RIJP at SemEval-2020 Task 1: Gaussian-based Embeddings for Semantic Change Detection

Ran Iwamoto
Keio University
Yokohama, Japan
iwamoto@ykw.elec.keio.ac.jp

Masahiro Yukawa
Keio University
Yokohama, Japan
yukawa@elec.keio.ac.jp

Abstract

This paper describes the model proposed and submitted by our RIJP team to SemEval 2020 Task1: Unsupervised Lexical Semantic Change Detection. In the model, words are represented by Gaussian distributions. For Subtask 1, the model achieved average scores of 0.51 and 0.70 in the evaluation and post-evaluation processes, respectively. The higher score in the post-evaluation process than that in the evaluation process was achieved owing to appropriate parameter tuning. The results indicate that the proposed Gaussian-based embedding model is able to express semantic shifts while having a low computational complexity.

1 Introduction

Words change their meaning over time. Long- and short-term semantic shifts relate to cultural and social changes (Blank and Koch, 1999; Grzega and Schöner, 2007; Garg et al., 2018). Semantic shifts have been actively studied as evidenced by recent comprehensive survey papers (Tahmasebi et al., 2018; Kutuzov et al., 2018) and the success of the workshop on lexical-semantic change held as part of the ACL 2019. In particular, many teams participated in the SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2014).

In this task, participants are required to devise an automated system for finding semantic shift in four languages: English, German, Swedish, and Latin. It consists of two subtasks: classification and ranking. In the classification task, words are categorized by the change in their meaning in two corpora, C_1 and C_2, for periods t_1 and t_2. In the ranking task, words are sorted by the degree of their semantic shift between periods t_1 and t_2.

The focus of this study is on detecting semantic shifts using word embedding. To tackle this task, our RIJP team represented words by Gaussian distributions, taking an inspiration from Gaussian embedding (Vilnis and McCallum, 2015). In Gaussian embedding, words are represented by one Gaussian distribution with mean vectors and covariance matrices. The mean vectors map the words in embedding space, while the covariance matrices represent about word hierarchies.

Instead of obtaining the variance using a trained model as suggested by existing studies, it is derived directly from the word frequency. It particular, two embeddings are created using each of the two corpora, and the semantic change is calculated based on the Kullback-Leiblar (KL) divergence. The proposed model achieved scores of 0.51 and 0.14 in the evaluation period for Subtasks 1 and 2, respectively. After parameter tuning in the post-evaluation process, scores of 0.70 and 0.41 were achieved in Subtask 1 and 2. Experimental results demonstrated that the proposed Gaussian-based embedding model can successfully detect various types of semantic shifts such as narrowing and widening. The code of the model will be releasing shortly.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 98–104
Barcelona, Spain (Online), December 12, 2020.

2 Related Work

Word senses generate and disappear over time (Basile et al., 2015). Distributed representations have made significant advances in the field of natural language processing. Word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) allow representing semantic patterns through vectors. Hamilton (2016) applied the word embedding method to several corpora, each covering ten year, to detect semantic shifts. Semantic changes of a word can be captured by the difference of collocation of other words. The majority of studies on semantic shift detection represent a word by a single vector, few studies use different embedding types such as Gaussian (Vilnis and McCallum, 2015) and Poincaré embeddings (Nickel and Kiela, 2017).

Semantic shifts are often measured in terms of the similarity of the word vectors between two periods with the cosine similarity and Euclidean distance being the most widely used metric. Only a few papers (Vilnis and McCallum, 2015; Athiwaratkun and Wilson, 2017) acknowledge similarity metrics suitable for Gaussian embedding.

3 Model Description

This study employs an improved structure of Gaussian embedding for detecting semantic shifts. According to this structure, each word is represented by a mean vector and a covariance matrix. To reduce the computational complexity, only the mean vectors are learned using the word2vec framework. The covariance matrices of words are obtained directly from their frequencies, i.e., they are not learned.

3.1 Types of Semantic Shifts

This study aims discovering those semantic shifts that are hard to find using only word2vec. Bloomfield's categorization (Bloomfield, 1934) dividing semantic shifts into nine types is used to illustrate how semantic changes can be detected using the proposed method. The mahority semantic shifts such as *metaphor* and *metonymy* can be identified based on word co-occurrence changes. The type of semantic change, *narrowing* or *widening*, is difficult to detect using only co-occurrence information. *Narrowing* and *widening* are phenomena, in which a word sense becomes embodied or abstracted from its original sense. For example, meat (mete in old German/English) has changed from “food” to “edible flesh”, and deer (dēor in old English) changed from “animal” to “deer”.

