TemporalTeller at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection with Temporal Referencing

Jinan Zhou *, **Jiaxin Li** *
Department of Computer Science and Engineering
The Chinese University of Hong Kong
{jinan.zhou, jiaxin.li}@link.cuhk.edu.hk

Abstract

This paper describes our TemporalTeller system for SemEval Task 1: *Unsupervised Lexical Semantic Change Detection* (Schlechtweg et al., 2020). We develop a unified framework for the common semantic change detection pipelines including preprocessing, learning word embeddings, calculating vector distances and determining threshold. We also propose Gamma Quantile Threshold to distinguish between changed and stable words. Based on our system, we conduct a comprehensive comparison among BERT, Skip-gram, Temporal Referencing and alignment-based methods. Evaluation results show that Skip-gram with Temporal Referencing achieves the best performance of 66.5% classification accuracy and 51.8% Spearman's Ranking Correlation.

1 Introduction

Task 1 in SemEval 2020 is Unsupervised Lexical Semantic Change Detection. Semantic change is defined as the changes in lexical meaning rather than grammatical usage of a language symbol (Bloomfield, 1933). It has become an increasingly common phenomenon under the influence of linguistic habits, social environments and many other factors. For example, the word *gay* previously meant being happy but now mostly means homosexual. Those changes are indicators of temporal ideologies as well as language development trends. Therefore, studying the evolution of word meanings can assist the research in human language, historical information retrieval and arguably other disciplines.

Recently, there have been many studies on this topic. One direction is word level change detection. Those approaches seek to represent each word as a single vector that reflects its overall meanings. Some works (Sagi et al., 2009; Gulordava and Baroni, 2011; Rodda et al., 2016; Kahmann et al., 2017) use co-occurrence matrices with SVD. Others (Kim et al., 2014; Rosenfeld and Erk, 2018; Hamilton et al., 2016a) adopt neural embeddings such as Skip-gram with negative sampling (SGNS) (Mikolov et al., 2013a). However, those word embeddings need to be aligned to the same vector space, which is shown to introduce noise (Dubossarsky et al., 2019). Another direction is detecting semantic changes from the sense level. These models investigate the different meanings of words, thus can compare in a fine grained manner. In this paper, we will focus on word level change detection.

We spotted two problems among those related works. First, no classification mechanisms have yet been discussed. All papers calculate the lexical semantic change, but none has proposed a systematic approach to classify between changed and stable words. The second problem is that the models may be potentially out-of-date. While the leaderboards of many NLP tasks such as reading comprehension are continuously refreshed by new models like Pre-training of Deep Bidirectional Transformers (BERT) (Devlin et al., 2018), the methods adopted in this topic are relatively conventional. Though it is not necessarily a problem, we are curious to know whether the latest models could bring extra benefits.

Aiming at the first problem, we propose Gamma Quantile Threshold (GQT). We propose to describe the cosine distances between word representations using Gamma distribution. The threshold for classification will be set as the quantile of the distribution. For the second problem, we introduce contextualized

* equal contributions

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 222–231
Barcelona, Spain (Online), December 12, 2020.

embeddings using BERT (Devlin et al., 2018). As a comparison, we also leverage temporal referencing (Dubossarsky et al., 2019), a promising method in this field. Finally, we designed TemporalTeller, a framework that integrates learning word representations, alignment and threshold calculation for semantic change detection. It can learn word representations via SGNS and BERT with or without temporal referencing. Experiment results show that GQT works well with the existing word embeddings methods, and temporal referencing with SGNS achieves the best performance in both classification and ranking subtasks.

Our contributions can be summarized as follows:

- We propose Gamma Quantile Threshold, which is the first algorithm to fill the gap between quantifying and classifying lexical semantic change.
- We introduce contextualized embeddings using BERT (Devlin et al., 2018) to the semantic change detection problem. We conducted a comprehensive comparison between BERT and other methods.
- We designed TemporalTeller, a framework that integrates the common semantic change detection pipelines. Under this framework, we performed a large scale study on the effects of hyperparameters.

