Analyzing Semantic Changes in Japanese Words Using BERT

Kazuma Kobayashi, Taichi Aida, Mamoru Komachi
Tokyo Metropolitan University
{kobayashi-kazuma1@ed., aida-taichi@ed., komachi@} tmu.ac.jp

Abstract

The usage and meaning of words may change over time. In this study, we compared two existing BERT-based methods of capturing semantic changes in Japanese words. Each method clusters word tokens obtained from BERT in different ways. One method uses example sentences from a dictionary, whereas the other method uses centroids from k-means clustering are used. We calculated the usage ratio for each period (25 years) according to the clustering results. The clustering-based method outperformed the dictionary-based method in almost all the cases. We also found that the dictionary-based method was sensitive to definitions and the specific choices of example sentences.

1 Introduction

The usage and meaning of words may change over time. For example, the word "gay" originally meant "carefree," but now it also means "homosexual." Such a phenomenon is not unique to English (Hou et al., 2020). Techniques for quantitatively analyzing the semantic changes of words over time, also known as diachronic changes, can positively contribute to diverse research domains including linguistics, lexicography, and sociology (Andrey et al., 2018).

Several distributed word representation-based methods have been proposed for capturing changes in the meanings of words(Kim et al., 2014; Hamilton et al., 2016; Yao et al., 2018). Although these methods can detect a meaning change, it can be difficult to interpret the specifics of the semantic change because these methods only assign a single representation to each word type.

An alternative to using word representation-based methods is to use a pre-trained language model such as BERT (Devlin et al., 2019), through which one can obtain a word vector according to the context in which each word token appears. As a result, one can then assess the meanings of words at the token level. Recently, two studies were conducted to capture the changes in meanings of words using BERT. One method used meanings from a dictionary (Hu et al., 2019), whereas the other used centroids from k-means clustering (Giulianelli et al., 2020) to create clusters. Although these two methods are similar, the influence of their different approaches on their performances are yet to be investigated.

Additionally, these studies have only been conducted on English words. One exception is Aida et al. (2021), wherein diachronic changes in the meanings of Japanese words were captured using word representation-based methods. However, no other studies, including those using BERT-based methods, have been conducted with Japanese words.

In this study, we applied the methods of Hu et al. (2019) and Giulianelli et al. (2020) to Japanese words, and then compared the results. The contributions of this study are as follows.

- We analyzed changes in meanings of Japanese words over time with BERT.
- We compared dictionary-based (Hu et al., 2019) and clustering-based (Giulianelli et al., 2020) methods.

2 Methodology

In this study, we obtained word vectors from the final layer of BERT. We input all sentences that included target words into BERT to obtain the contextual word vectors per token. If the target word consisted of multiple subwords, we averaged the vectors of the subwords and assigned them to the target words.

2.1 Word clustering

In the following sections, we define the set of word vectors acquired with BERT as S.

2.1.1 Dictionary-based clustering

First, we describe the dictionary-based clustering method (Hu et al., 2019). For a word with n senses, sense vectors $v_1, v_2, ..., v_n$ were the assigned vectors obtained from BERT as follows. We input the example sentences corresponding to each sense in the dictionary into BERT and obtained the vectors that corresponded to the target token. If multiple example sentences were listed for one sense, the average of the acquired word vectors was used as the sense vectors. Each element of S was assigned to the sense that had the highest cosine similarity with $v_x \in \{v_1, ..., v_n\}$. We henceforth refer to this method as the dictionary-based method.

2.1.2 k-means-based clustering

Next, we describe the k-means-based clustering method (Giulianelli et al., 2020). We used k-means clustering with the number of clusters ranging from 2 to 10 for S. Subsequently, we used the clustering results and the largest average of the silhouette scores (Peter, 1987) to determine the cluster to which the word vector belonged. The distance used in k-means clustering was the Euclidean distance. we henceforth refer to this method as the clustering-based method.

2.2 Semantic change detection

Figure 1 presents a simplified example of creating stacked bar graphs from clustering results. We divided a diachronic corpus into periods and created word vectors for each token by applying BERT to the sub-corpora, one sentence at a time. Then, we clustered the word vectors for all periods using the two methods described above. The clustering results are presented in scatter plots visualized by principal component analysis, as shown in Figure 1 (left). Next, we calculated the word usage ratio per period, and we represent the data as stacked bar graphs, as shown in Figure 1 (right). The legend for the dictionary-based method shows the dictionary senses, whereas the legend for the clustering-based method shows the most frequent meanings or usages that were confirmed in each cluster. We analyzed the results by comparing graphs with the corresponding examples.

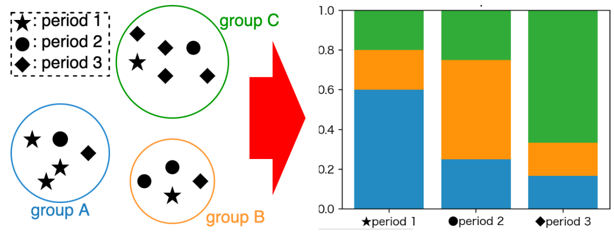


Figure 1: Example bar graph that shows the transitions in word meanings over time.

