University of Padova @ DIACR-Ita [*]

Benyou Wang and **Emanuele Di Buccio** and **Massimo Melucci**
Department of Information Engineering
University of Padova, Padova, Italy
{wang,dibuccio,melo}@dei.unipd.it

Abstract

Semantic change detection task in a relatively low-resource language like Italian is challenging. By using contextualized word embeddings, we formalize the task as a distance metric for two flexible-size sets of vectors. Various distance metrics like average Euclidean Distance, average Canberra distance, Hausdorff distance, as well as Jensen–Shannon divergence between cluster distributions based on K-means clustering and Gaussian mixture model are used. The final prediction is given by an ensemble of top-ranked words based on each distance metric. The proposed method achieved better performance than a frequency and collocation based baselines.

1 Introduction

Lexical Semantic Change detection aims at identifying words that change meaning over time; this problem is of great interest for NLP, lexicography, and linguistics. A semantic change detection task in English, German, Latin, and Swedish was proposed by Schlechtweg et al. (2020). Recently, Basile et al. (2020a) organized a lexical semantic change detection task in Italian called DIACR-Ita at EVALITA 2020 (Basile et al., 2020b). This technical report describes the methodology designed and developed by the University of Padova for the participation to DIACR-Ita.

Some previous approaches for semantic change modelling were based on static word embedding, where word vectors were trained for each time-stamped corpus and then were aligned, e.g. by orthogonal projections (Hamilton et al., 2016), vector initialization (Kim et al., 2014), and temporal teferencing (Dubossarsky et al., 2019). This work relies on contextualized word embeddings as the basic word representation component (Hu et al., 2019), since they have been shown to be effective in many NLP tasks including document classification and question answering. The methods relying on contextualized word embeddings performed worse than those based on static word embedding in Semantic Change detection tasks in many languages (Kutuzov and Giulianelli, 2020; Pömsl and Lyapin, 2020; Schlechtweg et al., 2020; Vani et al., 2020; Giulianelli et al., 2020; Giulianelli, 2019). However, it is our opinion that the use of contextualized word embeddings for this task is worth investigating because (1) they have highly expressive power as demonstrated in many downstream tasks e.g., document classification and question answering, and (2) they could handle fine-grained representations of individual context at the level of tokens.

By using contextualized word embedding, each word in a specific sentence is represented as a vector depending on the neighboring words which form the context of the word; a word appearing many times in a corpus is therefore represented as a set of vectors since one vector corresponds to each occurrence). In this paper, semantic change detection is addressed by computing the distance between two flexible-size sets consisting of vectors with respect to two time-stamped corpora. We investigated several distance metrics: average Euclidean Distance, average Canberra distance, and Hausdorff distance. Our methodology also relies on a clustering algorithm (e.g. K-means clustering and Gaussian Mixture Model) on the joint set and calculates a Jensen–Shannon divergence between cluster distributions in the two sub-corpora. We aggregate top-ranked words based on each distance metric as the final prediction. The proposed method achieved better performance than frequency and collocation based baselines and finally ranked the 8-th among 9 participanting teams.

2 Problem definition

Unlike the static word embedding like Word2vec (Mikolov et al., 2013) [1], contextualized word embeddings like ELMO (Peters et al., 2018) and BERT (Devlin et al., 2018) generate word representation based on the context of a word which does in this way not have a unique mapping with a fixed word vector.

