DiaSense at SemEval-2020 Task 1: Modeling sense change via pre-trained BERT embeddings

Christin Beck
University of Konstanz
christin.beck@uni-konstanz.de

Abstract

This paper describes DiaSense, a system developed for Task 1 'Unsupervised Lexical Semantic Change Detection' of SemEval-2020. In DiaSense, contextualized word embeddings are used to model word sense changes. This allows for the calculation of metrics which mimic human intuitions about the semantic relatedness between individual use pairs of a target word for the assessment of lexical semantic change. DiaSense is able to detect lexical semantic change in English, German, Latin and Swedish (accuracy = 0.728). Moreover, DiaSense differentiates between weak and strong change.

1 Introduction

Task 1 of SemEval-2020 (Schlechtweg et al., 2020) is concerned with the unsupervised detection of lexical semantic change (LSC) as reflected by word sense changes over time. More broadly, LSC refers to changes in meaning of a lexical item. A meaning change is manifested in the gain or loss of a particular meaning of a word which indicates an increase or decrease in polysemy (Traugott and Dasher, 2001; Traugott, 2006). A well-known example for LSC which is cited in Tahmasebi et al. (2018) is the historical evolution of the English word *hound* which changed from being the general word for 'dog' to referring to only a specific kind of 'dog' ('narrowing', cf. Traugott (2006)). Meanwhile, *dog* changed from describing a specific type of 'dog' to becoming the general term for 'dog' ('broadening', cf. Traugott (2006)). Moreover, senses can become obsolete and be lost overall, while cultural changes may drive the evolution of new senses. The main task in SemEval-2020 Task 1 is to identify and evaluate LSC in a set of target words between two text corpora stemming from two different time periods t_1 (earlier period) and t_2 (later period). The investigated languages are German, English, Swedish and Latin. The task is split into two subtasks: (i) a binary classification task, where it has to be determined whether a target word lost/gained senses between t_1 and t_2 or not, and (ii) a ranking task, where target words are ranked according to their degree of LSC (a higher rank indicates stronger change).

The system presented in this paper is called DiaSense and addresses the LSC tasks by modeling the different senses of a word via contextualized word embeddings using pre-trained BERT (Devlin et al., 2018).[1] The binary classification and the ranking task are approached by transferring the measures of change suggested by Schlechtweg et al. (2018), which are originally based on human annotated values for meaning relatedness, to change metrics calculated on the basis of differences between target word embeddings. DiaSense is able to detect LSC in the majority of cases in all four languages (average accuracy 0.728).[2] Although the results produced for the ranking task only show a weak correlation with the gold data (Spearman's ρ = 0.337), DiaSense is able to distinguish between strong and weak change.

[1] The source code is provided on https://github.com/christinschaetzle/DiaSense.

[2] DiaSense was substiantially improved in the post-evaluation phase. We focus on describing the improved version in this paper, but also provide details about the system in the evaluation phase, i.e., before the ground truth data was released.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 50–58
Barcelona, Spain (Online), December 12, 2020.

2 Related Work

Over recent years, research on LSC has seen an increasing use of computational methods (see Tahmasebi et al. (2018) and Kutuzov et al. (2018) for detailed overviews). The methods applied for LSC detection are manifold, but can be grouped into three classes according to the type of meaning representation involved (Schlechtweg et al., 2019): (i) semantic vector spaces, (ii) topic distributions, (iii) sense clusters.

In semantic vector space approaches, each target word is represented as a vector at each time stage. The vector representations are typically based on bag-of-words approaches and represent a co-occurrence statistics of a word with its context words. Common methods employed for computing vectors from co-occurrence statistics are Positive Pointwise Mutual Information (PPMI), which measures co-occurrence strength, and Singular Value Decomposition (SVD), for dimensionality reduction; see, e.g., Levy et al. (2015), Hamilton et al. (2016), Hellrich and Hahn (2017), Kahmann et al. (2017). Moreover, word embeddings as generated via the Skip-Gram with Negative Sampling (SGNS) technique (Mikolov et al., 2013), i.e., word2vec, and GloVe embeddings (Pennington et al., 2014) have been applied to LSC detection, e.g., Hamilton et al. (2016) and Hellrich and Hahn (2016). As measure of LSC, similarity across time periods is assessed via calculating the distance/similarity between vectors, using, e.g., cosine distance (Salton and McGill, 1983), or alternatively via computing differences in the contextual dispersion of the vectors (see, e.g., Schlechtweg et al. (2019)). In approaches where meaning is represented via topic distributions (Bamman and Crane, 2011; Lau et al., 2012; Cook et al., 2014), word senses are derived from topic models based on, e.g., Latent Dirichlet Allocation (LDA; Blei et al. (2003)) and Hierarchical Dirichlet Processes (HDP, Teh et al. (2006)). Furthermore, the dynamic topic model SCAN was specifically developed for the investigation of lexical change (Frermann and Lapata, 2016). With topic modeling, LSC is usually assessed via a frequency-based novelty score assigned to the senses. Sense clustering based approaches follow similar principles, but are used less often; e.g., Mitra et al. (2015).