3 Experimental settings

3.1 Corpora

We used the modern magazine corpus published as part of the "Corpus of Historical Japanese,"[1] and modern magazines "Chuokoron" and "Bungeishunju," which are part of the "Showa / Heisei Written Language Corpus."[2] They include magazines published from 1874 to 1997, and we used those published over a 100-year period, from 1898 to 1997, divided according to 25-year periods.

3.2 Dictionary

We used "Nihon kokugo daijiten,"[3] one of the largest Japanese dictionaries for the dictionary-based method. From this dictionary, we removed any example sentences that were written in classical Chinese, after which we removed any sense that had no example sentences.

[1] https://ccd.ninjal.ac.jp/chj/index.html
[2] https://ccd.ninjal.ac.jp/cmj/woman-mag/index.html
[3] https://japanknowledge.com/en/contents/nikkoku/

		Clustering-based method	
		Good	Bad
Dictionary-based method	Good	普段 “usually,” 障害 “disability,” 柔軟 “flexible,” 結構 “very well,” 要領 “way to deal,” ケース “particular situation,” 免許 “license,” 優勝 “take first prize,” 明細 “detailed statement,” 非常 “very,” 全然 “not at all,” 精々 “at best,” 渋滞 “traffic jam,” ポイント “essential matter,” 管制 “air traffic control,” 故障 “broken”	適当 “sloppy,” 教養 “broad knowledge,” 貴族 “privileged person,” 教授 “professor,” 心持ち “slightly”
	Bad	遊撃 “shortstop,” 風俗 “brothel,” カフェ “cafe,” 団塊 “baby boomer generation,” 普通 “normal,” 尋常 “ordinary (+ negation),” 愛人 “lover,” ボタン “button for machines,” 広告 “advertisement,” 了解 “approval,” 端末 “device,” 住居 “residence,” スーパー “supermarket”	自然 “nature,” 女性 “woman,” モデル ‘maquette,’ 設備 ‘facility,” 婦人 “married woman,” 情報 “information,” 主婦 “housewife,” こだわり “pursuing”

Table 1: Contingency table of the two methods. The English translation represents the contemporary most frequent sense. The dictionary method could not be applied to the underlined words.

3.3 Target words

We focused on 42 Japanese words that are widely known to have undergone semantic changes, extracted from a list obtained via Mabuchi and Ogiso (2021).

3.4 Pre-trained BERT model

In this study, we used a pre-trained BERT model. Specifically, we used Tohoku University’s Japanese version of BERT-base[4], which is trained on Japanese Wikipedia. The model consists of 12 layers, 768 dimensions of hidden states, and 12 attention heads. The model is trained with the same configuration as the original BERT(Devlin et al., 2019): 512 tokens per instance, 256 instances per batch, and 1M training steps. For training the masked language modeling objective, the model introduced whole word masking, in which all of the subword tokens corresponding to a single word (as tokenized by MeCab) are masked at once. The model trained with a v3-8 instance of Cloud TPUs provided by TensorFlow Research Cloud program[5]. The training took approximately 5 days for BERT-base. In tokenization, texts are first tokenized by MeCab with the Unidic 2.1.2 dictionary and then split into subwords by the WordPiece algorithm.

[4]https://github.com/cl-tohoku/bert-japanese

[5]https://sites.research.google/trc/about/

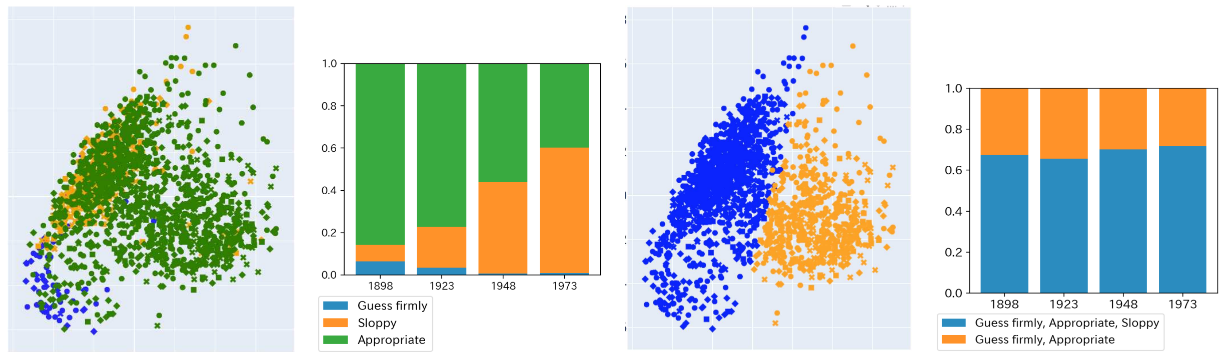
3.5 Evaluation

Hu et al. (2019) and Giulianelli et al. (2020) evaluated the method’s performance by comparing the datasets that were ranked by their degree of change based on human judgment, with the score calculated quantitatively.