Let us denote a corpus with m sentences as $\mathcal{C}$. In this paper, $\mathcal{C}$ is related to a time span t because of the task characteristics; however, the corpus can be tailored to any specific aspect, e.g. a specific domain such as news or books. For a word w_i appearing in $\mathcal{C}$, its contextualized word representation in the k-th sentence [2] is denoted by $e_{i,k}^{(\mathcal{C})}$. The word representation in the corpus is a set

$$\Phi_i^{\mathcal{C}} = \{e_{i,1}^{(\mathcal{C})}, e_{i,2}^{(\mathcal{C})}, \cdots e_{i,k}^{(\mathcal{C})}, \cdots, e_{i,m}^{(\mathcal{C})}\} \quad (1)$$

To examine whether a word w_i exhibits a semantic change between two corpora $\mathcal{C}_1$ (in t_1) and $\mathcal{C}_2$ (in t_2), we check the difference between two sets $\Phi_i^{\mathcal{C}_1}$ and $\Phi_i^{\mathcal{C}_2}$. Let l_i be a human-annotated label indicating the semantic change degree; l_i usually ranges from 0 to 1, where 1 denotes a full semantic change. Let D be the dimension of the word vector. We define the distance metric as a function

$$f : \{\mathbb{R}^D\}^m, \{\mathbb{R}^D\}^n \rightarrow \mathbb{R}. \quad (2)$$

to obtain a semantic change degree based on the representation of a word in two corpora denoted as $\Phi_i^{\mathcal{C}_1}, \Phi_i^{\mathcal{C}_2}$. When labels are binary, one may simply use a threshold on the values of $f(\cdot, \cdot)$ to predict the binary label. Let δ be a function to generate a binary output, e.g., based on a hand-crafted threshold. We can predict whether w_i exhibits a semantic change between $\mathcal{C}_1$ and $\mathcal{C}_2$ as follows

$$\bar{l}_i = \delta(f(\Phi_i^{\mathcal{C}_1}, \Phi_i^{\mathcal{C}_2})) \quad (3)$$

where $\bar{l}_i$ is the predicted binary label.

In conclusion, in our work the semantic change detection task is formalized as follows

$$\arg\max_{f,\delta} \sum_{w_i} \left(\delta(f(\Phi_i^{\mathcal{C}_1}, \Phi_i^{\mathcal{C}_2})) == l_i\right) \quad (4)$$

Since this is a closed task, we may not have enough annotated samples to train a f using gradient descent. Therefore, a well-selected f will be crucial.

3 Methodology

3.1 Contextualized Word Embedding

Using contextualized word embeddings like ELMO and BERT has be shown to improve performance in various downstream tasks due to its expressive power for words. In this paper, we use a multilingual-BERT[3]. Uncased models are adopted since we assume that semantic change detection is insensitive to word case. All models are in *base* settings with 12 layers, 12 heads, and a hidden state dimension of 768. Only last-layer output of BERT is used as word representation.

3.2 Measuring Semantic Change Degree

3.2.1 Distance-based methods

In this section, we introduce various methods to calculate the semantic change degree.

Average Geometric Distance. Average Geometric Distance (AGD) (also can be seen in (Kutuzov and Giulianelli, 2020; Giulianelli, 2019)) is defined as below:

$$\text{AGD}(\Phi_i^{\mathcal{C}_1}, \Phi_i^{\mathcal{C}_2}) = \frac{1}{mn} \sum_{\boldsymbol{x} \in \Phi_i^{\mathcal{C}_1}, \boldsymbol{y} \in \Phi_i^{\mathcal{C}_2}} d(\boldsymbol{x}, \boldsymbol{y})$$

The distance function $d(\cdot, \cdot)$ can be the *Euclidean Distance* [4], the *Canberra distance* (Lance and Williams, 1966) [5] or any distance function. In this paper, we also use the negative cosine similarity as a normalized distance metric.

Hausdorff distance. Hausdorff distance (Rockafellar and Wets, 2009) is denoted as HD in short and is generally used to measure the distance between two non-empty sets, namely,

$$\begin{aligned} \text{HD}(\Phi_i^{\mathcal{C}_1}, \Phi_i^{\mathcal{C}_2}) = \max(&\sup_{\boldsymbol{x} \in \Phi_i^{\mathcal{C}_1}} \inf_{\boldsymbol{y} \in \Phi_i^{\mathcal{C}_2}} ||\boldsymbol{x} - \boldsymbol{y}||_2, \\ &\sup_{\boldsymbol{x} \in \Phi_i^{\mathcal{C}_2}} \inf_{\boldsymbol{y} \in \Phi_i^{\mathcal{C}_1}} ||\boldsymbol{x} - \boldsymbol{y}||_2) \end{aligned} \quad (5)$$

[1] An overview on word vectors is in Wang et al. (2019).

