Frame Semantics Annotation Made Easy with DBpedia

Marco Fossati, Sara Tonelli, and Claudio Giuliano
{fossati, satonelli, giuliano}@fbk.eu

Fondazione Bruno Kessler - via Sommarive, 18 - 38123 Trento, Italy

Abstract. Crowdsourcing techniques applied to natural language processing have recently experienced a steady growth and represent a cheap and fast, albeit valid, solution to create benchmarks and training data. Nevertheless, some particularly complex tasks such as semantic role annotation have been rarely conducted in a crowdsourcing environment, due to their intrinsic difficulty. In this paper, we present a novel approach to accomplish this task by leveraging information automatically extracted from DBpedia. We show that replacing role definitions, typically meant for expert annotators, with a list of DBpedia types, makes the identification and assignment of role labels more intuitive also for non-expert workers. Results prove that such strategy improves on the standard annotation workflow, both in terms of accuracy and of time consumption.

Keywords: Natural Language Processing, Frame Semantics, Entity Linking, DBpedia, Crowdsourcing, Task Modeling

1 Introduction

Frame semantics [6] is one of the theories that originate from the long strand of linguistic research in artificial intelligence. A *frame* can be informally defined as an event triggered by some term in a text and embedding a set of participants. For instance, the sentence `Goofy has murdered Mickey Mouse` evokes the KILLING frame (triggered by `murdered`) together with the Killer and Victim participants (respectively `Goofy` and `Mickey Mouse`). Such theory has led to the creation of FrameNet [2], namely an English lexical database containing manually annotated textual examples of frame usage.

Annotating frame information is a complex task, usually modeled in two steps: given a sentence, annotators are first asked to choose the frame activated by a predicate (or *lexical unit, LU*, e.g. `murdered` in the example above evoking KILLING). Second, they assign the semantic roles (or *frame elements, FEs*) that describe the participants involved in the chosen frame. In this work, we focus on the second step, namely FEs recognition.

Currently, FrameNet development follows a strict protocol for data annotation and quality control. The entire procedure is known to be both time-consuming and costly, thus representing a burden for the extension of the resource [1]. Furthermore, deep linguistic knowledge is needed to tackle this annotation task, and the resource developed so far would not have come to light

without the contribution of skilled linguists and lexicographers. On one hand, the task complexity depends on the inherently complex theory behind frame semantics, with a repository of thousands of roles available for the assignment. On the other hand, these roles are defined for expert annotators, and their descriptions are often obscure to common readers. We report three examples below:

- Support: Support is a fact that lends epistemic support to a claim, or that provides a reason for a course of action. Typically it is expressed as an External Argument. (EVIDENCE frame)
- Protagonist: A person or self-directed entity whose actions may potentially change the mind of the Cognizer (INFLUENCE_OF_EVENT_ON_COGNIZER frame)
- Locale: A stable bounded area. It is typically the designation of the nouns of Locale-derived frames. (LOCALE_BY_USE frame)

Since we aim at investigating whether such activity can be cast to a crowd of non-expert contributors, we need to reduce its complexity by intervening on the FE descriptions. In particular, we want to assess to what extent more information on the role semantics coming from external knowledge sources such as DBpedia[1] can improve non-expert annotators' performance. We leverage the CrowdFlower platform,[2] which serves as a bridge to a plethora of crowdsourcing services, the most popular being Amazon's Mechanical Turk (AMT).[3]

We claim that providing annotators with information on the semantic types typically associated with FEs will enable faster and cheaper annotations, while maintaining an equivalent accuracy. The additional information is extracted in a completely automatic way, and the workflow we present can be potentially applied to any crowdsourced annotation task in which semantic typing is relevant.

2 Related Work

The construction of annotation datasets for natural language processing tasks via non-expert contributors has been approached in different ways, the most prominent being *games with a purpose* (GWAP) and *micro-tasks*. While the former technique leverages fun as the motivation for attracting participants, the latter mainly relies on a monetary reward. The effects of such factors on a contributor's behavior have been analyzed in the motivation theory literature, but are beyond the scope of this paper. The reader may refer to [10] for an overview focusing on AMT.

Games with a Purpose. Verbosity [17] was one of the first attempts in gathering annotations with a GWAP. Phrase Detectives [5,4] was meant to harvest a corpus with coreference resolution annotations. The game included a validation mode, where participants could assess the quality of previous contributions. A data unit, namely a resolved coreference for a given entity, is judged complete only if the agreement is unanimous. Disagreement between experts and

[1] http://dbpedia.org
[2] https://crowdflower.com
[3] https://www.mturk.com

the crowd appeared to be a potential indicator of ambiguous input data. Indeed, it has been shown that in most cases disagreement did not represent a poor annotation, but rather a valid alternative.