2 System Overview

2.1 Temporal Referencing (TR)

Temporal Referencing(TR) was created for term extraction (Ferrari et al., 2017), and later proven useful for lexical change detection (Dubossarsky et al., 2019). The idea is to tag each target word to indicate which corpus it comes from. For this task in particular, we add a suffix *_new* and *_old* to the target words from the modern and ancient corpus respectively. Then the word embeddings are learned from the merged corpus that includes both the modern and ancient one. After learning, the embeddings of *target_new* and *target_old* are in the same vector space, thus can be compared directly. TR is based on the assumption that the meanings of the majority of context words are stable, so that the change of target words can be highlighted in this way.

TR brings several advantages. First, TR eliminates alignment and thus reduces noise. Second, TR lowers the data requirement because all corpora can contribute to the word representations collaboratively. Finally, it is widely compatible with any embedding based methods. It can also be extended to the scenarios with multiple corpora from multiple periods.

2.2 Word Embeddings

2.2.1 Skip-Grams with Negative Sampling (SGNS)

SGNS tries to maximize the similarity between co-occurring words while minimize that of non-co-occurring words by optimizing

$$\arg\max_{v_w, v_c} \sum_{(w,c)\in D} \log \sigma(v_w \cdot v_c) + \sum_{(w',c')\in D'} \log \sigma(-v'_w \cdot v'_c) \tag{1}$$

where σ is the sigmoid function, v is the word vector, D is the whole set of word-context pairs, and D' is the randomly picked negative word-context pairs (Levy and Goldberg, 2014).

2.2.2 BERT Contextualized Embedding

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is a powerful language model trained on unlabeled text, and then fine-tuned for downstream tasks with a classifier added. As suggested by its authors, we use the sum of the last 4 layers to encode both word meaning and context information. For each target word, we input each occurrence with context to obtain its contextualized embedding. The representation of a word is the average of the embedding of all its occurrences. In order to reduce the noise from context, we propose to set a context window and remove stop words and rare words inside.

2.3 Alignment with Orthogonal Procrustes (OP)

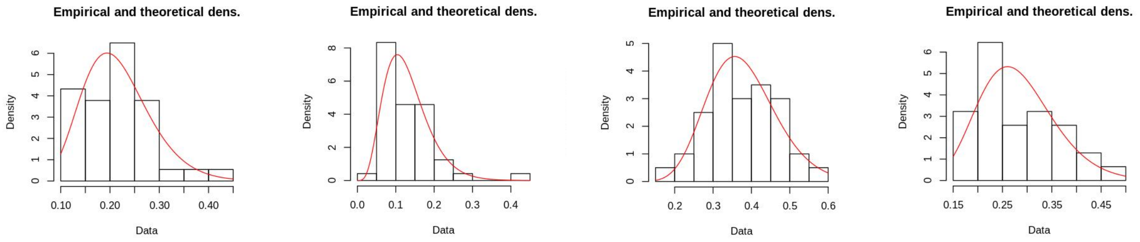
Orthogonal Procrustes (Hamilton et al., 2016b) is also based on the stable context assumption. OP maps word embeddings learned at year a, V^a, into V^aW such that $||V^aW - V^b||^2$ is minimized:

$$W^* = \underset{W^TW=I}{\arg\min} \sum_i \sum_j R_{i,j} ||V_{a,*i}W - V_{b,*j}||^2 \tag{2}$$

where R is a binary matrix with $R_{i,j} = 1$ if and only if the i^{th} word in the first corpus is the j^{th} word in the second one. The solutions can be obtained by first performing SVD $V_b^T RV_a = U\Sigma V^T$ and then $W^* = UV^T$ (Schlechtweg et al., 2019; Schönemann, 1966).

2.4 Gamma Quantile Threshold

We observe that the cosine distances (CD) between old and new word vectors of target words resemble Gamma distribution: the majority of data lies together, forming a peak, while larger CDs form a long tail at the right side.



(a) English (b) German (c) Latin (d) Swedish

Figure 1: Fitting gamma distribution on CD for 4 languages in Section 4.2

We assume those at the peak are stable, while those at the tail are changed. Thus, we propose to compute a global threshold by fitting the CDs to a Gamma distribution using maximum likelihood estimation. Based on the experiment results, we choose the 75% density quantile as the threshold. In Gamma distribution, 75% quantile cuts at its long right tail.

