Semantic Change Detection Based on Word Embedding

5525549

December 4, 2023

Abstract

Lexical semantic change, an important part of the evolution of language, has been observed and studied by linguists since about a century ago. It also attracted the attention of sociologists, as the changes of a language often reflect changes in society and mass psychology. The invention of word embedding models made it possible to explore the evolution of word meanings from large-scale diachronic corpora. Temporal word representations can be used capture real-life events, such as technological developments, changes in people's lifestyles, epidemics and armed conflicts, through which we may study the development of human society.

1 Lexical Semantic Change

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. Among them there is the "narrowing" of word meanings, cases where a word becomes more specific, for example, the Old English word 'mete' refers to all solid food, but its descendent 'meat' is narrowed down to the flesh of animals that are used as food. "Broadening", on the contrary, happens when a word becomes more general, such as 'dogge' used to refer to a specific breed of mastiff and bulldog, while nowadays 'dog' becomes a general term for the entire species. [Blo23] More recently, as the context and speaker's attitude are taken into consideration aside from the spoken or written text itself, a new class called 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]

1.1 Linguistic Approach

Studies on semantic change in the field of linguistics mainly looked at qualitative measures. Specific words are picked as examples to demonstrate various types of semantic change in the course of language evolution. With the rise of corpus linguistics, statistical tools are also introduced to perform some quantitative analysis, for example counting and comparing the how the frequency of a word's meanings or usages are distributed over time. However, to determine the meaning or usage of a word in a specific sentence within the corpus, linguists have to specify certain classification rules for each word and label the corpus manually. The size of these manually labelled corpus is very limited, and the accuracy relies greatly on the experience and judgment of the annotator.

1.2 Computational Approach

The advent of word embedding models frees researchers from searching for semantic change purely by observation. Now that word vectors can be extracted from co-occurrence patterns, we can represent word meanings using these vectors and compare their similarities. This also makes it possible to detect or "mine" semantic changes for massive diachronic corpora with less human intervention. The new methods can not only verify or challenge the phenomena and patterns observed and summarized by linguists in the past, they have also been used to brought forward new findings. For example, a recent work by Bastien et al. uses distributional models to assess the Law of Differentiation (LD) and the Law of Parallel Change (LPC), two contradictory hypotheses in linguistics that were both proposed more than a hundred years ago. [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]

1.3 Society and Culture

Changes within a language is often closely related to its surrounding social culture. For example, changes in the meaning and usage of the term 'gay' correlate with LGBT movements in reality, as well as the general attitude towards homosexual in society. Researchers in humanities and social sciences also made use of the mining of semantic change that reflect real-life events, like temporal information retrieval and detection of trending concepts [YSD+18].

2 Word Embedding and Semantic Change Detection

2.1 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 are often trained and compared across separate time periods, and the size of a slice may vary from a year to a decade. There are other approaches, however, such as Rosenfeld and Erk's model in which a word's meaning is represented as a function of the continuous variable, time. [RE18] 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. [KARPS14]. Hamilton et al. apply orthogonal Procrustes to align their embeddings [HLJ16], the method is accepted 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].

2.2 Contextualised Representation

The above models have a common flaw, however, that a single 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. With these models, vectors are generated to represent more abstract 'senses' rather than concrete words. Giulianelli et al. apply K-Means to partition a word's usage representations under different context into clusters and treat the clusters as 'senses', but there are still 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] Compared to the single representation models, contextualised temporal embeddings allows us to find out the precise frequency distribution of a word's senses, thus we may measure and compare the amount and speed of semantic change.

2.3 Evaluation of the Models

It is difficult to evaluate a temporal word embedding model and its findings, for there is no golden standard to tell when and how a word's meaning changed, or rather, did it even change at all? It might be a relief to find that a model has spotted the same phenomenon or arrived at the same conclusion as a human linguist, but what if they failed to agree? Do we have solid evidence to prove that the model is "more correct" than the human, or vice versa? One compromise is to use the temporal word analogy test designed by Szymanski. 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". [Szy17] 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 way 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. [YSD+18]

3 Future Work

Most works in temporal word embeddings and semantic change detection are done on English. A natural thought is that more languages should be included to increase the diversity and to test a method's transferablity. But it is more than taking a model that proved to work well on English and simply apply it on a different corpus. We take Chinese as an example to discuss what needs to be solved in order to evaluate semantic change in the language using temporal embeddings.

3.1 Tokenizer: Character or Word?

In English the morphemes (the smallest meaningful unit) are mostly words. For languages like Chinese and Japanese where a smaller unit, character exists, boundaries between 'words' are not marked by spaces, and word segmentation is required before we can apply any word embedding model. The quality of segmenter will greatly influence the quality of word vectors. The Chinese word segmenter used by Hamilton's work [HLJ16], for example, is not very satisfactory, so instead of grouping words with similar meaning together, characters that often cooccur in a word or a named entity were put into clusters, such as '红', '楼' and '梦' in the book title "红楼梦" (Dream of the Red Chamber). A segmenter should be carefully selected or designed for out task.

3.2 Building the Diachronic Corpus

Another issue is to construct the diachronic Chinese corpora. The difference between classical Chinese, vernacular Chinese (in the late modern era) and the standard Chinese (used today) are so huge that it would be unrealistically ambitious to even think of training, aligning and comparing word vectors on all of them. Similarly, due to the highly variable spelling in Old English, the longest time span researchers would attempt to measure is often restricted to after Early Middle Ages (1150 A.D.) [SKC12]. If one wishes to study the semantic change in Chinese from one of the periods mentioned above, they will also have to decide on an appropriate range of time during which changes of the language would be both recognisable and measurable.

3.3 Obstacles to Contextualisation

We have discussed the advantages of contextualised models over non-contextualised ones on dealing with polysemous words. However, we must also admit that the single-representation models are easier to use. To acquire the target sense representations, there must be one authoritative dictionary with (1) a comprehensive record of every existed meaning of a word, including the outdated ones that are no longer in use; (2) an adequate amount of sentences for each meaning. The Oxford English dictionary is a perfect example, as it offers 'quotations' for senses. Such dictionaries is not available in every language, and would take tremendous time and effort to compile.

3.4 Evaluation Standards

To evaluate the quality of a temporal model, we need to perform the analogy test. In English there are several existing test sets based on public knowledge, such as the US presidents, UK prime ministers, NFL superbowl champions, etc. [YSD+18] This type of tests are more easy to generate, while concepts like technology developments, infectious disease and pandemics, armed conflicts require more human effort to compare (and are still susceptible to some bias). The test sets cannot be directly translated and used on another language, for example, there may be enough occurrence in Chinese texts to tell who the US president is, but most likely it will no work for superbowl champions, thus each language must have its own test sets designed for evaluation.

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. A lot of work is needed to support the transfer, including the selection of an effective tokenizer, the construction of a diachronic corpus with a feasible time frame, and the design of a test set to evaluate the performance of the model.

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.

[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.

[KARPS14] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change, 2014.

[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.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

[RE18] Alex Rosenfeld and Katrin Erk. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 474–484, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[SKC12] Eyal Sagi, Stefan Kaufmann, and Brady Clark. *Tracing semantic change with Latent Semantic Analysis*, pages 161–183. De Gruyter Mouton, Berlin, Boston, 2012.

[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.