Micro-tasks. Design and evaluation guidelines for five natural language micro-tasks are described in [15]. Similarly to our approach, the authors compared crowdsourced annotations with expert ones for quality estimation. Moreover, they used the collected annotations as training sets for machine learning classifiers and measured their performance. However, they explicitly chose a set of tasks that could be easily understood by non-expert contributors. Similarly, [13] built a multilingual textual entailment dataset for statistical machine translation by developing an annotation pipeline to decompose the annotators' task into a sequence of activities. Finally, [8] exploited Google AdWords, a tool for web advertising, to measure message persuasiveness while avoiding subjects being aware of the experiments and being biased by external rewards.

Semantic Role Annotation. Manual annotation of semantic roles has been recently addressed via crowdsourcing in [9] and [7]. Furthermore, [1] highlighted the crucial role of recruiting people from the crowd in order to bypass the need for linguistics expert annotations. Uniformly to our contribution, the task described in [9] was modeled in a multiple-choice answers fashion. Nevertheless, the focus is narrowed to the frame discrimination task, namely selecting the correct frame evoked by a given LU. Such task is comparable to the word sense disambiguation one as per [15], although the difficulty seems augmented, due to lower inter-annotator agreement values. The authors experienced issues that are related to our work with respect to the quality check mechanism in CrowdFlower, as well as the complexity of the frame names and definitions. Outsourcing the task to the CrowdFlower platform has two major drawbacks: (a) the proprietary nature of the aggregated inter-annotator agreement value provided in the response data, and (b) the need to manually simplify FE definitions that generated high disagreement. In this respect, our previous work [7] was the first attempt to address item (b) by manually simplifying the way FEs are described. In this work, we further investigate this aspect by exploiting automatically extracted links to DBPedia.

3 Annotation Workflow

Our goal is to determine if crowdsourced annotation of semantic roles can be improved by providing non-expert annotators with information from DBpedia on the roles they are supposed to label. Specifically, instead of displaying the lexicographic definition for each possible role to be labeled, annotators are shown a set of semantic types associated with each role coming from FrameNet. Based on this, annotators should better recognize such roles in an unseen sentence. Evaluation is performed by comparing this annotation framework with a baseline, where standard FE definitions substitute DBpedia information.

Before performing the annotation task, we need to leverage the list of semantic types that best characterizes each FE in a frame. We extract these statistics by connecting the FrameNet database 1.5 [14] to DBpedia, after isolating a set

of sentences to be used as test data (cf. Section 4). The workflow to prepare the input for the crowdsourced task is based on the following steps.

Linking to Wikipedia. For each annotated sentence in the FrameNet database, we first link each textual span labeled as FE to a Wikipedia page W. We employ *The Wiki Machine*, a kernel-based linking system (details on the implementation are reported in [16]), which was trained on the Wikipedia dump of March 2010.[4] Since FEs can be expressed by both common nouns and real-world entities, we needed a linking system that satisfactorily processes both nominal types. A comparison with the state-of-the-art system *Wikipedia Miner* [12] on the ACE05-WIKI dataset [3] showed that The Wiki Machine achieved a suitable performance on both types (.76 F1 on real-world entities and .63 on common nouns), while Wikipedia Miner had a poorer performance on the second noun type (respectively .76 and .40 F1). These results were also confirmed in a more recent evaluation [11], in which The Wiki Machine achieved the highest F1 compared with an ensemble of academic and commercial systems, such as *DBpedia Spotlight*, *Zemanta*, *Open Calais*, *Alchemy API*, and *Ontos*.

The system applies an 'all word' linking strategy, in that it tries to connect each word (or multiword) in a given sentence to a Wikipedia page. In case a linked textual span (partially) matches a string corresponding to a FE, we assume that one possible sense of FE is represented in Wikipedia through W. The Wiki Machine also assigns a confidence score to each linked term. This confidence is higher in case the words occurring in the same context of the linked term show high similarity, because the system considers that the linking is likely to be more accurate.

We illustrate in Figure 1 the Wikipedia pages (and confidence score) that the Wiki Machine system associates with the sentence `Sardar Patel was assisting Gandhiji in the Salt Satyagraha with great wisdom`, an example sentence for the ASSISTANCE frame originally annotated with four FEs, namely Helper, Benefited_party, Goal and Manner. Since Wikipedia is a repository of concepts, which are usually expressed by nouns, we are able to link only nominal fillers.

