Tracing Lexical Semantic Change with Distributional Semantics: Change and Stability

Jing Chen(✉), Bo Peng, and Chu-Ren Huang

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Yuk Choi Road 11, Hung Hom, Kowloon, Hong Kong, China
jing95.chen@connect.polyu.hk, {peng-bo.peng,churen.huang}@polyu.edu.hk

Abstract. Recent studies suggest an increasing interest in detecting lexical semantic changes in the context of distributional semantics. However, most proposals have been implemented with English datasets but not much with Chinese data. This paper thus presents an exploratory study using the popular Skip-gram models and post-processing operations to obtain historical word embeddings, testing whether methods in fashion could capture lexical semantic change in Chinese historical texts. Our results demonstrate a positive answer to this question by suggesting interesting cases which may have undergone the process of meaning generalization and shown competence among homographs. Additionally, our analysis also indicates that social contexts play an important role in lexical semantic change.

Keywords: Lexical semantic change · Diachronic word embeddings · Social context

1 Introduction

Lexical semantic change, studying how lexical meanings have changed over time, has long been discussed in the linguistics community. Historical linguists are mostly concerned with how to conceptualize individual meaning changes into different types of mechanisms, [1–4], and also how to reveal the regularities and constraints in the process of lexical semantic change [5–7].

With the recent advance in the field of natural language processing and the availability of large historical corpora, recent studies situating lexical semantic change at the intersection of linguistics and computer science bring new insights to this ever-young topic. It also serves as a promising methodology to supplement traditional linguistic findings with more empirical evidence from large-scale data [8–13].

Over the past two decades, a growing number of proposals exploiting statistical and computational methods have been put forward and implemented to detect and evaluate lexical semantic change in historical texts [8,11,12,14,15]. Among these proposed methods, most make use of diachronic word embeddings

Q. Su et al. (Eds.): CLSW 2022, LNAI 13495, pp. 244–252, 2023.
https://doi.org/10.1007/978-3-031-28953-8_19

to detect the meaning changes [16–19]. However, current work is mainly conducted for English [11–13], while much fewer attempts have been made for Chinese data [20,21].

This paper serves as an exploratory study on Chinese lexical semantic change, testing whether the predominant method could work in Chinese texts. We trained static word embeddings on People's Daily to obtain word embeddings for four intervals from 1953 to 2003, respectively. The word embeddings are post-processed to align semantic spaces between two adjacent intervals and are then quantified by how far they have moved in the shared space. Our results suggest interesting cases that underwent meaning changes from 1953 to 2003, such as ‘推出 ’ and ‘加盟 ’, and that experienced competence among homographs for dominant usages, such as ‘帅 ’ and ‘’机制 . Besides, we briefly discussed the impact of social contexts on meaning change.

The remainder of this paper is organized as follows. Section 2 briefly summarizes studies exploiting computational approaches to detect lexical semantic changes and then situates our work with previous findings. In Sect. 3, we introduce our data and method. Section 4 describes and also discusses the preliminary results. This paper ends up with a brief conclusion and further steps for our work.

2 Related Work

We have witnessed an increasing number of papers statistically investigating lexical semantic change over the past twenty years. These contemporary works on semantic change roughly relied on two indicators: frequency and distribution. The frequency-based analysis holds a basic idea that changed meaning could be a possible reason for frequency fluctuation [15,16]. For example, the sudden frequency rise of *gay* in the 1980 s does have an underlying meaning change, from ‘happy’ to ‘homosexual’. Compared to frequency signals, distributional information provides more direct evidence to identify meaning change. Simply put, the distribution-based methods generally take the change of context information of target words in diachronic corpora as an approximation of meaning change [22,23].