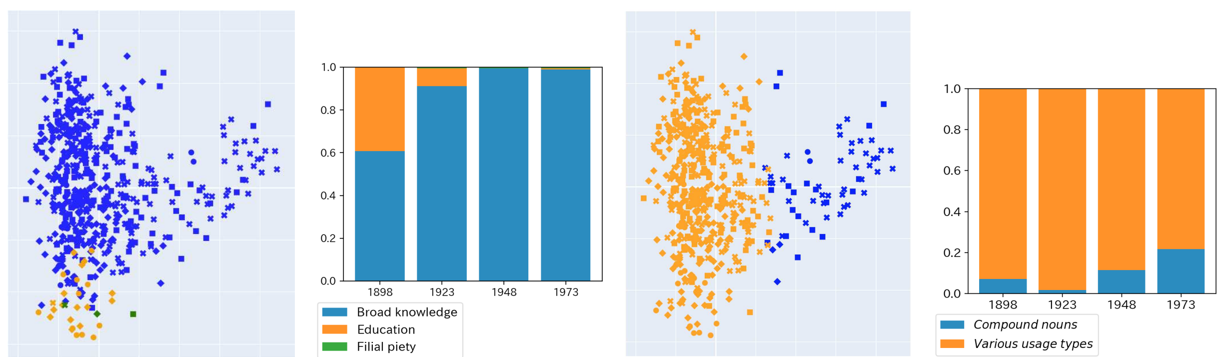
However, because there is no such data available in Japanese, our evaluation was based on the authors’ qualitative judgment. This was made by comparing the changes in the stacked bar graphs with the actual data corresponding to the clustering results. This qualitative analysis was conducted by focusing on two points: (1) whether the most frequent senses or usages of clusters correctly corresponded to each of the most frequent senses before and after the change and (2) whether the usage trends corresponding to clusters correctly decreased or increased.

4 Experimental results

Table 1 summarizes the experimental results for the two methods. The dictionary-based method was successful for 21 words, whereas the clustering-based method was successful for 29 words. The clustering-based method could be applied to any



(a) Scatter of the dictionary-based method for 適当 (b) Usage by the dictionary-based method for 適当 (c) Scatter of the clustering-based method for 適当 (d) Usage by the clustering-based method for 適当



(e) Scatter of the dictionary-based method for 教養 (f) Usage by the dictionary-based method for 教養 (g) Scatter of the clustering-based method for 教養 (h) Usage by the clustering-based method for 教養

Figure 2: Example results for 適当 and 教養.

word; however, the dictionary-based method that used the dictionary mentioned earlier was successfully applied to only 33 of the 42 words . This is because there were nine words that did not satisfy the application requirement that there must be at least two senses with example sentences in the dictionary. For example, there were words that had only a single sense, or words that had multiple senses but no corresponding example sentences. We argue that these words do not capture the semantic changes in the dictionary-based method.

In the scatterplots (Figures 2 to 7), the round symbols correspond to the years 1898–1922, the square symbols correspond to the years 1923–1947, the diamond symbols correspond to the years 1948–1972, and the X symbols correspond to the years 1973–1997.

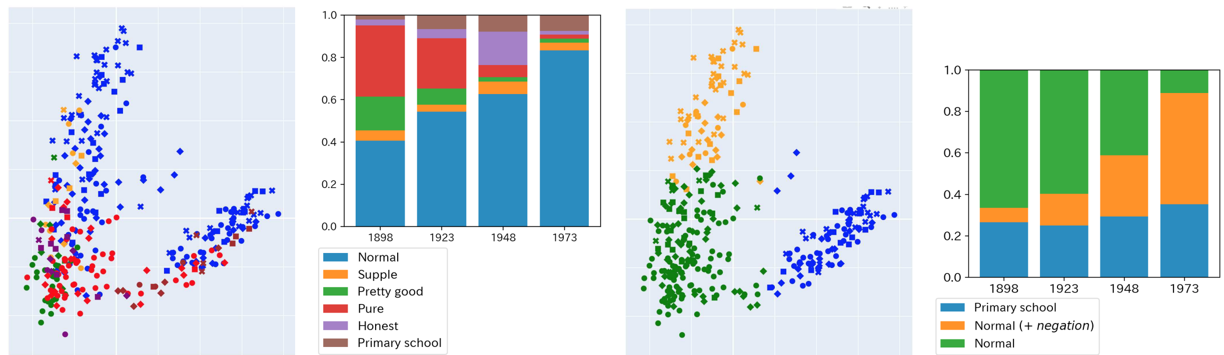
In the stacked bar graphs, the labels were as follows. For the dictionary-based method, we assigned senses from the dictionary to labels even if the clusters' components did not correspond to the label. By contrast, in the clustering-based method, we assigned corresponding meanings or explanations of the cluster to labels. Here, we used roman fonts when a label is a meaning and *italics* when the label is an explanation.