Linking to DBpedia. In order to obtain the semantic types that are typical for each FE, linking to Wikipedia is not enough. In fact, too many different pages would be connected to a FE, making it difficult to generalize over the Wikipedia pages (i.e. concepts). This emerges also from the example above, where the pages linked to `Sardar Patel`, `Gandhjii` and `Salt Satyagraha` do not provide information on the typical fillers of Helper, Benefited_party and Goal respectively. One possible option could be to resort to Wikipedia categories, which however are not homogenous enough to allow for a consistent extraction of FE semantic types.

We tackle this problem by using Wikipedia pages as a bridge to DBpedia. In fact, Wikipedia page URLs directly map to DBpedia resource URIs. Hence, for each linked FE, we query DBpedia for `rdf:type` objects. In this way, we are able to compute statistics on the most frequent semantic types associated with

[4] `http://download.wikimedia.org/enwiki/20100312`

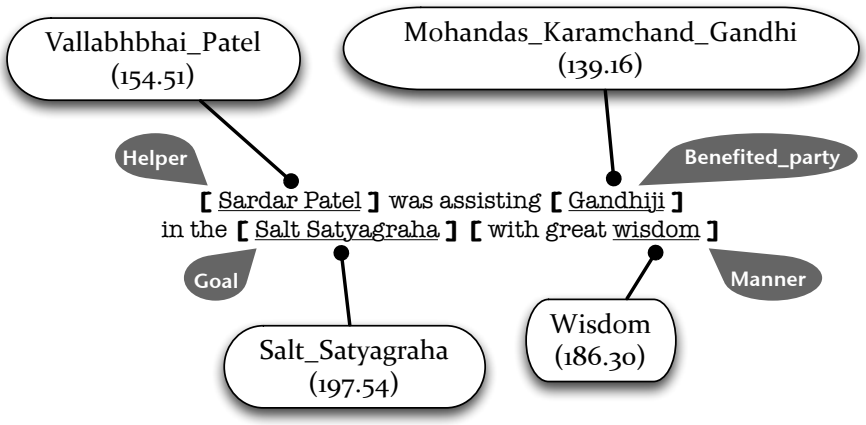

Fig. 1: Linking example with confidence score

a given FE from a given frame. For instance, the FE Victim from the KILLING frame has a top DBpedia type *Animal* with a frequency of 38. We aim at investigating whether such top-occurring types represent both valid generalizations and simplifications of a standard FE definition, and may thus substitute it. At the end of this pre-processing step, we create a repository where, for each FE, a set of DBpedia types is listed and ranked by frequency.

Posting the Annotation Task on CrowdFlower. We finally set up a crowdsourced experiment where, in each test sentence, annotators have to choose the most appropriate FE given the most frequent DBpedia types (proper task) or the standard FE definition (baseline). Details are reported in the following section.

4 Experiments

We first provide an overview of critical aspects underpinning a generic crowdsourced experiment. Subsequently, we describe the anatomy and the modeling of the tasks we outsourced to the CrowdFlower platform. Input data, full results, interface code and screenshots are available at `http://db.tt/iogsU7RI` .

Golden Data. Quality control of the collected judgements is a key factor for the success of the experiments. The essential drawback of crowdsourcing services relies on the cheating risk. Workers are generally paid a few cents for tasks which may only need a single click to be completed. Hence, it is highly probable to collect data coming from random choices that can heavily pollute the results. The issue is resolved by adding *gold* units, namely data for which the requester already knows the answer. If a worker misses too many gold answers within a given threshold, he or she will be flagged as untrusted and his or her judgments will be automatically discarded.

Worker Switching Effect. Depending on their accuracy in providing answers to gold units, workers may switch from a trusted to an untrusted status and vice

versa. In practice, a worker submits his or her responses via a web page. Each page contains one gold unit and a variable number of regular units that can be set by the requester during the calibration phase. If a worker becomes untrusted, the platform collects another judgment to fill the gap. If a worker moves back to the trusted status, his or her previous contribution is added to the results as free extra judgments. Such phenomenon typically occurs when the complexity of gold units is high enough to induce low agreement in workers' answers. Thus, the requester is constrained to review gold units and to eventually forgive workers who missed them. This has not been a blocking issue in our experiments, since we assessed a relatively low average percentage of missed judgments for gold units, namely 28%.

Cost Calibration. The total cost of a crowdsourced task is naturally bound to a data unit. This represents an issue in our experiments, as the number of questions per unit (i.e. a sentence) varies according to the number of frames and FEs evoked by the LU contained in a sentence. Therefore, we need to use the average number of questions per sentence as a multiplier to a constant cost per sentence. We set the payment per working page to 3 $ cents and the number of sentences per page to 3. Since most of the sentences in our annotation task have 3 FEs, the average cost per FE results in 0.325 $ cent (see Table 2 below).

