Interim Report: Analysing Lexical Semantic Change in Chinese Language Using Word Embeddings

5525549

July 17, 2024

1 Introduction and Background

1.1 Semantic Change, Society and Culture

Semantic change, also semantic shift, describes changes in the meaning and usage of words across time, or as defined by Bloomfield, "innovations which change the lexical meaning rather than the grammatical function of a form" [Blo23]. Early theoretical research on lexical semantic change involved recording and classifying different types of changes, such as "narrowing" where a word becomes more specific, "broadening" where it becomes more general, etc. [Blo23] More recently, cultural shifts are proposed in contrast to linguistic drifts: "culturally determined changes in associations of a given word" versus "slow, regular changes in core meaning of words" [KØSV18]. The boundaries can be obscure, however, as changes within a language are closely related to its surrounding social culture. For example, changes in the meaning and usage of the term 'gay' correlate with LGBT movements and people's attitudes 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 The Computational Approach

Traditionally, studies on semantic change are mostly qualitative, though some quantitative work has been done on relatively small human-annotated datasets. The motivation for introducing a computational approach is to automatically detect semantic changes, perform case studies, and statistically analyse patterns, for example, the law of conformity and the law of innovation: words that are less frequent and more polysemous have higher rates of semantic change [HLJ16].

1.2.1 Temporal Word Embeddings

Static word embedding models, represented by Mikolov's famous Word2Vec [MSC+13], are 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 are reflected through its collocational patterns [Hil08], it is possible to measure semantic changes using temporal or diachronic word embeddings. Due to the randomness in training neural networks, however, word vectors trained on different time slices would fall into different vector spaces and must be fitted into a unified coordinate system for 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 artefacts 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].

1.2.2 Contextualised Sense Clustering

The single representation models have a common flaw, in that a word can only be represented as one vector in each period, which is not precise for polysemous terms with various senses. Pretrained contextualised models are more 'expressive' in representing polysemous words, as they can produce different representations for the same word according to its contexts. Giulianelli et al. apply K-Means to partition a word's usage representations under different contexts into senses clusters [GDTF20]. Other clustering methods include DBSCAN, AP (Affinity Propagation), etc.

1.3 Semantic Change in Chinese Language

Most works in semantic change detection have been done in English. Obstacles to applying existing methods to a new language include adjustments of preprocessing steps, construction of diachronic corpus and adequate evaluation datasets. Our research aims to build upon the limited groundwork laid by predecessors and make further attempts in the aspects mentioned above.

2 Progress and Appraisal

2.1 Data Collection and Preprocessing

To construct our diachronic Chinese corpus, we collected 4.88 GB of raw news text between the years 1946 and 2023 from the Chinese newspaper 人民日报 *(People's Daily)*.

We used the jieba tokenizer to perform word segmentation and remove stopwords according to the list published by the Harbin Institute of Technology. The news was then divided in three ways: two slices (before and after the Chinese economic reform in 1978), 1-year and 5-year.

2.2 Word2Vec Based Models

We constructed Word2Vec-based word embeddings on time slices using Hamilton's orthogonal Procrustes alignment [HLJ16] and Di Carlo's frozen atemporal compass [CBP19].

2.2.1 Alignment Based

Hamilton's alignment method is based on two assumptions, 1) the meanings of most words remain roughly the same over time, and 2) embedding vectors of these words differ by a global rotation over time slices, so that forced alignment of the vector space can be achieved by computing this rotation. Given $\mathbf{W}^{(t)} \in \mathbb{R}^{d \times |\mathcal{V}|}$ as the embedding matrix at time slice t, where d is the vector size and $|\mathcal{V}|$ the vocabulary size, to align two adjacent time slices, t and $t+1$, we optimize

$$\mathbf{R}^{(t)} = \arg \min_{\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}} \left\| \mathbf{W}^{(t)} \mathbf{Q} - \mathbf{W}^{(t+1)} \right\|_F \tag{1}$$

which can be solved using SVD [Sch66]. For multiple consecutive time slices t_1, t_2, t_3, ..., t_n, we first "rotate" $\mathbf{W}^{(t_2)}$ to fit into $\mathbf{W}^{(t_1)}$'s vector space, then $\mathbf{W}^{(t_3)}$ into $\mathbf{W}^{(t_2)}$'s, and similarly for the subsequent slices. Note that each time we perform such "rotation" to achieve alignment between t and $t+1$, only the shared vocabulary is kept for $t+1$, so any new words at $t+1$ that did not exist at t would be discarded, thus vocabulary size would "shrink" over time.