[2] If a word appears in a sentence more than once, we take the average.

[3] https://storage.googleapis.com/bert_models/2018_11_03/multilingual_L-12_H-768_A-12.zip.

[4] Euclidean Distance : $d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_2$

[5] Canberra distance is a normalized version of the Manhattan distance, $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{D} \frac{|x_i - y_i|}{|x_i| + |y_i|}$

3.2.2 Clustering-based Methods

By clustering the union set between $\Phi_i^{\mathcal{C}_1}$ and $\Phi_i^{\mathcal{C}_2}$ in K clusters/categories, we obtained the category distributions $\boldsymbol{p}, \boldsymbol{q}$ for $\Phi_i^{\mathcal{C}_1}$ and $\Phi_i^{\mathcal{C}_2}$, respectively. We adopted two commonly used clustering methods: the K-means clustering method and the Gaussian Mixture Model method. As for the distance between distributions, we adopted the Jensen–Shannon Divergence (JSD), which is a symmetrized and smoothed version of the Kullback–Leibler divergence:

$$\text{JSD} = \frac{1}{2}\text{KL}(\boldsymbol{p}, \boldsymbol{q}) + \frac{1}{2}\text{KL}(\boldsymbol{q}, \boldsymbol{p})$$

where $\text{KL}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{K} \boldsymbol{p}_i \log \frac{\boldsymbol{p}_i}{\boldsymbol{q}_i}$.

3.3 Threshold and Ensemble

We took the top-K ranked target words of each metric and aggregated them for the final submission. The K was decided when the aggregated target words reached the half of total words numbers, since we assumed that the annotated labels are balanced. See (Schlechtweg et al., 2020) for detailed discussions about thresholds.

4 Experiments

4.1 Dataset and Evaluation Methodology

DIACR-Ita is the first task on lexical semantic change for Italian. DIACR-Ita aims to automatically detect whether a word semantically change over time. The task is to detect if a set of words, called target words, change their meaning across two periods, t_1 and t_2, where t_1 precedes t_2. Participants are provided with two corpora $\mathcal{C}_1$ and $\mathcal{C}_2$ (corresponding to t_1 and t_2, respectively), and a set of target words. For instance, the meaning of the word ‘imbarcata’ has changed from t_1 to t_2; originally, the word referred to an ‘acrobatic manoeuvre of aeroplanes’, but it is nowadays used to refer to the state of being deeply in love (Basile et al., 2020a) although the latter meaning is much less used than the former meaning. The task is formulated as a closed task, namely, models must be trained solely on the provided data. The occurrence about target words is reported in Table 1.

Labels in this task are binary and the task is considered as a binary classification problem. The evaluation is based on accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

word	# corpus $\mathcal{C}_1$	# corpus $\mathcal{C}_2$
egemonizzare	11	37
lucciola	64	226
campanello	109	628
trasferibile	7	60
brama	17	93
polisportiva	74	134
palmare	19	88
processare	39	594
pilotato	34	285
cappuccio	60	198
pacchetto	274	5690
ape	123	252
unico	4524	29620
discriminatorio	110	262
rampante	26	462
campionato	3918	11871
tac	88	438
piovra	30	621

Table 1: ‘#1’ and ‘#2’ denote the number of sentences where the target word occurs in two time-stamped corpora $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively.

methods	accuracy
Frequencies	0.50
Collocations	0.61
Aggregated results (submitted)	**0.67**
Average negative cosine similarity	0.67
Average distance with Euclidean distance	0.61
Average distance with Canberra distance	0.61
Hausdorff distance	0.50
JS divergence with K-means Clustering	0.61
JS divergence with Gaussian Mixture Model	0.61

Table 2: Results of the proposed methods.