Recently, efforts have been made towards developing evaluation standards and datasets for LSC (Hamilton et al., 2016; Frermann and Lapata, 2016; Schlechtweg et al., 2018; Dubossarsky et al., 2019; Schlechtweg and im Walde, 2020). For example, Schlechtweg and im Walde (2020) generate simulations of LSC on the basis of synchronic data, providing a testbed for diachronic LSC, while Schlechtweg et al. (2019) provide human annotations via the Diachronic Usage Relatedness (DURel) dataset. The current pitfall of the existing works on LSC is the lack of a common state-of-the-art evaluation task which makes the comparison of methods difficult. This shortcoming is addressed by SemEval-2020 Task 1.

3 System description

DiaSense measures change by combining word sense representations generated via BERT with LSC metrics which are based on the calculation of cosine distance as detailed in the following.

3.1 Word sense representations

DiaSense makes use of a semantic vector space approach to represent the lexical semantic content of the target words. In contrast to previous approaches, which employ static word embeddings as generated by, e.g., SGNS and GloVe, DiaSense is based on the state-of-the-art contextualized word embeddings provided by BERT. With static word embeddings, each word is represented via a single vector for each time period, which is shared by all senses of a polysemous word. Although some contextual information is captured, it is difficult to differentiate between the senses involved. This problem is alleviated by contextualized vector representations, where each vector is a function of a whole input sentence, keeping different word senses apart.

A further advantage of using BERT is that we can leverage the pre-trained models released by Google AI (Devlin et al., 2018) which spares us the task of training models by ourselves. Pre-trained contextualized embeddings have proven to be almost as effective as corresponding state-of-the-art models in linear NLP probing tasks such as part-of-speech tagging (Liu et al., 2019). Wiedemann et al. (2019) furthermore showed that pre-trained BERT allows for the disambiguation of polysemic words. Being able to use a pre-trained model is beneficial when working with historical data which is sparse by nature, with a lesser amount of training data available for the longer-standing past than for more recent time stages. Pre-trained static embeddings exist (e.g., fasttext[3]), but are less applicable to historical data since they usually do not provide for out-of-vocabulary (OOV) words. Since the vocabulary in historical documents might differ substantially from the modern vocabulary used for the generation of word embeddings, historical documents are likely to contain a large amount of OOV words. The token-based approach employed by BERT on the other hand is designed to include OOV words, handling them via sub-word embeddings. Moreover, the pre-trained multilingual BERT embeddings allow for a language-independent approach to LSC, without having to scale to new languages by calculating new embedding matrices and parameters.

Our system is based on bert-as-service (Xiao, 2018), a Python library which uses Google's BERT model as sentence encoder, hosting it as service via ZeroMQ[4]. Bert-as-service is easy to implement and allows for the mapping of sentences into fixed-length BERT embeddings with just two simple lines of code. In DiaSense, LSC is assessed separately for each language, but we feed the same model to bert-as-service, i.e., the cased multilingual 12-layer BERT-base model.[5] By default, bert-as-service works on the second-to-last layer. Bert-as-service makes provision to get contextualized ('ELMo-like', cf. Peters et al. (2018)) word representations from the sentence embeddings. In DiaSense, we compute a sentence embedding for each sentence a target word occurs in via bert-as-service and take the corresponding target word embeddings to be representations of the target word's senses. This is done separately for t_1 and t_2. If a target word has been tokenized into several subword units by BERT, we average over all subword embeddings which belong to the target word, taken from the corresponding sentence embedding.