The distribution-based methods later become predominant and a variety of distributional proposals have been implemented to model meaning change over time [11,14,19]. Earlier work exploits co-occurrence matrices to record co-occurrence patterns and then directly makes comparisons between target intervals [24,25]. Later, the arrival of word2vec revolutionized the method of word representations with significant improvements in precision and efficiency [9] and also spurred the study of using distributional models to detect semantic change. The workflow of applying static word embeddings to semantic change typically is 1)first training individual word embeddings on diachronic corpora for time-sliced intervals, and 2) then post-processing word embeddings by projecting embeddings living in different periods onto the same space for further comparison [11,16,18,19].

Table 1. Overview of four subcorpora: *C1, C2, C3, and C4*

Periods	Word tokens(million)	Word types(million)
C1	164	1.15
C2	98	0.58
C3	155	1.26
C4	176	1.28

Word2vec models generally train one vector for a single word form, which does not differentiate the possible senses of each word. Meanwhile, some proposals, such as state-of-the-art BERT models [26], take this task to a sense-level detection by training different embeddings for different senses. For example, [27] obtain representations for each occurrence of target words using BERT models [26], then clustered them into different usage types making use of the K-Means clustering algorithm, and then measured semantic change with multiple metrics. Other models, but less widely used, such as topic-based analysis, K nearest neighbor analysis, and ensemble models, have also been put forward in existing studies [28–30]. However, the recent SemEval2020 shared task1 working on unsupervised lexical semantic change detection surprisingly indicated that static neural embeddings outperformed other paradigms in both two subtasks: whether a word's meaning changed and to what extent a word's meaning changed [13].

3 Methodology

Building on the insightful findings from previous studies, this paper, as an exploratory study, investigates Chinese lexical semantic change using static neural embeddings, firstly.

Corpora. Detecting lexical semantic change requires the target corpus containing temporal information. Our dataset is a 50-year dataset collected from *People's Daily*, one of the most popular newspapers in China, spanning from 1953 to 2003. To our knowledge, this is the largest dataset with the longest time span that is publicly free to all texts. To continuously monitor lexical semantic change and to cover the most critical milestones in the Chinese history of these 50 years, we split the dataset into four roughly equal periods: from 1953 to 1966 (*C1*), from 1967 to 1978 (*C2*), from 1979 to 1991 (*C3*), from 1992 to 2003 (*C4*).

Preprocessing. We first conduct data cleaning, such as removing all website links, blank lines, and other non-character signs, like '#, $, %, *'. We then exploit *Thulac*[1] package for word segmentation. The statistics of the preprocessed corpora are presented in Table 1.

[1] Thulac, THU lexical analyzer for Chinese. More information could be accessed via https://github.com/thunlp/THULAC-Python.

Shared Vocabulary. The varying sizes for each subcorpora and errors introduced by word segmentation necessitate a normalization of raw frequencies into measures such as instances per million. We assume that words with a normalized frequency larger than 1 are adopted in the lexicon for each subcorpus. To track the change in word usage over the four periods, we first built a shared vocabulary V by intersecting four wordlists for four subcorpora[2], which has a capacity of 12,548 words (see Table 2).

Table 2. Normalized words in each period and the shared vocabulary

	C1	C2	C3	C4
Words	20, 378	18, 576	24, 564	25, 028
Shared vocabulary	12, 548			

Training Models. In line with previous findings, we investigated lexical semantic change with static neural word embeddings in this paper. Such a detection system generally takes three components: a model for semantic representations, an alignment technique, and a change measure [13]. We first train static word embeddings for each subcorpus using the two most popular models, both CBOW and Skip-gram models [9]. We then evaluate the quality of word embeddings using the biggest Chinese word similarity benchmark dataset *COS960* [31]. The results suggest that word embeddings trained by Skip-gram models have higher correlation scores, around 0.5[3] (p <0.01), which are supposed to provide a better basis for further processing.

Post-Processing. We then selected word embeddings trained by Skipgram models for post-processing by aligning every two consecutive models into a shared space using the Orthogonal Procrustes algorithm [18,19], such as projecting the C2 model onto C1's space, and making vectors between two adjacent intervals comparable. C1 and C2 intervals are also aligned as we expect that longer time intervals would capture more statistically salient meaning changes.