T, F refers to ‘True’ and ‘False’, P, N refers to ‘positive’ and ‘negative’. For example, TP is the number of Truly-predicted Positive samples.

The task735680 organizers provided two baselines: **Frequencies**: the absolute value of the difference between the words’ frequencies is computed; **Collocations**: for each word, it computes the cosine similarity between two Bag-of-Collocations (BoCs) vector representations related to $\mathcal{C}_1$ and $\mathcal{C}_2$. In both baseline models, a threshold is used to predict if the word has changed its meaning.

4.2 Experimental Results

Experimental results are reported Table 2 and show that the proposed method achieved better performance than frequency and collocation based baselines.

4.3 Post-hoc Analysis

In this section, we will provide a bi-dimensional visualization of word representation to intuitively understand how the contextualized word vectors work. For each word, we get all contextualized word vectors (with a dimension of 768) based on its context. To visualized word in a 2D plane, we used a typical dimension reduction algorithm called T-SNE (Maaten and Hinton, 2008) to reduce word vectors from 768 to 2. Red and blue points denote the low dimensional representation of vectors when considering the two time-stamped corpora $\mathcal{C}_1$ (blue) and $\mathcal{C}_2$ (red).

For example, 'rampante' and 'palmare' are the predicted positive samples while 'cappuccio' and 'campanello' are predicted negative samples. As shown in Figure 1, the predicted semantically-shifted words exhibit a clear difference between red points an blue points with respect to two time-stamped corpora. For the predicted semantically-unshifted words (see Figure 2), it looks slightly indistinguishable.

5 Limitations

In (Schlechtweg et al., 2020), semantic representations are mainly divided to two categories: average embeddings ('type embeddings') and contextualized embeddings ('token embeddings'). Schlechtweg et al. (2020) illustrated the performance of token-based models are much lower than type-based embedding models. In this section, we will discuss some limitations of currently-used contextualized embedding based methods for semantic change detection.

There are typically two kinds of methods to use contextualized embeddings for semantic change detection: *embedding-based distance metrics* and *clustering-based distance metrics* (Schlechtweg et al., 2020; Vani et al., 2020; Giulianelli et al., 2020; Giulianelli, 2019). The former are directly calculated on the raw contextualized word embeddings while the latter are based on the clustering results of contextualized word embeddings.

5.1 Embedding-based Distance Metrics

Can distance metrics distinguish semantic shift patterns? Many typical patterns of semantic shifts have been investigated (Grossmann and Rainer, 2013; Basile et al., 2020a): 1) pejoration or amelioration (when word meanings become more negative or more positive); 2) broadening or narrowing (when it evolves as a generalized/extended object or a restricted or specialized one); 3) adding/deleting a sense; 4) totally shifted. The patterns of semantic change are multifaceted and we are questioning that a single distance metric could precisely distinguish all the above typical semantic shift patterns.

Normalization. Most of distance metrics are not normalized except for negative cosine similarity. Absolute values of unnormalized distance metrics may differ a lot among individual words; they are sometimes unexpectedly affected by the number of samples, leads to that the values of metrics may not be comparable among words.

Outliers. Some distance metrics (e.g., Hausdorff distance) are sensitive to outliers. For example, since the calculation of Hausdorff distance is based on infimum and supremum, an outlier point may largely affect the final Hausdorff distance. As seen in Table 3, frequently-appearing words e.g., 'campionato' and 'unico' have the highest Hausdorff distance between $\mathcal{C}_1$ and $\mathcal{C}_2$, this is probably biased by the fact that the two words appear frequently (see Table 1) and therefore likely have more unexpected outliers.