3.2 LSC metrics

DiaSense was altered substantially in the post-evaluation phase, i.e., after the publication of the ground truth, with respect to the metrics employed for assessing LSC. We report on the metrics used in the evaluation and the post-evaluation phase in the following, but focus on the post-evaluation metrics, since these immensely improved the system's overall performance (see Section 5).

Evaluation phase In the evaluation phase, the binary classification task was approached via clustering. We clustered the target word embeddings generated via BERT using the KMeans algorithm as implemented in scikit-learn (Pedregosa et al., 2011), generating 'sense clusters' (with $k = 2$ as default). Change was then measured on the basis of a frequency threshold. That is, a word was classified as changing, when a cluster consisted of at least 90% of embeddings from one corpus only. The ranking task was addressed by calculating an average embedding for each target word in each corpus. Then, we measured the degree of change by computing the cosine distance between the average embedding from t_1 and the average embedding from t_2 of each target word. Cosine distance ($cosine$) between two (non-zero) vectors $\vec{x}$ and $\vec{y}$ is defined on the basis of their cosine similarity (sim), which corresponds to the dot product of the vectors divided by the product of their Euclidean lengths (Manning et al., 2008):

$$cosine(\vec{x},\vec{y}) = 1 - sim(\vec{x},\vec{y}) = 1 - \frac{\vec{x}\cdot\vec{y}}{||\vec{x}||\,||\vec{y}||} \tag{1}$$

A cosine distance value close to 0 indicates a low difference and a value close to 1 a high difference. Thus, we interpreted a large distance as high degree of change in the ranking task in the evaluation phase.

Post-evaluation phase To detect and measure LSC in the post-evaluation phase, DiaSense calculates several different metrics on the basis of the target word embeddings. The metrics are based on the measures provided by Schlechtweg et al. (2018) for the assessment of LSC change with respect to DURel annotations: ΔLATER and COMPARE. DURel contains gold standard annotations for 22 target words with respect to diachronic LSC in German. The annotations rest upon meaning relatedness scores assigned to sentence pairs in which a specific word occurs ('use pairs'), ranging from 1 (unrelated) to 4 (identical). The scores are inspired by Blank's (1997) semantic proximity continuum (proximity increases): homonymy > polysemy > context variance > identity. Thus, a high mean relatedness value between use pairs indicates

[3] https://fasttext.cc/

[4] https://zeromq.org/

[5] https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

meaning identity or context variance and a low value indicates polysemy or homonymy. According to this rationale, in a scenario of innovative meaning change from t_1 to t_2 (emergence of a new meaning), the meaning relatedness in t_2 should be lower than in t_1, and vice versa when reductive meaning change (loss of a meaning) takes place.

ΔLATER of a word *w* captures these intuitions and measures changes in the degree of mean relatedness by substracting *w*'s mean value in t_1 (*earlier*) from the mean value in t_2 (*later*): $\Delta\text{LATER}(w) = Mean_l(w) - Mean_e(w)$. A high positive ΔLATER value shows an increasing relatedness over time and can be interpreted as reductive meaning change. A low negative ΔLATER in turn indicates innovative meaning change. In contrast to ΔLATER, COMPARE directly measures the relatedness of a word between t_1 and t_2, via the mean value of relatedness scores assigned to use pairs which consist of a sentence from t_1 and a sentence from t_2 (with $\text{COMPARE}(w) = Mean_c(w)$). COMPARE measures the degree of change (Schlechtweg et al., 2018), with a high value indicating weak and a low value indicating strong change.

Instead of assigning relatedness ranks to use pairs, DiaSense captures relatedness between target word embeddings via cosine distance. In doing so, a high cosine distance between target word embeddings can be interpreted as low meaning relatedness, while a low cosine distance value indicates a high meaning relatedness. To compute $\Delta\text{LATER}(w)$, we calculate the cosine distances between all embeddings of a target word in t_2 and assign the mean value of these distances to $Mean_l(w)$. We proceed similarly for t_1 to compute $Mean_e(w)$ and calulate $\Delta\text{LATER}(w)$ analogously to Schlechtweg et al. (2018) as difference between $Mean_l(w)$ and $Mean_e(w)$. Using cosine distance allows for the intuitive interpretation of a high positive value for ΔLATER as innovative meaning change and a low negative value as reductive meaning change. $\text{COMPARE}(w)$ is computed by calculating the mean of cosine distances between all use pairs where one embedding is from t_1 and the other is from t_2. In turn, the values must be interpreted inversely on the basis of cosine distance: a high COMPARE value indicates strong change, while a low value is an indicator of weak change.