2.2.2 Compass Based

Di Carlo's idea of an "atemporal compass" is based on the hypothesis that most words do not go through semantic changes, and that for a word that actually changed, most words occurring in its context would stay the same. The model involves two stages. First, the original Word2Vec model is applied to the entire corpus; then the weights of the hidden layer (compass) are kept frozen while we update the output layer on each time slice. We take the CBOW based Word2Vec model as an example: given $\langle w_k, \gamma(w_k) \rangle$ where $\gamma(w_k) = \langle w_{j_1}, \ldots, w_{j_M} \rangle$ are the M words in the context of word w_k at time t, the optimization problem for this single training example is

$$\max_{\mathbf{C}^t} \log P(w_k \mid \gamma(w_k)) = \sigma\left(\vec{u}_k \cdot \vec{c}^{\,t}_{\gamma(w_k)}\right) \tag{2}$$

where $\vec{u}_k \in \mathbf{U}$ is the atemporal target embedding of w_k, and $\vec{c}^{\,t}_{\gamma(w_k)} = \frac{1}{M}\left(\vec{c}^{\,t}_{j_1} + \cdots + \vec{c}^{\,t}_{j_M}\right)^T$ is the mean of the temporal context embeddings of $\gamma(w_k)$. Intuitively, during the second step w_k is predicted by combining the global target embedding $\mathbf{U}$ with the local context $\gamma(w_k)$, so that the temporal context embedding $\vec{c}^{\,t}_{j_m}$ of w_k's "neighbour" w_{j_m} is pulled towards the atemporal target embedding $\vec{u}_k$ of w_k, and the resulting $\mathbf{C}^t$ is the output sense representation at time t. The model has an apparent advantage compared to training separately on different time slices and rotate-align afterwards, for new words that did not appear in older time slices can be preserved.

2.3 Evaluation

For evaluation, we have four tasks. The time-independent word similarity test and word analogy test are used to evaluate the quality of word embeddings within each time slice, the diachronic word similarity test to assess the models' ability to measure the degree of semantic change, and the temporal word analogy test to measure the alignment quality between semantic spaces.

2.3.1 Synchronic Word Similarity/Relatedness

The Chinese word similarity datasets wordsim-240 and wordsim-297 [CXL+15] reflect the within-time-period quality of word embeddings. We compute the Spearman correlation ρ (×100) between human-rated relatedness scores and cosine similarity for each word pair. A SoTA system VCWE published by Sun et al. combines character compositionality scores of 57.81 and 61.29 respectively on these two testsets. For alignment-based Word2Vec, the scores on wordsim-240 range from 58.63 to 18.07, and on wordsim-297 they fall between 66.29 and 28.36. For compass-based Word2Vec we have wordsim-240 between 59.65 and 29.93 and wordsim-297 from 57.97 to 42.40.

Empirically, the compass-based model is more robust in preserving the quality of word embeddings within each period. Other factors remain to be studied, for example, how the recency or quantity of text data affects the embedding quality, for we found that later time slices tend to score higher than earlier ones, but it must not be overlooked that over the years, the annual volume of the newspaper experienced fluctuation, showing an overall increase, and by around 2020, the total length of news text per year was twice as much as it was around 1980.