Model Fine-tuning. The contextualized word embedding that is based on pre-trained language models like BERT achieved much better results compared to static word embedding with a two-stage training paradigm, where the two stages are pre-training in language model (e.g., mask language model) and fine-tuning in downstream tasks (e.g., classifications). However, in the semantic change detection task, fine-tuning in downstream tasks is currently impossible because the annotated labels are insufficient to this aim; to some extent, the lack of fine-tuning stage may harm the performance of the pre-trained language models.

5.2 Clustering-based Distance Metrics

After clustering, we used the Jensen–Shannon divergence (JSD) which is affected by the issues mentioned in Section 5.1 like other distance metrics. Plus, the clustering algorithm may introduce some errors of semantic change detection. First, typical clustering algorithms may not necessarily converge to an identical clustering result when the seed centroids are changed. Moreover, the number of clusters is crucial since the optimal number of clusters cannot easily be decided before clustering.

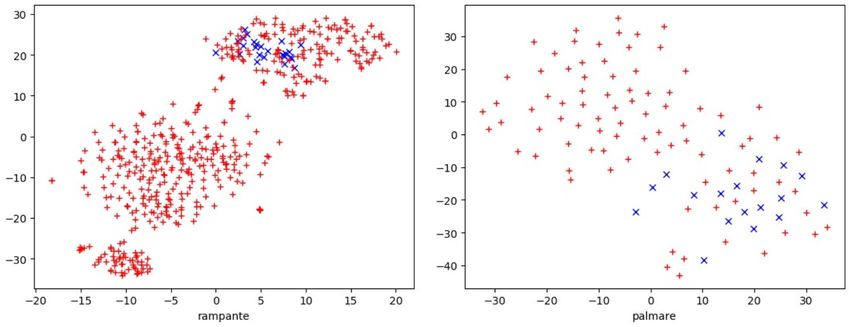

Figure 1: Examples (i.e., ‘rampante’ and ‘palmare’) of predicted ”semantically-shifted” words. Red and blue points denote dimensionally-reduced vectors of two time-stamped corpora respectively.

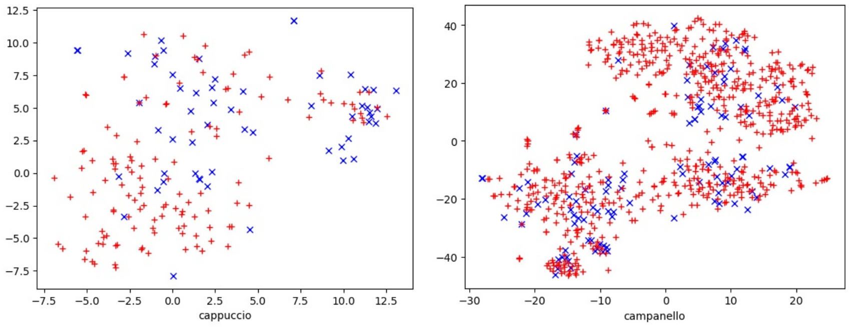

Figure 2: Examples (i.e., ‘cappuccio’ and ‘campanello’) of predicted ”semantically-unshifted” words. Red and blue points denote dimensionally-reduced vectors of two time-stamped corpora respectively.

6 Conclusions

This paper formalizes semantic change detection as a distance metric between two variable-sized sets of vectors. The final prediction is based on an ensemble of different distance metrics. The proposed method outperformed weak frequency and collocation baselines, but it performed less well than SOTA baselines. As a future work, this task may be largely improved via a supervised task in a unified multi-lingual framework; thus, any human-annotated labels in other languages could be used in this task since currently the number of annotated semantically-shift words in a single language is limited.

Acknowledgments

This work is supported by the Quantum Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union‘s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 721321.

A Appendix

Table 3 reports the predictions based on various distance metrics.

