Unsupervised Embedding-based Detection of Lexical Semantic Changes

Ehsaneddin Asgari[†]**, Christoph Ringlstetter**[†]**, Hinrich Schütze**[◇]

[†] NLP Expert Center, Data:Lab, Volkswagen AG, Munich, Germany
[◇] Center for Information and Language Processing, LMU Munich, Germany

{ehsaneddin.asgari, christoph.ringlstetter}@volkswagen.de
inquiries@cislmu.org

Abstract

This paper describes EmbLexChange, a system introduced by the "Life-Language" team for SemEval-2020 Task 1, on unsupervised detection of lexical-semantic changes. EmbLexChange is defined as the divergence between the embedding based profiles of word w (calculated with respect to a set of reference words) in the source and the target domains (source and target domains can be simply two time frames t_1 and t_2). The underlying assumption is that the lexical-semantic change of word w would affect its co-occurring words and subsequently alters the neighborhoods in the embedding spaces. We show that using a resampling framework for the selection of reference words, we can reliably detect lexical-semantic changes in English, German, Swedish, and Latin. EmbLexChange achieved second place in the binary detection of semantic changes in the SemEval-2020.

1 Introduction

SemEval 2020 Task 1 is defined on the unsupervised detection of word sense changes over time in German, English, Swedish, and Latin. In particular, this challenge focused on detection and quantification of the sense changes of word w in the transition from the time period t_1 to the time period t_2 in the above mentioned four languages, where the input for each language are the text corpora dating from t_1 and t_2. This challenge involved two sub-tasks:

i. Classification: The goal of the classification task is the binary detection of lexical semantic change (from t_1 to t_2) for the given word w.

ii. Ranking: This sub-task involves the ranking of lexical-semantic change for a given list of words $(w_1, w_2, \ldots, w_M)$ by assigning scores quantifying relative changes of the word senses.

To measure the two sub-tasks, the participating systems are evaluated against a ground truth corpus annotated by native speakers or scholars of the respective languages (Schlechtweg et al., 2020).

Human languages constantly change due to cultural, technological, and social drift. Lexical semantic changes of human languages can materialize in the form of introducing/borrowing new words, or for the existing words can involve acquiring/losing some word senses. Computational methods for automated detection of semantic changes can be extremely helpful in the study of historical texts or corpora spanning a very long period of time, e.g., in the design of the OCR algorithm for text digitization, or in designing an information retrieval system incorporating the semantic changes (Tahmasebi et al., 2018). Applications in the study of historical texts aside, the proposed methods detect lexical-semantic drift also in the same time period for different domains. This can be useful for compiling glossaries and specific training material in certain industries where new senses are introduced for words as compared to their standard usage e.g. to facilitate a more efficient training for new employees.

In the past decade, a variety of methods were introduced in the literature for automatic detection of lexical-semantic changes (Tahmasebi et al., 2018), where we only can refer to a subset of work, including but not limited to (i) co-occurrence-based methods (Sagi et al., 2009; Basile et al., 2016), (ii) embedding-based approaches (Kim et al., 2014; Kulkarni et al., 2015; Asgari and Mofrad, 2016; Asgari, 2019; Asgari et al., 2020), (iii) topic-models-based (Frermann and Lapata, 2016), and (iv) alignment-based (Bamman and Crane, 2011) approaches. In this paper, we extend our recently introduced DomDrift embedding-based approach for the detection of semantic changes (Asgari et al., 2020) introduced for the extension of the computational analyses on 1000+ languages (Asgari and Schütze, 2017). DomDrift works based on a comparison of relative distances of words in the embedding spaces of the source and the target domains. To increase the stability of detection in the DomDrift, we extend DomDrift to EmbLexChange by the following modifications: (i) instead of creating word profile against all common words between the source and target domain, we use only a subset of pivot words, which are frequent words with unchanged relative frequencies. (ii) We create multiple word profiles by resampling from a set of pivot words. We show that the EmbLexChange can reliably detect the lexical-semantic changes in English, German, Swedish, and Latin achieving an average accuracy of 0.686 as second best system of the competition where the first place system achieved an accuracy of 0.687.

