knowledge helps to draw the decision between the two warrants. Contrary, '+WD' performs equal to the baseline for three or more annotations.

The performance of the '+FN' model against the number of the frame annotations is plotted in Figure 6.4. Whilst the difference between '+FN' and baseline varies more, we can observe a similar tendency: once some semantic annotations are available the enriched model outperforms the baseline, whereas with the growing number of frames the difference in performance decreases. This hints at shorter sentences profiting more from the (few) frame annotations than longer sentences having many frame annotations. Longer sentences are more likely to directly contain the semantic background needed for comprehension, whilst shorter sentences rely on the semantic world knowledge of the reader in order to be understood. These are the cases where semantic annotations linked to external semantic world knowledge are beneficial to automatic approaches.

Both annotation tools, the WD entity linker as well as the FN frame identifier, introduce some noise: for the entity linker, Sorokin and Gurevych (2018a) report 0.73 F-score and the frame identifier has an accuracy of 0.89 (Botschen et al., 2018a). We perform a manual error analysis on 50 instances of the test set to understand the effect of the noisy WD annotation.[13] In 44% of errors, no WD annotation was available and in 52%, the annotations were (partially) incorrect. Only 4% of errors occur despite correct entities being available to the model. Notably, in 65% of the cases a correct entity annotation leads to a correct prediction.

Taken together, for instances with little context and therefore only some annotations with frames or entities the semantic enrichment helps to capture a broader context of the argumentation chain which in turn benefits the classification. However, for instances with more context and therefore more annotations with frames or entities the benefit is turned down by a worse precision of the annotation tools. Interestingly, the effect of improved performance only for shorter sequences with less

[13] Judging if a predicted frame is correct requires deep linguistic expertise and special training on the FrameNet guidelines. Therefore, we excluded FN from this first study.

annotations is in line with findings of research on information retrieval (Müller et al., 2008), where the trade-off between some annotations that increase the accuracy and more annotations that can hurt the performance is known as the precision-recall balance (Riedel et al., 2010; Manning et al., 2008).

Qualitative Analysis. In the previous paragraph on error analysis, we show that our approach is of help by successfully enriching the context with semantics for shorter instances; and in this paragraph on qualitative analysis, we elaborate on why our approach is too limited to solve some key challenges of the Argument Reasoning Comprehension task.

When manually inspecting the annotations of frames and entities, it becomes questionable whether these actually have the potential of contributing to a clear distinction between the two warrants. The following example shows two instances of the Argument Reasoning Comprehension corpus with FN and WD annotations.

Example 1 - companies and trust:

(a) title: *Can companies be trusted?*

(b) claim: *Companies* $^{Q4830453}_{Business}$ *can't be* *trusted* $_{Certainty}$

(c) reason: *Corporations* $^{Q167037}_{Business}$ *have only one* *goal* $_{Purpose}$ *:*
to *make* $_{Intentionally\_Create}$ *a* *profit* $^{Q26911}_{Earnings\_and\_losses}$

(i) warrant: *they do not have to satisfy customers to make a profit*

(ii) warrant: *they have to satisfy customers to make a profit*

Example 2 - police and use of force:

(a) title: *Do Police use deadly force too often?*

(b) claim: *Police* ^{Q35535} *is too* *willing* $_{Willingness}$ *to* *use* $_{Using}$ *force* $_{Military}$

(c) reason: *Police* ^{Q35535} *use* $_{Using}$ *the* *excuse* $_{Thwarting}$ *of* *fear* $_{Fear}$ *for life*
to *abuse* $_{Abusing}$ *use of force* $^{Q971119}_{Military}$

(i) warrant: *the excuse is rarely warranted*

(ii) warrant: *the excuse is sometimes warranted*

Both annotation layers contribute useful information about the world, which is not contained in the textual input. For instance, 'companies' and 'corporations' are correctly disambiguated and linked to the same frame and the phrase 'use of force' is mapped to the entity Q971119 for a legal concept. Nevertheless, when manually inspecting the annotations of frames and entities it becomes apparent that the provided background knowledge is not sufficient to draw the distinction between the two warrants. In the first example with annotations (Example 1 - companies and trust), the key difference between the two warrants is negation (and similar in the second example).

Even if our approach performs a semantic enrichment of the context, the crucial capability of performing reasoning is still missing. This means, our input representation is semantically enriched, but is not parsed into a logic-based representation. Thus, this seems to be a promising research direction for future work: to combine semantic enrichment with logic-based representations.

To sum up, we start from the hypothesis of the evaluators of the shared task about world knowledge being essential for the Argument Reasoning Comprehension task and we show the potential of semantic enrichment of the context for shorter instances. We integrate world knowledge from FrameNet and Wikidata into the task of commonsense argumentative reasoning and achieve an improvement in performance compared to the baseline approach. Based on the experiments and the manual analysis we conclude that external world knowledge might be helpful but not be enough to gain significant improvements in argumentative reasoning, and we rather point towards logical analysis. The results offer a first perspective on using external resources for the Argument Reasoning Comprehension task. More broadly, the approaches '+FN' (situations and actions) and '+WD' (relations) showcase the contribution of semantic enrichment to high-level tasks requiring commonsense knowledge. FrameNet and Wikidata are open-domain resources and the enrichment approach is task-independent. Consequently, we encourage utilizing knowledge about situations and actions and about relational triples in further language understanding tasks, e.g., Story Cloze (Mostafazadeh et al., 2016) or Semantic Textual Similarity (Agirre et al., 2012), and, crucially, to combine it with logical analysis if the tasks requires this.

We conclude with the key challenge of Argument Reasoning Comprehension not being the lexical-semantic gap between warrants but rather different phenomena such as negation. We suggest that this challenge is to be resolved with logical analysis on top of the world knowledge.

Conclusion of Applying Frame Information in Semantic Similarity and Commonsense Reasoning. In retrospection of applying frame embeddings in higher-level tasks, we find evidence for the hypothesis of FrameNet's knowledge contributing to a semantic understanding in the context of semantic similarity and commonsense reasoning.

For future work further exploring frame-to-frame relations, we formulate the following suggestions. We can see two major groups of frame-to-frame relations with respect to the application in high-level tasks: relations for more detailed specifications and relations for broader context knowledge. The *inheritance*- and *using*-relations mark special cases or 'is-a'-relations. Thus, they seem promising in tasks such as Semantic Textual Similarity or Summarization. In contrast, the *precedes*- and *subframe*-relations correspond to narrative schemata. Thus, they seem promising in tasks involving an understanding scripts, plans, and goals.

To conclude, on the one hand, we find evidence for knowledge about frames in terms of embeddings to be helpful in both applications, and on the other hand, our analyses hint towards frame information to be of more help in tasks where a semantic gap needs to be bridged and of less help in tasks where logical reasoning is the key.

Conclusion of Applying Frame Identification and Frame Embeddings. From the previous two sections, we gain insights about the impact of frame knowledge, summarized in Table 6.4 for an overview.

Regarding summarization, Zopf et al. (2018a) demonstrate the gain in evaluation when using feature-based systems: different features are contrasted with respect to

application	Section	Frame Knowledge	Insights concerning Enrichment with Frame Knowledge
Summarization	6.1.1	frame labels	frames help in ensemble classifier and in isolation when maximizing recall
Evaluation of Summaries	6.1.2	frame labels	frames cannot improve overall performance
Motif Construction	6.1.3	frame labels	frames are useful foundation for metamotifs
Semantic Textual Similarity	6.2.1	frame labels, embeddings	frames as textual embeddings help, inheritance relation helps (tendency)
Argument Reasoning Comprehension	6.2.2	frame labels, embeddings	frames as structured embeddings help, no further improvement with facts

Table 6.4: Overview of applications and insights.

performance and ablation studies show the impact of one feature in an ensemble. In the context of summary evaluation, Peyrard et al. (2017) could not find frames to be of additional help. Subsequent to this, we point to more ambiguous and heterogeneous scenarios as an environment for leveraging frame information. In the context of motif construction with frames, Arnold (2018) demonstrates the gain in evaluation when using frame labels: in addition to quantitative results, they also elaborate on qualitative results where they discuss the influence of different frame labels. Concerning the tasks of Semantic Textual Similarity and Argument Reasoning Comprehension, Zelch (2018) as well as Botschen et al. (2018b) find an advantage by extending baseline approaches with frame knowledge, even if the implementation of frame knowledge differs (textual versus structured embeddings, respectively).

Going one step further in retrospection, the benefit of frame information coming with an effort of preprocessing can be questioned by end-to-end approaches that are independent of any preprocessing. We address this question in the next Section 6.3 where we elaborate on the potential of frame knowledge versus end-to-end approaches.

6.3 Potential of Frame Knowledge versus End-To-End Approaches

In this section, we reflect on the advantages and disadvantages of including frame knowledge in terms of features or embeddings versus implementing end-to-end models that are independent of any pre-annotated linguistic knowledge.

This mirrors the ongoing debate on relative importance of two seemingly opposed camps: deep learning versus linguistics.[14]

On the one hand, deep learning systems are proposed to be designed and trained in an 'end-to-end' way (Collobert et al., 2011) and are shown to be successful in

[14] Also see Jacob Eisenstein's 2018 draft on 'Natural Language Processing' (under contract with MIT Press, shared under CC-BY-NC-ND license): `https://github.com/jacobeisenstein/gt-nlp-class/tree/master/notes`

various NLP-tasks, as for example sequence labeling (Ma and Hovy, 2016), relation extraction (Miwa and Bansal, 2016) and speech recognition (Bahdanau et al., 2016). Here, 'end-to-end' means that the input to a neural network is raw text (or tokenized text mapped to word embeddings (cf. Chapter 3)) and this is directly processed to obtain an output according to a task, as for example a summary or a frame prediction. Thus, an end-to-end approach is not dependent on linguistic knowledge to pre-process raw text into linguistically informed labels. The end-to-end approach requires the neural network to automatically build internal representations that encode those aspects of meaning from the input text which are important for solving the respective task. The main effort here is in developing neural network architectures and fine-tuning hyperparameters. For setups with large amounts of training data, end-to-end approaches are highly competitive, see for example end-to-end speech recognition with a long-short term memory network requiring minimal preprocessing (Graves and Jaitly, 2014). However, Glasmachers (2017) points out the technical limits and inefficiencies of end-to-end learning. Moreover, in end-to-end setups, it is difficult to analyze where and why mistakes happen. This line is the extreme in *bottom-up* approaches to modeling meaning.

On the other hand, the extreme line in *top-down* approaches to modeling meaning puts the main effort on finding linguistic structures for the raw input text. Annotations such as part-of-speech tags or syntactic dependency labels are used to build parsable grammar formalisms such as combinatory categorial grammar (Steedman, 1987) or abstract meaning representations (Banarescu et al., 2013). These linguistically informed structures are then, in theory, general and holistic so that any task can profit from them. This approach does not depend on large amounts of training data for solving a task, but it requires the linguistic annotations as input to a rule-based decision system or a traditional machine learning classifier.

Applied to our research focus, this refines the perspective of the *bottom-up* camp into:

> RQ: *What is the potential of frame knowledge versus end-to-end approaches and what are the reasons for leveraging frame knowledge?*

In the following, we contrast such frame-informed approaches with end-to-end approaches, finally converging to an integrative view where the *bottom-up* knowledge of meaning learned by deep learning systems on large amounts of data is enriched with *top-down* knowledge from linguistic annotations. This is in line with the intuition by Young et al. (2018) to combine 'internal memory (bottom-up knowledge learned from the data)' with 'external memory (top-down knowledge inherited from a KB)'.

End-to-end Approaches. In this section, we review end-to-end approaches for a selection of tasks for which we were discussing the potential of frames or which are related to those. The end-to-end approaches are free of preprocessing routines and engineering hand-crafted features, by directly processing input embeddings with neural network architectures to obtain a prediction.

Regarding the tasks of Summarization, Semantic Textual Similarity and Argument Reasoning Comprehension, we reported about extending existing end-to-end approaches with frame information (see Sections 6.1.1, 6.2.1 and 6.2.2, respectively).

Additionally, we broaden the view for reporting about end-to-end approaches towards the tasks of Question Answering and Semantic Role Labeling as they are well-known semantic tasks where current end-to-end systems are tested on.

Summarization. Concerning the task of Summarization as reported in Section 6.1.1, the end-to-end system then extended with frame information is the summarization system `CPSum` (Zopf et al., 2016a). This system learns the importance or contextual preference of bigrams on a background corpus, and is then evaluated in the context of summarization. Thus, at the moment being applied to summarization, `CPSum` does not require any additional information and performs well even in heterogeneous test scenarios.

In their work, Zopf et al. (2018a) investigate into the potential of linguistically informed features such as frames when extending the feature-free baseline system `CPSum`. They analyze the impact of different labels conveying linguistic information in isolation and in an ensemble.

Furthermore, Kedzie et al. (2018) critically look at deep learning models for content selection in Summarization in an empirical study where they point out that currently the abilities of deep learning models to learn robust and meaningful content features is over-estimated. Interestingly, they draw the attention towards new forms of sentence representations or external knowledge sources for the task of Summarization.

Semantic Textual Similarity. Concerning the task of Semantic Textual Similarity as reported in Section6.2.1, the end-to-end system then extended with frame information is `InferSent` (Conneau et al., 2017). Once the sentence embeddings are pre-trained on the SNLI corpus (Bowman et al., 2015), `InferSent` does not require any additional information and reaches best performance amongst the unsupervised systems. However, when considering all systems, `InferSent` is outperformed by the current winners of the shared task STS-2017, `ECNU` (Tian et al., 2017) and `BIT` (Wu et al., 2017a)).