Word co-occurrences do not change significantly through *narrowing* and *widening*. They are based on the distributional inclusion hypothesis (Geffet and Dagan, 2005), according to which the hyponym appears in the similar contexts of its hypernym. Therefore, such semantic shifts can be detected using the frequency information as covariance matrices.

3.2 Gaussian Embedding

According to the proposed method, words are modeled using Gaussian distributions as follows:

$$\mathcal{N}\left(x, \mu, \Sigma\right) = \frac{1}{\sqrt{2\pi^d|\Sigma|}} \exp -\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}\left(x-\mu\right), \tag{1}$$

where μ is the mean vector of dimension d, and Σ is the covariance matrix of d rows and columns. The idea is to use the mean vector to detect major semantic shift and the variance to detect word hierarchy shifts.

A previous study (Vilnis and McCallum, 2015) reported high accuracy in the word similarity task by learning both the mean vector and variance. However, the method has a high computational complexity and is sensitive to initial values. The variance differs greatly depending on the learning time and initial values of the variance and mean vector. The authors suggested that their method is suitable for small dataset and a small number of parameters. Therefore, only the mean vector is learned using word2vec in this study, while the word variance is obtained from frequency differences.

It has been pointed out that word embeddings and frequencies are useful for detecting semantic changes. The normalized frequency difference is used in this study as a baseline of the task. Mean vectors and

method	dimension	similarity threshold	F1
frequency	100	1.40	**0.726**
		no	0.690
learning	100	1.37	**0.735**
		no	0.699

Table 1: Hypernym detection with BLESS dataset using Gaussian embeddings. The method using variances generated from word frequencies achieves an accuracy comparable to another method by using variances obtained from the learning process.

covariance matrices are combined, while the KL divergence is used for measuring semantic shifts with Gaussian embedding.

3.3 Preliminary experiment

Before applying the proposed model to semantic shift detection, the importance of the frequency information for determining the hypernym-hyponym relations was investigated first, In particular, the performance of the word similarity metric on hierarchical datasets was evaluated in a preliminary experiment.

In the experiment, words were represented by mean vectors, and variances learned based on the Text8 [1] corpus. The mean vectors were learned using the word2vec algorithm, and evaluated on SimLex-999 (Hill et al., 2015) and WordSim353 (Finkelstein et al., 2001). The scores were 0.260 and 0.613 for SimLex-999 and WordSim353, respectively. The word vectors were used as mean vectors while they captured the word relevance sufficiently. Two variances of the proposed model were explored, one generated from frequency counts and the other initialized with frequencies and learned using the loss function of Gaussian embeddings. The BLESS dataset (Baroni and Lenci, 2011) was used as an evaluation dataset for hypernym-hyponym detection. Hypernym-hyponym relationships were assumed to exist only when two words had related meanings to each other. A word w_i is a hypernym of a word w_j if (1) the variance of w_i is greater than the variance of w_j , and (2) the Euclidean distance of the mean vectors of w_i and w_j is smaller than a given threshold.

The results of the preliminary experiment are shown in Table 3.4. It can be noticed from the table that the F1 score with thresholds is higher than that without thresholds. In particular, the words are assumed not to have a hierarchical relationship when the Euclidean distance between them is large. An existing method used concatenated ukWaC and WaCkypedia corpora (Baroni et al., 2009) with 3 billion tokens, while this study employed Text8 with 17 million tokens. A small corpus was used in the preliminary experiment as large corpora with semantic shift are often unavailable. The results demonstrate that the model can detect word hierarchies using variances without learning.

3.4 Similarity measure

While the preliminary experiment focused on capturing the hierarchical relations of words, the purpose of the next experiment was to evaluate the performance of the proposed model in detecting various semantic changes. The KL divergence was employed as the similarity measure of word embedding between two periods. Let the Gaussian distributions of a word learned from corpora C_1 and C_2 be $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_1}, \Sigma_1)$ and $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_2}, \Sigma_2)$, respectively. The KL divergence for these two multivariate Gaussian distributions can be expressed as

$$\begin{aligned} D_{KL}\left(\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_1}, \Sigma_1) | \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_2}, \Sigma_2)\right) &= \int_{\mathbb{R}} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_1}, \Sigma_1) \log \frac{\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_2}, \Sigma_2)}{\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_1}, \Sigma_1)} dx \\ &= \frac{1}{2}\left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{Tr}\left({\Sigma_2}^{(-1)} \Sigma_1\right) + (\boldsymbol{\mu_2} - \boldsymbol{\mu_1})^{\mathrm{T}} {\Sigma_2}^{-1} (\boldsymbol{\mu_2} - \boldsymbol{\mu_1})\right]. \end{aligned} \tag{2}$$

The KL divergence is zero when the two distributions are equal, and it increases as the difference between the two distributions becomes larger.

