9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Quantifying Diachronic Language Change via Word Embeddings: Analysis of Social Events using 11 Years News Articles in Japanese and English

Keywords: Semantic shift, Word embeddings, word2vec, Social events, and COVID-19

Extended Abstract

Languages are constantly changing, partially because of social events. For example, the worldwide spread of COVID-19 has caused various changes in daily life, and the meanings of some words have changed [1]. Such changes in words are referred to as *semantic shifts* and have attracted considerable attention in linguistics, sociology, and information science. Semantic shift has long been studied from a variety of perspectives, and efforts to investigate semantic shift using word embeddings have accelerated as large corpora have become available.

Kutuzov et al. [3] distinguished two main types of research on semantic shift through word embeddings, including linguistic explorations with long-term corpora and works on the detection and analysis of social events with relatively short-term corpora. The latter includes studies on the prediction of the occurrence of armed conflicts, an analysis of Sinophobia (feelings of fear of China), and an analysis of the impact of COVID-19 [1]. Studies on the detection and analysis of social events have often focused primarily mainly on a single event, and relatively few works have discussed the effects of multiple events. We consider that the development of a quantification method robust to different times and corpora could lead to more comprehensive detection and analysis.

In this study, we quantified diachronic semantic shifts caused by social events using a corpus of news articles published in Japanese and English over a period of 11 years. First, we describe a method for quantifying the semantic shift each year by extending a previous study [1] that compared word embeddings. The method has the advantage of identifying words that exhibited a significant semantic shift. Second, we conducted case studies in Japanese and English. Our hypotheses were that 1) the semantic shift caused by COVID-19 is larger over the past 11 years, and 2) the trends of change in Japan and English are similar because social events affect the entire world. Here we focused only on social events as factors in the semantic shift.

The process flow to quantify diachronic language changes using word embeddings is shown in Figure 1. Following many previous works, we used word2vec to acquire word embeddings. First, the vector spaces of two trained word2vec models were rotated and transformed into the same space (this is referred to as *mapping*). Here, based on the idea that the rate of semantic shift follows a negative power of word frequency, we assume that the meaning of frequently appearing words does not change over time and that local structure is preserved. To summarize, the method takes two trained word2vec models as input and derives a rotation matrix to align their coordinate axes. The stability $stab(w)$ of a word w in comparing two models i and j is defined as follows:

$$stab(w) = \frac{sim_{ij}(w) + sim_{ji}(w)}{2} \quad sim_{ij}(w) = cossim(R^{ji}R^{ij}V_w^i, V_w^i)$$

Note that *cossim* denotes cosine similarity; R^{ji} is a rotation matrix used for mapping from model j to i; and V_w^i denotes the word embeddings of a word w in model i. We considered that the degree of change of the entire model was correlated with the degree of change of the words in the model. We refer to the average value of *stab* of words as *semantic shift stability*, and adopted this as a representative value. The smaller this value, the greater the degree of shift in the overall model. The results confirm that the semantic shift stability calculated from models i and j is correlated with the performance difference of text classifiers using models i and j and may be considered valid as a representative value [2]. Let W be the vocabulary commonly included in models i and j, and N be the size of W.

$$\text{semantic shift stability} = \frac{1}{N} \sum_{w \in W} stab(w)$$

The annual change in semantic shift stability is shown in Figure 2. In the experiments, we used a corpus of news articles published in Japanese and English from 2011 to 2021. The Japanese content was obtained from the Nikkei Online Edition[1] (*Nikkei*), and the English content was collected from News on the Web[2] (*NOW*). Both corpora were divided into segments covering one year each, and word2vec models were trained. For example, the degree of change between 2011 and 2012 was calculated by comparing the word2vec model with the training corpus for 2011 and 2012. The semantic shift stability for 2019-2020 was observed to be the lowest for both Nikkei and NOW, i.e., the degree of change was the greatest. The correlation coefficient between Nikkei and NOW was calculated to be 0.66, indicating a similar trend. The overall value of the semantic shift stability of NOW was smaller than that of Nikkei.

The top 10 words with the lowest *stab* in Nikkei in 2019-2020 were infection, spread, corona, vaccine, virus, mask, infected, North Korea, vaccination, and epidemic. Those in NOW were king, Scott, de, virus, masks, wear, mask, pi, q, and wearing. In both cases, words related to COVID-19 appeared at the top of the lists. The top three synonyms of *corona* in the Nikkei indicate that words related to COVID-19 appeared after 2020. A similar trend was observed in the top five synonyms of *virus* in the NOW. Note that the analysis for 2015-2016 implied the impact of the U.S. presidential election. For example, the meaning of the word *trump* changed in the Nikkei.