4 Results and Discussion

By projecting word embeddings for target intervals onto the same semantic space, we could speculate semantic change by measuring cosine similarities between word embeddings for the same word living in different time intervals. The larger the cosine similarity, the more consistent its contexts across the compared intervals, and possibly the more stable its meaning. On the contrary, the

[2] Nouns referring to institutions and places, numbers, quantifiers, and exclamation words are removed as stop words.

[3] Considering the scale of raw data and the loss of word pairs in each subcorpus, as well as the relation between 'cosine similarity' and 'true similarity', we assume the correlation score here is reasonable.

smaller its cosine similarity, the more possible a word that has changed its meaning.

Our work builds on the intuition that radical social events may bring about significant lexicon changes [32]. Relying on this intuition, we design four roughly equal parts, covering the most important milestones from 1953 to 2003, for comparison and to estimate whether significant semantic changes occurred in resonance with substantial social changes. For our analysis, we first evaluate this question by assessing lexical semantic change between every two consecutive intervals by calculating cosine similarities for each word in the shared vocabulary. Since semantic change is in nature cumulative, we also align the first interval *C1* approximately representing word usage in the 50s with the last interval *C4* approximately representing word usage in the 90s for comparison, which enables us to identify words whose meanings have changed, statistically.

4.1 Interval Comparisons

We compare words in terms of semantic change across every two adjacent intervals by calculating and ranking their cosine similarities. Statistics is presented in Table 3.

If words whose cosine similarities are less than 0.5 are deemed as having significant meaning changes across the discussed intervals, results indicate that most words in the shared vocabulary stay relatively stable across the adjacent intervals, as much fewer words have a score within the range of 0.0 to 0.5, with 1.1% of words for the C1C2 interval, 1.89% for the C2C3 interval, and 0.2% for the C3C4 interval.

Among these three compared groups, the C2C3 interval seems to have more significant changes, while the C3C4 interval has fewer words statistically showing meaning changes. The C2C3 interval roughly represents the time of the Cultural Revolution (the C2 period) and the first 10 years of the Reform and Opening-up (the C3 period), which could be reasonably regarded as a time that underwent radical social changes in the history of Modern China. We also note that the division between C3 and C4 is out of practical considerations, keeping the length of each interval as much equal as possible, and there are no as significant milestones as C2 and C3 in the C4 interval. Therefore, the C3C4 interval is regarded as a period of history with fewer major social changes.

From the statistics here, we speculate that there is a higher correlating possibility between radical social changes and significant lexical semantic changes, which means that social contexts have significant impacts on lexical meaning changes, even for a common vocabulary.

4.2 Word Comparisons

To satisfy the curiosity of whether historical word embeddings capture any meaning changes sensitive to native speakers, we take the C1C4 interval for comparison with the intuition that the longer the time, the easier the changed words could be identified.

Table 3. The number of words on different scales of cosine similarities

	C1C2	C2C3	C3C4	C1C4
[0.0, 0.3)	1	3	0	8
[0.3, 0.4)	9	29	3	62
[0.4, 0.5)	128	204	19	305
[0.5, 0.6)	659	1060	136	1296
[0.6, 0.7)	3246	4324	965	4490
[0.7, 1.0]	8505	6928	11425	6387

In the C1C4 group, we first take words whose similarity score is lower than 0.3 as target words for further investigation. At the risk of oversimplifying, we extract the top nearest neighbors of these target words for preliminary discussions(see Table 4).