Different from the quantile of data, quantile of distribution relies less on the selection of target words. In Figure 1a, for example, the number of large CDs lying at the tail is clearly smaller than what its Gamma density function expects. As a result, the fraction of English target words classified as changed is the least among the four languages.

2.5 System Framework

Figure 2 shows our TemporalTeller framework. The system integrates TR preprocessing, learning word embeddings, alignment, calculating distance and threshold.

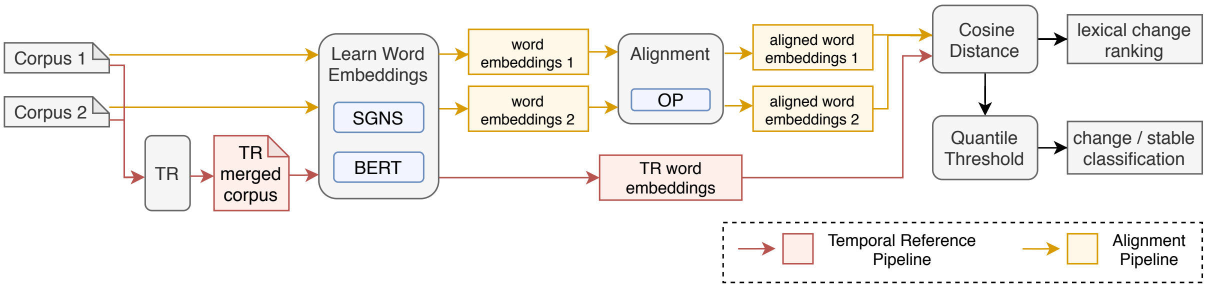


Figure 2: Our system framework integrates the common semantic change detection pipelines

3 Experiment Setup

3.1 The SemEval 2020 shared task

The dataset of this task covers English, German, Swedish and Latin (Schlechtweg et al., 2020). For each language, there are two corpora collected from ancient and modern documents respectively. The details are shown in Table 1. There are two subtasks. The first is distinguishing between change and stable words. It is evaluated by the accuracy (ACC) against the truth annotated by human. The second is ordering the target words with respect to their extent of change. It is evaluated by Spearman's rank-order correlation coefficient (SPR) against the manually annotated rankings. We use the average performance over the four languages as the formal measure for each subtask.

Languages	Ancient Period	Modern Period	Number of Target Words
English	1810 - 1860	1960 - 2010	37
German	1800 - 1899	1946 - 1990	48
Latin	-200 - 0	0 - 2000	40
Swedish	1790 - 1830	1895 - 1903	31

Table 1: Information about the corpora in SemEval 2020 task 1

3.2 Hyperparemeters and Implementation Details

Our implementation of SGNS with OP and TR is based on the repository of Dubossarsky et al (2019)[1]. We set the context distribution smoothing factor $\alpha = 0.75$ and number of negative samples $k = 5$ as suggested by previous works (Mikolov et al., 2013b; Levy et al., 2015). The word embeddings are trained for 10 epochs with the words occurring less than twice omitted. We will study the effects of different context window sizes l and embeddings dimensions d. Our implementation of BERT is based on the Transformers library (Wolf et al., 2019)[2]. Pretrained BERT base models are used for each language and are fine tuned on the SemEval dataset for 10 epochs. We retain the original embedding size $d = 768$ in BERT.

4 Experiment Results[3]

4.1 Experiment 1: TR with SGNS

Dubossarsky et al.(2019) suggests that SGNS+TR introduces less noise than PPMI+TR, SGNS+OP and PPMI+CI. Therefore, we firstly test this method and perform a hyperparameter search. We study the effects of two hyperparameters: context window size l (number of words in each direction) and word embedding dimension d. Specifically, we test $l \in \{2, 5, 10\}$ and $d \in \{50, 100, 150, 200, 300, 400\}$.

l	10						5			2		
d	50	100	150	200	300	400	50	100	150	50	100	150
ACC	0.663	**0.713**	0.676	0.623	0.600	0.590	0.696	0.649	0.573	0.662	0.625	0.600
SPR	0.430	0.444	0.415	0.396	0.421	0.386	**0.499**	0.488	0.445	0.464	0.450	0.425

Table 2: Performance of TR + SGNS with different context sizes and embedding dimensions