Additionally, Schlechtweg et al. (2018) suggest for future research to normalize COMPARE with respect to polysemy in order to be able to differentiate between context variation and real diachronic change. Therefore, they propose to substract the mean relatedness value of the earlier time period, i.e., $Mean_e(w)$, from COMPARE and calculate ΔCOMPARE in this way. However, this only captures the variation in the earlier period, not accounting for the whole variation present in the two corpora. Thus, instead of substracting the mean value of the earlier period, we propose to calculate ΔCOMPARE by substracting the mean cosine distance between all target word embeddings from C_1 and C_2, without differentiating between periods. In this way, we can capture the amount of within variation across both corpora.[6]

4 Experimental setup

Data For each language, two corpora C_1 and C_2 (for t_1 and t_2) and a set of target words were provided in the task. The corpora were pre-processed in that punctuation, empty and one word sentences were removed. Additionally, all sentences were lemmatized and are randomly shuffled within each corpus. This is meant to mimic the challenging nature of historical linguistic data, where incomprehensible and incomplete data is the norm rather than the exception. For English, C_1 and C_2 consist of data from the CCOHA corpus (Alatrash et al., 2020), representing the time stages 1810-1860 (t_1) and 1960-2010 (t_2). C_1 for German contains texts from 1800 to 1899 taken from the DTA corpus (Deutsches Textarchiv, 2017), and combines data from two newspaper corpora (Berliner Zeitung[7], Neues Deutschland[8]) for C_2, with data from 1946 to 1990. For Latin, data was taken from the LatinISE corpus (McGillivray and Kilgarriff, 2013). C_1 features data from the beginning of the second century to the end of the first century BCE, while C_2 contains data from the beginning of the first century to the end of the twenty-first century CE. The Swedish corpora are based on data from KubHist (Asedam et al., 2019), with data from 1790-1830 for C_1 and from 1895-1903 for C_2. Overall, the application scenario is broad with corpora covering four languages, whilst spanning over time stages which differ in terms of their chronological depth and length.

[6] We experimented with both ways of normalizing COMPARE and achieved better results with our version of ΔCOMPARE.

[7] http://zefys.staatsbibliothek-berlin.de/index.php?id=155

[8] http://zefys.staatsbibliothek-berlin.de/index.php?id=156

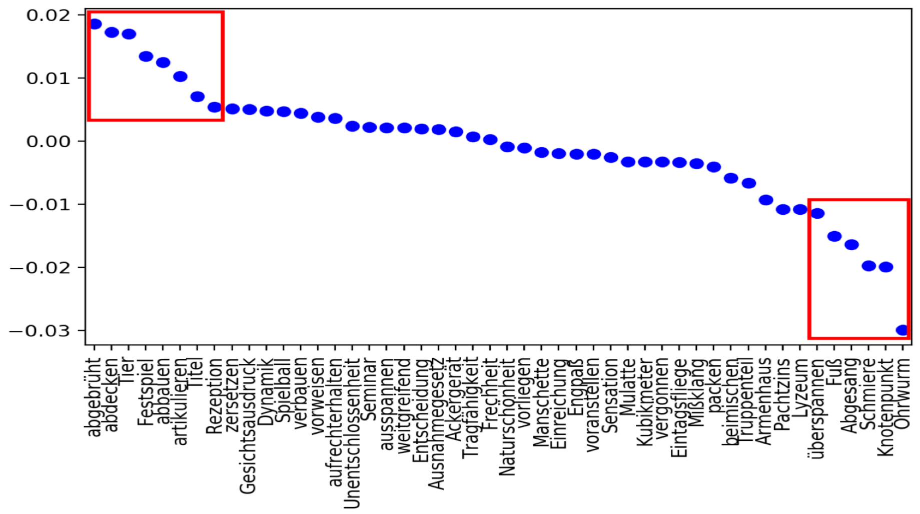

Figure 1: Ranked ΔLATER values for all German target words.