2.3.2 Synchronic Word Analogy

The word analogy task consists of questions in the form of "男人(man) : 女人(woman) :: 父亲(father) : ?". By adding and subtracting within the vector space, we expect to find the word whose embedding vector is closest to vec(女人) - vec(男人) + vec(父亲) to answer the question. Instead of accuracy, that is, how many of such questions are answered correctly, we look at two other metrics, the Mean Reciprocal Rank (MRR) which is defined as

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}[i]} \tag{3}$$

where rank$[i] = j$ if the target answer is ranked as the jth closet result of the ith query. If the answer is not found in the list of results (here we take top-10) for the query, set $\frac{1}{\text{rank}[i]} = 0$. We also take the Mean Precision at K (MP@K) with $K = 1$, 5 and 10, which is defined as

$$MP@K = \frac{1}{N}\sum_{i=1}^{N}(P@K[i]) \tag{4}$$

where P@K$[i] = 1$ if the target word is among the top-K results of the query and 0 otherwise. The word analogy testset [CXL+15] consists of three analogy types: 1) capital-country (687 pairs); 2) state/province-city (175 pairs); and 3) family relationships (240 pairs). We test the embeddings from each time slice on separate categories and also with all three types of analogy combined. A SoTA model JWE by Yu et al. [YJXS17] scored 0.91 (capital), 0.93 (city), 0.62 (family) and 0.85 (total) on accuracy, which can be compared with our MP@1, 0.88-0.11 (capital), 0.98-0.05 (city), 0.63-0.01 (family) and 0.79-0.08 (total). For this test, it is more evident that corpus size would affect the word embedding's ability to draw such analogical inferences, for embeddings trained on 5-year slices tend to outperform those trained on 1-year slices. It is also understandable that embeddings trained on early time slices had lower accuracy, for historically the capital of countries/states/provinces have changed over time, and the testset is constructed based on recent facts. For example, for analogy questions on capital-country relationships like "北京(Beijing) : 中国(China) :: 莫斯科(Moscow) : ?", word embeddings trained on 1960s news text would give the technically not "wrong" but outdated answer "苏联(The Soviet Union)".

Full results (on MRR, MP@5, MP@10) are under analysis and will be given in the final report.

2.3.3 Diachronic Word Similarity

Word similarity/relatedness can not only be calculated and compared across different words at the same time, but also for the same word at different times. ChiWUG [CCS+23], a graph-based evaluation dataset, is built upon the framework DURel [SSiWE18] using over 61,000 human semantic relatedness judgments. To evaluate the temporal word embeddings' ability to capture whether a word's meaning has changed and measure to what extent the meaning has changed, we compare the cosine similarity of a word's embeddings from two decades to the CHANGE score in ChiWUG through their Spearman correlation. The decades are separated according to ChiWUG's setting, 1949-1978 versus 1979-2003. The results are shown in Table 1. The alignment-based model outperformed the compass-based, and for the inner structure of Word2Vec, SGNS (Skip-Gram with Negative Sampling) worked better than CBOW (Continues Bag of Words).

alignment		compass	
SGNS	CBOW	SGNS	CBOW
0.5126	0.4382	0.4796	0.4078

Table 1: Spearman correlation scores between cosine similarities and ChiWUG CHANGE

We also noticed a significant difference in the model's performance between single-character (Table 2) and two-character (Table 3) words, with the former being far inferior to the latter, which might be due to issues with word segmentation. When we look at the example sentences used for human annotation, there are cases like "热", a polysemy character that can appear in "热门" (popular), "热电站" (thermal power plant), and "热心家" (warm-hearted person), where the character is not a Chinese "word" unit on its own but part of a compound word, so we argue that one-character "words" are not suitable for this task and should be excluded from the testset.

alignment		compass	
SGNS	CBOW	SGNS	CBOW
0.0699	0.2098	0.0490	0.2517

Table 2: Spearman correlation scores for single-character words

alignment		compass	
SGNS	CBOW	SGNS	CBOW
0.7586	0.6685	0.7148	0.6392

Table 3: Spearman correlation scores for two-character words

2.3.4 Temporal Word Analogy

The temporal word analogy is designed to capture "pairs of words which occupy the same semantic space at different points in time" [Szy17]. The tests can be based on publicly recorded knowledge for particular roles and positions, such as the U.S. president, "Reagan : 1987 :: Clinton : 1997", or emerging technologies and major events, for example "Walkman : 1987 :: iPod : 2007" [YSD+18]. For our Chinese temporal analogy dataset, we only use the first type, more specifically, the names of politicians, which is more reliable since technology innovations often take place asynchronously in different countries, and therefore it may not be accurate if we simply translate the American English testsets. We have selected nine roles or positions, the supreme leader[1] or premier of China, president and secretary of states of the U.S., prime minister of the U.K., president of France, premier of Germany, premier of Japan, and president of South Korea.