From Table 2, we can see that SGNS+TR achieves decent result, but is relatively sensitive to hyperparameters. It yields the best classification accuracy when $d = 100, l = 10$ and best SPR when $d = 50, l = 5$. This shows that by looking at a wider range of neighbours with a slightly larger embedding dimension, SGNS could better distinguish stable words from changed ones. In comparison, a medium context window $l = 5$ lead to better ranking. It also shows that $l = 2$ may miss some useful contexts while

[1] https://github.com/Garrafao/TemporalReferencing

[2] https://github.com/huggingface/transformers

[3] Please be noted that the results presented in this section are retrieved from post-evaluation unless otherwise specified.

$l = 10$ may include too many irrelevant contexts that impacts ordering. This experiment suggests that the setting of fixing d at 300 in some previous works (Schlechtweg et al., 2019; Hamilton et al., 2016b; Dubossarsky et al., 2019) may be suboptimal. It also contrasted the finding that $l = 2$ yields the best results (Levy et al., 2015).

Under the optimal hyperparameters learned above, we compare performance of SGNS+OP and SGNS+TR. Results in Table 3 show that TR is a clear winner. It also confirms that alignment introduces extra noise, and eliminating alignment yields better performance.

Methods	SGNS + OP		SGNS + TR	
Hyperparameters	d=50, l=5	d=100, l=10	d=50, l=5	d=100, l=10
ACC	0.659	0.652	0.696	**0.713**
SPR	0.430	0.444	**0.499**	0.444

Table 3: Comparison of SGNS+OP and SGNS+TR

4.2 Experiment 2: TR with BERT

We propose to use BERT with two add-ons in this task. The first one is contextualization. Normally, the sum of the last 4 layers of the fine-tuned BERT is used as the word embedding. To further exploit the context information, we propose to input each occurrence of a target word along with its neighbors into the fined-tuned model and retrieve the embedding of that particular instance. Then the overall word embedding is obtained by averaging its all instance embeddings. Furthermore, we propose to remove unimportant words from the context window for more concise and refined information. For all languages, we define unimportant words as the those occur less than 30 times in the merged corpus and those are stop words according to Natural Language Toolkit (Bird and Loper, 2004).

One thing to notice is that temporal referencing is still meaningful when using BERT with contextualization. With the temporal tags, all the target words are treated as brand-new words whose representations are to be learned from scratch. It blocks the impact of the semantic information from BERT pre-training materials (which are mostly modern corpus), thus better highlights the difference between ancient and modern senses.

After BERT is fine-tuned on the merged corpus, experiment (a) follows the normal setting without any add-ons. In experiment (b), we add contextualization. In (c), we further remove unimportant words from the contexts at the inference stage.

Experiment Index	a	b	c
Contextualization?	✗	✓	✓
Unimportant word removal?	✗	✗	✓
ACC	0.617	0.657	**0.664**
SPR	0.327	0.359	**0.392**

Table 4: Performance of TR + BERT with ablation study

Table 4 presents a notable improvement from (a) to (b), which proves the effectiveness of contextualization. The comparison between (b) and (c) reveals that removing unimportant words also helps. However, BERT is not as good as SGNS+TR. Reasons might include the fact that SGNS is more focused on generating word vectors, which gives it a privilege in terms of semantic representation. It is also noticed that fine tuning BERT on this dataset is anything but easy. On one hand, the target words with temporal referencing tags are new vocabularies that need more iterations to learn. On the other hand, the corpora are not suitable for BERT as they are noisy and relatively short. According to our experiments, overfitting BERT model on this dataset can lead to catastrophically worsened performance. As a result, it is hard to balance between learning new words and reusing the linguistic information in pretrained models.

In addition, we also test the quality of thresholds under different quantiles. It is observed again that 75% is the optimal choice.

quantile	0.65	0.70	0.75	0.80
ACC average	0.658	0.658	**0.664**	0.645

Table 5: Comparison of different quantiles in experiment (c)

4.3 Performance and Ranking

Table 6 summaries the scores and rankings of our system by the time the paper is submitted.