2   System overview

Here we detail the steps of the EmbLexChange system, where the overview is depicted in Figure 1. The EmbLexChange framework is developed based on the following assumptions:
H1: frequent words change at slower rates (Hamilton et al., 2016). **H2:** the relative frequency of unchanged words is not dramatically different in different time periods/domains. **H3:** changes of the word sense change the context and consequently alter the neighbors in the embedding space. Thus, the relative drift of a query word (a words which we target to investigate its lexical semantic change) with respect to unchanged words in the embedding space can characterize the lexical-semantic change.
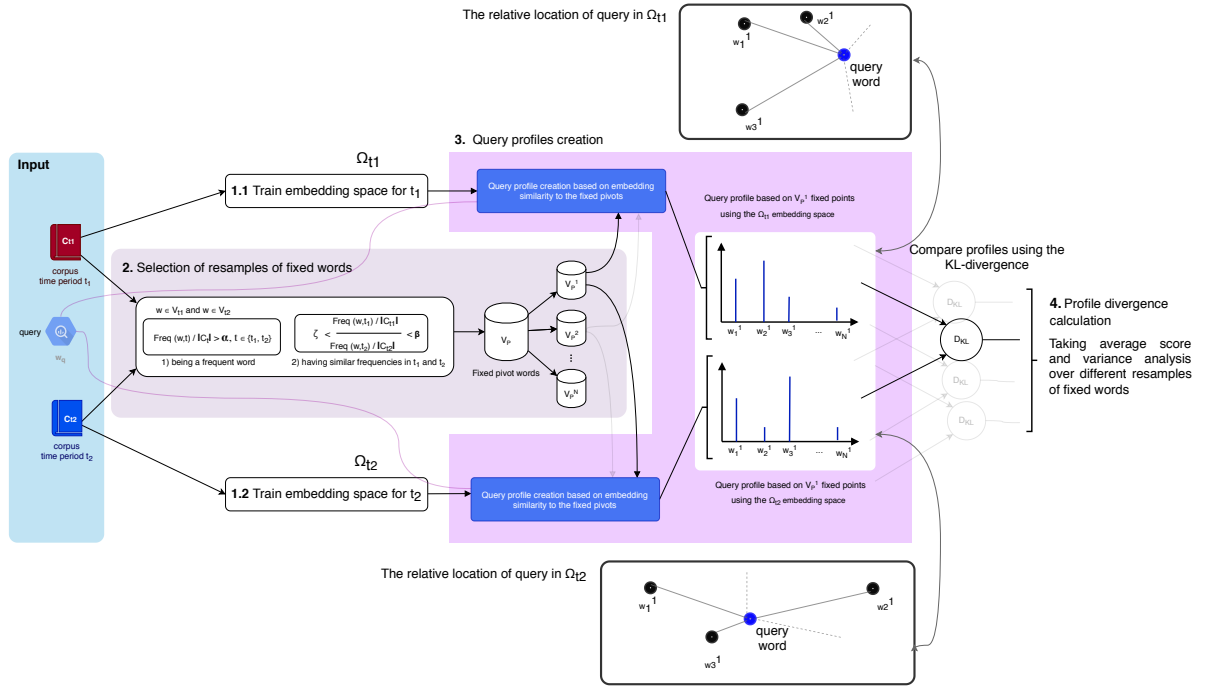

Figure 1:   The overview of EmbLexChange system for unsupervised detection of lexical-semantic changes. The steps are detailed in the §2.1.