Interestingly, neither `ECNU` nor `BIT` are end-to-end systems, they belong to the category of feature engineered and mixed systems. `ECNU` implements feature engineered and deep learning models and afterwards averages and combines the single scores in an ensemble. In particular, the feature engineered part comprises more than 60 hand-crafted features for sentence pairs and single sentences, requiring syntactic dependency parses for example. `BIT` (Wu et al., 2017a) uses the WordNet hierarchical taxonomy to build a semantic information space for a sentence. Two sentences are compared by the overlap of their information content in the information space and also in terms of `word2vec` embeddings.

Both `ECNU` and `BIT` are supervised systems with complex architectures. Crucially, both system show that combining deep learning modules or unsupervised embedding modules with linguistically informed feature engineered modules is advantageous.

In their work, Zelch (2018) investigates into the potential of frame informed features when extending the unsupervised baseline system `InferSent`. By avoiding combinations of complex modules, she can analyze the impact of frame information directly on the embedding level.

Argument Reasoning Comprehension. Concerning the task of Semantic Textual Similarity as reported in Section 6.2.2, the end-to-end system then extended with frame information is the baseline provided by Habernal et al. (2018). It operates on `Word2Vec` embeddings. The baseline is outperformed by another end-to-end system that operates on embeddings pretrained on (SNLI, Bowman et al., 2015) and then transfers inference knowledge to the new task: the shared task winner GIST (Choi and Lee, 2018). GIST profit from the external inference knowledge contained in the pre-trained embeddings and significantly outperforms all other systems on this particular test data, as reported by Habernal et al. (2018). Interestingly, our extension of the baseline with frame information (cf. Section 6.2.2) ranks second when compared to the systems submitted to the shared task, of which several are elaborate end-to-end neural approaches.

Question Answering. Regarding the task of Question Answering connected to knowledge bases, Yin et al. (2016) propose a neural end-to-end approach for generating answers to simple factoid questions. Furthermore, Gated Graph Neural Networks encoding the graph structure of the semantic parse involving multiple entities and relations from a knowledge base are reported to be strong compared to models that do not explicitly model the structure (Sorokin and Gurevych, 2018b).

With respect to Question Answering for reading comprehension, an interesting trend is to observe for widely used datasets such as the Stanford Question Answering Dataset (Rajpurkar et al., 2016, SQuAD, common abbreviation). On the one hand, end-to-end neural approaches outperform simple models with manually crafted features (Wang and Jiang, 2016). On the other hand, these questions are either easy to answer or obviously unanswerable – calling for an extension which makes the task more challenging. Thus, Rajpurkar et al. (2018) introduce SQuAD 2.0 which is augmented with challenging unanswerable questions that are not obvious to identify. On this more challenging version, the performance of strong neural approaches drops by 20 percentage points on F1-score. This demonstrates that current neural systems are still far away from true language understanding and challenging datasets help to further develop advanced systems. To this end, it is open to research to what extend the incorporation of background knowledge is beneficial.

Semantic Role Labeling. In the following, we review research on Semantic Role Labeling, which is all performed on PropBank data, not FrameNet data. The breakthrough on end-to-end Semantic Role Labeling without further syntactic input was opened by Zhou and Xu (2015), who propose a neural end-to-end approach integrated into recurrent neural networks. Furthermore, He et al. (2017) introduce a deep highway bidirectional long-short term memory architecture for Semantic Role Labeling, showing the success of deep models at recovering long-distance dependencies, but at the same time arguing for syntactic parsers to improve upon the deep-only results. He et al. (2018) propose an end-to-end approach for jointly predicting predicates, arguments spans, and argument labels setting a new state of the art on PropBank data without gold predicates.

Following up on the success of end-to-end Semantic Role, Strubell et al. (2018) point out the gap between current state-of-the-art models leveraging deep neural networks with no explicit linguistic features, whilst prior work shows the benefit of

syntax trees towards Semantic Role Labeling (Roth and Lapata, 2016). Strubell et al. (2018) introduce a linguistically-informed self-attention model (short: LISA) for Semantic Role Labeling. Their neural network model integrates multitask learning across dependency parsing, part-of-speech tagging, predicate detection and Semantic Role Labeling. By integrating syntactic information and thereby combining deep learning and linguistic formalisms, they achieve new state-of-the-art results.

Most advances reported on Semantic Role Labeling focus on PropBank data (see above), however some also include FrameNet data. With respect to incorporating linguistic knowledge into neural models, Swayamdipta et al. (2018) achieve competitive performance on the tasks of PropBank semantics, frame semantics, and coreference resolution by training in a multitask setup with simple syntax-based helper-tasks. With respect to linguistic knowledge in FrameNet studied at a distributional level, Kleinbauer and Trost (2018) present an exploration of similarities between frame semantics and distributional semantics. They see potential for enhancing current embedding approaches with the rich frame semantic structures.

Conclusion on the Potential of Frame Knowledge versus End-To-End Approaches. In the previous paragraph, we could see that the best systems involve linguistically informed modules on top of or in addition to deep learning modules or unsupervised embedding modules. Moreover, the relevance of linguistic structure in neural architectures is an emerging field of research (Strubell and McCallum, 2018). In their work, Strubell and McCallum (2018) address the most recent advances of embedding learning such as contextualized language models (Peters et al., 2018) by showing that these still profit from additional linguistic structures.

Thus, the question arises whether the gain of including linguistically informed modules compensates the effort (with respect to human or computational resources) of obtaining the annotations in a pre-processing step. The 'gain' can be of different nature.

First, this can be a large gain in overall performance. However, in the tasks we looked at, the overall gain is not considered to be large.

Second, this can be a gain in performance with respect to specific cases, which are regarded as important or interesting. Interestingly, this is the case for the following tasks we looked at: Summarization, Semantic Textual Similarity and Argument Reasoning Comprehension.

Third, this can be judged independent of any concrete gains in performance, but from a theoretical perspective. In our setup, we are interested in finding out about the potential of frame information. Our investigations are to pinpoint the cases and setups where frame information can be advantageous and we point to ambiguous setups with semantic gaps to be filled. These insights can be used for future approaches to further tasks to decide whether to include frame knowledge.

Finally, the next open question is about the ability of end-to-end approaches to involve linguistic meaning. On the one side, it is suggested to combine the *bottom-up* knowledge of meaning learned on large amounts of data with *top-down* knowledge from linguistic annotations (Young et al., 2018). On the other side, the linguistic potential and impact on discoveries in the sciences of end-to-end approaches is an open question. Whilst being confirmed that end-to-end approaches perform well

on NLP-tasks with enough training data, it is open to discussion[15] whether these approaches learn anything about linguistic meaning and if yes, whether the gained knowledge can be applied in the context of other tasks.

Outlook: Lifelong Learning Systems. Learning a set of skills in the context of one task and then applying them or even developing them further in the context of the next task points into the direction of transfer learning (Pan and Yang, 2010) and multitask learning (Caruana, 1997) or even lifelong machine learning in the sense of continuously learning to solve tasks of incremental complexity (Silver et al., 2013; Ruvolo and Eaton, 2013a,b; Chen and Liu, 2018). Transfer learning aims at improving the performance of a new task by transferring domain knowledge learned from previous tasks. In a multitask learning setting, the aim is to improve the performance of each task by jointly optimizing the learning on all of them. However, both transfer learning and multitask learning are not an ongoing learning process which would build up a memory. Lifelong learning systems, which could also be described as online multitask systems, learn to acquire knowledge which they can abstract from and use this for different tasks they will be confronted with later. In contrast to traditional knowledge bases storing relational triples about the world, the knowledge that is built up by a lifelong learning system is rather about how to solve task than about looking up facts. Connecting to earlier work, lifelong learning system can be seen as an extension of long short-term memory models (Hochreiter and Schmidhuber, 1997) whilst following the idea of learning in a lifelong context with minimal training data (Thrun, 1996).

There is a need for developing and exploring lifelong learning strategies in the field of Natural Language Processing. The idea of connecting neural networks to an external memory which can be written to and read from was implemented on the one hand in memory networks (Weston et al., 2015) which were applied to Question Answering and on the other hand it was implemented in the neural Turing machine (Graves et al., 2014) which was applied to lower level tasks. It is developed further in a differentiable neural computer (Graves et al., 2016) which is a neural network with access to an external memory for the purpose of representing learned knowledge, performing well on reasoning and inference problems in natural language.

This seems useful in any learning setting where the interest is not just in a system performing well on one particular task, but in a system performing well on several related tasks over the course of its lifespan.

As frame knowledge offers a semantically motivated repertoire of knowledge about situations and actions and meta-knowledge about interactions of these which showed to be beneficial to high-level tasks, it could contribute semantic background knowledge in a lifelong learning setup.

[15] see Twitter discussion about learning meaning in Natural Language Processing: `https://twitter.com/jacobandreas/status/1023246560082063366`

6.4 Summary of the Chapter

This chapter provides an extrinsic evaluation of our Frame Identification system `UniFrameId` and our pre-trained textual and structured frame embeddings, applied in the context of higher-level applications.

We provide a summary of this chapter in terms of bullet points in the following box below:

APPLICATIONS IN A NUTSHELL

Applications of **Frame Identification** showcase the potential of knowledge about situations, actions and their interactions for higher-level language understanding tasks.

- Frames are useful abstractions in preprocessing for higher-level tasks:
  Summarization and Evaluation thereof and Motif Construction

Applications of **frame embeddings** showcase the potential of FrameNet's knowledge for higher-level language understanding tasks.

- Frame embeddings are useful abstractions for semantic tasks:
  Semantic Textual Similarity and Argument Reasoning Comprehension

→ We find the highest potential for frame information to be in short sentences where frames can enrich the semantic context to fill a semantic gap.

Chapter 7

Outlook – Multimodal Challenges and Trend

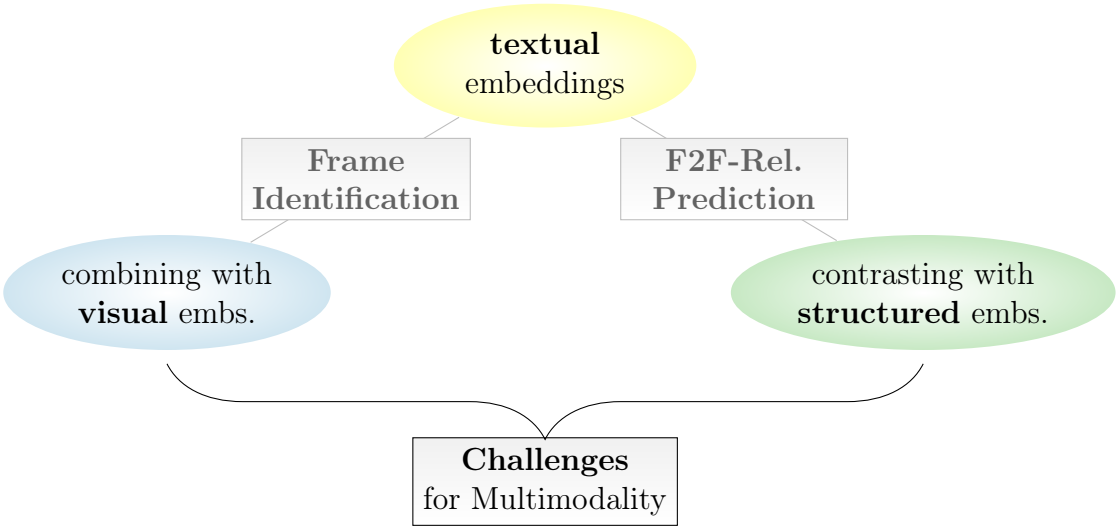


Figure 7.1: Chapter 7. Lowest gray box: outlook on challenges for grounded language processing involving embeddings from several modalities.

This chapter is to take a look ahead towards which next challenges and trends we can see given this thesis (as indicated in Figure 7.1). First, we determine future challenges for multimodal processing and explore the potential of using images of verbs, which are more difficult to grasp in pictures than entities are (Section 7.1). Second, we discuss the impact on Natural Language Processing (Section 7.2). Taken together, this chapter rounds off the thesis with broadening the view to future challenges including several modalities to infer representations from.

7.1 Challenges for Grounded Language Processing

Across many disciplines such as linguistics, cognitive science or artificial intelligence, multimodal grounding of language has been a longstanding goal. The discussion of grounded language processing has been fueled by recent research in learning multimodal embeddings, mostly by combining language and vision. Whilst most eval-

uations point out the beneficial impact of multimodal embeddings for a variety of tasks, explanatory analyses of this effect are still in a developing state.

In this section, we discuss open challenges that arise from existing work. This discussion is based on the literature review on combining and selecting information from different modalities as published in our survey (Beinborn et al., 2018).[1] In this survey, we propose for future work in language understanding to examine multimodal grounding beyond concrete nouns and adjectives, starting with the grounding of verbs. Consequently, this calls for larger multimodal datasets and also for a wider range of word classes.

Finally, we propose to broaden the view for implementing the grounding of language and not to purely focus on the vision modality but also to think about further modalities.

Multimodal Grounding of Verbs

In Section 4.2 we have seen that the visual information in terms of embeddings for noun synsets is useful to Frame Identification, especially in difficult settings with rare frames. Whilst visual embeddings complementing textual ones for concrete nouns, it is questionable whether the visual domain beneficially adds information to abstract nouns, verbs, or stop words. Referring back to Section 4.2.4, we explore the textual and visual grounding of highly embodied verbs (Section 4.2.4.1) and suggest to develop embeddings for verbs specifically that incorporate multimodal information.