References

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020a. DIACR-Ita @ EVALITA2020: Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In *EVALITA 2020*, Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro (Eds.). CEUR.org, Online.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020b. EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro (Eds.). CEUR.org, Online.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust

word	AGD-cosine	AGD-euclidean	AGD-canberra	Hausdorff distance	JSD-GMM	JSD-Kmeans
matematica	0.996	1.02	86.6	10.0	0.004	0.025
dettagliato	**0.895**	**6.09**	**290.9**	7.5	**0.693**	**0.693**
sanità	0.990	1.86	130.8	**10.9**	0.025	0.052
senatore	0.997	0.79	79.1	7.7	0.009	0.002
istruzione	**0.854**	**6.14**	**333.7**	**14.4**	**0.275**	**0.279**
egemonizzare	0.988	1.62	136.6	5.6	0.003	0.033
lucciola	0.970	2.58	187.3	8.4	**0.414**	**0.154**
campanello	0.990	1.13	131.7	10.8	0.003	0.003
trasferibile	**0.873**	**4.25**	**300.7**	7.2	0.059	0.073
brama	**0.830**	**5.80**	**346.2**	8.3	**0.420**	**0.406**
polisportiva	**0.921**	**4.42**	**285.7**	7.5	**0.293**	**0.291**
palmare	0.955	2.55	220.5	8.0	0.130	**0.154**
processare	0.986	1.76	159.9	6.9	0.105	0.067
pilotato	0.970	2.27	198.9	**12.1**	0.108	0.128
cappuccio	0.973	1.78	183.6	**12.2**	0.015	0.016
pacchetto	0.984	1.67	149.6	10.5	0.011	0.009
ape	0.953	2.09	216.7	**15.3**	0.033	0.031
unico	0.985	1.89	149.9	**16.2**	0.035	0.032
discriminatorio	0.987	1.56	150.5	10.2	0.007	0.007
rampante	**0.888**	**4.78**	**302.7**	6.5	**0.293**	**0.299**
campionato	0.978	2.51	183.1	**16.0**	0.074	0.071
tac	**0.815**	**5.25**	**366.2**	9.9	**0.301**	**0.391**
piovra	0.976	2.27	189.6	9.7	0.033	0.033

Table 3: Calculated scores of various distance metrics. Top ranked scores are in bold.

modeling of lexical semantic change. *arXiv preprint arXiv:1906.01688* (2019).

Mario Giulianelli. 2019. Lexical semantic change analysis with contextualised word representations. *Unpublished master's thesis, University of Amsterdam, Amsterdam* (2019).

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. *arXiv preprint arXiv:2004.14118* (2020).

Maria Grossmann and Franz Rainer. 2013. *La formazione delle parole in italiano*. Walter de Gruyter.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *ACL*. 1489–1501.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *ACL*. 3899–3908.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. *ACL 2014* (2014), 61.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. *arXiv preprint arXiv:2005.00050* (2020).

Godfrey N Lance and William T Williams. 1966. Computer programs for hierarchical polythetic classification ("similarity analyses"). *Comput. J.* 9, 1 (1966), 60–64.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* 9, Nov (2008), 2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*. 2227–2237.

Martin Pömsl and Roman Lyapin. 2020. CIRCE at SemEval-2020 Task 1: Ensembling Context-Free and Context-Dependent Word Representations. *arXiv preprint arXiv:2005.06602* (2020).

R Tyrrell Rockafellar and Roger J-B Wets. 2009. *Variational analysis*. Vol. 317. Springer Science & Business Media.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. *arXiv preprint arXiv:2007.11464* (2020).

K Vani, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. 2020. SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces. *arXiv preprint arXiv:2010.00857* (2020).

Benyou Wang, Emanuele Di Buccio, and Massimo Melucci. 2019. Representing Words in Vector Space and Beyond. In *Quantum-Like Models for Information Retrieval and Decision-Making*. Springer, 83–113.