Phase	Evaluation				Post-Evaluation			
Subtask	Classification		Ranking		Classification		Ranking	
Language	ACC	Rank	SPR	Rank	ACC	Rank	SPR	Rank
Average	0.66	3	0.52	2	0.713	4	0.499	4
English	0.65	4	0.32	5	0.676	3	0.295	14
German	0.73	4	0.72	2	0.771	3	0.735	2
Latin	0.70	1	0.44	3	0.600	5	0.388	10
Swedish	0.58	19	0.59	2	0.806	1	0.579	5

Table 6: Evaluation Results and Ranking in SemEval 2020 Task 1

5 Conclusion

In this paper, we systematically studied unsupervised lexical semantics change detection from a computational perspective under our TemporalTeller framework. We proposed the Gamma Quantile Threshold, which is the first algorithm to fill gap between quantifying semantic change and classifying change status. We exploited the latest advances in temporal referencing. Furthermore, we introduced BERT into this semantic change detection task. We conducted a comprehensive comparison among these methods in terms of hyperparameters and performance. We found that the combination of SGNS with temporal referencing yields the best performance. However, it is sensitive to hyperparameters including embedding dimension and context window size. A suboptimal setting could lead to considerable performance degeneration. BERT can give comparable results with base model hyperparameters, but the uncontrollable training and a diverged focus prevents it from being the champion. Even with contextualized embeddings and unimportant words removal, it fails to prevail over SGNS with TR. It shows that the advantages of powerful general language models like BERT does not naturally extends to this task.

Acknowledgements

The authors would like to thank Prof. Leung Kwong-Sak, Prof. Lam Wai and Sunny Lai from The Chinese University of Hong Kong for providing helpful discussions throughout the project. We are also deeply grateful to the task organizers and reviewers for their constructive comments.

References

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July. Association for Computational Linguistics.

Leonard Bloomfield. 1933. *Language*. The University of Chicago Press.

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.

Alessio Ferrari, Beatrice Donati, and Stefania Gnesi. 2017. Detecting domain-specific ambiguities: an nlp approach based on wikipedia crawling and word embeddings. In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, pages 393–399. IEEE.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 67–71.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. 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.

Christian Kahmann, Andreas Niekler, and Gerhard Heyer. 2017. Detecting and assessing contextual change in diachronic text documents using context volatility. *arXiv preprint arXiv:1711.05538*.

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.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

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

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Martina Rodda, Marco Senaldi, and Alessandro Lenci. 2016. Panta rei: Tracking semantic change with distributional semantics in ancient greek. In *Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 258–262. Accademia University Press.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. 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 474–484.

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, 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.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Appendix

A A case study in SGNS + TR

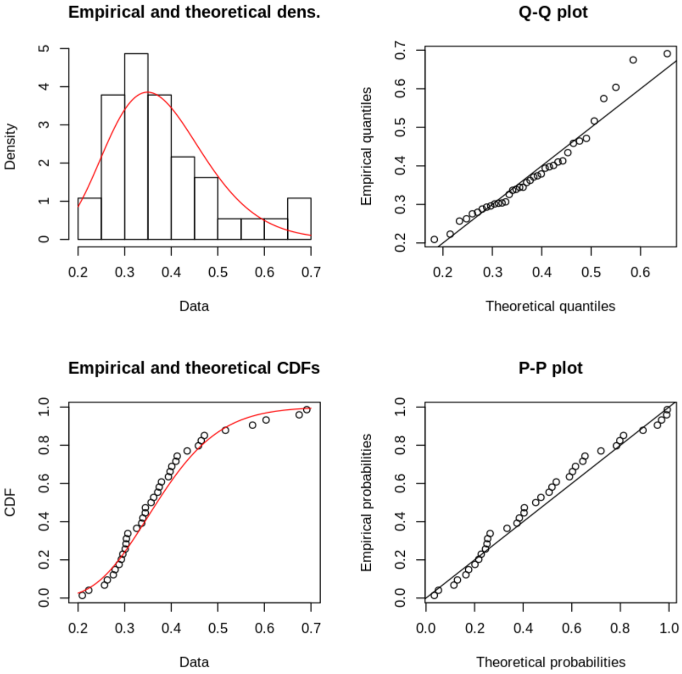


Figure 3: Fitting CDs to Gamma distribution in English (SGNS + TR, l=10, d=100)