4.1 Successful cases of the dictionary-based method

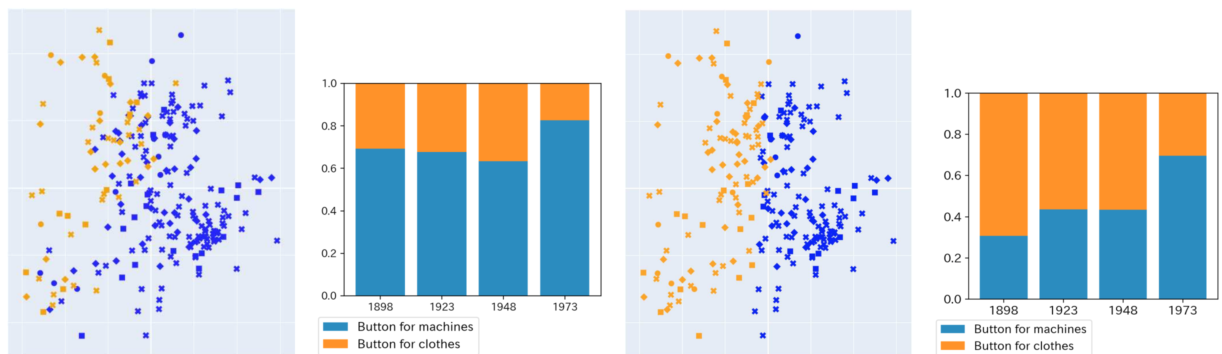
The successful words were 適当, 教養, 貴族, 教授, and 心持ち. These are words with large differences in usage frequencies for the meanings before and after the change. In contrast to the clustering-based method, the dictionary-based method successfully captured the less frequent meanings. We introduce 適当 and 教養 as examples.

適当 The most frequent sense of this word has changed from “appropriate” to “sloppy.”

First, we discuss the dictionary-based method. In Figure 2(a), sense 2 (orange) and sense 3 (green)



(a) Scatter of the dictionary-based method for 尋常 (b) Usage by the dictionary-based method for 尋常 (c) Scatter of the clustering-based method for 尋常 (d) Usage by the clustering-based method for 尋常



(e) Scatter of the dictionary-based method for ボタン (f) Usage by the dictionary-based method for ボタン (g) Scatter of the clustering-based method for ボタン (h) Usage by the clustering-based method for ボタン

Figure 3: Example results for 尋常 and ボタン.

are the most frequent senses before and after the change, respectively. Figure 2(b) indicates that usage of sense 3 is gradually decreasing, whereas that of sense 2 is gradually increasing.

Next, we discuss the results of the clustering-based method. The clustering-based method did not capture the semantic change of 適当. In Figure 2(c), there is no clear difference between cluster 1 (blue) and cluster 2 (orange), and both clusters contain a variety of examples. Furthermore, Figure 2(d) shows that there is almost no change in usage frequency between time periods.