Finally, for action- and motion-verbs, we suggest to explore the sensomotoric domain and to orient on how humans connect the understanding of actions and motions to the actual performance of those.

Visual Frame-Relation-Induction

In Sections 5.1 and 5.2 we have shown how textual and structured frame embeddings differ in modeling and prediction frame-to-frame relations. Referring back to Section 5.2.3, we see a large potential of leveraging the visual modality to extend structured frame embeddings for Frame-to-Frame Relation Prediction. In Chapter 4 we have shown how the frame-semantic task of Frame Identification can benefit from additional visual information. This leads to the question whether the task of predicting frame-to-frame relations could also benefit from additional visual information. However, for Frame-to-Frame Relation Prediction, visual embeddings for frames are needed, but these are not trivial to provide. In lack of images for frames with a broad coverage, we explore the potential of multimodal approaches to traditional Knowledge Base Completion (Section 5.2.3.1).

Finally, we suggest to develop approaches that incorporate visual information about frames to benefit Frame-to-Frame Relation Prediction and also frame induction (short frame-relation-induction). This could be explored in the context of the

[1] The literature review on combining and selecting information from different modalities is a joint contribution of myself together with my co-author Lisa Beinborn, for which we refer to our survey (Beinborn et al., 2018).

next international workshop on semantic evaluation (SemEval-2019)[2] where task 2 poses the challenge of unsupervised lexical semantic frame induction.

Combining and Selecting Information

The main challenge for jointly processing information from multiple modalities (cf. Section 2.3.2.3) is to efficiently combine information: complementary information shall be integrated whilst overlapping information can by summarized. In human language understanding, this process is performed naturally (Crocker et al., 2010), but the underlying mechanisms of human multimodal representations are not yet fully understood.

On the one hand, different modalities can capture different information as they are sensitive to different cues. To give an example, for a biker on a bike lane it is crucial to *(a)* see the traffic in front and *(b)* hear whether a car approaches from behind. Here, the information is complementary and needs to be combined. We will elaborate on this further in the first paragraph.

On the other hand, different modalities can capture the same information from different points of view in order to backup each other in case one modality is hindered. To give an example, a biker on a bike lane might perceive a crossing police car *(a)* by emergency lights and *(b)* by the siren; however, when the view is obstructed by buildings, the biker might just hear the siren and is still able to react if the auditory channel *(b)* is preferred over the visual channel *(a)*. Here, the information is either overlapping or even contradicting; thus the most relevant information needs to be selected. The second paragraph elaborates further on this. We close with a broadened view about the grounding of language.

Combining Complementary Information. During information processing and the interpretation of language, different modalities contribute qualitatively different conceptual information that can complement each other.

On the one hand, Bruni et al. (2014) argue that linguistic models rarely capture prominent visual properties because these are too obvious to be explicitly described in text (e.g., birds have wings, companies are built in cities). In Section 4.2, we follow the same motivation for including visual information into Frame Identification. On the other hand, textual models are superior to visual ones in terms of capturing taxonomic and functional relations between concepts which are not obvious on images (Collell and Moens, 2016). Thus, it would be ideal if multimodal representations could integrate these complementary perspectives.

This is questioned by Louwerse (2011) who sceptically state that perceptual information is already sufficiently encoded in textual cues. Following Louwerse (2011), the success of multimodal embeddings that has been found would mainly be due to a more robust representation of highly redundant information. However, in our experiments on multimodal Frame Identification, we find a benefit by specifically combining textual and visual embeddings, not by combining textual embeddings with structure-enhanced or random embeddings (cf. Section 4.2.2). Furthermore, with experiments on textual and visual similarity, Silberer and Lapata (2014) and

[2] SemEval-2019: `http://alt.qcri.org/semeval2019`

Hill et al. (2014) empirically support that textual embeddings better model textual similarities and visual embeddings better model visual similarities and, crucially, multimodal models are superior on both similarity tasks. Interestingly, with respect to the successful multimodal models, both evaluations show that a straight-forward concatenation ($\frown$, cf. Equation 3.19) of the two modalities is competitive already.

Finally, it remains an open research question to which extent visual embeddings can contribute complementary information when combined with textual embeddings.

Selective Multimodal Grounding. During information processing, different modalities back up each other in order to ensure processing of as much information as possible.

Regarding frame identification, we obtain multimodal sentence representations by including visual embeddings of noun synsets into the average of embeddings for each word; and this procedure improves the performance for the task compared to unimodal embeddings (cf. Section 4.2). This indicates that the superior performance is mainly due to a better representation of concepts. Consequently, we ask the question whether multimodal grounding should be performed on selected words only.

For determining whether visual information should be included into judging sentence similarities, Glavaš et al. (2017) leverage the image dispersion score (Kiela et al., 2014) to compute the concreteness of a concept and to then include visual information for the concrete concepts. The image dispersion score calculates the average pairwise cosine distance (d_{cos}, cf. Equation 3.2) in a set of images. Thus, the score is high for a diverse set of images, i.e. for an abstract concept like *freedom*, and it is low for a homogeneous set of images, i.e. for a concrete concept like *cup*. Further measures for concreteness, which are based on the same idea, are proposed by Lazaridou et al. (2015) and Hessel et al. (2018). However, these measures depend on the set of images delivered by a retrieval algorithm, which in turn might be optimized towards obtaining a diverse or homogeneous set.

Finally, it remains an open research question how to determine which modalities to select and whether this decision should be done on a very fine-grained level, such as the word-level. To give an example on word-level other than concrete versus abstract nouns or highly or little embodied verbs, some functional words (e.g., locative expressions) might benefit from multimodal information, but it remains open to research how words with syntactic functions (e.g., coordinating expressions) should be represented visually.

Broadened View about Grounding of Language. Current research on multimodal grounding of language (including this chapter) focuses on the visual modality to complement traditional text-based approaches. Most multimodal research is driven by the vision and language community; still, we also report about research including the auditory (Kiela and Clark, 2015, 2017) and olfactory channel (Kiela et al., 2015) (cf. Section 2.3.1). From our perspective, it is important to also work on the inclusion of further modalities.

Especially for the grounding of highly embodied verbs, we think the next step is the exploration of the motion modality. This means in order to learn embeddings for embodied verbs, information could be inferred from the trace of the motion

when performing the verb. The cognitive theory on mental simulations is supported by evidence from somatotopic activations of the motor cortex (Hauk et al., 2004; Pulvermüller, 2005; Barsalou, 2008), thus, grounding verb embeddings in motion data can be understood as an implementation of this theory.

For embedding abstract concepts, we think the view on modalities should be broadened. Whilst exploring further human modalities (such as motion for verbs), we consider the inclusion of 'modalities' that do not directly correspond to the human sensomotoric inventory as promising, too. In our work, we experiment with only one aspect: knowledge bases that store information about abstract concepts. Other aspects could be the inclusion of simple tables (as for example to explain statistics and results of experiments) or of more complex interaction data (as for example to model social networks by interactions online).

We see a large potential in the integration of multimodal information to improve the understanding of meaning in language.

7.2 Trend for the Role of Natural Language Processing

This section is to round off the thesis with some final comments on the trend for the role of Natural Language Processing. After this thesis, we argue for Natural Language Processing not to be treated as an isolated field but in collaboration with other disciplines.

In order to automatically understand language as humans do naturally (Jurafsky and Martin, 2017), algorithms are required to learn and think like people and to acquire meta-level skills (Lake et al., 2017). Consequently following the theory of mental simulations for grounded cognition and embodied sentence comprehension (Barsalou, 2008), approaches are developed to simulate interactions in the real world.

McClelland et al. (2016) outline the challenge of teaching an artificial agent that lives in a simulated two-dimensional world to explore and acquire mathematical abilities with embodied learning and cognition so that, eventually, it could pass a high school test in geometry. To enable experiences similar to those of a human learner, they suggest to allow the agent to explore and to manipulate its environment via a simple simulated hand. The underlying research question asks about the nature of human knowledge and about the mechanisms (e.g., experience, culture) we use to learn these concepts.

Lazaridou et al. (2017) introduce a framework for language learning that relies on multi-agent communication instead of large amounts of text. Their framework implements referential games where a sender and a receiver communicate about images. They use this setup as a testbed for evolving languages.

Furthermore, simulations of physical actions are currently receiving attention: dynamic physical manipulations are embedded computationally to approach tool-use and motion planning problems (Toussaint et al., 2018).

Whilst in this thesis, we work on two complementary branches of knowledge and experiment with the integration of the visual modality into text- or structure-based approaches, the future trend for understanding language is arising in the direction of embodied embedding learning as outlined above. Embodied embedding learning

grounds the acquisition of representations in actions or experiences of agents that interact in simulated environments. This requires the integration of foundations and methods from several disciplines.

7.3 Summary of the Chapter

We provide a summary of this chapter in terms of bullet points in the box below:

OUTLOOK IN A NUTSHELL

We determine **challenges** for grounded language processing:

- Multimodal grounding of verbs *(e.g., highly embodied 'jumping')*
- Visual embeddings for frames *(e.g., imSitu and VerSe datasets)*
- Combining complementary information *(e.g., text and images in comic)*
- Selective multimodal grounding *(e.g., focus on audio when vision is obscured)*

We see the **role** of Natural Language Processing as contributor in or aggregator of interdisciplinary research:

- Understanding of meaning in language involves different modalities, thus Natural Language Processing is widening its input format from text-only to multiple modalities
- Natural Language Processing is not an isolated field of research, but one facet to combine with others for grounded approaches to problems in Artificial Intelligence

Chapter 8

Summary

This chapter is to reflect upon the work we have done in the context of this thesis. We summarize our contributions and insights and discuss future research questions that arise.

The ability to communicate meaning plays an integral role to human intelligence. When understanding words, sentences, or texts, complex ambiguities need to be resolved in order to approximate the underlying meaning: words are linked to their referents in a large multimodal repertoire of knowledge. Automatic language understanding is a challenging problem for Artificial Intelligence in general and especially for Natural Language Processing (NLP). The field of NLP aims to model and to analyze language as used by humans as means of communication and the higher-order goal is to automatically understand language as humans do naturally.

The challenge of capturing meaning and world knowledge in language can be split into the two branches of (a) *knowledge about situations and actions* and (b) *structured relational knowledge* and can be studied with different kinds of embeddings, textual, structured and visual or multimodal, for operationalizing different aspects of knowledge.

To approach the challenges, we choose to closely rely upon the lexical-semantic knowledge base FrameNet as it addresses both branches of capturing world knowledge whilst taking into account the linguistic theory of frame semantics which orients on human language understanding. FrameNet provides frames for categories of meaning and frame-to-frame relations for interactions between concepts, which are central to the tasks of Frame Identification and Frame-to-Frame Relation Prediction, respectively.

In this thesis, we focus on aspects of automatic language understanding and leverage current embedding methods where human concepts, such as words in language, or artefacts from the world, such as objects depicted on images, are modeled as high-dimensional vectors. The fundamental assumption of this thesis is about holistic language understanding requiring different aspects of world knowledge – which is inspired by human language understanding.

In the following, we look back on the thesis in order to summarize our work on modeling frame semantics with different kinds of representations.

Retrospection

In retrospection of this thesis, we review our contributions and insights along the sketch in Figure 8.1.

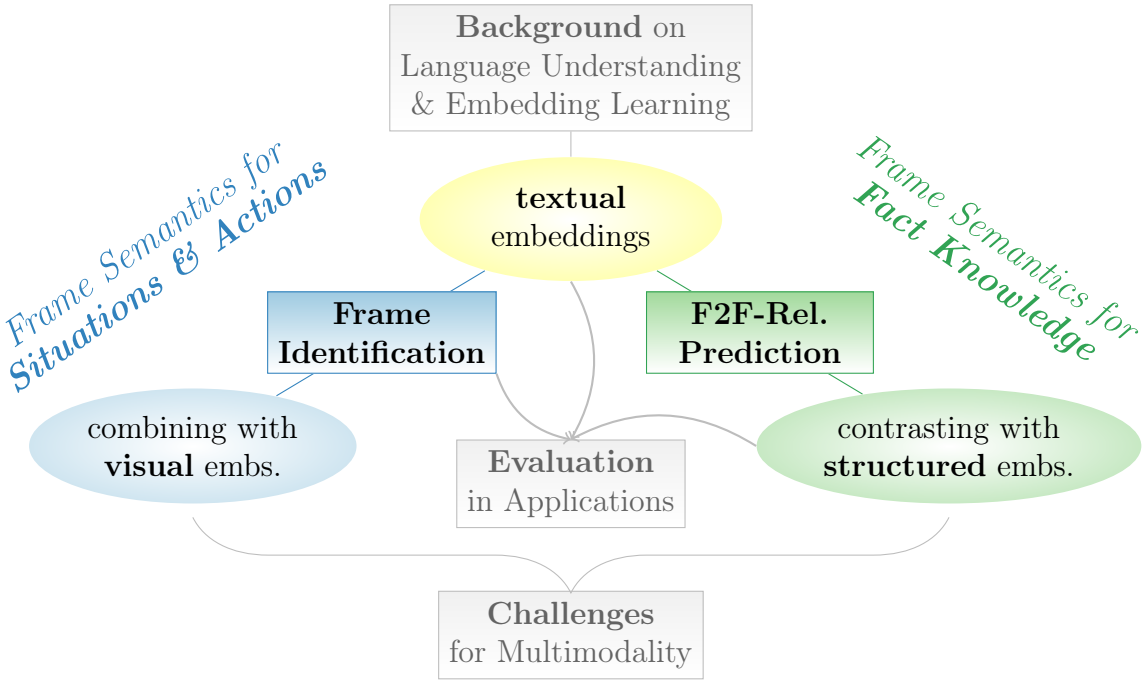


Figure 8.1: Retrospection of thesis structure. Upper gray box (Ch. 2, 3): theoretical and methodological background. Left blue branch (Ch. 4): knowledge about situations and actions with textual and visual word embeddings for Frame Identification. Right green branch (Ch. 5): knowledge about facts with textual versus structured frame embeddings for Frame-to-Frame Relation Prediction. Middle gray box (Ch. 6): evaluation of frame knowledge in applications. Lower gray box (Ch. 7): outlook on challenges for grounded language processing.

