Research Proposal: Analysing Lexical Semantic Change in Chinese Language Using Word Embeddings

5525549

February 28, 2024

Abstract

Lexical semantic change is part of the overall evolution of a language, which has been a part of conventional linguistic studies for nearly a century. It has also attracted considerable attention in the field of sociology, as the evolution of languages tends to reflect changes in the social and mass psychology. The emergence of word embedding models has enabled the mining of lexical semantic changes from large-scale diachronic corpora, but little work have been done on the Chinese language. This study plans to use word embedding models to study lexical semantic change in Chinese on the People's Daily news corpus from 1946 to 2023.

1 Introduction

1.1 Semantic Change, Society and Culture

Semantic change, also semantic shift, describes changes in the meaning and usage of words across time. One of its earliest, widely accepted definitions is given by Bloomfield, that semantic change is a class of "innovations which change the lexical meaning rather than the grammatical function of a form" [Blo23]. Most theoretical research since then focused on recording, classifying and analysing different types of semantic change. Bloomfield classified them into nine categories, such as "narrowing" where a word becomes more specific, "broadening" where it becomes more general, etc. [Blo23] More recently, a new class of cultural shifts is proposed in contrast to linguistic drifts: Kutuzov et al. summerized the former as "culturally determined changes in associations of a given word", and the latter "slow and regular changes in core meaning of words". [KØSV18] These two types are not always distinct, as changes within a language is closely related to its surrounding social culture. For example, changes in the meaning and usage of the term 'gay' correlate with LGBT movements and the general attitude towards homosexuality. Researchers in humanities and social sciences can make use of semantic change to study the development of society and solve tasks like temporal information retrieval and detection of trending concepts [YSD+18].

1.2 Semantic Change in Chinese Language

Most works in temporal word embeddings and semantic change detection are done on English. Though more languages should be introduced to increase the diversity and transferablity, the obstacles of apply existing methods to a new language include adjustments of the preprocessing techniques (for example, lemmatisation would be replaced with word segmentation), construction of a diachronic corpora, and collecting human judgements for evaluation. In the relatively unexplored domain of Chinese semantic change detection, this research aims to build upon the limited groundwork laid by predecessors and make further attempts in the aspects mentioned above.

2 Methodology and Design

2.1 Diachronic Chinese Corpus

The dataset used for training and aligning temporal word embeddings come from the Chinese newspaper *People's Daily*, 人民日报图文数据库, which contains news articles from 1946 to 2024, 4.88 GB of local data in total. We also plan to use CCL (Center for Chinese Linguistics PKU) Corpus for acquiring contexualised word embeddings base on a pre-trained language model. CCL provides an online website, CCL语料库检索系统, that allows querying all occurrences of a word in the corpus within a specified time frame and returns the corresponding context windows.

2.2 Temporal Word Embedding

Static word embedding models, represented by Mikolov's famous Word2Vec [MSC$^+$13], is designed upon Firth's distributional hypothesis that ”a word is characterized by the company it keeps” [Fir57]. Based on a further assumption that changes in a word's meaning and usage is reflected through its collocational patterns [Hil08], it is a natural attempt to measure such changes using a temporal or diachronic word embedding. The word vectors can be trained and compared across separate time periods, but due to the randomness in training neural networks, word vectors trained on different time slices would fall into different vector spaces, and must be fitted into a unified coordinate system before comparison. [KARPS15]. Hamilton et al. apply Orthogonal Procrustes to align their embeddings [HLJ16], the method is accepted as a benchmark by some of the peers, but there are also criticisms, for example, that it is hard to “distinguish artifacts of the approximate rotation from a true semantic drift” [BM17]. Other approaches to bypass the rotation issue include using a frozen pre-trained atemporal target embedding as a ‘compass’, so that word vectors from all time slices naturally lie in a shared coordinate system [CBP19].

We plan to compare the effectiveness of these two alignment methods on the People's Daily database. First, we evaluate the word embeddings obtained on each time slice as if static, and test the quality of each model using the Chinese word similarity dataset (COS960) [HQY$^+$19]. We then use ChiWUG [CCS$^+$23], A Graph-based Evaluation Dataset built upon the framework DURel [SSiWE18] using over 61,000 human semantic relatedness judgments, to evaluate how well the aligned embeddings can capture (whether a word's meaning has changed) and measure (to what extent has the meaning changed) semantic change against human judgement metrics. This can be done, for example, by comparing the cosine similarity of a word's embeddings from two decades to the CHANGE score in ChiWUG through their Spearman correlation score.