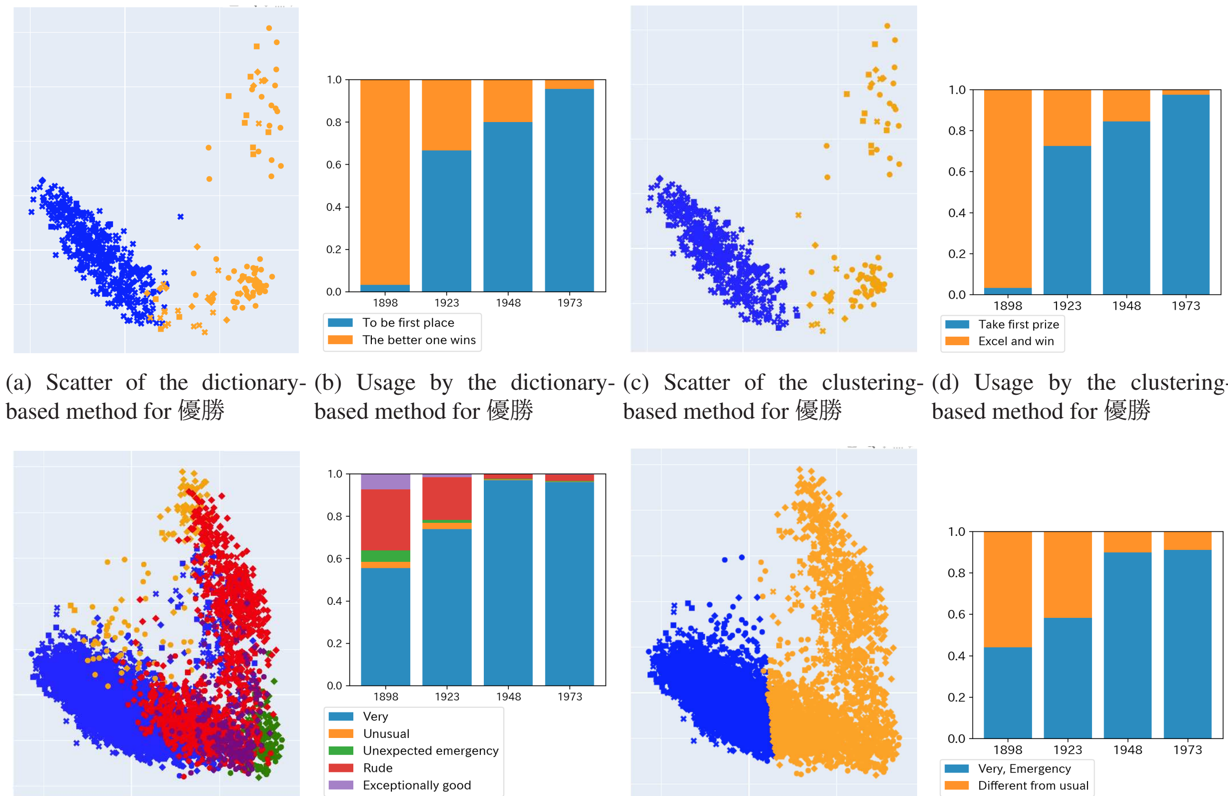
教養 The most frequent sense of this word has changed from “teaching or educating” to “broad knowledge.”

First, let us discuss the dictionary-based method. In Figure 2(e), sense 2 (orange) is the most frequent sense before the change, and sense 1 (blue) is the most frequent sense after the change. Figure 2(f) indicates that the example used in sense 2 is no longer used.

Next, we discuss the clustering-based method. The clustering-based method successfully captured the semantic changes in 教養. Cluster 1 (blue) in Figure 2(g) is a cluster of compound nouns, i.e., of which “教養 + noun,” 教養学部 “College of Liberal Arts” and 教養課程 “academic course” are typical examples. Cluster 2 (orange) is a mixed cluster of examples different from those from cluster 1. In other words, each cluster does not capture the semantic features before and after the change. No semantic change can be identified from Figure 2(h).

4.2 Successful cases of the clustering-based method

Next, the words that worked well in the clustering-based method but not in the dictionary-based method were 遊撃, 風俗, カフェ, 団塊, 普通, 尋常, 愛人, ボタン, 広告, 了解, 端末, 住居, and スーパー. These words presented problems for



(a) Scatter of the dictionary-based method for 優勝
(b) Usage by the dictionary-based method for 優勝
(c) Scatter of the clustering-based method for 優勝
(d) Usage by the clustering-based method for 優勝
(e) Scatter of the dictionary-based method for 非常
(f) Usage by the dictionary-based method for 非常
(g) Scatter of the clustering-based method for 非常
(h) Usage by the clustering-based method for 非常

Figure 4: Example results for 優勝 and 非常.

the dictionary-based method, as they often lacked the corresponding senses for the semantic changes and lacked sufficient example sentences, both qualitatively and quantitatively, for each sense. We introduce 尋常 and ボタン as examples.

尋常 The most frequent sense of this word changed from "normal" to it being used as an idiomatic expression, i.e., "尋常 + negation."

First, we discuss the dictionary-based method. The dictionary-based method detected semantic changes but did not capture the differences in meanings before and after the changes because of the lack of corresponding senses. Sense 1 (blue) in Figure 3(a) is the corresponding sense before the change. The sense corresponding to the idiomatic expression with negation is not found in the dictionary and is instead merged with sense 1. Figure 3(b) shows that the usage frequency of sense 1 increased over time, but it does not capture the desired semantic change.

Next, we describe the clustering-based method. Cluster 3 (green) in Figure 3(c) is a typical example of the most frequently used meaning before the change, and cluster 2 (orange) is a typical example of the most frequently used meaning after the change. Figure 3(d) indicates that cluster 3 gradually decreases, whereas cluster 2 gradually increases.

ボタン This is a word for which the most frequent sense has changed from "button for clothes" to "button for machines."

We discuss the dictionary-based method. The dictionary-based method did not capture the semantic changes in ボタン, even though the dictionary sense was appropriate because of the lack of sufficient example sentences. In Figure 3(e), sense 2 (orange) corresponds to the sense before the change, and sense 1 (blue) corresponds to the meaning after the change. No semantic changes can be observed in Figure 3(f).

Next, we describe the clustering-based method. In Figure 3(g), cluster 1 (blue) is a typical example of the meaning after the change, and cluster 2 (orange) is a typical example of the meaning before the change. Cluster 2 gradually decreased, and cluster 1 gradually increased, as seen in Figure 3(h).