Table 4. Changed words with their cosine similarities and nearest neighbors in each period

Word	Cosine similarities	Nearest neighbours	
		C1	C4
越共	0.2049	越盟；共军；叛乱	农德孟；杜梅；阮文
推出	0.2110	推出手；推开；推出去	面世；面市；问世
南越	0.2758	印度支那；败局；泰国；西贡	亡墓；辽代；谒；观音堂
机制	0.2786	链霉素；合霉素；青霉素；人造胶	机制型；新机制；体制；制度化
帅	0.2899	尧舜；挂帅；穆桂英；诸葛	陈子华；程潜；关生；王子健
加盟	0.2910	边疆区；乌克兰共和国；土库曼共和国	现加盟；加盟站；加盟制

Our results demonstrated that the first sixth words of the eight target words, '越共', '推出', '南越', '机制', '帅', '加盟', have shown interesting usage shifts from 1953 to 2003, which further suggest two possible tendencies: generalization, and competence among homographs from dominant usages.

Neighboring words of '推出' are more concrete action verbs in *C1*, while its neighbors are more abstract in *C4* mostly referring to the launch of new products. Similarly, the neighboring words of '加盟' in the *C1* period mostly are countries, and its usage predominantly relates to accessing an influential organization. However, its dominant usage drifted to *franchise-related* concepts, such as stores, in the*C4* period.

The detected words, '南越', '越共', '机制', '帅',are homographs that are competing for dominance in two periods, respectively. For the term '南越', the dominant meaning is highly correlated with the Vietnam War (1955–1975) in the *C1* period. More crucially, the need to refer to Vietnam in terms of two separate entities by the north and south is, in fact, the consequence of the geopolitical situation during the Vietnam war, i.e., Vietnam was divided into two countries to the north and south. The duration of the Vietnam War coincides with the period between C1 and C2. However, the southern Vietnam government in Saigon no longer exists in C4, which cannot compete for the original meaning of ‘南越’, a historical state in Chinese history.

This is corroborated by the changes of ‘越共’, ‘the Vietcong, the Vietnam communist party’. The C1 period is during the height of the Vietnam War, and the collocates of ‘越共’ are historically situated, such as its predecessor ‘越盟’, ‘Viet Minh’, and its role in the war such as ‘共军’, ‘叛乱’, which are underlying its roles as a belligerent in the Vietnam War. During the C4 period, the fight for independence and the Vietnam War were both past histories. Therefore, the term keeps the basic referential meaning of the party and hence is often used close to the characteristics of the party, such as the names of its leaders, ‘农德孟’ and ‘杜梅’. These are interesting cases of usage drifts that indicate the necessity of contextualizing changes in social and historical environments.

Also, the word ‘机制’, is used more frequently for the synthesis products, such as penicillin, in period *C1*. However, it refers more generally to the operating system of an organization or the market in *C4*. The term ‘帅’ was predominantly used as the supreme commander in the period *C1* but was used more frequently as an adjective *handsome* in *C4*.

This is an open question of how social contexts influence word meaning in historical texts. Another related and more fundamental issue is now how to define nominal meanings. Note that nouns having endurant meanings refer to continuants and their denoting meanings do not change over time.

5 Conclusions

This paper exploits Skip-gram models with post-processing operations to obtain diachronic word embeddings to detect lexical semantic changes from 1953 to 2003. As an exploratory study, we briefly discussed the impact of social changes on meaning changes based on this 50-year newspaper dataset.

While we haven’t extensively discussed how meaning has changed in a finer-grained time interval, our results still demonstrate a positive answer to the question of whether the predominant diachronic word embeddings could capture any meaning changes in Chinese data.

Due to the lack of evaluation datasets for the detection task, we leave our detection results under quantitative evaluation at the current stage. In addition, this paper generally focuses on whether and to what extent a word has changed, but does not step into the question of which sense(s) has changed. These questions are served as research tasks in the near future.