2.1 EmbLexChange

1. Training language-model-based embedding spaces: The training of word embeddings using language modeling objective (e.g., skip-gram) has shown to preserve the syntactic and the semantic regularities in the vector space (Mikolov et al., 2013; Pennington et al., 2014). Semantic changes impact the neighborhoods in the embedding space ($H3$). Thus, the first step is to train embeddings separately for the text corpora in time period t_1 and t_2 (steps 1.1 and 1.2 in 1). In order to generate the embedding space Ω_t, the only necessary resource is the raw text. For embedding creation, we use `fasttext` (Bojanowski et al., 2017) which leverages subword information within the skip-gram architecture. Using sub-word information minimizes the query terms Out-OF-Vocabulary (Bojanowski et al., 2017). The result of this step would be separate embedding spaces Ω_{t_1} and Ω_{t_2} for the time periods t_1 and t_2, $\Omega_{t_x} : V_{t_x} \rightarrow \mathbb{R}^{h_{t_x}}, t_x \in \{t_1, t_2\}$ mapping their vocabulary $V_{l_{t_x}}$ to continuous vector representations in $\mathbb{R}^{h_{t_x}}$.

2. Selection of fixed words and prepare pivot sets: To measure the degree of semantic change for the given query words in Ω_{t_1} and Ω_{t_2}, we need some fixed points, called pivot set V_P comprising words with the property that their semantics are not dramatically changed and their relative positions in Ω_{t_1} and Ω_{t_2} remain almost constant (step 2 in 1). For this purpose, based on $H1$ and $H2$, we propose the use of frequent words with their relative frequency higher than α in both time periods of t_1 and t_2. Secondly, we filter this set by removing words whose relative frequency has changed between t_1 and t_2, resulting in V_P. These fixed points are then used to create query profiles in t_1 and t_2. In order to increase the reliability and make variance analysis possible, we execute N resamples each containing M words from V_P: $V_P^{(1)}, V_P^{(2)}, \ldots, V_P^{(N)}$.

3. Query profiles creation: In the next step, for each query word, we create t_1 and t_2 profiles based on the pivot resamples (step 3 in 1). The profile in time t is an $l1$ normalized embedding similarity vector of the query to the terms in $V_P^{(i)}$:

$$P(w_q, V_P^{(i)}, \Omega_t)_j = \frac{\exp{(\phi cos(\overrightarrow{w_q}, \overrightarrow{w_j^{(i)}}))}}{\sum_{w_k^{(i)} \in V_P^{(i)}} \exp{(\phi cos(\overrightarrow{w_q}, \overrightarrow{w_k^{(i)}}))}},$$

where $V_P^{(i)}$ is the ith resample from the pivot set (created in step 2), w_q is the query word, $w_k^{(i)}$ is the kth word in the ith resample, $\overrightarrow{w}$ is the vector representation of word w in the embedding space Ω_t, ϕ is the temperature of the softmax function (used as a hyper parameter). Hence, for each $V_P^{(i)}$ we can create one profile in t_1 and one profile in t_2.

4. Profile divergence calculation: Next, for each resample $V_P^{(i)}$ we calculate the divergence between the profile in the time period t_1 and the time period t_2 using KL-divergence:

$$\lambda_i = D_{\mathrm{KL}}(P(w_q, V_P^{(i)}, \Omega_{t_1}) \| P(w_q, V_P^{(i)}, \Omega_{t_2}))$$

We average the λ_i's over N resamples as the measure of semantic change for the query word w_q. Since D_{KL} does not have an upper-bound, we estimate an upper-bound based on λ_i's on resamples of a large set of randomly selected words $V_{explore}$ including V_P and a set of words with a change in their relative frequency. We draw K resamples of size M' words from $V_{explore}$ and calculate the λ_k 's (λ's of words in k^{th} resample of $V_{explore}$). We select the average of 90th percentile over K resamples as the upper bound and the average of 10th percentile as the lower bound of the λ to scale any calculated λ_i for a query word to $\hat{\lambda}_i$ ($0 \leq \hat{\lambda}_i \leq 1$). Considering a threshold of h, we assign $\bar{\lambda} > h$ to the category of lexical semantic change, which can be adjusted as a hyperparameter on a validation set.