4.3 Cases where both methods were successful

Both methods worked well for 普段, 障害, 柔軟, 結構, 要領, ケース, 免許, 優勝, 明細, 非常, 全然, 精々, 渋滞, ポイント, 管制 and 故障. These words do not have the problems we mentioned earlier. We introduce 優勝 and 非常 as examples.

優勝 The most frequent sense changed from "excel and win" to "take first prize." There were almost no differences between the two methods for 優勝.

First, we discuss the dictionary-based method. Figure 4(a) shows that sense 2 (orange) corresponds to the most frequently used sense before the change, and sense 1 (blue) corresponds to the most frequently used sense after the change. Figure 4(b) shows that the usage of sense 2 is decreasing and that of sense 1 is increasing.

Next, we discuss the clustering-based method. In Figure 4(c), cluster 2 (orange) mainly consists of the most frequently meaning before the change, and cluster 1 (blue) mainly consists of the most frequently used meaning after the change. Figure 4(d) confirms that cluster 2 is decreasing and cluster 1 is increasing.

非常 The most frequent sense changed from "different from usual" to meaning "very" as an adverb indicating degree. The results of the clustering-based and dictionary-based methods were nearly identical for some words, such as 優勝, and the clustering results were clearer. However, for others, such as 非常, the results were slightly different, and the clustering was less clear but still captured the semantic change.

First, we discuss the dictionary-based method. Figure 4(e) shows that sense 1 (blue) was used the most after the change, and sense 2 (orange) was used the most before the change; the other senses do not correspond to either one of these meanings. Figure 4(f) shows that usage of sense 1 increased and that of other senses, especially sense 3, decreased.

Next, we discuss the clustering-based method. Figure 4(g) shows that the typical usage of cluster 2 (orange) was the most frequent before the change and that of cluster 1 (blue) was the most frequent after the change. Figure 4(h) reveals a decrease in cluster 2 and increase in cluster 1.

4.4 Cases where neither method was successful

Finally, neither method could successfully capture well the semantic changes in 自然, 女性, モデル, 設備, 婦人, 情報, 主婦 and こだわり. 自然 has both of the problems mentioned in Sections 4.1 and 4.2. Other words had problems in the target corpus. For example, there were no (or few) sentences pertaining to the target senses. We introduce 自然 as an example.

自然 The most frequently used sense changed from "unintentional" to "nature." The difference in usage frequency of the sense before and after the change in the corpus was large.

First, we discuss the dictionary-based method. In Figure 5(a), sense 1 (blue) is the closest to the most frequently used senses before the change, whereas the other senses are the most frequently used sense after the change. Most of all sentences use senses 2 (orange) and 9 (yellow ocher); thus, they do not capture the most frequently used sense after the change. Figure 5(b) shows a slight upward trend of the usage of sense 1, but the change is not clearly captured.

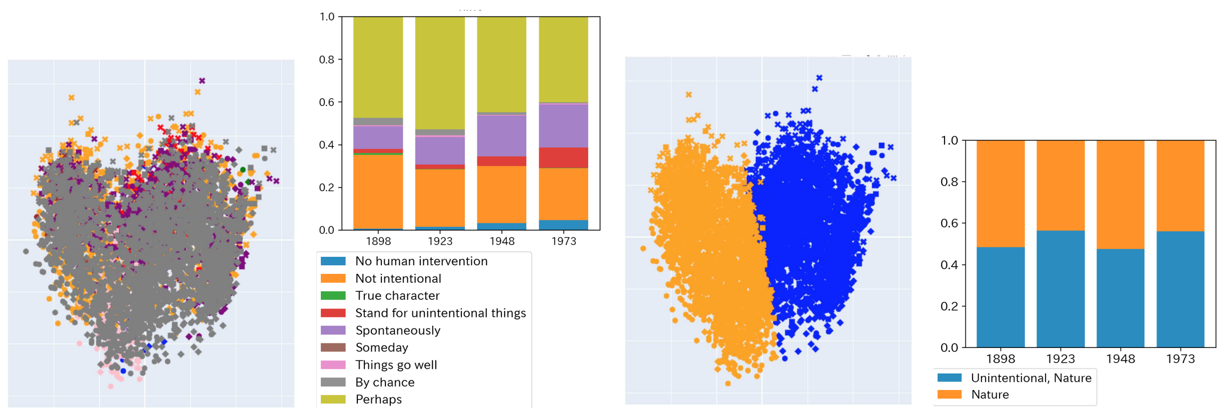
Next, we describe the results of the clustering-based method. Cluster 1 (blue) in Figure 5(c) contained sentences with the most frequently used meaning after the change, but a majority of the sentences used the meanings before the change. Cluster 2 (orange) was represented by sentences with the most frequently used sense before the change. Figure 5(d) clearly indicates that this method does not capture the semantic change.