References

1. Bloomfield, L.: Language. Rinehart & Winston, Holt, New York (1933)
2. Ullmann, S.: The Principles of Semantics. Glasgow University Publications, Edinburgh

3. Brèal, M., Cust, N., Postgate, J.P.: Semantics: Studies in the Science of Meaning
4. Geeraerts, D.: Diachronic Prototype Semantics: A Contribution to Historical Lexicology. Oxford Studies in Lexicography, Oxford (1997)
5. De Saussure, F.: Course in General Linguistics. Columbia University Press, Columbia (2011)
6. Traugott, E.C., Dasher, R.B.: Regularity in Semantic Change. Cambridge Studies in Linguistics, Cambridge (2002)
7. Zhao, Q., Huang, C.-R., Long, Y.: Synaesthesia in Chinese: a corpus-based study on gustatory adjectives in mandarin. Linguistics **56**(5), 1167–1194 (2018)
8. Michel, J., et al.: Quantitative analysis of culture using millions of digitized books. Science **331**(6014), 176–182 (2011)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Pre-training of deep bidirectional transformers for language understanding, Bert (2019)
11. Tahmasebi, N., Borin, L., Jatowt, A.: Survey of computational approaches to lexical semantic change (2019)
12. Kutuzov, A., Øvrelid, L., Szymanski, T., Velldal, E.: Diachronic word embeddings and semantic shifts: a survey (2018)
13. Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., Tahmasebi, N.: SemEval-2020 task 1: unsupervised lexical semantic change detection. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, December 2020. International Committee for Computational Linguistics (2020)
14. Sagi, E., Kaufmann, S., Clark, B.: Semantic density analysis: comparing word meaning across time and phonetic space. In: Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics, pp. 104–111, March 2009
15. Hilpert, M., Gries, S.: Assessing frequency changes in multistage diachronic corpora: applications for historical corpus linguistics and the study of language acquisition. Literary Linguist. Comput. **24**, 385–401 (2009)
16. Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S.: Statistically significant detection of linguistic change (2014)
17. Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., Petrov, S.: Temporal analysis of language through neural language models (2014)
18. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Cultural shift or linguistic drift? comparing two computational measures of semantic change. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Association for Computational Linguistics, November 2016
19. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change (2018)
20. Tang, X., Qu, W., Chen, X.: Semantic change computation: a successive approach. In: Cao, L., et al. (eds.) BSI/BSIC -2013. LNCS (LNAI), vol. 8178, pp. 68–81. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-04048-6_7
21. Tang, X., Qu, W., Chen, X.: Semantic change computation: a successive approach. World Wide Web **19**, 375–415 (2016). https://doi.org/10.1007/s11280-014-0316-y
22. Harris, Z.S.: Distributional structure. Word **10**(2–3), 146–162 (1954)
23. Firth, J.R.: A synopsis of linguistic theory, 1930–1955 (1957)
24. Gulordava, K., Baroni, M.: 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, Edinburgh, UK, Association for Computational Linguistics, July 2011

25. Rodda, M.A., Senaldi, M., Lenci, A.: Panta rei: tracking semantic change with distributional semantics in ancient Greek. Italian J. Comput. Linguist. **3**, 11–24 (2017)
26. Devlin, J., Chang, M., Lee, K., Toutanova, K.: 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), Minneapolis, Minnesota, Association for Computational Linguistics, June 2019
27. Giulianelli, M., Del Tredici, M., Fernández, R.: Analysing lexical semantic change with contextualised word representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, July 2020
28. Wijaya, D.T., Yeniterzi, R.: Understanding semantic change of words over centuries. In: Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural DiversiTy on the Social Web, DETECT 2011, pp. 35–40, New York, Association for Computing Machinery (2011)
29. Gonen, H., Jawahar, G., Seddah, D., Goldberg, Y.: Simple, interpretable and stable method for detecting words with usage change across corpora. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, July 2020
30. Gruppi, M., Adali, S., Chen, P.: Schme at semeval-2020 task 1: a model ensemble for detecting lexical semantic change (2020)
31. Huang, J., Qi, F., Yang, C., Liu, Z., Sun, M.: COS960: a Chinese word similarity dataset of 960 word Pairs. arXiv preprint arXiv:1906.00247 (2019)
32. Diao, Y.: The Development and Reform of Mainland Chinese in the New Era. Hung Yeh Publishing, Taibei (1995)