To start with, we provide the theoretical and methodological background on language understanding and on embedding learning (upper gray box). In Chapter 2 we introduce the facets of language understanding: textual semantics for situations and actions, structured relational knowledge and grounded language understanding. In Chapter 3 we review methods of representation learning which we apply to our data: textual, structured, visual and multimodal embedding approaches.

To study world knowledge as conceptualized by frame semantics and viable by embeddings, we branch out into two directions.

On the one hand (left **blue** branch), in Chapter 4, we model knowledge about situations and actions with textual word embeddings and in combination with visual ones for the task of Frame Identification. First, we develop a state-of-the-art Frame Identification `UniFrameId` system that operates on FrameNets of two languages, namely English and German. The underlying assumption is about context knowledge being necessary for abstracting from single words to categories of meaning in terms of frames. We find that taking the context words into account in terms of textual embeddings in a straight-forward neural network architecture yields state-

of-the-art results for English as well as for German data. Second, we extend our unimodal Frame Identification system to a use-case with multimodal embeddings, `MultiFrameId`, which improves the performance on English data. The underlying hypothesis is that language understanding requires implicit commonsense knowledge which is rarely expressed textually but can be extracted from images. We find that additional information from images is beneficial to Frame Identification and also to further NLP-tasks. To further advance the performance of the multimodal approach, we suggest to develop embeddings for verbs specifically that incorporate multimodal information.

On the other hand (right **green** branch), in Chapter 5, we contrast textual and structured frame embeddings to model knowledge about relations in the task of Frame-to-Frame Relation Prediction. First, we introduce Frame-to-Frame Relation Prediction as a Knowledge Base Completion task for FrameNet. Second, we contrast the performance of different kinds of frame embeddings: textual versus structured. The underlying research question is whether frame-to-frame relations can be directly inferred from text. We point out the limitations of textual embeddings in mirroring frame-to-frame relations, and also we point out the advantage of structured embeddings in correctly predicting relations between frame pairs. As textual and structured frame embeddings differ, they can provide different kind of frame knowledge when applied as features in the context of further tasks. Our best-performing system of our `StruFFRel` approach can be used to generate recommendations for annotations with relations. To further advance the performance of Frame-to-Frame Relation Prediction and also of the induction of new frames and relations (short frame-relation-induction), we suggest to develop approaches that incorporate visual information.

Subsequently, in Chapter 6, we extrinsically evaluate frame knowledge (from the two branches) in high-level tasks (middle gray box) by reporting about applications of our unimodal Frame Identification system and of our textual and structured frame embeddings. The applications are: Summarization, Summary Evaluation, Motif Construction, Semantic Textual Similarity, and Argument Reasoning Comprehension. Across these applications, we see a trend that frame knowledge is particularly beneficial in ambiguous and short sentences.

Finally, in Chapter 7, we resume with an outlook on the directly implied next challenges for grounded language processing (lower gray box). Other than the development of multimodal verb embeddings and the integration of visual information for frame-relation-induction, we see the need to automatically learn how to combine complementary information and select relevant information from different modalities. We see a trend for Natural Language Processing to mutually benefit from the expertise in other sub-fields of Artificial Intelligence, such as, but not limited to, Computer Vision.

Conclusion and Future Research Directions

After our investigations and insights gained in this thesis, we see:

- Reasoning over several modalities makes one more robust
- Incorporating frame knowledge helps one in tasks asking to bridge a semantic gap

By this we mean, in more detail, that NLP-tasks need to be approached from different perspectives. This is in line with our suggestion to (I) develop embeddings for verbs specifically that incorporate multimodal information for Frame Identification, and (II) develop approaches that incorporate visual information for frame-relation-induction. An important challenge within NLP-tasks targeting facets of language understanding is to find out which modalities or kinds of knowledge complement each other in a beneficial way in the context of a respective task. For different tasks and domains, multimodal combinations of embeddings need to adjust the degree of involvement of the single modalities, respectively. To illustrate this, we touch three simple explanatory examples.

First, on the task-level: the tasks of Frame Identification and Knowledge Base Completion are of different nature. The first one focuses on the meaning of words in the context of situations and actions – where textual embeddings are better suited; and the second one focuses on relations between concepts – structured embeddings are better suited.

Second, regarding the text domain: fact-based texts versus figurative texts are better captured by different kinds of embeddings. Structured embeddings better incorporate relational triples whereas visual embeddings better capture figures or metaphors.

Third, with respect to embeddings: visual embeddings complement textual ones for concrete nouns, but it is questionable whether the visual domain beneficially adds information to abstract nouns, verbs, or stop words. Thus, for action- and motion-verbs, we suggest to explore the sensomotoric domain.

Furthermore, we see the caveat of frame knowledge being beneficial for higher-level applications but training data for Frame Identification or Frame-to-Frame Relation Prediction being sparse. Thus, we encourage developing approaches for extending training data to acquire frame knowledge. Besides manually extending FrameNet, there are promising automatic approaches. In our work, we experiment with supervised and supervision-less approaches for Frame-to-Frame Relation Prediction and we combine knowledge from different modalities for Frame Identification. As direct implication of our work, we determine visual knowledge to be integrated next into frame-relation-induction. On the one hand, we point out the advantage of structured frame embeddings over textual ones for predicting frame-to-frame relations. On the other hand, we point out an overall advantage of multimodal embeddings for Frame Identification together with their considerable advantage on rare predicate classes. Combining the insights from both aspects, we see a large potential of multimodal embeddings to advance the task of Frame-to-Frame Relation

Prediction and also of frame-relation-induction.

In addition to the challenges for grounded language processing as described in Section 7.1, we see the following research directions subsequent to this thesis.

We identify the comparison of different representations as an interesting follow-up direction. More detailed, this includes the development of methods for comparing embedding spaces beyond qualitatively comparing word pair similarities (which are the closest neighbors to a word and do these neighbors make sense?) or clusters (which groups of words are close to each other and do these clusters make sense?). To the best of our knowledge, there is no established method to determine the similarity and the complementarity of embedding spaces – which could indeed be helpful to decide what aspects of which embedding space adds new complementary information to an approach of interest. Recent work by Sikos and Padó (2018b) suggests making use of the neighborhood structure within frame embeddings in order to apply frame semantics cross-lingually, for example on the pair of FrameNet and SALSA. Interestingly, when further elaborating on their IMAGINED method, Collell and Moens (2018), propose a new similarity measure to find out whether the neighborhood structure (i.e., pairwise similarities) of the mapped embeddings rather resemble the initial embedding space or the target embedding space. Following this, it would be useful to automatically learn how to select qualitatively different aspects of embedding spaces with respect to a task of interest. This means, we want to be able to judge the complementarity of embedding spaces with respect to specific NLP-tasks and finally, this knowledge of complementarity of embedding spaces could be leveraged by weighing them accordingly.

Regarding the cognitive point of view, this thesis investigates language processing involving information or experience from different (human) modalities – however, the other way round, it is an open field of research: to what extent human knowledge and experience are shaped by language? We see this as an important question to study as research is advancing in the field of automatically understanding and generating language.

List of Figures

List of Tables

Bibliography

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre: 'SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity', in: *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval, held in conjunction with NAACL)*, pp. 385–393, Association for Computational Linguistics, Montréal, Canada, June 2012.

Roni Ben Aharon, Idan Szpektor, and Ido Dagan: 'Generating Entailment Rules from FrameNet', in: *Proceedings of the 48th Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 241–246, Association for Computational Linguistics, Uppsala, Sweden, July 2010.

Simon Ahrendt and Vera Demberg: 'Improving Event Prediction by Representing Script Participants', in: *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 546–551, Association for Computational Linguistics, San Diego, California, USA, June 2016.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio: 'Visually Grounded and Textual Semantic Models Differentially Decode Brain Activity Associated with Concrete and Abstract Nouns', *Transactions of the Association for Computational Linguistics (TACL)* 5 (1): 17–30, 2017.

Thomas O. Arnold: *Advanced Motif Analysis on Text Induced Graphs*, Dissertation, Technische Universität Darmstadt, 2018.

Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli: 'Multimodal Fusion for Multimedia Analysis: A Survey', *Multimedia Systems* 16 (6): 345–379, 2010.

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio: 'Neural Machine Translation by Jointly Learning to Align and Translate', *arXiv preprint arXiv:1409.0473* 2014.

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio: 'End-to-End Attention-Based Large Vocabulary Speech Recognition', in: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4945–4949, IEEE, 2016.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe: 'The Berkeley FrameNet Project', in: *Proceedings of the 36th Annual Meeting of the Association for Com-*

putational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL), pp. 86–90, Association for Computational Linguistics, Montréal, Canada, August 1998.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency: 'Multimodal Machine Learning: A Survey and Taxonomy', *Transactions on Pattern Analysis and Machine Intelligence* 2018.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider: 'Abstract Meaning Representation for Sembanking', in: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (held in conjunction with ACL)*, pp. 178–186, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

Lawrence W. Barsalou: 'Perceptual Symbol Systems', *Behavioral and Brain Sciences* 22 (4): 637–660, 1999.

Lawrence W. Barsalou: 'Grounded Cognition', *Annual Review of Psychology* 59: 617–645, 2008.

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych: 'Multimodal Grounding for Language Processing', in: *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pp. 2325–2339, Association for Computational Linguistics, Santa Fe, New Mexico, USA, August 2018.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor: 'Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge', in: *Proceedings of the International Conference on Management of Data (ACM SIGMOD)*, pp. 1247–1250, 2008.

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio: 'Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing', in: *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pp. 127–135, 2012.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko: 'Translating Embeddings for Modeling Multi-relational Data', in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 2787–2795, 2013.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio: 'Learning Structured Embeddings of Knowledge Bases', in: *Proceedings of the 25th Conference of Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 301–306, AAAI Press, San Francisco, California, USA, 2011.

Teresa Botschen, Iryna Gurevych, Jan-Christoph Klie, Hatem Mousselly Sergieh, and Stefan Roth: 'Multimodal Frame Identification with Multilingual Evaluation', in: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 1481–1491, Association for Computational Linguistics, New Orleans, Louisiana, USA, June 2018a.

Teresa Botschen, Hatem Mousselly-Sergieh, and Iryna Gurevych: 'Prediction of Frame-to-Frame Relations in the FrameNet Hierarchy with Frame Embeddings', in: *Proceedings of the 2nd Workshop on Representation Learning for NLP (RepL4NLP, held in conjunction with ACL)*, pp. 146–156, Vancouver, Canada, August 2017.

Teresa Botschen, Daniil Sorokin, and Iryna Gurevych: 'Frame- and Entity-Based Knowledge for Common-Sense Argumentative Reasoning', in: *Proceedings of the 5th International Workshop on Argument Mining (ArgMin, held in conjunction with EMNLP)*, pp. 90–96, Association for Computational Linguistics, Brussels, Belgium, November 2018b.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning: 'A Large Annotated Corpus for Learning Natural Language Inference', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 632–642, Association for Computational Linguistics, Lisbon, Portugal, September 2015.

Will Bridewell and Paul F. Bello: 'A Theory of Attention for Cognitive Systems', *Advances in Cognitive Systems* 4: 1–16, 2016.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran: 'Distributional Semantics in Technicolor', in: *Proceedings of the 50th Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 136–145, Association for Computational Linguistics, Jeju, Republic of Korea, July 2012.

Elia Bruni, Nam-Khanh Tram, and Marco Baroni: 'Multimodal Distributional Semantics', *Journal of Artificial Intelligence Research* 49: 1–47, 2014.

Elia Bruni, Giang Binh Tran, and Marco Baroni: 'Distributional Semantics from Text and Images', in: *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (GEMS, held in conjunction with EMNLP)*, pp. 22–32, Association for Computational Linguistics, Edinburgh, Scotland, UK, July 2011.

Luana Bulat, Stephen Clark, and Ekaterina Shutova: 'Modelling Metaphor with Attribute-Based Semantics', in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 523–528, Association for Computational Linguistics, Valencia, Spain, April 2017b.

Luana Bulat, Stephen Clark, and Ekaterina Shutova: 'Speaking, Seeing, Understanding: Correlating Semantic Models with Conceptual Representation in the Brain', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1081–1091, Association for Computational Linguistics, Copenhagen, Denmark, September 2017a.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal: 'The SALSA Corpus: A German Corpus Resource for Lexical Semantics', in: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, May 2006.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal: 'Using FrameNet for the Semantic Analysis of German: Annotation, Representation, and Automation', in: *Multilingual FrameNets in Computational Lexicography*, pp. 209–244, Mouton de Gruyter, New York City, USA, 2009.