Parameters of change For each language and each target word, we compute ΔLATER, COMPARE and ΔCOMPARE. The BERT embeddings for each target word per time period are based on a maximum number of 500 sentences.[9] The binary classification task is addressed via ΔLATER since the metric measures changes in the mean relatedness of words over time (Schlechtweg et al., 2018). The target words are ranked according to ΔLATER and we take the top ranked target words, i.e., the highest positive and lowest negative values, to undergo LSC, see, e.g., Figure 1 for German. The thresholds for the binary classification were experimentally defined on the basis of these ranks and vary across languages. That is, we plotted the ranked target words as shown in Figure 1 and defined the thresholds on the basis of the points in the plot where the distribution begins to become skewed to the left and right respectively. The results for the ranking task are based on the absolute ΔCOMPARE values, i.e., the normalized version of the COMPARE measure which differentiates between weak and strong change (Schlechtweg et al., 2018), thus measuring the degree of change. Moreover, we calculate the standard deviation (*sd*) of cosine distances in the earlier and in the later group to provide a measure of the context variation in each corpus.

Evaluation In SemEval-2020 Task1, the system is evaluated with respect to its performance on the binary classification and the ranking task. The evaluation of the binary classification output is based on accuracy measured against the true binary classification as annotated by humans. The output of the system for the ranking task is evaluated using Spearman's rank correlation coefficient (ρ) by calculating the correlation between the produced values and the true ranks as annotated by humans. For a detailed description of the gold data please see Schlechtweg et al. (2020).

5 Results

DiaSense has been significantly improved after the publication of the ground truth for SemEval-2020 Task 1. Before this, the system showed a comparably low performance in the evaluation phase, with an average accuracy of 0.554 in the classification task (rank 17 of 21) and $\rho = 0.234$ in the ranking task (rank 14 of 21); see Schlechtweg et al. (2020) for the full rankings. We attribute the low performance in the binary classification to two factors: For one, $k = 2$ might not have been the optimal parameter for KMeans clustering for all target words. We experimented with approaches to automatically determine k, but did not arrive at a suitable solution. For another, cluster initialization turned out to be difficult with the pre-trained BERT embeddings, since distances between the embeddings are generally low (cf. Reimers and Gurevych (2019) on clustering issues with pre-trained BERT embeddings). In addition, the frequency threshold employed for identifying change in the clustering was arbitrarily defined. The low performance in the ranking task could be the result of averaging the embeddings, where context variation

[9] This number was sufficient to reliably model the metrics, while maintaining a manageable computing time. However, occurrence frequencies of target words vary across target words and corpora. For example, in the German C_2, *Entscheidung* 'decision' occurs over 8 000 times, *Ohrwurm* 'earwig/catchy tune' appears only roughly 100 times.

might substantially bias the resulting vectors. Given these shortcomings, we decided to opt for alternative ways of measuring LSC, experimenting with ΔLATER and ΔCOMPARE instead.

Currently (post-evaluation), the system ranks third in the binary classification with an overall average accuracy of 0.728 (English: 0.649, German: 0.771, Latin: 0.750, Swedish: 0.742).[10] In the ranking task, DiaSense occupies rank 14 with $\rho = 0.337$ (English: 0.293, German: 0.414, Latin: 0.343, Swedish: 0.300). Exemplary, we discuss the results for German in the following.[11]

For German, the words with the highest positive ΔLATER (innovative meaning change) are *abgebrüht* 'boiled out/indifferent', *abdecken* 'cover/unroof/blanket', *Tier* 'animal', *Festspiel* 'festival', *abbauen* 'win (mining)/reduce', *artikulieren* 'articulate/enunciate', *Titel* 'title' and *Rezeption* 'reception' (see Figure 1-left). In the ground truth, *abgebrüht, Tier, Festspiel,* and *Titel* are not classified as change. We can confirm *Tier* and *Titel* as false positive. *Tier* and *Titel* show high standard deviations in both t_1 and t_2 (with $sd >$ 0.04), indicating that they exhibit a high context variation overall instead of undergoing a meaning change. However, standard deviation does not provide insights into whether *Festspiel* and *abgebrüht* are false positives. Instead, *Festspiel* shows a large difference between t_1 and t_2 based on a frequency effect (51 occurrences in C_1, > 500 occurrences in C_2). Yet, since data sparsity is an inherent problem of historical corpora, we can not exclude *Festspiel* as easily. Moreover, *abgebrüht* indeed shows LSC on the basis of our data: In t_1, *abgebrüht* is almost exclusively used as participle of the verb *abbrühen* 'boil out', while it occurs mostly as adjective with the more figurative meaning 'indifferent' in t_2.