2.3 Contextualised Representation

The above single representation models have a common flaw, however, that a word can only be represented as one single vector in each time period, which is not precise for polysemous terms with various senses. Pretrained contextualised models are naturally more ’expressive’ in representing polysemous words, as they can produce different representations for the same word according to their contexts. Giulianelli et al. apply K-Means to partition a word's usage representations under different contexts into clusters and treat the clusters as ‘senses’, but there remain difficulties in the selection of k and the interpretation of the results. [GDTF20] Another possible solution is to calculate target sense representations with example sentence from authoritative dictionaries and use them as a benchmark, grouping a word's contextualised usage to the target sense with the highest similarity for each occurrence. [HLL19] Ideally, the second method would be more reliable, for it makes use of human knowledge (compiled dictionaries) and guarantees more interpretable sense embeddings. However, to acquire the target sense representations, there must be an authoritative dictionary with (1) a comprehensive record of every existed meaning of a word, including outdated ones; (2) an adequate amount of sentences for each meaning. We do not have such example sentence bases in Chinese, so clustering would be more applicable than classification.

Taking the same target words from the ChiWUG dataset, we plan to search for their occurrence in the CCL Corpus and use the context windows on Google's pre-trained Chinese BERT [DCLT18] to acquire the contextualised word sense embeddings. We then perform clustering on these embedding vectors and look at how the clusters vary between time slices. The objective then is to answer the same questions (1) whether there is a semantic change; and (2) if so, how significant is the change; based on how the number of clusters and the number of occurrences in each cluster change cross time. Our hypothesis to be tested is that contextualised embeddings should be able to give answers closer to human judgement than using single representation vectors.

2.4 Temporal Word Analogy Test

Test sets that requires a large amount of human-annotated data, for example, rating on how similar or different two words' meanings are, are often limited to few (mostly just two) time periods and a dozen of words, small, and also subject to bias of the annotators. One compromise is to use the temporal word analogy tests [Szy17] [YSD$^+$18]. While traditional word analogies aim to describe the relationship “word w1 is to word w2 as word w3 is to word w4”, temporal word analogy looks for “word w1 at time t1 is like word w2 at time t2”. For example we have “Ronald Reagan in 1987 is like Bill Clinton in 1997” and “Walkman in 1987 is like iPod in 2007”. Though the change of a word's connotation or denotation according to the outside word is “less interesting” than changes within language itself from the linguistic perspective, the question "who is the US president in 1987?" or "what do people listen to music with in 2007" are much more objective and easy to answer than what a certain word 'means' in a specific year, so the query test is very useful in evaluation on, for example, the alignment quality of the models. In English there are several existing test sets based on public knowledge, such as the US presidents, NFL Superbowl champions, etc. We may construct a similar test set based on Chinese society changes, with analogy tests like "习近平*(Xi Jingping)* in 2023 is to 毛泽东*(Mao Zedong)?* in 1949", "手机*(mobile phone)* in 2023 is to 呼机*(beeper, pager)?* in 2000", or "微信*(Wechat, a messenger software)* in 2023 is to 短信*(text message)?* in 2000, 传真*(fax)?* in 1995 or 信函*(letter)?* in 1970". Note that this test set would be suitable for testing temporal word embeddings, as it mainly involves finding the most similar word in an aligned vector space, but makes less sense to contextualised embeddings.

2.5 Further Exploration

Once we have proved the effectiveness and chosen a best pipeline for using single representation and contextualised embeddings respectively, we can further use them to test some linguistic theories or hypothesis. For example, we may quantitatively assess the Law of Differentiation (LD) and the Law of Parallel Change (LPC), two contradictory hypotheses in linguistics that argues on whether the meaning of synonyms tend to diverge or change in parallel over time. [LKD23] Moreover, Hu et al. compares the evloution of word meanings to the evolution in ecology, proposing the concepts of sense competition and sense cooperation, which is modeled and proved through their representation of fine-grained word senses based on deep contextualized word embedding. [HLL19] Based on this ecological view, we may also propose that "competition" not only happens between senses of the same word: for example, if one meaning of a word dies out, another word with the meaning may be used more frequently in that context to take over its "niche". This can be analysed through the change of the (normalised) size of sense clusters using contextualised word embeddings.