Rich Caruana: 'Multitask Learning', *Machine Learning* 28 (1): 41–75, 1997.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia: 'SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation', in: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval, held in conjunction with ACL)*, pp. 1–14, Association for Computational Linguistics, Vancouver, Canada, August 2017.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman: 'Return of the Devil in the Details: Delving Deep into Convolutional Nets', in: *Proceedings of the British Machine Vision Conference (BMVC)*, BMVA Press, Nottingham, Great Britain, 2014.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei: 'Neural Natural Language Inference Models Enhanced with External Knowledge', in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2406–2417, Association for Computational Linguistics, Melbourne, Australia, July 2018.

Zhiyuan Chen and Bing Liu: 'Lifelong Machine Learning', *Synthesis Lectures on Artificial Intelligence and Machine Learning* 12 (3): 1–207, 2018.

Pengxiang Cheng and Katrin Erk: 'Implicit Argument Prediction with Event Knowledge', in: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 831–840, Association for Computational Linguistics, New Orleans, Louisiana, USA, June 2018.

Kyunghyun Cho, Bart van Merriënboer, Gülçehre Çağlar, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio: 'Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Association for Computational Linguistics, Doha, Qatar, October 2014.

HongSeok Choi and HyunJu Lee: 'GIST at SemEval-2018 Task 12: A Network Transferring Inference Knowledge to Argument Reasoning Comprehension Task', in: *Proceedings of The 12th International Workshop on Semantic Evaluation (SemEval, held in conjunction with NAACL)*, pp. 773–777, Association for Computational Linguistics, New Orleans, Louisiana, USA, June 2018.

Noam Chomsky: *Knowledge of Language: Its Nature, Origin, and Use*, Praeger Scientific, New York, 1986.

Guillem Collell and Marie-Francine Moens: 'Is an Image Worth More than a Thousand Words? On the Fine-Grain Semantic Differences Between Visual and Linguistic Representations', in: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pp. 2807–2817, Association for Computational Linguistics, Osaka, Japan, December 2016.

Guillem Collell and Marie-Francine Moens: 'Do Neural Network Cross-Modal Mappings Really Bridge Modalities?', in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 462–468, Association for Computational Linguistics, Melbourne, Australia, July 2018.

Guillem Collell, Ted Zhang, and Marie-Francine Moens: 'Imagined Visual Representations as Multimodal Embeddings', in: *Proceedings of the 31st Conference of Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 4378–4384, AAAI Press, 2017.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa: 'Natural Language Processing (Almost) from Scratch', *Journal of Machine Learning Research* 12 (August): 2493–2537, 2011.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes: 'Supervised Learning of Universal Sentence Representations from Natural Language Inference Data', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 670–680, Association for Computational Linguistics, Copenhagen, Denmark, September 2017.

Bob Coyne and Owen Rambow: 'LexPar: A Freely Available English Paraphrase Lexicon Automatically Extracted from FrameNet', in: *Proceedings of the 3rd IEEE International Conference on Semantic Computing (ICSC)*, pp. 53–58, IEEE, 2009.

Matthew W. Crocker, Pia Knoeferle, and Marshall R. Mayberry: 'Situated Sentence Processing: The Coordinated Interplay Account and a Neurobehavioral Model', *Brain and Language* 112 (3): 189–201, 2010.

Walter Daelemans, Anja Höthker, and Erik T. K. Sang: 'Automatic Sentence Simplification for Subtitling in Dutch and English', in: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pp. 1045–1048, Lisbon, Portugal, May 2004.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith: 'Frame-Semantic Parsing', *Computational Linguistics* 40 (1): 9–56, 2014.

Dipanjan Das and Noah A. Smith: 'Semi-Supervised Frame-Semantic Parsing for Unknown Predicates', in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pp. 1435–1444, Association for Computational Linguistics, Portland, Oregon, USA, June 2011.

Ernest Davis and Gary Marcus: 'Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence', *Communications of the ACM* 58 (9): 92–103, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei: 'ImageNet: A Large-Scale Hierarchical Image Database', in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, IEEE, Miami, Florida, USA, June 2009.

Timothy Dozat: 'Incorporating Nesterov Momentum into Adam', in: *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, 2016.

Michael Ellsworth, Katrin Erk, Paul Kingsbury, and Sebastian Padó: 'PropBank, SALSA, and FrameNet: How Design Determines Product', in: *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora (held in conjunction with LREC)*, Lisbon, Portugal, May 2004.

David Embick and David Poeppel: 'Towards a Computational(ist) Neurobiology of Language: Correlational, Integrated and Explanatory Neurolinguistics', *Language, Cognition and Neuroscience* 30 (4): 357–366, 2015.

Nicolai Erbs, Torsten Zesch, and Iryna Gurevych: 'Link Discovery: A Comprehensive Analysis', in: *Proceedings of the 5th IEEE International Conference on Semantic Computing (ICSC)*, pp. 83–86, IEEE, 2011.

Katrin Erk: 'Vector Space Models of Word Meaning and Phrase Meaning: A Survey', *Language and Linguistics Compass* 6 (10): 635–653, 2012.

Katrin Erk and Sebastian Pado: 'Shalmaneser - A Flexible Toolbox for Semantic Role Assignment', in: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, May 2006.

Michael Färber, Basil Ell, Carsten Menne, and Achim Rettinger: 'A Comparative Survey of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO', *Semantic Web* 1: 1–5, 2015.

Parvin S. Feizabadi and Sebastian Padó: 'Automatic Identification of Motion Verbs in WordNet and FrameNet', in: *Workshop Proceedings of the 11th Conference on Natural Language Processing (KONVENS)*, pp. 70–79, Vienna, Austria, September 2012.

Christiane Fellbaum: 'English Verbs as a Semantic Net', *International Journal of Lexicography* 3 (4): 278–301, 1990.

Christiane Fellbaum: 'WordNet: An Electronic Lexical Database', *Language* 76 (3): 706–708, 2000.

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou: 'Applying Deep Learning to Answer Selection: A Study and An Open Task', in: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 813–820, IEEE, 2015.

Yansong Feng and Mirella Lapata: 'Visual Information in Semantic Representation', in: *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 91–99, Association for Computational Linguistics, Los Angeles, California, USA, June 2010.

Charles J. Fillmore: 'The Case for Case', in Emmon Bach and Robert T. Harms (Eds.): *Proceedings of the Texas Symposium, on Language Universals*, pp. 1–88, Holt, Rinehart & Winston, April 1967.

Charles J. Fillmore: 'Frame Semantics and the Nature of Language', *Annals of the New York Academy of Sciences* 280 (1): 20–32, 1976.

Charles J. Fillmore: 'Frames and the Semantics of Understanding', *Quaderni di Semantica* 6 (2): 222–254, 1985.

Charles J. Fillmore: 'Encounters with Language', *Computational Linguistics* 38 (4): 701–718, 2012.

Charles J. Fillmore and Collin F. Baker: 'Frame Semantics for Text Understanding', in: *Proceedings of the WordNet and Other Lexical Resources Workshop (held in conjunction with NAACL)*, June 2001.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck: 'Background to FrameNet', *International Journal of Lexicography* 16 (3): 235–250, 2003.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin: 'Placing Search in Context: The Concept Revisited', *ACM Transactions on Information Systems (TOIS)* 20 (1): 116–131, January 2002.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das: 'Semantic Role Labeling with Neural Network Factors', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 960–970, Association for Computational Linguistics, Lisbon, Portugal, September 2015.

Lucie Flekova and Iryna Gurevych: 'Supersense Embeddings: A Unified Model for Supersense Interpretation, Prediction, and Utilization', in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2029–2041, Association for Computational Linguistics, Berlin, Germany, August 2016.

Jerry A. Fodor: 'Précis of the Modularity of Mind', in: *Readings in Cognitive Science*, pp. 73–77, Elsevier, 1988.

Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov: 'DeViSE: A Deep Visual-Semantic Embedding Model', in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 2121–2129, 2013.

Spandana Gella, Mirella Lapata, and Frank Keller: 'Unsupervised Visual Sense Disambiguation for Verbs Using Multimodal Embeddings', in: *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 182–192, San Diego, California, USA, June 2016.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen: 'SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2173–2182, Association for Computational Linguistics, Austin, Texas, USA, November 2016.

Daniel Gildea and Daniel Jurafsky: 'Automatic Labeling of Semantic Roles', *Computational linguistics* 28 (3): 245–288, 2002.

Tobias Glasmachers: 'Limits of End-to-End Learning', in: *Proceedings of the 9th Asian Conference on Machine Learning*, pp. 17–32, 2017.

Goran Glavaš, Ivan Vulić, and Simone P. Ponzetto: 'If Sentences Could See: Investigating Visual Information for Semantic Textual Similarity', in: *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*, Montpeiller, France, 2017.

Max Glockner, Vered Shwartz, and Yoav Goldberg: 'Breaking NLI Systems with Sentences that Require Simple Lexical Inferences', in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 650–655, Association for Computational Linguistics, Melbourne, Australia, July 2018.

Yoav Goldberg: 'A Primer on Neural Network Models for Natural Language Processing', *Journal of Artificial Intelligence Research* 57: 345–420, 2016.

Alex Graves and Navdeep Jaitly: 'Towards End-to-End Speech Recognition with Recurrent Neural Networks', in: *International Conference on Machine Learning*, pp. 1764–1772, 2014.

Alex Graves, Greg Wayne, and Ivo Danihelka: 'Neural Turing Machines', *arXiv preprint arXiv:1410.5401* 2014.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou et al.: 'Hybrid Computing Using a Neural Network with Dynamic External Memory', *Nature* 2016.

Jiang Guo, Wanxiang Che, Haifeng Wang, Ting Liu, and Jun Xu: 'A Unified Architecture for Semantic Role Labeling and Relation Classification', in: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pp. 1264–1274, Association for Computational Linguistics, Osaka, Japan, December 2016.

Saurabh Gupta and Jitendra Malik: 'Visual Semantic Role Labeling', *arXiv preprint arXiv:1505.04474* 2015.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein: 'The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants', in: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 1930–1940, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018.

Hans van Halteren, Jakub Zavrel, and Walter Daelemans: 'Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems', *Computational linguistics* 27 (2): 199–229, 2001.

Zellig S. Harris: 'Distributional Structure', *WORD* 10 (2-3): 146–162, 1954.

Silvana Hartmann and Iryna Gurevych: 'FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection', in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1363–1373, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych: 'Out-of-domain FrameNet Semantic Role Labeling', in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 471–482, Association for Computational Linguistics, Valencia, Spain, April 2017.

Joshua K. Hartshorne, Claire Bonial, and Martha Palmer: 'The VerbCorner Project: Findings from Phase 1 of Crowd-Sourcing a Semantic Decomposition of Verbs', in: *Proceedings of the 52nd Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 397–402, Association for Computational Linguistics, Baltimore, Maryland, USA, June 2014.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Vol. 1, Springer-Verlag, 2009.

Olaf Hauk, Ingrid Johnsrude, and Friedemann Pulvermüller: 'Somatotopic Representation of Action Words in Human Motor and Premotor Cortex', *Neuron* 41 (2): 301–307, 2004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun: 'Deep Residual Learning for Image Recognition', in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, 2016.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer: 'Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling', in: *Proceedings of the 56th Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 364–369, Association for Computational Linguistics, Melbourne, Australia, July 2018.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer: 'Deep Semantic Role Labeling: What Works and What's Next', in: *Proceedings of the 55th Annual*

Conference of the Association for Computational Linguistics (ACL), pp. 473–483, Association for Computational Linguistics, Vancouver, Canada, August 2017.

Karl M. Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev: 'Semantic Frame Identification with Distributed Word Representations', in: *Proceedings of the 52th Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 1448–1458, Association for Computational Linguistics, Baltimore, Maryland, USA, June 2014.

Karl M. Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom: 'Teaching Machines to Read and Comprehend', in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1693–1701, 2015.

Jack Hessel, David Mimno, and Lillian Lee: 'Quantifying the Visual Concreteness of Words and Topics in Multimodal Datasets', *arXiv preprint arXiv:1804.06786* 2018.

Felix Hill and Anna Korhonen: 'Learning Abstract Concept Embeddings from Multi-Modal Data: Since you Probably Can't See What I Mean', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 255–265, Association for Computational Linguistics, Doha, Qatar, October 2014.

Felix Hill, Roi Reichart, and Anna Korhonen: 'Multi-Modal Models for Concrete and Abstract Concept Meaning', *Transactions of the Association for Computational Linguistics (TACL)* 2: 285–296, 2014.

Felix Hill, Roi Reichart, and Anna Korhonen: 'Simlex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation', *Computational Linguistics* 41 (4): 665–695, 2015.

Sepp Hochreiter and Jürgen Schmidhuber: 'Long Short-Term Memory', *Neural computation* 9 (8): 1735–1780, 1997.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli: 'SensEmbed: Learning Sense Embeddings for Word and Relational Similarity', in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing*, pp. 95–105, Association for Computational Linguistics, Beijing, China, 2015.

Anders Johannsen, Héctor M. Alonso, and Anders Søgaard: 'Any-Language Frame-Semantic Parsing', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 2062–2066, Lisbon, Portugal, 2015.

Biing-Hwang Juang and Lawrence R. Rabiner: 'Automatic Speech Recognition – A Brief History of the Technology Development', *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara* 1: 67, 2005.

Daniel Jurafsky and James H. Martin: *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, online draft, 2017.