The target words *überspannen* 'span/overstretch/straddle', *Fuß* 'foot', *Abgesang* 'last verse (minnesong)/swansong', *Schmiere* 'grease/lookout', *Knotenpunkt* 'junction/intersection', *Ohrwurm* 'earwig/catchy tune' have the lowest negative values (reductive meaning change), see Figure 1-right. Similarly to *abgebrüht*, *Fuß* is not classified as undergoing LSC in the gold data, but can in principle be identified as change: While in t_1 *Fuß* is still frequently used as measure unit, this meaning only occurs scarcely in t_2. Overall, the system performs well when it comes to the identification of large changes. For example, *Ohrwurm*, which shows the highest (absolute) ΔLATER value in German, changed quite drastically from being mainly used in the meaning of 'earwig' in t_1 to almost exclusively denoting a 'catchy tune' in t_2. However, the system fails to identify smaller scale changes such as, e.g., *Manschette* 'sleeve/cuff', where meanings are close and occur in similar contexts.

DiaSense performs less well in the ranking task as in the classification. However, although the correct ranking could not be identified, DiaSense puts the target words into similar regions, i.e., words with high ΔCOMPARE values (e.g., *Ohrwurm*, *abgebrüht*) generally rank high, indicating strong change and vice versa. Moreover, without normalizing COMPARE, *Titel* and *Tier* ranked highest – an error which was avoided by using ΔCOMPARE instead.

6 Conclusion

In this paper, we presented DiaSense, a system developed for SemEval-2020 Task 1. Based on contextualized word embeddings, DiaSense is able to identify change in English, German, Latin and Swedish, while also differentiating between weak and strong change. Our approach leverages the strength of pre-trained BERT embeddings for modeling word senses language-independently and avoids the necessity of large amounts of training data which is beneficial for historical linguistic work. Moreover, we were able to translate metrics developed to capture human intuitions about meaning relatedness into automated measures of LSC. Still, DiaSense was not able to predict the correct ranking in terms of degrees of LSC. We will address this issue in future research, experimenting with further measures and techniques.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Project D02). We thank Aikaterini-Lida Kalouli for the detailed and helpful feedback on this work.

[10] The organizers decided to leave the submission board open (more participants): `https://competitions.codalab.org/competitions/20948#results`. The system has been submitted via team 'cbk'. Last access: 21st July, 2020.

[11] This is the author's native language.

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical american english. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC)*, Marseille. European Language Resources Association (ELRA).

Yvonne Asedam, Dana Dannells, and Nina Tahmasebi. 2019. Exploring the quality of the digital historical newspaper archive KubHist. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN)*, pages 9–17, Copenhagen, Denmark.

David Bamman and Gregory Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, page 1–10, New York, NY, USA. Association for Computing Machinery.

Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Niemeyer, Tübingen.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Deutsches Textarchiv. 2017. *Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften.*

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*.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy, July. Association for Computational Linguistics.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Johannes Hellrich and Udo Hahn. 2017. Exploring diachronic lexical semantics with JeSemE. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Christian Kahmann, Andreas Niekler, and Gerhard Heyer. 2017. Detecting and assessing contextual change in diachronic text documents using context volatility. *Proceedings of the 9th International Joint Conference on Knowledge Discovery*, pages 135–143.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France, April. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transaction of the Association for Computational Linguistics*, 3:211–225.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. In *New Methods in Historical Corpus Linguistics*. Narr, Tübingen.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw–Hill Book Company, New York.

Dominik Schlechtweg and Sabine Schulte im Walde. 2020. Simulating lexical semantic change from sense-annotated data. In A. Ravignani, C. Barbieri, M. Martins, M. Flaherty, Y. Jadoul, E. Lattenkamp, H. Little, K. Mudd, and T. Verhoef, editors, *The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13)*.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana, June. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy, July. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *To appear in Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, abs/1811.06278.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

Elizabeth Closs Traugott and Richard B. Dasher. 2001. *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge University Press.

Elizabeth Closs Traugott. 2006. Semantic change: Bleaching, strengthening, narrowing, extension. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 124–131. Elsevier.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of KONVENS 2019*, Erlangen, Germany.

Han Xiao. 2018. bert-as-service. `https://github.com/hanxiao/bert-as-service`.