3 Work Plan

1. March: Collecting, cleaning and preprocessing data from the *People's Daily* website, choosing an adequate tokenizer, as it would have a great impact on the quality of splitting Chinese characters into "words", and thus affecting the performance of word embedding models. We may choose from pkuseg [LXZ+19] or THULAC [LS09], two toolkits for Chinese word segmentation with good reputation in academia. (two weeks)

   Training and aligning word embeddings on time slices using Hamilton's method of Orthogonal Procrustes [HLJ16] and Carlo's frozen atemporal compass approach [CBP19], test them on the COS960 [HQY+19] and ChiWUG [CCS+23] to find an optimal pipeline (tokenizer + model + aligning method). Also find out how the number of slices or the size of each size influence the quality of each temporal word embedding, choose a balanced number so that the embeddings would be robust enough for analysing word meanings, while preserving the ability to study more fine-grained language changes on short time scales. (four weeks)

2. April: Some qualitative analysis: use cosine similarity to filter words with potentially significant changes in meaning, then use the neighbourhood of words to observe and verify such changes, as well as to plot the transition paths of the word vectors. Compare the findings to common knowledge or intuition, and to the observations of linguists. Pick out words that may be interesting for case studies in the contextualised embeddings stage. (two weeks)

3. May: Perform clustering on the contextualised word/sense embedding of words at different time periods based on Google's pre-trained Chinese BERT, measure changes in the frequency of meanings and usages among the clusters, design another metric to represent the level of language change, test it again with the ChiWUG [CCS+23] dataset to see if this new metric beats cosine similarity for the aligned time slices. (three weeks)

   Perform case study on specific words chosen from previous stages to find whether they had experienced a specific type of change, such as broadening (sense cluster grows larger (more variation inside) or new clusters appear), narrowing (sense cluster grows smaller or old clusters disappear), etc; whether there are patterns between different sense clusters of the same word or similar sense clusters of different words (competition or cooperation); and try to locate exact the period of time in which the changes took place. (three weeks)

4. June: Design temporal word analogy tasks from two perspectives, intra-language change and reflections of changes in the external world. Explore the potential of using diachronic word representations to capture real-world events such as technological developments, changes in people's lifestyles, infectious disease pandemics, armed conflicts, etc. (two weeks)

5. July and onwards: Work on the dissertation report. If applicable, try to build a website to share the results with people (especially linguistics enthusiasts with little or none computing background) potentially interested in the topic. For example, based on the aligned temporal word embeddings, let user query a word and visualise the transition paths of the word and its neighbours in the vector space; based on the contextualised word embeddings, visualise how its sense clusters and their normalised frequency of occurrence change over time.

4 Conclusion

The invention and progression of word embedding models made it possible to explore the evolution of word meanings from large-scale diachronic corpus, which has already been widely explored and tested on English corpora. Consider its potential in assisting the discoveries in linguistics and sociology, it is tempting to apply the method on other languages that have not been explored yet, such as Chinese. In this research, we aim to extend the application of word embedding models to the Chinese language, leveraging their capability to capture semantic shifts over time. Through qualitative and quantitative analyses, we will identify and categorize semantic changes, analysing linguistic phenomena that may have implications for sociolinguistics and cultural studies. Additionally, we will explore the potential of temporal word analogies to draw connections between linguistic changes and external socio-cultural transformations. By extending the application of word embedding models to the Chinese language, this research seeks to contribute to the broader field of diachronic linguistics, introducing cross-cultural insights and advancing our understanding of how language evolves in more diverse linguistic environments.

References

[Blo23] Leonard Bloomfield. *Language.* George Allen & Unwin Ltd, london, 1923.

[BM17] Robert Bamler and Stephan Mandt. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389. PMLR, 06–11 Aug 2017.

[CBP19] Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. Training temporal word embeddings with a compass. In *AAAI Conference on Artificial Intelligence*, 2019.

[CCS$^+$23] Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore, December 2023. Association for Computational Linguistics.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Fir57] J. R. Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32, 1957.

[GDTF20] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 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 2020. Association for Computational Linguistics.

[Hil08] Martin Hilpert. *Germanic Future Constructions: A usage-based approach to language change.* John Benjamins, Amsterdam, Netherlands, 2008.

[HLJ16] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 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 2016. Association for Computational Linguistics.

[HLL19] Renfen Hu, Shen Li, and Shichen Liang. 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 2019. Association for Computational Linguistics.

[HQY+19] Junjie Huang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, and Maosong Sun. Cos960: A chinese word similarity dataset of 960 word pairs. *ArXiv*, abs/1906.00247, 2019.

[KARPS15] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[KØSV18] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[LKD23] Bastien Lietard, Mikaela Keller, and Pascal Denis. A tale of two laws of semantic change: Predicting synonym changes with distributional semantic models. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics*, pages 338–352, Toronto, Canada, July 2023. Association for Computational Linguistics.

[LS09] Zhongguo Li and Maosong Sun. Punctuation as implicit annotations for Chinese word segmentation. *Computational Linguistics*, 35(4):505–512, December 2009.

[LXZ+19] Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. Pkuseg: A toolkit for multi-domain chinese word segmentation. *CoRR*, abs/1906.11455, 2019.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[SSiWE18] Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Szy17] Terrence Szymanski. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[YSD+18] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining.* ACM, February 2018.