This study provides a framework for the quantitative analysis of semantic shifts across multiple corpora and social events. The results of this case study using Japanese and English news articles published over an 11-year period showed that the trends of the semantic shift were similar in both languages and that COVID-19 had a significant impact.

References

[1] Y. Guo, C. Xypolopoulos, and M. Vazirgiannis. How COVID-19 is changing our language: Detecting semantic shift in twitter word embeddings. *arXiv:2102.07836*, Feb. 2021.

[2] S. Ishihara, H. Takahashi, and H. Shirai. Semantic shift stability: Efficient way to detect performance degradation of word embeddings and pre-trained language models. In *Proceedings of the AACL-IJCNLP 2022*, pages 205–216, Online only, Nov. 2022.

[3] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the COLING 2018*, pages 1384–1397, Santa Fe, New Mexico, USA, Aug. 2018.

[1] `https://aws.amazon.com/marketplace/seller-profile?id=c8d5bf8a-8f54-4b64-af39-dbc4aca94384`

[2] `https://www.english-corpora.org/now/`

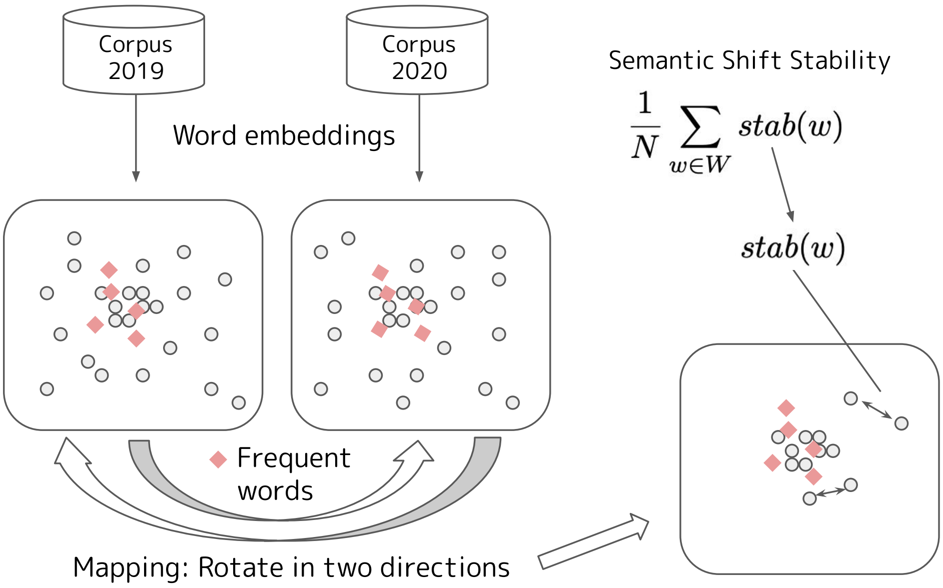

Figure 1: A method for quantifying diachronic language change using word embeddings. The movement of a word w in the space is calculated as $stab(w)$, and the average of all words is considered as the degree of change between the word embeddings.

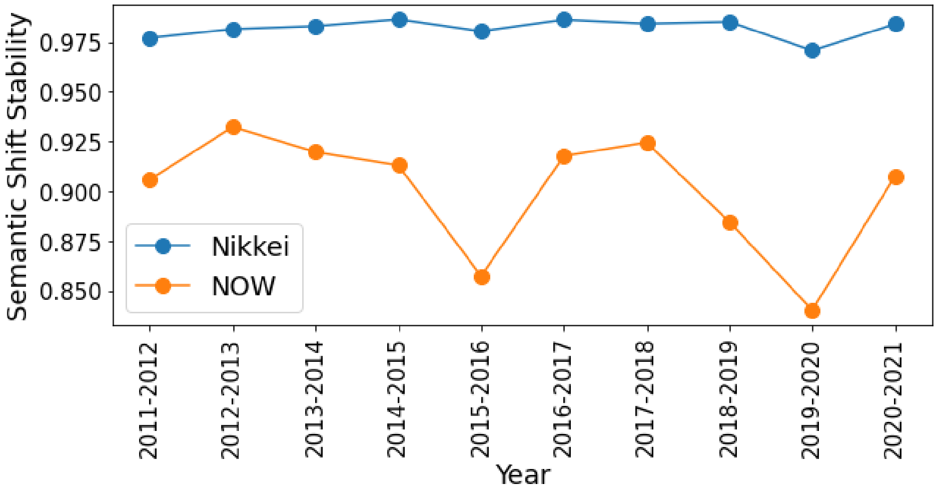

Figure 2: Annual changes in semantic shift stability with the lowest values in 2019-2020 for both Nikkei and NOW and similar transitions with a correlation coefficient of 0.66.