Alexandre Kabbach, Corentin Ribeyre, and Aurélie Herbelot: 'Butterfly Effects in Frame Semantic Parsing: Impact of Data Processing on Model Ranking', in: *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pp. 3158–3169, Association for Computational Linguistics, Santa Fe, New Mexico, USA, August 2018.

Andrej Karpathy, Armand Joulin, and Li Fei-Fei: 'Deep Fragment Embeddings for Bidirectional Image Sentence Mapping', in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1889–1897, 2014.

Chris Kedzie, Kathleen McKeown, and Hal Daume III: 'Content Selection in Deep Learning Models of Summarization', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1818–1828, Association for Computational Linguistics, Brussels, Belgium, November 2018.

Maurice G. Kendall: 'A New Measure of Rank Correlation', *Biometrika* 30 (1/2): 81–93, 1938.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth: 'Question Answering as Global Reasoning over Semantic Abstractions', in: *Proceedings of the 32nd Conference of Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 1905–1914, AAAI Press, 2018.

Douwe Kiela and Léon Bottou: 'Learning Image Embeddings Using Convolutional Neural Networks for Improved Multi-Modal Semantics', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 36–45, Doha, Qatar, 2014.

Douwe Kiela, Luana Bulat, and Stephen Clark: 'Grounding Semantics in Olfactory Perception', in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing*, pp. 231–236, Association for Computational Linguistics, Beijing, China, 2015.

Douwe Kiela and Stephen Clark: 'Multi- and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2461–2470, Association for Computational Linguistics, Lisbon, Portugal, September 2015.

Douwe Kiela and Stephen Clark: 'Learning Neural Audio Embeddings for Grounding Semantics in Auditory Perception', *Journal of Artificial Intelligence Research* 60: 1003–1030, 2017.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark: 'Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More',

in: *Proceedings of the 52nd Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 835–841, Association for Computational Linguistics, Baltimore, Maryland, USA, June 2014.

Douwe Kiela, Anita Verő, and Stephen Clark: 'Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 447–456, Association for Computational Linguistics, Austin, Texas, USA, November 2016.

Diederik Kingma and Jimmy Ba: 'Adam: A Method for Stochastic Optimization', *arXiv preprint arXiv:1412.6980* 2014.

Thomas Kleinbauer and Thomas A. Trost: 'Comparing Distributional and Frame Semantic Properties of Words', pp. 60–68, September 2018.

Jan-Christoph Klie: *Deep Learning for FrameNet Semantic Role Labeling*, Master thesis, Technische Universität Darmstadt, October 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton: 'Imagenet Classification with Deep Convolutional Neural Networks', in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.

Sandra Kübler, Ryan McDonald, and Joakim Nivre: 'Dependency Parsing', *Synthesis Lectures on Human Language Technologies* 1 (1): 1–127, 2009.

Kostiantyn Kucher and Andreas Kerren: 'Text Visualization Techniques: Taxonomy, Visual Survey, and Community Insights', in: *Pacific Visualization Symposium (PacificVis)*, pp. 117–121, IEEE, 2015.

Maciej Kula: 'Metadata Embeddings for User and Item Cold-start Recommendations', in: *Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems (held in conjunction with ACM Conference on Recommender Systems)*, Vol. 1448, pp. 14–21, CEUR Workshop Proceedings, Vienna, Austria, September 2015.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman: 'Building Machines that Learn and Think Like People', *Behavioral and Brain Sciences* 40, 2017.

George Lakoff and Mark Johnson: *Metaphors We Live By*, University of Chicago Press, 1980.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni: 'Is This a Wampimuk? Cross-Modal Mapping Between Distributional Semantics and the Sisual World', in: *Proceedings of the 52nd Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 1403–1414, Association for Computational Linguistics, Baltimore, Maryland, USA, June 2014.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni: 'Multi-Agent Cooperation and the Emergence of (Natural) Language', in: *International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

Angeliki Lazaridou, Nghia T. Pham, and Marco Baroni: 'Combining Language and Vision with a Multimodal Skip-Gram Model', in: *Proceedings of the 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 153–163, Association for Computational Linguistics, 2015.

Quoc Le and Tomas Mikolov: 'Distributed Representations of Sentences and Documents', in: *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pp. 1188–1196, 2014.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner: 'Gradient-Based Learning Applied to Document Recognition', *Proceedings of the IEEE* 86 (11): 2278–2324, 1998.

Chee W. Leong and Rada Mihalcea: 'Going Beyond Text: A Hybrid Image-Text Approach for Measuring Word Relatedness', in: *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 1403–1407, 2011.

Omer Levy and Yoav Goldberg: 'Dependency-Based Word Embeddings', in: *Proceedings of the 52nd Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 302–308, The Association for Computer Linguistics, Baltimore, Maryland, USA, June 2014a.

Omer Levy and Yoav Goldberg: 'Neural Word Embedding as Implicit Matrix Factorization', in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 2177–2185, 2014b.

Omer Levy, Yoav Goldberg, and Ido Dagan: 'Improving Distributional Similarity With Lessons Learned from Word Embeddings', *Transactions of the Association for Computational Linguistics (TACL)* 3: 211–225, 2015a.

Omer Levy, Steffen Remus, Chris Biemann, Ido Dagan, and Israel Ramat-Gan: 'Do Supervised Distributional Methods Really Learn Lexical Inference Relations?', in: *Proceedings of the 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 970–976, 2015b.

Yi Li, Timothy M. Hospedales, Yi-Zhe Song, and Shaogang Gong: 'Free-Hand Sketch Recognition by Multi-Kernel Feature Learning', *Computer Vision and Image Understanding* 137: 1–11, 2015.

Percy Liang: 'Learning Executable Semantic Parsers for Natural Language Understanding', *Communications of the ACM* 59 (9): 68–76, 2016.

Chin-Yew Lin: 'Rouge: A Package for Automatic Evaluation of Summaries', in: *Proceedings of the Workshop on Text Summarization Branches Out (held in conjunction with ACL)*, pp. 74–81, The Association for Computer Linguistics, Barcelona, Spain, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick: 'Microsoft COCO: Common Objects in Context', in: *European conference on computer vision*, pp. 740–755, Springer-Verlag, 2014.

Yankai Lin, Zhiyuan Liu, Huan-Bo Luan, Maosong Sun, Siwei Rao, and Song Liu: 'Modeling Relation Paths for Representation Learning of Knowledge Bases', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 705–714, Association for Computational Linguistics, Lisbon, Portugal, September 2015a.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu: 'Learning Entity and Relation Embeddings for Knowledge Graph Completion', in: *Proceedings of the 29th Conference of Association for the Advancement of Artificial Intelligence (AAAI)*, AAAI Press, 2015b.

Huan Ling and Sanja Fidler: 'Teaching Machines to Describe Images via Natural Language Feedback', in: *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Chi-kiu Lo, Karteek Addanki, Markus Saers, and Dekai Wu: 'Improving Machine Translation by Training Against an Automatic Semantic Frame-Based Evaluation Metric', in: *Proceedings of the 51st Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 375–381, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

John L. Locke and Barry Bogin: 'Language and Life History: A New Perspective on the Development and Evolution of Human Language', *Behavioral and Brain Sciences* 29 (3): 259–280, 2006.

Max M. Louwerse: 'Symbol Interdependency in Symbolic and Embodied Cognition', *Topics in Cognitive Science* 3 (2): 273–302, 2011.

Will Lowe: 'Towards a Theory of Semantic Space', in: *In Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Vol. 23, pp. 576–581, 2001.

Xuezhe Ma and Eduard Hovy: 'End-to-End Sequence Labeling Via Bi-Directional LSTM-CNNs-CRF', in: *Proceedings of the 54th Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 1064–1074, Association for Computational Linguistics, Berlin, Germany, August 2016.

Laurens van der Maaten and Geoffrey Hinton: 'Visualizing Data Using t-SNE', *Journal of Machine Learning Research* 9 (Nov): 2579–2605, 2008.

David MacKay: *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz: 'Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images', in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, IEEE, December 2015.

Arun Mallya and Svetlana Lazebnik: 'Recurrent Models for Situation Recognition', in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 455–463, IEEE, 2017.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze: *Introduction to Information Retrieval*, Cambridge University Press, New York, USA, 2008.

Gary Marcus: 'Deep Learning: A Critical Appraisal', *arXiv preprint arXiv:1801.00631* 2018a.

Gary Marcus: 'Innateness, AlphaZero, and Artificial Intelligence', *arXiv preprint arXiv:1801.05667* 2018b.

André Markard: *Deep Learning for Extraction of Predicate-Argument Structures*, Bachelor thesis, Technische Universität Darmstadt, August 2018.

Teresa Martin, Fiete Botschen, Ajay Nagesh, and Andrew McCallum: 'Call for Discussion: Building a New Standard Dataset for Relation Extraction Tasks', in: *Proceedings of the 5th Workshop on Automated Knowledge Base Construction (AKBC, held in conjunction with NAACL)*, pp. 92–96, San Diego, California, USA, June 2016.

James L. McClelland, Kevin Mickey, Steven Hansen, Arianna Yuan, and Qihong Lu: 'A Parallel-Distributed Processing Approach to Mathematical Cognition', *Manuscript, Stanford University* 2016.

Warren S. McCulloch and Walter Pitts: 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *The bulletin of mathematical biophysics* 5 (4): 115–133, 1943.

Rada Mihalcea: 'Using Wikipedia for Automatic Word Sense Disambiguation', in: *Proceedings of the 9th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 196–203, 2007.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean: 'Efficient Estimation of Word Representations in Vector Space', *arXiv preprint arXiv:1301.3781* 2013a.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig: 'Linguistic Regularities in Continuous Space Word Representations', in: *Proceedings of the 13th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, Vol. 13, pp. 746–751, Association for Computational Linguistics, May 2013b.

George A. Miller: 'Nouns in WordNet: A Lexical Inheritance System', *International Journal of Lexicography* 3 (4): 245–264, 1990.

George A. Miller: 'WordNet: A Lexical Database for English', *Communications of the ACM* 38 (11): 39–41, 1995.

Marvin Minsky: 'A Framework for Representing Knowledge', in: *Readings in Cognitive Science*, pp. 156–189, Elsevier, 1988.

Makoto Miwa and Mohit Bansal: 'End-to-End Relation Extraction Using LSTMs on Sequences and Tree Structures', in: *Proceedings of the 54th Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 1105–1116, Association for Computational Linguistics, Berlin, Germany, August 2016.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho: 'Multimodal Named Entity Recognition for Short Social Media Posts', in: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 852–860, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi: 'Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web', in: *Proceedings of the 13th international conference on multimodal interfaces*, pp. 169–176, ACM Press, 2011.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen: 'A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories', in: *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 839–849, Association for Computational Linguistics, San Diego, California, USA, June 2016.

Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth: 'A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning', in: *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM, held in conjunction with NAACL)*, pp. 225–234, Association for Computational Linguistics, New Orleans, Louisiana, USA, June 2018.

Hatem Mousselly-Sergieh and Iryna Gurevych: 'Enriching Wikidata with Frame Semantics', in: *Proceedings of the 5th Workshop on Automated Knowledge Base Construction (AKBC, held in conjunction with NAACL)*, pp. 29–34, 2016.

Christof Müller, Iryna Gurevych, and Max Mühlhäuser: 'Closing the Vocabulary Gap for Computing Text Similarity and Information Retrieval', *International Journal of Semantic Computing* 2 (02): 253–272, 2008.

Vinod Nair and Geoffrey E. Hinton: 'Rectified Linear Units Improve Restricted Boltzmann Machines', in: *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 807–814, Haifa, Israel, 2010.

Roberto Navigli: 'Word Sense Disambiguation: A Survey', *ACM computing surveys (CSUR)* 41 (2): 10, 2009.

Roberto Navigli and Simone P. Ponzetto: 'BabelNet: Building a Very Large Multilingual Semantic Network', in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 216–225, Association for Computational Linguistics, Uppsala, Sweden, July 2010.

Roberto Navigli and Simone P. Ponzetto: 'BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network', *Artificial Intelligence* 193: 217–250, 2012a.

Roberto Navigli and Simone P. Ponzetto: 'Multilingual WSD With Just a Few Lines of Code: The BabelNet API', in: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 67–72, Association for Computational Linguistics, Jeju, Republic of Korea, July 2012b.

Ani Nenkova and Kathleen McKeown: 'Automatic summarization', *Foundations and Trends in Information Retrieval* 5 (2–3): 103–233, 2011.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng: 'Multimodal Deep Learning', in: *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 689–696, 2011.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova: 'An Assessment of the Accuracy of Automatic Evaluation in Summarization', in: *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pp. 1–9, Association for Computational Linguistics, Montréal, Canada, 2012.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi: 'Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features', in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 528–540, Association for Computational Linguistics, Melbourne, Australia, July 2018.

Alexis Palmer and Caroline Sporleder: 'Evaluating FrameNet-Style Semantic Parsing: The Role of Coverage Gaps in FrameNet', in: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 928–936, Beijing, China, 2010.

Martha Palmer, Daniel Gildea, and Paul Kingsbury: 'The Proposition Bank: An Annotated Corpus of Semantic Roles', *Computational linguistics* 31 (1): 71–106, 2005.

Sinno Jialin Pan and Qiang Yang: 'A Survey on Transfer Learning', *IEEE Transactions on knowledge and data engineering* 22 (10): 1345–1359, 2010.

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme: 'FrameNet+: Fast Paraphrastic Tripling of FrameNet', in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing*, pp. 408–413, Association for Computational Linguistics, Beijing, China, 2015.