5 Discussions

Based on these results, we found that the clustering-based method was more effective and useful for dealing with Japanese words.

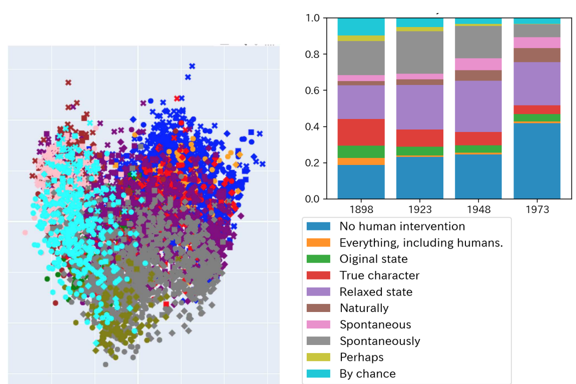
However, it is not clear whether a particular choice of experimental settings affects the performance. Thus, we further explored dictionaries and clustering methods. We conclude this section by analyzing the word vectors obtained from BERT.

The choice of dictionary is important for the dictionary-based method. The dictionary-based



(a) Scatter of the dictionary-based method (b) Usage by the dictionary-based method (c) Scatter of the clustering-based method (d) Usage by the clustering-based method

Figure 5: Example results for 自然.



(a) Scatter of the dictionary-based method (b) Usage by the dictionary-based method

Figure 6: Results for 自然 by Digital Daijisen.

method had limitations when choosing target words and depended too heavily on the quality of the dictionary. This is not a problem in English words because in English, the Oxford English Dictionary (OED) has a sufficient number of senses and example sentences for most words. In our experiment, we used "Nihon kokugo daijiten," one of the largest Japanese dictionaries; however, the quality and quantity of the sense inventory and example sentences provided by this dictionary may have negatively affected the results for some words.

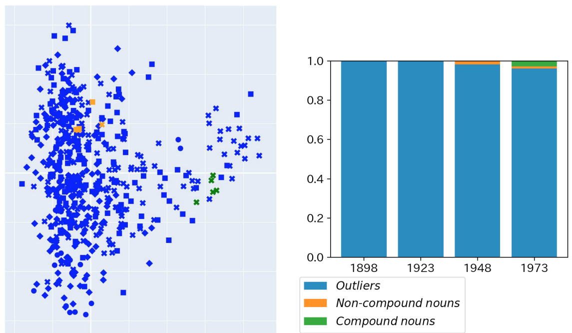
Therefore, to confirm the influence of the dictionary, we compared the results using "Digital Daijisen."[6] Table 2 summarizes the experimental results for the two methods. In Table 2, words that did not meet the requirements mentioned in Section 3.3 in both dictionaries are removed. Table 2 reveals that different dictionaries produced different results, even when using the same method. Furthermore, "Nihon kokugo daijiten" outperformed "Digital Daijisen."

We introduce 自然 as an example. The dictionary-based method using "Digital Daijisen"

[6] https://japanknowledge.com/

		Nihon	
		Good	Bad
Jisen	Good	要領, 優勝, 免許, 明細, 非常, 適当, 精々, 障害, 柔軟, 故障, 結構, 教養, 教授, 心持ち	自然, スーパー
	Bad	普段, 全然, ケース, ポイント, 渋滞, 管制, 貴族,	風俗, 尋常, 愛人, モデル, 普通, 遊撃, ボタン, カフェ, 団塊, 普通, 広告, 情報, こだわり

Table 2: Contingency table of the two methods. Words that could only be applied in "Nihon kokugo daijiten" are marked with a wavy line. Words that could only be applied in "Digital Daijisen" are underlined. Nihon is "Nihon kokugo daijiten," and Jisen is "Digital Daijisen."



(a) Scatter of the DBSCAN-based method (b) Usage by the DBSCAN-based method

Figure 7: Results for 教養 by DBSCAN.

		k-means	
		Good	Bad
DBSCAN	Good	スーパー, 住居, 渋滞, 障害, 非常, 普通, 免許, 優勝, 要領, 了解, 全然	
	Bad	カフェ, 愛人, ケース, 管制, ポイント, 故障, ボタン, 結構, 柔軟, 端末, 団塊, 普段, 風俗, 明細, 遊撃, 尋常, 精々, 広告	こだわり, モデル, 貴族, 教授, 教養, 自然, 主婦, 女性, 設備, 適当, 婦人, 情報, 心持ち

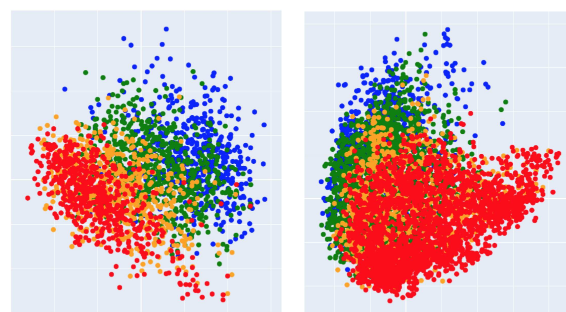
Table 3: Contingency table of the two clustering methods.