Figure 3 shows that Gamma distribution fits the cosine distances of English words very well in SGNS + TR. Figure 4 presents a case study of the target word *risk_nn_old* and *risk_nn_new* projected to 2D space using Principal Component Analysis (PCA). We can see that the distance between *risk_nn_old* and *risk_nn_new* is clearly larger than that between *face_nn_old* and *face_nn_new*. Indeed, the cosine distance of *risk_nn* is 0.571 and CD of *face_nn* is 0.321. The former is classified as changed and the latter stable. Looking into their nearest neighbours, we can find that *risk_nn* moves closer to medical concepts such as *arrythmia*, *diabetes*, *premenopausal* and *lumpectomy*, while *face_nn* sticks to the meaning of *front of head*.

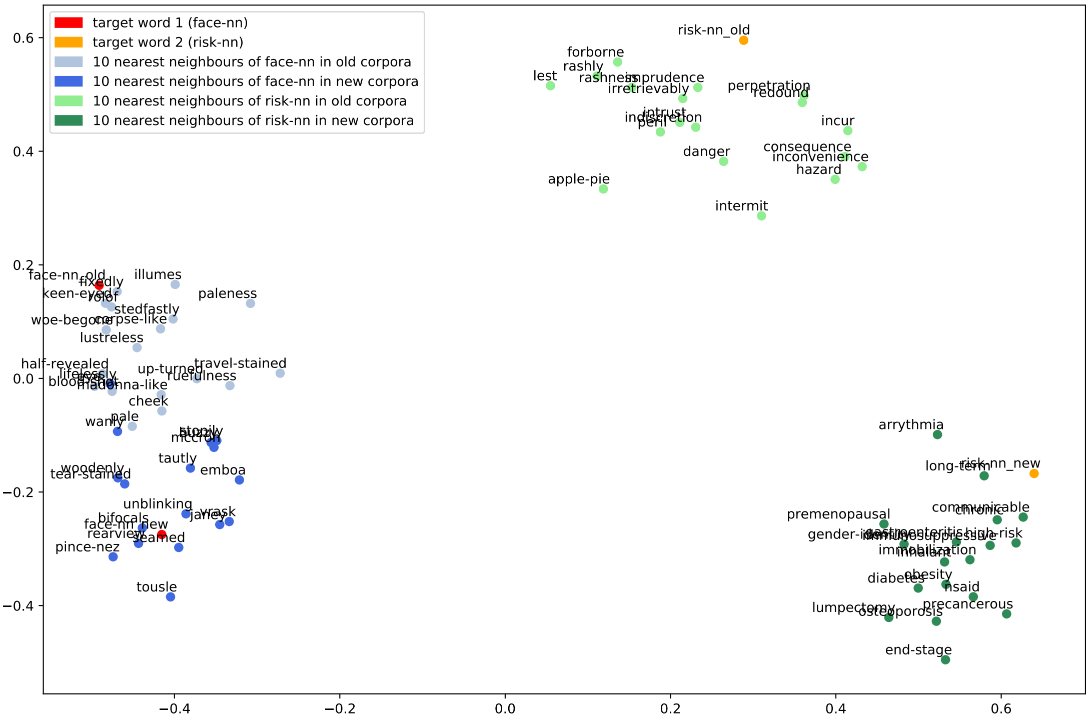


Figure 4: Visualizing *face_nn* and *risk_nn* and their 10 nearest neighbours in English in SGNS + TR