We test the analogies on 1-year slices. For years during which more than one person served in the position (for example, in the year of election and transition), any one of the names would be considered a correct answer to the query. However, only those years with one single occupation throughout the time would be used for the query. Similar to the static word analogy test, we report the MRR, MP@1, MP@5 and MP@10 with various maximum time depths. Recall that the alignment-based model always cuts down to a common vocabulary, most of the query names here would have been discarded, therefore we test on the compass-based model only.

[1] during different eras in China, this role may or may not be accompanied by the title "chairman", for example, 邓小平(Deng Xiaoping) wield political power without officially holding any of the "highest" party or government positions, he was nonetheless considered the paramount leader during his era, 1978-1989.

We can see from Figure 1 that the CBOW version of Word2Vec outperformed SGNS, and despite that performance declined as the maximum time depth of the query increased for both models, the CBOW model remained more robust. It is also observed that for certain positions including Chinese premier, German premier and US president, the model performed well with maximum time depth over 70 years, with accuracy (MP@1) above 0.80, while for others like Chinese supreme leader and Korean president, the model performs well within a 20-year time span, but then its accuracy hastily dropped beyond that range. The reasons for such differences remain to be explored.

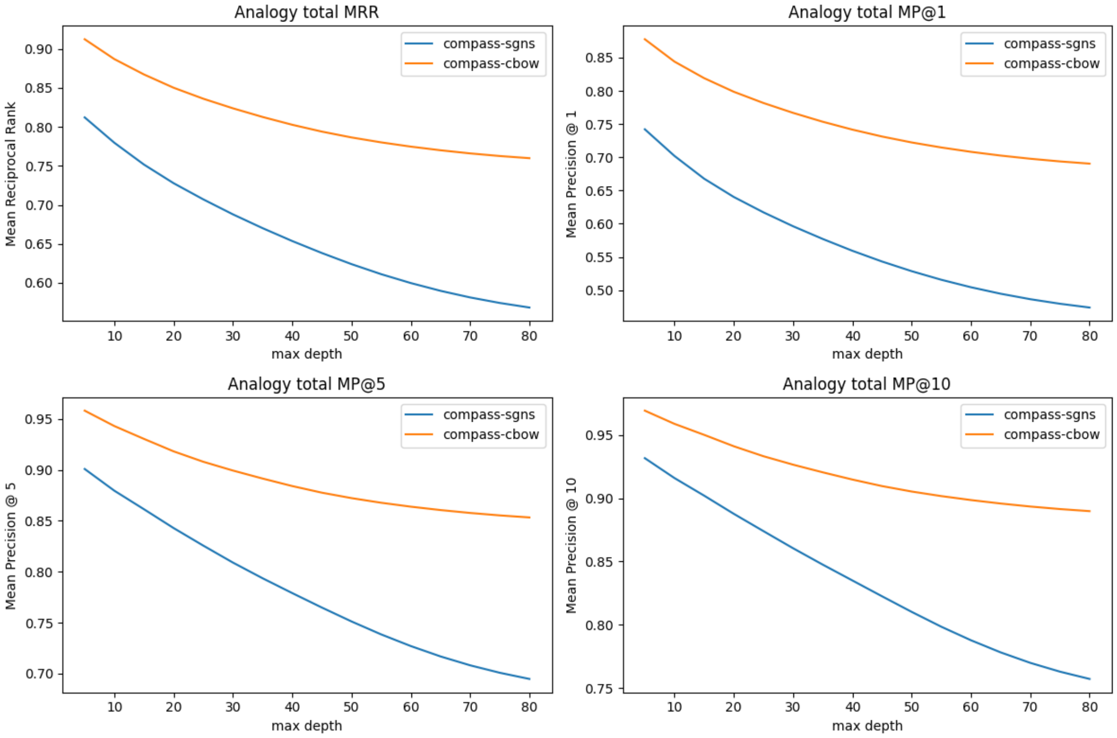

Figure 1: Temporal analogy test on all examples

2.4 Lessons and Reflection

2.4.1 Missing Data

Due to restrictions on access (requires renting of certain library accounts) of the newspaper's official online achieve, 人民日报图文数据库, we collected part of the data (1946-2003) from the non-profit information archive website laoziliao, 老资料网. However, we observed the abruptly low score on word similarity testsets around 1991-1992 as shown in Figure 2, and thus found out that for these two years the data was incomplete on laoziliao, each with only 3-5 month of news data. To resolve the issue, we plan to switch to another resource for the same newspaper, which GitHub user prnake collected, merged from various sources and uploaded to huggingface.