Pre-processing of FrameNet Data for DBpedia Types Extraction. Table 1 provides some statistics of the processed FrameNet data that were leveraged to extract DBpedia types (cf. Section 3). More specifically:

1. From the FrameNet 1.5 database, the Wiki Machine managed to link 77% of the total number of FE instances. Hence, unlinked data is skipped for the next step.
2. DBpedia provided type information for 42% of the total number of linked FE instances. Types occurring once are ignored, as they reflect the content of a single sentence and are likely to convey misleading suggestions. The too generic `owl#Thing` type is filtered as well.

Table 1: FrameNet data processing details

Workflow step	**FE instances**
Raw FrameNet	148440
Linking to Wikipedia	114242
DBpedia types extraction	47732

Test Data Preparation. Before linking the FrameNet database to DBpedia, we isolate a subset to be used as test data. From 500 randomly chosen sentences, we select those in which the number of FEs per frame is between 3 and 4.

This small dataset serves as input for our experiments. Table 2 details the final settings. We hand-pick six sentences and for each of them we mark one question as gold for quality check. Almost all sentences contain three FEs with few exceptions (cf. the average value in Table 2). We extract the five most

frequent DBpedia types from the statistics and assign them to the corresponding FEs in our input. Since not all FEs have exactly five associated types (cf. the average value in Table 2), we provide workers with variable suggestion sets. Finally, we ensure all workers are native English speakers.

Table 2: Experimental settings

Sentences	43
Gold	6
Frames	24
Lexical Units	41
Average FEs per sentence	3.07
Average cost per FE ($ cents)	.325
Average DBpedia types per FE	4.66
Workers nationality	United States

Modeling. Data units are delivered to workers via a web interface. Our task is illustrated in Figure 2 and is presented as follows:

(a) Workers are invited to read a sentence and to focus on the bolded word appearing as a title above the sentence (e.g. `taste` in the screenshot).
(b) A question concerning each FE is then shown together with a set of answers corresponding to the sentence chunks that may express the given FE. For instance, in Figure 2, the question `Which is the Perceiver Passive?` is coupled with multiple choices taken from the given sentence.
(c) For each question, a suggestion box displays the top types retrieved from DBpedia and connected to the given FE (cf. Section 3 for details). This should help annotators in choosing the text chunk that better fits the given FE.
(d) Finally, workers match each question with the proper text chunk.

On the other hand, the baseline differs from our strategy in that (i) it does not display the suggestion box and (ii) questions are replaced with the FE definition extracted from FrameNet. For instance, in Figure 2, the question about the Perceiver Passive would be replaced with `This FE is the being who has a perceptual experience, not necessarily on purpose`. The baseline is more compliant with the standard approach adopted to annotate FEs in the FrameNet project.

5 Results

Our main purpose is to evaluate the validity of the proposed approach against the conventional FrameNet annotation procedure. We leverage expert-annotated sentences and are thus able to directly measure workers' accuracy. Specifically, we compute 2 values:

– *Majority vote.* An answer is considered correct only if the majority of judgments are correct.

Fig. 2: Worker interface unit screenshot

– *Absolute*. The total number of correct judgments divided by the total number of collected judgments.

The results of our experiments are detailed in Table 3. The number of untrusted judgments may be considered as a shallow indicator of the overall task complexity. In fact, we tried to maximize objectivity and simplicity when choosing gold units. Moreover, the input dataset (and gold units as well) is identical in both experiments. Therefore, we can infer that the number of workers who missed gold is directly influenced by the question model, which is the only variable parameter. We compute the execution time as the interval between the first and the last judged unit.

Table 3: Overview of the experimental results

Measure	Baseline	DBpedia
Majority vote accuracy	.763	.803
Absolute accuracy	.646	.720
Untrusted judgments	90	82
Time (minutes)	160	106

Our approach outperformed the baseline both in terms of accuracy and time. While majority vote accuracy values differ slightly, absolute accuracy clearly favors our strategy. Such measure can be seen as a further indicator of the task complexity. A higher score implies a higher number of correct judgments, which may designate a better inter-worker agreement, thus a more straightforward task. This claim is not only supported by the moderate decrease of untrusted judgments, but also by the dramatic reduction of the execution time. Consequently, the results we obtained demonstrate that entity linking techniques combined with DBpedia types simplify FEs annotation.