[1] http://mattmahoney.net/dc/textdata.html

For our experiments, we set the covariance matrix to $\Sigma = \sigma I$ for some $\sigma > 0$. The variances σ_1 and σ_2 for Σ_1 and Σ_2 must be positive, and if the meaning does not change, σ_2 should be close to σ_1. Semantic changes were categorized in various types (Traugott, 2017) as described in the previous section. It is difficult to determine which changes affect the annotation score, core meaning (mean vector) and word abstractness. The ratio of mean vectors and variances should be adjusted depending on the data so that both of them have an appropriate effect on the KL divergence.

3.5 Learning process

According to the proposed method, mean vectors and covariance matrices are computed separately. Mean vectors are learned using the Kim (2014)'s method. First, distributed representations are learned using only corpus C_1. Distributed representations are obtained using the corpus C_2 with pretrained embeddings as the initial vectors.

For training word embeddings, word2vec was used to obtain d-dimensional normalized vectors. Since word2vec updates word vectors in each iteration, the difference between the word vectors obtained from the two corpora is non-zero even if the words are assumed to have no semantic change. This does not matter because the purpose of this experiment is to measure the degree of the semantic shift compared to other words.

To obtain the variance, the frequency of the word is calculated and normalized according to the size of corpora C_1 and C_2 fir eacg period t_1 and t_2. There is no need to compute the variance for each word; it should be calculated only for target words.

Given the normalized frequencies f_1 and f_2 of a target word, the covariance matrices can be calculated as

$$\Sigma_1 = \left(1 + \alpha \log\left(\beta + \frac{f_1}{f_2}\right)\right) I, \Sigma_2 = I. \tag{3}$$

The logarithm of the normalized frequency is taken to make the variance be positive and not impact the KL divergence too much. Based on the formula of the KL divergence in Equation 2, Σ_2 is set to the identity matrix to reduce the range of the KL divergence. To simplify the comparison of semantic changes across two periods, t_1 and t_2, if the variance of one of them (period t_2) is set to be fixed.

4 Experimental Setup

In the experiment, corpora C_1 and C_2 were used to create the Gaussian embedding, while the KL divergence was adopted as the criterion for measuring the semantic shift. The obtained scores were sorted for Subtask1: ranking and the classification task was solved with appropriate thresholds for each language.

Parameters allowing to achieve the highest accuracy in the post-evaluation period were employed in the proposed method. For learning mean vectors, the embedding size d was set to 300, while the window size was set to 5. The other parameters were set to the default values of the gensim library. The settings for Equation 3 were $\alpha = 0.02$ and $\beta = 3$.

Two other methods were employed for comparison with the proposed method. In the first method, only the mean vectors were used and the cosine similarity was employed to measure the semantic shift. In the second method, only covariance matrices were used, while the semantic shift was measured based on the difference in covariances between t_1 and t_2. The same parameters were used in all methods. The threshold of subtask 1 was set to maximize the classification score. While the threshold was set for each language separately, the number or parameters can be reduced the parameters by considering moving the seed words (e.g. country names) as future work.

5 Results

The results are shown in Table 5. We show our proposed model, compared models and the highest model. We supplement here that various models achieves higher scores in the post-evaluation process than the models in the evaluation process. There are no significant differences between the frequency-based method and other methods including word2vec and proposed (E) in Subtask 1. In contrast, the frequency-based

Subtask	Method	Measure	average	English	German	Latin	Swedish
1	proposed (PE)	KL	0.701	0.676	0.729	0.650	0.741
	UWB (E)	-	0.687	0.622	0.750	0.700	0.677
	word2vec	cosine	0.686	0.676	0.708	0.650	0.710
	frequency	difference	0.581	0.568	0.625	0.550	0.581
	proposed (E)	KL	0.511	0.541	0.500	0.550	0.452
	subtask2 gold	threshold	0.848	0.811	0.854	0.825	0.903
2	UG_Student_Intern(E)	-	0.527	0.422	0.725	0.412	0.547
	proposed (PE)	KL	0.410	0.358	0.578	0.329	0.373
	word2vec	cosine	0.402	0.358	0.576	0.336	0.337
	proposed (E)	KL	0.087	0.157	0.099	0.065	0.028
	frequency	difference	0.028	0.070	-0.049	0.157	-0.067

Table 2: Results for Unsupervised Semantic Shift Detection. PE is post evaluation process, and E is evaluation process. - is unknown measurement. The word embedding is useful to capture semantic shift, and adding frequency information to it improve the performance slightly.

method does not capture the degree of semantic shifts in Subtask 2. As a comparison, the results are listed using the gold data from Subtask 2 for Subtask 1. Even if the degree of a semantic shift is perfectly predicted (i.e., even if the model achieves a correlation of 1.000 in Subtask 2), the result of the model is 0.848 in Subtask 1, which means 1.000 cannot be achieved in both subtasks using the same model. In addition, the bias of the classification labels caused a significant difference in the results between Subtask 1 and 2. As described in the task description paper, the results between Subtask 1 and 2 do not have strong relationship. Therefore, we focus on the results obtained for Subtask 2 given that the word ranking labels are more informative than classification labels.