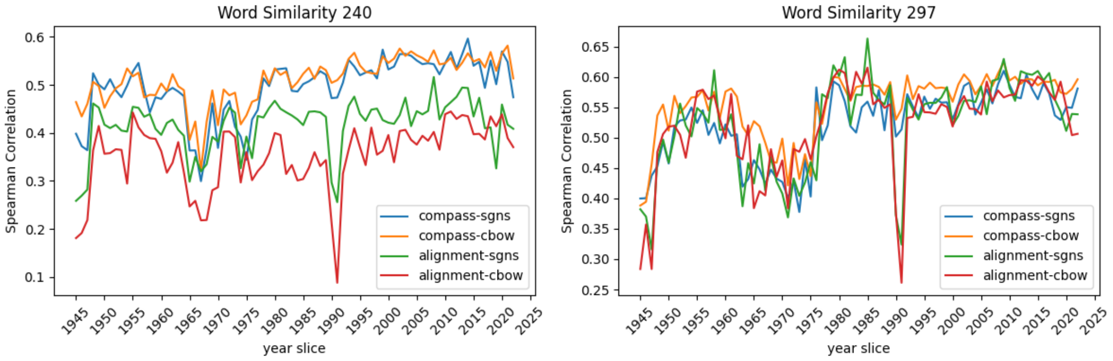

Figure 2: Word similarity test on 1-year time slices

With the new data source, we plan to describe the statistical features of each year first, looking at the number of news articles per month, average length of articles, total word count, etc. to verify whether the dataset is complete and examine how the numbers change over time. This is a lesson learned, to always study the basic features of the dataset before starting the main task.

2.4.2 Version Control and Package Management

We only started using GitHub for version control and Python virtual environment for managing packages at a later stage of the project, when we discovered the compassed-based model relied on a very specific version of gensim Word2Vec: to keep the hidden embeddings frozen as the "compass" during the second step of training, it is most convenient to set the model parameter "learn_hidden" to "False", however, later updates of the gensim library have removed the parameter.

3 Plans and Project Management

For the remaining one and a half months, we plan to start writing the final report while running and evaluating the models again on the new data source. We also plan to carry out some more exploratory case studies with the current models. If time permits, we would like to experiment with the contextualised BERT-based sense clustering as well as provide a website for queries. However, these are not the focus of this current project and may need to be carried on later.

3.1 Exploration

Since the diachronic word similarity and temporal word analogy tasks have proven the model's liability to measure the degree of semantic change and draw inferences on temporal analogical relationships, we may further look at cases that are not covered by the testsets.

3.1.1 Degree of Change

We plan to rank words above a certain frequency (for example, more than 1,000 occurrences per year) by the cosine similarity between their word embedding vectors on the earliest and the latest time slices, and go through the list in ascending order to see if the top words had gone through semantic changes. We may look at a word's occurrences in the corpus as well as its neighbours in the semantic space to manually determine whether or how it has changed.

3.1.2 More Temporal Analogies

We plan to study queries with concepts that are more obscure and difficult to verify, for example technological developments. Table 4 shows the analogy words for mobile messenger app "微信" (WeChat) in 2022, from which we can infer that the semantic space for "means of communication", is occupied by "短信" (text messages) in the early 21st century, "传真" (fax) in the 90s, "电话" (telephone) in the 80s and "信函/书信" (letter) from 1950 to 1970s, etc.

Year	Analogy
1951-1979	信函, 书信 letter
1980-1993	电话 telephone
1994-1999	传真 fax
2002-2007	短信 text
2014-2022	微信 WeChat

Table 4: Analogy word for "微信" (WeChat) 2022

Similarly, we can use query words like "微博" (social media and news platform Weibo), "新冠" (Coronavirus), "俄乌" (Russia-Ukraine) to explore their equivalence, reflecting real-life events like changes in people's lifestyles, infectious disease pandemics, armed conflicts, etc.