6 Discussion and Conclusions

In this work, we present a novel approach to annotate frame elements in a crowdsourcing environment using information extracted from DBpedia. The task is simplified for non-expert annotators by replacing FE definitions, usually meant for linguistic experts, with semantic types obtained from DBpedia. This is accomplished without manual simplification, in a completely automatic fashion.

Results prove that such method improves on the standard annotation workflow, both in terms of accuracy and of time consumption. Although the interconnection between FEs and DBpedia is semantically not perfect, extracting frequency statistics from the whole FrameNet database and considering only the most occurring types from DBpedia make the procedure quite robust to wrong links.

Possible issues may arise when two or more frame elements in the same frame share the same semantic type. For instance, the *Goal* and *Place* FEs in the ARRIVING frame are both likely to be filled by elements describing a location. We also expect that our approach is less accurate with FEs that can be filled both by nouns and by verbs, for instance the *Activity* FE in the ACTIVITY_FINISH frame. In such cases, information extracted from DBpedia would probably be inconsistent. Besides, DBpedia statistics are reliable when several annotated sentences are available for a frame, while they may be misleading if extracted from few instances. We plan to investigate these issues and to explore possible solutions to cope with data sparseness.

Additional future work will involve the following aspects:

- Evaluation of an ad-hoc strategy for the extraction of semantic types, namely providing workers with suggestions by matching information that are dynamically derived from each given sentence with DBpedia types.
- Clustering of similar semantic types with respect to the meaning they convey and to the frequency, e.g. `Place` and `Location_Underspecified`.

Finally, the overall effectiveness of our approach depends both on the performance of the entity linking system and on the coverage of the knowledge base. Hence, long term research will focus on enhancing The Wiki Machine precision and recall, and extending DBpedia type coverage.

References

1. Baker, C.F.: FrameNet, current collaborations and future goals. Language Resources and Evaluation pp. 1–18 (2012)
2. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet Project. In: Proceedings of the 17th international conference on Computational linguistics-Volume 1. pp. 86–90. Association for Computational Linguistics (1998)
3. Bentivogli, L., Forner, P., Giuliano, C., Marchetti, A., Pianta, E., Tymoshenko, K.: Extending English ACE 2005 Corpus Annotation with Ground-truth Links to Wikipedia. In: 23rd International Conference on Computational Linguistics. pp. 19–26 (2010)
4. Chamberlain, J., Kruschwitz, U., Poesio, M.: Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In: Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources. pp. 57–62. Association for Computational Linguistics (2009)
5. Chamberlain, J., Poesio, M., Kruschwitz, U.: Phrase detectives: A web-based collaborative annotation game. Proceedings of I-Semantics, Graz (2008)
6. Fillmore, C.: Frame semantics. Linguistics in the morning calm pp. 111–137 (1982)

7. Fossati, M., Giuliano, C., Tonelli, S.: Outsourcing FrameNet to the Crowd. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. pp. 742–747. Sofia, Bulgaria (August 2013), http://www.aclweb.org/anthology/P13-2130
8. Guerini, M., Strapparava, C., Stock, O.: Ecological Evaluation of Persuasive Messages Using Google AdWords. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. ACL2012 (July 2012)
9. Hong, J., Baker, C.F.: How Good is the Crowd at "real" WSD? In: Proceedings of the 5th Linguistic Annotation Workshop. pp. 30–37 (2011)
10. Kaufmann, N., Schulze, T., Veit, D.: More than fun and money. Worker motivation in crowdsourcing – A study on Mechanical Turk. In: Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, MI (2011)
11. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems. pp. 1–8. I-Semantics '11, ACM, New York, NY, USA (2011)
12. Milne, D., Witten, I.H.: Learning to link with Wikipedia. In: CIKM '08: Proceedings of the 17th ACM conference on Information and knowledge management. pp. 509–518. ACM, NY, USA (2008)
13. Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D., Marchetti, A.: Divide and conquer: crowdsourcing the creation of cross-lingual textual entailment corpora. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 670–679. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
14. Ruppenhofer, J., Ellsworth, M., Petruck, M.R., Johnson, C.R., Scheffczyk, J.: FrameNet II: Extended Theory and Practice. Available at http://framenet.icsi.berkeley.edu/book/book.html (2006)
15. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 254–263. Association for Computational Linguistics (2008)
16. Tonelli, S., Giuliano, C., Tymoshenko, K.: Wikipedia-based WSD for multilingual frame annotation. Artificial Intelligence 194, 203–221 (2013)
17. Von Ahn, L., Kedia, M., Blum, M.: Verbosity: a game for collecting common-sense facts. In: Proceedings of the SIGCHI conference on Human Factors in computing systems. pp. 75–78. ACM (2006)