3 Data

The dataset used in this shared task includes corpora of English, German, Latin, and Swedish texts. For each language, the text corpora of two time periods are given. More details on the on the extact time frames and data sizes are provided in Table 1.

Table 1: The English, German, Swedish, and Latin datasets used in the SemEval shared task.

Languages	t1: time period (# tokens)	t2: time period (# tokens)	Reference
English	1810-1860 (6.6M)	1960-2010 (6.8M)	CCOHA (Alatrash et al., 2020)
German	1800-1899 (70.2M)	1946-1990 (72.4M)	DTA, BZ, and ND corpora
Latin	-200-0 (1.8M)	0-2000 (9.4M)	LatinISE (McGillivray and Kilgarriff, 2013)
Swedish	1790-1830 (71.1M)	1895-1903 (110.8M)	KubHist (Adesam et al., 2019)

4 Experiment

The goal of SemEval task 1 is to detect the words with a change in their semantics in the transition from the time period t_1 to the time period t_2 in English, German, Swedish, and Latin languages. We closely follow the steps described in §2.

1. Language-model-based embedding setup: We train `fasttext` (Bojanowski et al., 2017) embedding using the skip-gram architecture for each pair of language and time period separately. In the training of `fasttext`, we set the window size to $c = 7$ and the embedding size of $d = 100$. In the presence of a validation set, both c and d can be optimized as the hyper-parameters for each setting. It is known that a larger c is favorable for semantics representation of words and a smaller c for a syntax-related representation.

2. Pivot resamples creation: We firstly prepare a set of frequent words existing in both t_1 and t_2 for each language considering the α (relative freq.) as a way to select the top $10\% - 20\%$ frequent words. Next, we filter this set to keep the words with the property that the ratio of their normalized frequency is not substantially changed in t_1 and t_2, $\frac{2}{3} < \frac{freq_{t1}(w)}{freq_{t2}(w)} < \frac{3}{2}$ resulting in our V_P set. Subsequently, we draw $N = 10$ resamples from V_P with the size of $M = 5000$ for each language.

3. Query profiles creation: In the next step, as presented in §2 step 3, for each query word we create t_1 and t_2 profiles for each $N = 10$ pivot resamples as in the previous step.

4. Profile divergence calculation: Next, for each of N resamples $V_P^{(i)}$, we calculate the λ_i and scale them to $\hat{\lambda}_i$ using $K = 5$ resamples of size $M' = 5000$ words from $V_{explore}$. Subsequently, for the binary detection of changes, we apply different thresholds h over the average of scaled divergences (the average of $\hat{\lambda}_i$'s = $\bar{\lambda}$) of $V_P^{(i)}$'s and assign $\bar{\lambda} > h$ to the category of lexical semantic change.

Evaluation: For evaluation purpose, in the case of binary detection the accuracy metric is used to compare the given ground-truth and the predicted lexical semantic changes. For the ranking setting, we report both pearson correlation coefficient and Kendall-τ to measure the correspondence between calculated divergences ($\bar{\lambda}$'s) and the provided ground-truth scores.

5 Results

The results of EmbLexChange in the detection and the quantification of changes in the lexical semantics in English, German, Swedish, and Latin are provided in Table 2. After the competition, we had the chance to perform further optimizations of the hyperparameters leading to the current results, slightly improved from those submitted to the competition leaderboard. The EmbLexChange scores of the test set for all languages are available at http://language-lab.info/emblexchange/.

Binary detection: EmbLexChange could detect the semantic changes in English, German, Swedish, and Latin with the accuracy of 70.3%, 75.3%, 77.4%, 60% respectively. The selected h value, the thresholds to assign the positive or the negative class for each language, are also provided in Table 2.