captured the semantic change. In Figure 6(a), senses 1 (blue), 2 (orange), and 3 (green) correspond to the most frequently used sense after the change, and the others correspond to the most frequently used sense before the change. Figure 6(b) shows that the usage of senses 1, 2, and 3 gradually increased over time. As a result, we confirmed that the results differed significantly, depending on the sense inventory or the example sentences.

k-means clustering is simple but effective for this task. We found that the k-means clustering-based method did not capture the semantic changes when either of the target senses was rare in the corpus. However, it captured the semantic changes for most other words. We adopted the k-means method as the clustering method in this study; however, it was less effective when there were significant differences in the frequency of sentences that included the meaning before or after the change. We suppose that this is because k-means clustering is not sensitive to the density of clusters.

To assess the influence of other clustering methods, we conducted experiments with a density-based clustering method, DBSCAN (Ester et al., 1996). We tuned hyperparameters to maximize the silhouette score.

Table 3 summarizes the experimental results for two methods. DBSCAN was successful for 11 words, whereas the k-means method was successful for 29 words. In many cases, DBSCAN removed a large amount of data as outliers, which is problematic for semantic change detection. As a result, we found that DBSCAN is not suitable for capturing semantic changes, and k-means is simpler and more effective for the clustering-based method.



(a) Scatter of 全然 for each period (b) Scatter of 教授 for each period

Figure 8: Results for 全然 and 教授 for each period. Blue symbols correspond to the years 1898–1922, orange symbols correspond to the years 1923–1947, green symbols correspond to the years 1948–1972, and red symbols correspond to the years 1973–1998.

We introduce 教養 as an example. In Figure 7(a), the blue data are outliers[7], cluster 2 (orange) corresponds to the usage of non-compound nouns, and cluster 3 (green) corresponds to the usage of compound nouns. Although clusters 2 and 3 consist of the same syntactic feature, most of the data were assigned to special cluster 1 as noise. A large amount of noise data is problematic when trying to capture semantic change.

[7] This is the noise generated by the clustering process of DBSCAN. They are the instances that do not form any clusters.

Vectors obtained with BERT. We confirmed that the clustering-based method forms clusters with metaphors or compound words, as has been discussed in a previous study (Giulianelli et al., 2020), in Japanese. This may be due to the characteristics of the vectors obtained with BERT.

For some words, the distributed word representations obtained with BERT captured the transitions of the distribution over time. We introduce 全然 and 教授 as examples. In Figure 8, the data ranked from oldest to most recent are in the order of red, yellow, green, and blue. The distributions of both words gradually shifted over time. This shows that BERT may capture the contexts that change over time.

6 Conclusion

In this study, we compared two BERT-based methods of capturing semantic changes in Japanese words. We found that the clustering-based method, in particular k-means clustering, performed better than the dictionary-based method. We conducted experiments with different dictionaries and different clustering methods for further analysis. Consequently, we found that the dictionary-based method was sensitive to the dictionary's sense inventories and example sentences, whereas k-means clustering was more effective than DBSCAN.

We also confirmed that the distributed word representations obtained with BERT reflect not only the semantic features but also the syntactic features. Furthermore, the distributions of these representations shift over time. To perform more precise analyses of this time shift, we plan to fine-tune BERT with the target corpus.

Acknowledgments

The work reported in this article was partly supported by the NINJAL collaborative research project 'The Construction of Diachronic Corpora and New Developments in Research on the History of Japanese' and JSPS KAKENHI Grant Number 19H00531.

References

Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. 2021. A Comprehensive Analysis of PMI-based Models for Measuring Semantic Differences. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, November. PACLIC Steering Committee.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

M Ester, H P Kriegel, J Sander, and Xu Xiaowei. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226−231, Portland. AAAI Press.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, July. Association for Computational Linguistics.

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.

Renkui Hou, Chu-Ren Huang, and Kathleen Ahrens. 2020. Language change in Report on the Work of the Government by Premiers of the People's Republic of China. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 100–106, Hanoi, Vietnam, October. PACLIC Steering Committee.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy, July. 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.

Yoko Mabuchi and Toshinobu Ogiso. 2021. An Attempt to Construct a Dataset of Words for Semantic Change Analysis of Modern Japanese. In *Proceedings of the Twenty-seventh Annual Meeting of the Association for Natural Language Processing*, pages 1166–1170, Online, March. Association for Natural Language Processing.

Rousseeuw Peter. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. In *Journal of Computational and Applied Mathematics, Volume 20*, pages 53–65.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic Word Embeddings for Evolving Semantic Discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, page 673–681, New York, USA. Association for Computing Machinery.