B A case study in BERT + TR

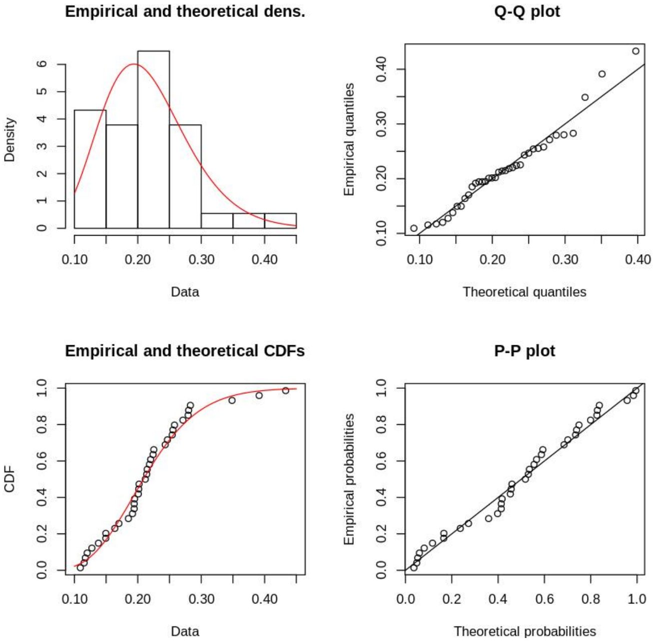


Figure 5: Fitting CDs in BERT experiment (c) to Gamma distribution in English

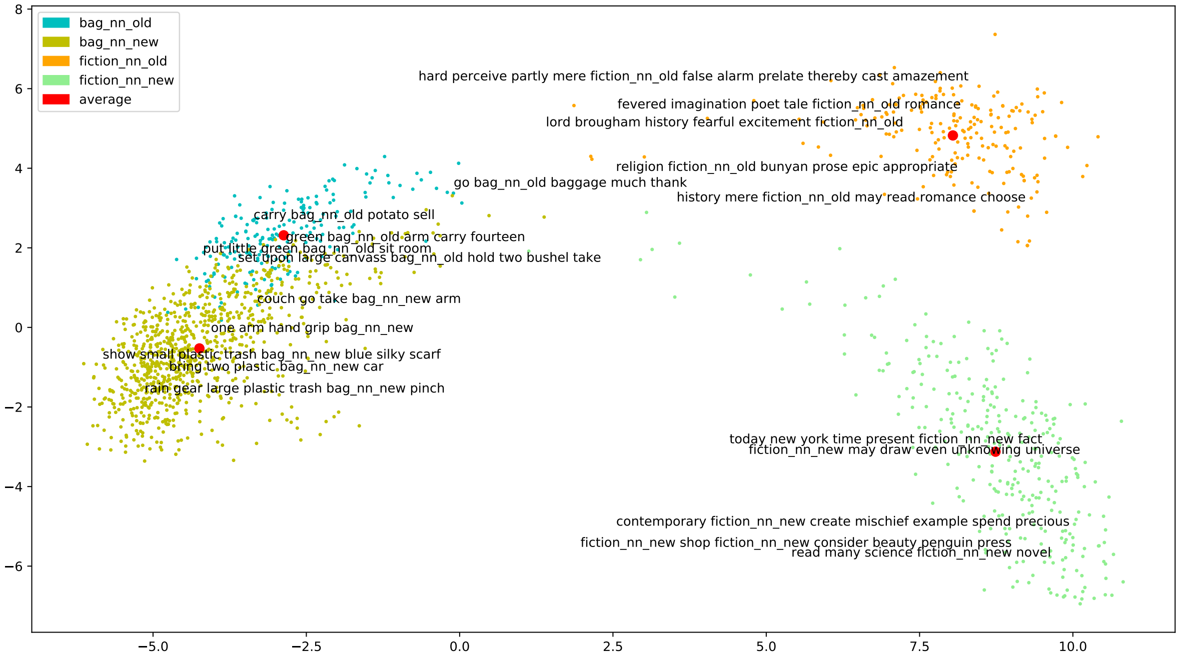


Figure 6: Visualizing *bag_nn* and *fiction_nn* in BERT experiment (c)

Figure 5 shows Gamma distribution also fits the cosine distances very well in BERT + TR. In Figure 6 (embeddings projected by PCA), the distance between *bag_nn_old* and *bag_nn_new* is smaller and their contextualized embeddings overlap. Looking at their contexts, we can see both *bag_nn_old* and *bag_nn_new* refer to *a container used for carrying things*. Some relatively faraway occurrences of *bag_nn_new* is about *plastic*, which dose not overstep too much. On the contrary, the distance between *fiction_nn_new* and *fiction_nn_old* is much larger. For *fiction_nn*, its modern meaning is close to *science fiction* and *contemporary magazines*, while its ancient meaning is more related to *romantic*. Indeed, *bag_nn* is classified as stable and *fiction_nn* changed.