Ranking: The Kendall-τ p-values for English, German, and Swedish show that there is a significant correspondence between the EmbLexChange scores and the ground truth scores in those languages. The Pearson correlation is also calculated for all languages, with an average of 0.306 over four languages. The case of Latin has been more challenging in both binary and graded prediction of lexical semantic change.

Table 2: The summary of results for the detection and quantification of lexical-semantics changes in English, German, Swedish, and Latin in SemEval 2020.

Languages	Accuracy in binary detection	h	Pearson correlation	Kendall-τ	Kendall-τ-p-value
English	70.3	0.4	0.314	0.292	0.034
German	75.0	0.5	0.432	0.315	0.009
Swedish	77.4	0.4	0.316	0.383	0.0115
Latin	60.0	0.15	0.162	0.161	0.222
Average	70.7		0.306		

6 Discussion and Conclusions

In this paper, we proposed EmbLexChange, a framework for the detection of lexical semantic changes in an unsupervised manner. We defined EmbLexChange as the divergence between the embedding-based profiles of word w (calculated for a set of pivot words) in the source and the target domains (or between two time-frames). With the selection of pivot words by a resampling framework, we raise the reliability of this divergences. The underlying assumption of our method is that the changes in lexical semantics of word w would affect its co-occurring words and subsequently alters the neighborhoods in the embedding spaces.

We showed that EmbLexChange can reliably detect lexical-semantic changes in English, German, Swedish, and Latin achieving the second place in the binary detection of semantic changes in the SemEval-2020. The detection of semantic changes in Latin has been more challenging than other languages. One reason behind this can be the imbalance of embedding training instances for Latin t_1 and Latin t_2 as well as the overall smaller corpora for Latin in comparison to the other languages (shown in Table 1). Another reason can be the split of time frames, where t_2 in Latin spans a large period of 2000 years.

The SemEval overall results show that EmbLexChange works better in the binary detection of semantic changes versus its performance in the ranking problem setting (Schlechtweg et al., 2020). However, we should note that the manual creation of ranking ground truth is a much more challenging task than the creation of binary classification ground truth. Thus, we believe that the classification results might be more reliable than ones for the ranking.

The EmbLexChange requires only the raw texts in the time-frames/domains of interest. Then the semantic changes can be detected based on the divergence between the embedding-based profiles of words between the source and the target domain. One advantage of using an embedding-based profile is that by increasing the window size in the embedding training we can move from syntactic changes toward semantic changes which can be investigated in more depth as a future direction.

References

Yvonne Adesam, Dana Dannélls, and Nina Tahmasebi. 2019. Exploring the quality of the digital historical newspaper archive kubhist. In *DHN*, pages 9–17.

Reem Alatrash, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020. Clean corpus of historical american english (ccoha). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2020)*. European Language Resources Association (ELRA).

Ehsaneddin Asgari and Mohammad R.K. Mofrad. 2016. Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (weld) as a quantitative measure of language distance. In *In Proceedings of the NAACL-HLT Workshop on Multilingual and Cross-lingual Methods in NLP, San Diego, CA*, pages 65–74. Association for Computational Linguistics.

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ehsaneddin Asgari, Fabienne Braune, Benjamin Roth, Christoph Ringlstetter, and Mohammad RK Mofrad. 2020. Unisent: Universal adaptable sentiment lexica for 1000+ languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2020)*. European Language Resources Association (ELRA).

Ehsaneddin Asgari. 2019. *Life Language Processing: Deep Learning-based Language-agnostic Processing of Proteomics, Genomics/Metagenomics, and Human Languages*. Ph.D. thesis, UC Berkeley.

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*, pages 1–10.

Pierpaolo Basile, Annalina Caputo, Roberta Luisi, and Giovanni Semeraro. 2016. Diachronic analysis of the italian language exploiting google ngram. *CLiC it*, page 56.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA, June. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of latin. *New Methods in Historical Corpus Linguistics*, (3):247–257.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111. 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 lexical semantic change. *arXiv preprint arXiv:1811.06278*.