3.2 BERT Based Models (further future)

Despite the evident flaw of non-contextualised embeddings that they can only detect how a word changes in its dominant sense, there are obstacles in making use of contextualised embeddings.

One of the biggest issues is interpretability. For the two most commonly used clustering methods, K-Means and AP, the former requires specifying the number of clusters, which is unreasonable since we do not know how many senses there are beforehand, and the latter tends to capture changes in contextual variance (word usages) rather than lexicographic senses (word meanings), often resulting in a much larger number of "usage" clusters rather than sense representations [PM24]. Due to these factors, we chose to use static word embeddings as the main model to study. But it is nonetheless worthwhile to try contextualised embeddings in Chinese semantic change, to provide some new insight into evolving sense clusters and see if the same issues exist as in English.

3.3 Website for Query (further future)

If applicable, we would like to build a website to share the results with potentially interested readers (especially linguistics enthusiasts with little or no computing background) potentially interested in the topic. For example, based on the temporal word embeddings, let the user input a word and visualise the transition paths of the word and its neighbours in the vector space; input a query and return the analogy for each year as a list, etc. Figure 3 shows the path of "导师" (mentor) as we can see its meaning shifts from the metaphoric political/spiritual guide, in the neighbourhood of "马克思(Marx), "列宁" (Lenin), "毛泽东" (Mao Zedong) to the position in academia, in the neighbourhood of "教授" (professor), "博士" (PhD) and "研究生" (graduate student).

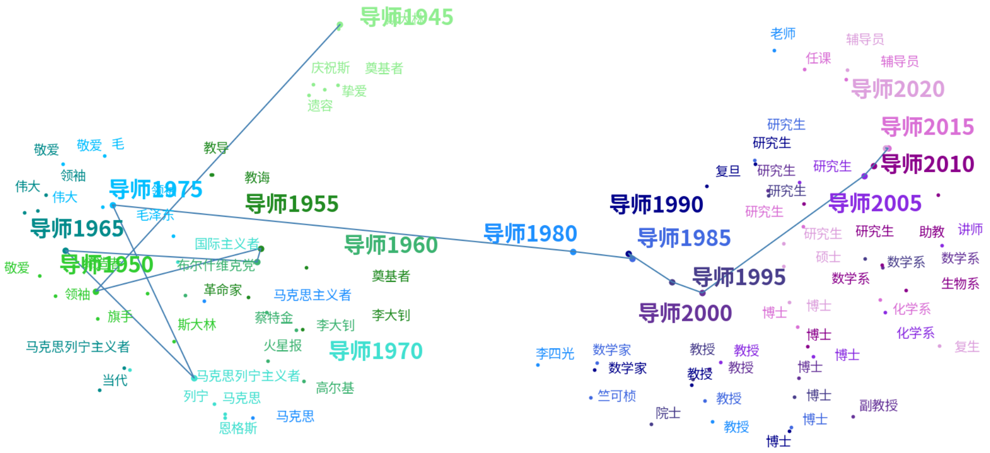

Figure 3: Transition path of "导师" (mentor) in its neighbourhoods

3.4 Project Management

The report is expected to take a month for writing, polishing and a week for final revision.

- Week 1 (22nd Aug): writing background and literature review for the final report; running and evaluating current models on the new data source.
- Week 2 (29th Aug): writing the technical content (model and evaluation designs);
- Week 3 (5th Aug): result analysis, exploration and case studies;
- Week 4 (12th Aug): exploration and case studies continued; writing the conclusion;
- Week 5 (26th Aug): revising and polishing the final report.

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.

[CXL+15] Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. Joint learning of character and word embeddings. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 1236–1242. AAAI Press, 2015.

[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.

[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.

[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.

[PM24] Francesco Periti and Stefano Montanelli. Lexical semantic change through large language models: a survey. *ACM Computing Surveys*, 56(11):1–38, June 2024.

[Sch66] Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10, 1966.

[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.

[YJXS17] Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 286–291, Copenhagen, Denmark, September 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.