It can be noticed from Table 5 that the obtained distributed representation (using word2vec in this case) roughly captures the semantic shifts. The result indicates that the KL divergence works as a criterion for semantic shift detection. When comparing the proposed embedding method with Kim's (2014) method, it can be concluded that including the frequency information can yield a similar or even higher Spearman's rank-order correlation coefficient.

The relationship between the frequency information and Spearman's coefficient is analyzed. We focused on German and English because the proposed method achieved a higher Spearman's coefficient on the German dataset but not the English dataset. For example, the obtained embedding could detect the semantic change of a noun Knotenpunkt from "junction" to "hub" in German. This noun was first used to mean *junction* but then changed its meaning to *hub*, and was used with that letter meaning more often compared to the earlier meaning. Thus, the frequency information is useful for identifying semantic shifts. The verb abdecken"uncover" is an example, where both the word2vec and the proposed methods were able to detect a semantic shift. The old meaning of this verb referred to taking the skin off an animal carcass, whereas now it refers to covering up. If the old meaning disappeared, it would be possible to detect the meaning using only word2vec.

The Spearman's coefficients obtained using the word2vec and proposed method did not change for the English dataset. There were few examples in the dataset that could be identified as changes in the hypernym-hyponym semantic shift of words. Word2vec and the proposed embedding could detect semantic shifts for the noun "graft", which was first used to mean a *grafting tree* but later was used to mean an *implant*. This is an example of the development of another meaning of a word.

The proposed method was found to be able to capture semantic shifts of *narrowing* and *widening* with variances. It would be desirable to find a way to reducing the number of parameters when no test data is available, which was the case in the evaluation process of this study. The majority of existing machine learning methods require parameter tuning. Only a few of them play a major role in the task of detecting semantic shift, where the correct answer is unknown in advance.

6 Conclusion

This paper described the systems submitted to SemEval 2020 Task1: Unsupervised Lexical Semantic Change Detection. We designed to detect semantic shifts based on an improved Gaussian embedding method. According to the proposed method, the mean vectors are trained using the word2vec algorithm, while the variances are obtained from word frequencies. The computational complexity and parameter sensitivity to initial values of the proposed method is reduced compared to the traditional Gaussian embedding method.

The proposed method that used mean vectors and variances achieved higher Spearman's coefficients for several languages in Subtask 2 compared to the technique that used only mean vectors (word2vec). This result indicates that variances are effective for detetcting the changes of word abstractness. As part of our future work, we plan to focus on semantic shift detection across multiple languages and detailed classification of semantic shift types.

Acknowledgments

We gratefully acknowledge Kwangjin Jeong for valuable discussions and the anonymous reviewers for useful comments.

References

Ben Athiwaratkun and Andrew Wilson. 2017. Multimodal Word Distributions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1645–1656.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43:209–226.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2015. Temporal Random Indexing: A System for Analysing Word Meaning over Time. *Italian Journal of Computational Linguistics*, 1:55–68, 12.

Andreas Blank and Peter Koch. 1999. *Historical Semantics and Cognition*. De Gruyter Mouton.

Leonard Bloomfield. 1934. *Language*. Holt.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing Search in Context: The Concept Revisited. In *Proceedings of the 10th International Conference on World Wide Web*, page 406–414.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Maayan Geffet and Ido Dagan. 2005. The Distributional Inclusion Hypotheses and Lexical Entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 107–114.

Joachim Grzega and Marion Schöner. 2007. *English and General Historical Lexicology : Materials for Onomasiology Seminars*, volume 1. Katholische Universität Eichstätt-Ingolstadt.

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*, pages 1489–1501.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

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.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré Embeddings for Learning Hierarchical Representations. In *Advances in Neural Information Processing Systems 30*, pages 6338–6347.

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*, pages 1532–1543.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2014. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 27th International Conference on Computational Linguistics (SemEval)*.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Lexical Semantic Change.

Elizabeth Closs Traugott. 2017. Semantic Change.

Luke Vilnis and Andrew McCallum. 2015. Word Representations via Gaussian Embedding. In *3rd International Conference on Learning Representations*.