Hao Peng, Sam Thomson, Swabha Swayamdipta, and Noah A. Smith: 'Learning Joint Semantic Parsers from Disjoint Data', *arXiv preprint arXiv:1804.05990* 2018.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning: 'GloVe: Global Vectors for Word Representation', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Association for Computational Linguistics, Doha, Qatar, October 2014.

Wiebe R. Pestman: *Mathematical Statistics: An Introduction*, Mouton de Gruyter, Berlin, Germany, 1st edition, 1998.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer: 'Deep Contextualized Word Representations', in: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 2227–2237, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych: 'Learning to Score System Summaries for Better Content Selection Evaluation', in: *Proceedings of the Workshop "New Frontiers in Summarization" (held in conjunction with EMNLP)*, pp. 74–84, Association for Computational Linguistics, Copenhagen, Denmark, September 2017.

David Premack: 'Is Language the Key to Human Intelligence?', *Science* 303 (5656): 318–320, 2004.

Friedemann Pulvermüller: 'Brain Mechanisms Linking Language and Action', *Nature Reviews Neuroscience* 6 (7): 576–582, 2005.

Friedemann Pulvermüller, Olaf Hauk, Vadim V. Nikulin, and Risto J. Ilmoniemi: 'Functional Links Between Motor and Language Systems', *European Journal of Neuroscience* 21 (3): 793–797, 2005.

Pranav Rajpurkar, Robin Jia, and Percy Liang: 'Know What You Don't Know: Unanswerable Questions for SQuAD', in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 784–789, Association for Computational Linguistics, Melbourne, Australia, July 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang: 'SQuAD: 100,000+ Questions for Machine Comprehension of Text', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, Association for Computational Linguistics, Austin, Texas, USA, November 2016.

Pushpendre Rastogi and Benjamin Van Durme: 'Augmenting FrameNet Via PPDB', in: *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pp. 1–5, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal: 'Grounding Action Descriptions in Videos', *Transactions of the Association of Computational Linguistics (TACL)* 1: 25–36, 2013.

Ines Rehbein, Josef Ruppenhofer, Caroline Sporleder, and Manfred Pinkal: 'Adding Nominal Spice to SALSA – Frame-Semantic Annotation of German Nouns and Verbs', in: *Workshop Proceedings of the 11th Conference on Natural Language Processing (KONVENS)*, pp. 89–97, Vienna, Austria, September 2012.

Radim Řehůřek and Petr Sojka: 'Software Framework for Topic Modelling with Large Corpora', in: *Proceedings of the Workshop on New Challenges for NLP Frameworks (held in conjunction with LREC)*, pp. 45–50, Valletta, Malta, May 2010.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych: 'GermEval-2014: Nested Named Entity Recognition with Neural Networks', in: *Workshop Proceedings of the 12th Conference on Natural Language Processing (KONVENS)*, pp. 117–120, Hildesheim, Germany, 2014.

Nils Reimers and Iryna Gurevych: 'Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 338–348, Association for Computational Linguistics, Copenhagen, Denmark, September 2017.

Sebastian Riedel, Limin Yao, and Andrew McCallum: 'Modeling Relations and Their Mentions without Labeled Text', in: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 148–163, Barcelona, Spain, 2010.

Matteo R. Ronchi and Pietro Perona: 'Describing Common Human Visual Actions in Images', in: *Proceedings of the British Machine Vision Conference (BMVC)*, BMVA Press, Swansea, Wales, 2015.

Frank Rosenblatt: 'The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain', *Psychological Review* 65 (6): 65–386, 1958.

Michael Roth and Mirella Lapata: 'Context-Aware Frame-Semantic Role Labeling', *Transactions of the Association for Computational Linguistics (TACL)* 3: 449–460, 2015.

Michael Roth and Mirella Lapata: 'Neural Semantic Role Labeling with Dependency Path Embeddings', in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1192–1202, Association for Computational Linguistics, Berlin, Germany, August 2016.

Sascha Rothe and Hinrich Schütze: 'AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes', in: *Proceedings of the 52nd Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 1793–1803, Association for Computational Linguistics, Beijing, China, July 2015.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams: 'Learning Representations by Back-Propagating Errors', *Nature* 323: 533–536, 1986.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk: *FrameNet II: Extended Theory and Practice*, International Computer Science Institute, Berkeley, USA, revised november 1, 2016 edition, 2016.

Stuart J. Russell and Peter Norvig: *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.

Paul Ruvolo and Eric Eaton: 'Active Task Selection for Lifelong Machine Learning', in: *Proceedings of the 27th Conference of Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 862–868, AAAI Press, 2013a.

Paul Ruvolo and Eric Eaton: 'ELLA: An Efficient Lifelong Learning Algorithm', in: *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Vol. 28, pp. 507–515, 2013b.

Pedro B. Santos, Lisa Beinborn, and Iryna Gurevych: 'A Domain-Agnostic Approach for Opinion Prediction on Speech', in Malvina Nissim, Viviana Patti, and Barbara Plank (Eds.): *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pp. 163–172, Osaka, Japan, 2016.

Roger C. Schank: 'Conceptual Dependency: A Theory of Natural Language Understanding', *Cognitive psychology* 3 (4): 552–631, 1972.

Roger C. Schank and Robert P. Abelson: *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*, Psychology Press, 2013.

Lenhart K. Schubert: 'Semantic Representation', in: *Proceedings of the 29th Conference of Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 4132–4139, AAAI Press, 2015.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard: 'Black Holes and White Rabbits: Metaphor Identification with Visual Features', in: *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 160—-170, Association for Computational Linguistics, San Diego, California, USA, 2016.

Ekaterina Shutova, Andreas Wundsam, and Helen Yannakoudakis: 'Semantic Frames and Visual Scenes: Learning Semantic Role Inventories from Image and Video Descriptions', in: *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM, held in conjunction with ACL)*, pp. 149–154, Association for Computational Linguistics, 2017.

David M. Sidhu, Rachel Kwan, Penny M. Pexman, and Paul D. Siakaluk: 'Effects of Relative Embodiment in Lexical and Semantic Processing of Verbs', *Acta psychologica* 149: 32–39, 2014.

Jennifer Sikos and Sebastian Padó: 'FrameNet's Using Relation as a Source of Concept-Based Paraphrases', *Constructions and Frames* 10 (1): 38–60, 2018a.

Jennifer Sikos and Sebastian Padó: 'Using Embeddings to Compare FrameNet Frames Across Languages', in: *Proceedings of the 1st Workshop on Linguistic Resources for Natural Language Processing (held in conjunction with ACL)*, pp. 91–101, 2018b.

Carina Silberer and Mirella Lapata: 'Grounded Models of Semantic Representation', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1423–1433, Association for Computational Linguistics, Jeju Island, Korea, July 2012.

Carina Silberer and Mirella Lapata: 'Learning Grounded Meaning Representations with Autoencoders', in: *Proceedings of the 52nd Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 721–732, Association for Computational Linguistics, Baltimore, Maryland, USA, June 2014.

Daniel L. Silver, Qiang Yang, and Lianghao Li: 'Lifelong Machine Learning Systems: Beyond Learning Algorithms', in: *AAAI Spring Symposium: Lifelong Machine Learning*, pp. 49–55, Citeseer, 2013.

Karen Simonyan and Andrew Zisserman: 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *arXiv preprint arXiv:1409.1556* 2014.

Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Ng: 'Zero-Shot Learning through Cross-Modal Transfer', in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 935–943, 2013.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y Ng: 'Grounded Compositional Semantics for Finding and Describing Images with Sentences', *Transactions of the Association for Computational Linguistics (TACL)* 2 (1): 207–218, 2014.

Daniil Sorokin and Iryna Gurevych: 'Mixing Context Granularities for Improved Entity Linking on Question Answering Data across Entity Categories', in: *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM, held in conjunction with NAACL)*, pp. 65–75, Association for Computational Linguistics, New Orleans, Louisiana, USA, June 2018a.

Daniil Sorokin and Iryna Gurevych: 'Modeling Semantics with Gated Graph Neural Networks for Knowledge Base Question Answering', in: *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pp. 3306–3317, Association for Computational Linguistics, Santa Fe, New Mexico, USA, August 2018b.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov: 'Dropout: A Simple Way to Prevent Neural Networks from Overfitting', *Journal of Machine Learning Research* 15: 1929–1958, 2014.

Nitish Srivastava and Ruslan Salakhutdinov: 'Learning Representations for Multimodal Data with Deep Belief Nets', in: *Representation Learning Workshop (held in conjunction with ICML)*, Vol. 79, 2012.

Mark Steedman: 'Combinatory Grammars and Parasitic Gaps', *Natural Language & Linguistic Theory* 5 (3): 403–439, 1987.

Emma Strubell and Andrew McCallum: 'Syntax Helps ELMo Understand Semantics: Is Syntax Still Relevant in a Deep Neural Architecture for SRL?', in: *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP (RelNLP, held in conjunction with ACL)*, pp. 19–27, Association for Computational Linguistics, 2018.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum: 'Linguistically-Informed Self-Attention for Semantic Role Labeling', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5027–5038, Association for Computational Linguistics, Brussels, Belgium, November 2018.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza: 'Learning to Rank Answers to Non-Factoid Questions from Web Collections', *Computational linguistics* 37 (2): 351–383, 2011.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith: 'Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold', *arXiv preprint arXiv:1706.09528* 2017.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith: 'Syntactic Scaffolds for Semantic Structures', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3772–3782, Association for Computational Linguistics, Brussels, Belgium, November 2018.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi: 'Inception-V4, Inception-Resnet and the Impact of Residual Connections on Learning', in: *Proceedings of the 31st Conference of Association for the Advancement of Artificial Intelligence (AAAI)*, Vol. 4, p. 12, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich: 'Going Deeper with Convolutions', in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, IEEE, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna: 'Rethinking the Inception Architecture for Computer Vision', in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, IEEE, 2016.

Ming Tan, Bing Xiang, and Bowen Zhou: 'LSTM-Based Deep Learning Models for Non-Factoid Answer Selection', *arXiv preprint arXiv:1511.04108* 2015.

Sebastian Thrun: 'Is Learning the n-th Thing Any Easier than Learning the First?', in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 640–646, 1996.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu: 'ECNU at SemEval-2017 Task 1: Leverage Kernel-based Traditional NLP features and Neural Networks to Build a Universal Model for Multilingual and Cross-Lingual Semantic Textual Similarity', in: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval, held in conjunction with ACL)*, pp. 191–197, Association for Computational Linguistics, Vancouver, Canada, August 2017.

Marc Toussaint, Kelsey R. Allen, Kevin A. Smith, and Joshua B. Tenenbaum: 'Differentiable Physics and Stable Modes for Tool-Use and Manipulation Planning', in: *Proceedings of the 14th Robotics: Science and Systems conference (RSS)*, Pittsburgh, Pennsylvania, USA, June 2018.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon: 'Representing Text for Joint Embedding of Text and Knowledge Bases', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1499–1509, Association for Computational Linguistics, Lisbon, Portugal, September 2015.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen: 'Literal and Metaphorical Sense Identification through Concrete and Abstract Context', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 680–690, Association for Computational Linguistics, Edinburgh, UK, July 2011.

Ramakrishna Vedantam, Lawrence C. Zitnick, and Devi Parikh: 'CIDEr: Consensus-Based Image Description Evaluation', in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, IEEE, 2015.

Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville: 'Modulating Early Visual Processing by Language', in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 6597–6607, 2017.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo: 'Knowledge Graph Embedding: A Survey of Approaches and Applications', *IEEE Transactions on Knowledge and Data Engineering* 29 (12): 2724–2743, 2017.

Sai Wang, Ru Li, Ruibo Wang, Zhiqiang Wang, and Xia Zhang: 'SXUCFN-Core: STS Models Integrating FrameNet Parsing Information', in: *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM, held in conjunction with NAACL)*, pp. 74–79, Association for Computational Linguistics, Atlanta, Georgia, USA, June 2013.

Shuohang Wang and Jing Jiang: 'Machine Comprehension Using Match-LSTM and Answer Pointer', *arXiv preprint arXiv:1608.07905* 2016.

Su Wang, Greg Durrett, and Katrin Erk: 'Modeling Semantic Plausibility by Injecting World Knowledge', in: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 303–308, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen: 'Knowledge Graph Embedding by Translating on Hyperplanes', in: *Proceedings of the 28th Conference of Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 1112–1119, AAAI Press, 2014.

Jason Weston, Samy Bengio, and Nicolas Usunier: 'WSABIE: Scaling Up to Large Vocabulary Image Annotation', in: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2764–2770, AAAI Press, Barcelona, Catalonia, Spain, 2011.

Jason Weston, Sumit Chopra, and Antoine Bordes: 'Memory Networks', in: *International Conference on Learning Representations (ICLR)*, San Diego, California, USA, 2015.

Olivia Winn and Smaranda Muresan: ''Lighter' Can Still Be Dark: Modeling Comparative Color Descriptions', in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 790–795, Association for Computational Linguistics, Melbourne, Australia, July 2018.

Ludwig Wittgenstein: 'Tractatus Logico-Philosophicus', *Bertrand Russell, translated by Charles K. Ogden* 1922.

Ludwig Wittgenstein: 'Philosophical Investigations', *London, Basil Blackwell, translated by Gertrude E.M. Anscombe* 1953.

Hao Wu, Heyan Huang, Ping Jian, Yuhang Guo, and Chao Su: 'BIT at SemEval-2017 Task 1: Using Semantic Information Space to Evaluate Semantic Textual Similarity', in: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval, held in conjunction with ACL)*, pp. 77–84, Association for Computational Linguistics, Vancouver, Canada, August 2017a.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel: 'Visual Question Answering: A Survey of Methods and Datasets', *Computer Vision and Image Understanding* 163: 21–40, 2017b.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun: 'Image-Embodied Knowledge Representation Learning', in: *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3140–3146, AAAI Press, 2017.

Jin Xu, Yubo Tao, and Hai Lin: 'Semantic Word Cloud Generation Based on Word Embeddings', in: *Pacific Visualization Symposium*, pp. 239–243, IEEE, Taipei, Taiwan, April 2016.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio: 'Show, Attend and Tell: Neural Image Caption Generation with Visual Attention', in: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2048–2057, Lille, France, 2015.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Y Chai: 'Grounded Semantic Role Labeling', in: *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 149–159, San Diego, California, USA, 2016.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi: 'Situation Recognition: Visual Semantic Role Labeling for Image Understanding', in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5534–5542, IEEE, 2016.

Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li: 'Neural Generative Question Answering', in: *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 36–42, Association for Computational Linguistics, San Diego, California, USA, June 2016.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier: 'From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions', *Transactions of the Association for Computational Linguistics (TACL)* 2: 67–78, 2014.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria: 'Recent Trends in Deep Learning Based Natural Language Processing', *IEEE Computational Intelligence Magazine* 13 (3): 55–75, 2018.

Ines Zelch: *Frame Semantic Based Approach for Semantic Textual Similarity*, Bachelor thesis, Technische Universität Darmstadt, June 2018.

Rowan Zellers and Yejin Choi: 'Zero-Shot Activity Recognition with Verb Attribute Induction', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 946–958, Association for Computational Linguistics, Copenhagen, Denmark, September 2017.

Heiga Zen, Keiichi Tokuda, and Alan W. Black: 'Statistical Parametric Speech Synthesis', *Speech Communication* 51 (11): 1039–1064, 2009.

Michael Zhai, Johnny Tan, and Jinho D. Choi: 'Intrinsic and Extrinsic Evaluations of Word Embeddings', in: *Proceedings of the 30th Conference of Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 4282–4283, AAAI Press, 2016.

Jie Zhou and Wei Xu: 'End-to-End Learning of Semantic Role Labeling Using Recurrent Neural Networks', in: *Proceedings of the 53rd Annual Conference of the Association for Computational Linguistics (ACL)*, pp. 1127–1137, Association for Computational Linguistics, Beijing, China, July 2015.

Markus Zopf, Teresa Botschen, Tobias Falke, Benjamin Heinzerling, Ana Marasović, Todor Mihaylov, Avinesh P.V.S, Eneldo Loza Mencía, Johannes Fürnkranz, and

Anette Frank: 'What's Important in a Text? An Extensive Evaluation of Linguistic Annotations for Summarization', in: *Proceedings of the 5th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Valencia, Spain, October 2018a.

Markus Zopf, Eneldo Loza Mencía, and Johannes Fürnkranz: 'Beyond Centrality and Structural Features: Learning Information Importance for Text Summarization', in: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pp. 84–94, Berlin, Germany, August 2016a.

Markus Zopf, Eneldo Loza Mencía, and Johannes Fürnkranz: 'Which Scores to Predict in Sentence Regression for Text Summarization?', in: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, Vol. 1, pp. 1782–1791, New Orleans, Louisiana, USA, June 2018b.

Markus Zopf, Maxime Peyrard, and Judith Eckle-Kohler: 'The Next Step for Multi-Document Summarization: A Heterogeneous Multi-Genre Corpus Built with a Novel Construction Approach', in: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pp. 1535–1545, Association for Computational Linguistics, Osaka, Japan, December 2016b.

Krunoslav Zubrinic, Damir Kalpic, and Mario Milicevic: 'The Automatic Creation of Concept Maps from Documents Written Using Morphologically Rich Languages', *Expert Systems with Applications* 39 (16): 12709–12718, 2012.

Rolf A. Zwaan and Carol J. Madden: 'Embodied Sentence Comprehension', *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*, pp. 224–245, 2005.

Ehrenwörtliche Erklärung[1]

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades "Doktor der Naturwissenschaften (Dr. rer. nat.)" mit dem Titel "Uni- and Multimodal and Structured Representations for Modeling Frame Semantics" selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 29. November 2018

Teresa Isabel Botschen (geb. Martin), M.Sc.

[1] Gemäß §9 Abs. 1 der Promotionsordnung der TU Darmstadt

Wissenschaftlicher Werdegang der Verfasserin[2]

09/10 – 07/11	Studium Generale am Leibniz Kolleg Tübingen
10/11 – 08/14	Studium der Kognitionswissenschaft, Bachelor of Science, an der Eberhard Karls Universität Tübingen, Bachelorarbeit *'Cognitive Simulation Theory: A Study on the Association between Language and Experience'*
09/14 – 09/15	Studium von Computational Statistics and Machine Learning, Master of Science, am University College London, Masterarbeit *'Automatic Recognition of Protective Behaviour in Body Movement of Chronic Back Pain Patients'*
02/16 – 04/16	Gastwissenschaftlerin am Fachgebiet Information Extraction and Synthesis (IESL-Lab) der University of Massachussets Amherst
10/15 – 12/18	Doktorandin im Graduiertenkolleg Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) und am Fachgebiet Ubiquitous Knowledge Processing (UKP-Lab) der Technischen Universität Darmstadt

[2] Gemäß §20 Abs. 3 der Promotionsordnung der TU Darmstadt

Publikationsverzeichnis der Verfasserin

Teresa Botschen*, Daniil Sorokin*, and Iryna Gurevych: 'Frame- and Entity-Based Knowledge for Common-Sense Argumentative Reasoning', in: *Proceedings of the 5th International Workshop on Argument Mining (ArgMin, held in conjunction with EMNLP)*, pp. 90–96, Brussels, Belgium, November 2018. (* equal contribution)

Markus Zopf, **Teresa Botschen**, Tobias Falke, Benjamin Heinzerling, Ana Marasović, Todor Mihaylov, Avinesh P.V.S, Eneldo Loza Mencía, Johannes Fürnkranz, and Anette Frank: 'What's important in a text? An extensive evaluation of linguistic annotations for summarization.', in: *Proceedings of the 5th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Valencia, Spain, October 2018.

Lisa Beinborn*, **Teresa Botschen***, and Iryna Gurevych: 'Multimodal Grounding for Language Processing', in: *Proceedings of the 27th International Conference on Computational Linguistics: Technical Papers (COLING)*, pp. 2325–2339, Santa Fe, USA, August 2018. (* equal contribution)

Teresa Botschen, Iryna Gurevych, Jan-Christoph Klie, Hatem Mousselly-Sergieh, and Stefan Roth: 'Multimodal Frame Identification with Multilingual Evaluation', in: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 1481–1491, New Orleans, USA, June 2018.

Hatem Mousselly-Sergieh, **Teresa Botschen**, Iryna Gurevych, and Stefan Roth: 'A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning', in: *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (StarSem, held in conjunction with NAACL)*, pp. 225–234, New Orleans, USA, June 2018.

Teresa Botschen, Hatem Mousselly-Sergieh, and Iryna Gurevych: 'Experimental study of multimodal representations for Frame Identification - How to find the right multimodal representations for this task?', in: *Language-Learning-Logic Workshop (3L)*, London, UK, September 2017.

Maxime Peyrard, **Teresa Botschen**, and Iryna Gurevych: 'Learning to Score System Summaries for Better Content Selection Evaluation', in: *Proceedings of the Workshop "New Frontiers in Summarization" (held in conjunction with EMNLP)*, pp. 74–84, Copenhagen, Denmark, September 2017.

Teresa Botschen, Hatem Mousselly-Sergieh, and Iryna Gurevych: 'Prediction of Frame-to-Frame Relations in the FrameNet Hierarchy with Frame Embeddings', in: *Proceedings of the 2nd Workshop on Representation Learning for NLP (RepL4NLP, held in conjunction with ACL)*, pp. 146–156, Vancouver, Canada, August 2017.

Silvana Hartmann, Ilia Kuznetsov, **Teresa Martin**, and Iryna Gurevych: 'Out-of-domain FrameNet Semantic Role Labeling', in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 471–482, Valencia, Spain, April 2017.

Michael Bugert, Yevgeniy Puzikov, Andreas Rücklé, Judith Eckle-Kohler, **Teresa Martin**, Eugenio Martínez Cámara, Daniil Sorokin, Maxime Peyrard, and Iryna Gurevych: 'LSDSem 2017: Exploring Data Generation Methods for the Story Cloze Test', in: *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem, held in conjunction with EACL)*, pp. 56–61, Valencia, Spain, April 2017.

Teresa Martin, Fiete Botschen, Ajay Nagesh, and Andrew McCallum: 'Call for Discussion: Building a New Standard Dataset for Relation Extraction Tasks', in: *Proceedings of the 5th Workshop on Automated Knowledge Base Construction (AKBC, held in conjunction with NAACL)*, pp. 92–96, San Diego, USA, June 2016.

Nicholas Dingwall, Alan Chalk, **Teresa Martin**, Catherine Scott, Carla Semedo, Quan Le, Eliza Orasanu, Jorge Cardoso, Andrew Melbourne, and Neil Marlow: 'T2 relaxometry in the extremely-preterm brain at adolescence', in: *Journal of Magnetic Resonance Imaging*, pp. 508–514, May 2016.

Julia Bahnmüller, Florian Faehling, Daniel Markus, **Teresa Martin**, and Markus Stöger: 'Sprache und Freiheit. Sprachfähigkeit als Grundvoraussetzung zur Selbstbestimmung des Willens', in: *Sprache und Kognition: Theory of Mind, Emergenz, Neue Medien, Freiheit, Grenzen. Interdisziplinäre Forschungsarbeiten am Forum Scientiarum, Bd. 7, hrsg. von Niels Weidtmann*, pp. 163–199, November 2015.

Anmerkungen zum Umgang mit Forschungsdaten

Gemäß der "Leitlinien zum Umgang mit Forschungsdaten" der Deutschen Forschungsgemeinschaft[3] wurden alle im Zusammenhang mit dieser Dissertation entstandenen Forschungsdaten langfristig archiviert und sofern möglich öffentlich zugänglich gemacht. Folgende Forschungsdaten wurden frei verfügbar gemacht:

- Software
  - Das in Abschnitt 4.1.3 beschriebene `SimpleFrameId`-System steht unter der Apache-Lizenz 2.0 unter `https://github.com/UKPLab/eacl2017-oodFrameNetSRL` zur Verfügung.
  - Das in Abschnitt 4.1 beschriebene `UniFrameId`-System und das in Abschnitt 4.2 beschriebene `MultiFrameId`-System stehen unter der Apache-Lizenz 2.0 unter `https://github.com/UKPLab/naacl18-multimodal-frame-identification` zur Verfügung.
  - Die für die in Abschnitt 4.2.4.1 beschriebenen Experimente notwendige Software steht unter der Apache-Lizenz 2.0 unter `https://github.com/UKPLab/coling2018-multimodalSurvey` zur Verfügung.
  - Die für die in Abschnitt 5.2.3.1 beschriebenen Experimente notwendige Software steht unter der Apache-Lizenz 2.0 unter `https://github.com/UKPLab/starsem18-multimodalKB` zur Verfügung.
  - Die für die in Abschnitt 6.2.2 beschriebenen Experimente notwendige Software steht unter der Apache-Lizenz 2.0 unter `https://github.com/UKPLab/emnlp2018-argmin-commonsense-knowledge` zur Verfügung.
- Embeddings
  - Die in Abschnitt 4.2 beschriebenen visuellen und imaginierten Vektorrepräsentationen stehen unter der MIT-Lizenz unter `https://public.ukp.informatik.tu-darmstadt.de/naacl18-multimodal-frame-identification/` zur Verfügung.
  - Die in Abschnitt 4.2.4.1 beschriebenen textuellen und visuellen Vektorrepräsentationen stehen unter der MIT-Lizenz unter `https://public.ukp.informatik.tu-darmstadt.de/coling18-multimodalSurvey/` zur Verfügung.
  - Die in den Abschnitten 5.1 und 5.2 beschriebenen textuellen und strukturierten Vektorrepräsentationen stehen unter der Creative Commons 3.0-Lizenz unter `https://public.ukp.informatik.tu-darmstadt.de/repl4nlp17-frameEmbeddings/` zur Verfügung.

[3] `http://dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf`

  - Die in Abschnitt 5.2.3.1 beschriebenen textuellen, strukturierten und visuellen Vektorrepräsentationen stehen unter der MIT-Lizenz unter `https://public.ukp.informatik.tu-darmstadt.de/starsem18-multimodalKB/WN9-IMG/` zur Verfügung.

- Erweiterungen von Korpora
  - Die in Abschnitt 4.1.4.1 beschriebene Einteilung des SALSA Korpus in Training-, Developmet- und Test-Teil steht unter der MIT-Lizenz unter `https://public.ukp.informatik.tu-darmstadt.de/naacl18-multimodal-frame-identification/` zur Verfügung.

- Forschungsergebnisse
  - Alle im Zusammenhang mit dieser Dissertation stehenden Publikationen sind in der ACL Anthology (`https://aclanthology.coli.uni-saarland.de/`) verfügbar.
  - Alle Forschungsergebnisse sind zudem auch in dieser Dissertation selbst dokumentiert, die von der Universitäts- und Landesbibliothek Darmstadt zur Verfügung gestellt wird.