UNIVERSITY OF WARWICK

DEPARTMENT OF COMPUTER SCIENCE

Analysing Lexical Semantic Change in Chinese Language Using Word Embeddings

Hanzhi Zhang

MSc Computer Science

Supervisor: Dr. Anna Guimarães

Submission Date: 3. September 2024

Abstract

Lexical semantic change has been a subject of linguistics as a part of the broader topic, language evolution, for nearly a century. It also attracted considerable attention in sociology, as language changes often reflect changes in society as well as mass psychology. Traditionally, works on lexical semantic change would typically involve a couple of handpicked case studies, where changes in a word's meaning and usage are manually proposed, observed and analysed.

Since the invention of word embeddings that could capture semantic relationships among words, it became possible to statistically explore the evolution of word meanings from large-scale diachronic corpora. Word embeddings can capture the meaning of words, thus changes in word sense can be detected and evaluated by comparing the embedding vectors of words at different times. The field has great potential for cross-disciplinary theoretical research (e.g., sociolinguistics, digital humanities, etc.), as well as text-mining-based event discovery and prediction.

This dissertation constructs a diachronic Chinese corpus using People's Daily News text between 1946 and 2023. Temporal word embedding models are then applied and compared on the corpus. Static and temporal word similarity and analogy tasks are designed to test the quality of alignment for these embeddings. The compass-based approach proves to be more robust and balanced across different tasks, therefore it is chosen as our final model for more exploratory work.

With our final model, words with potentially significant changes in meaning are filtered out based on the cosine distance between temporal embeddings across time. We then identify and visualise semantic changes by plotting the trace of a word in its evolving neighbourhood, and manually group the examples into categories: changes regarding part of speech, word sense, domain and collocation, etc. The distances between adjacent time slices are then analysed to identify key periods when a word's meaning changes the most rapidly, which is consistent with the results presented in the trace map. Finally, we discuss how changes in the external world could be reflected in the language by exploring the temporal analogy of roles and concepts. This is demonstrated in the ability of temporal word representations to capture real-life events, including technological innovations, changes in people's lifestyles, infectious diseases and epidemics, armed conflicts, as different words and name entities are observed to occupy the corresponding semantic space during specific eras.

Keywords: Word Embeddings, Diachronic Corpus, Semantic Change Detection.

Contents

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, or as defined by Bloomfield, “innovations which change the lexical meaning rather than the grammatical function of a form” (1923). Early theoretical research on lexical semantic change involved recording and classifying different types of changes, such as “narrowing” where a word becomes more specific, e.g. from ***mete*** in old English for any solid, edible food to ***meat*** for animal flesh; “broadening” where the word becomes more general, e.g. from old English word ***dogge***, a specific breed of mastiff and bulldog to ***dog***, the entire species (Bloomfield 1923). 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” (Hamilton et al. 2016a). The boundaries can be obscure, however, as changes within a language are closely related to the social culture. For example, changes in meaning and usage of the term ***gay*** correlate with LGBT movements and people’s attitudes towards homosexuality. Researchers in humanities and social sciences can use semantic change to study the development of society, solving tasks like temporal information retrieval and detection of trending concepts (Yao et al. 2018).

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 annotation cost limits the size of such corpora, while annotators’ experience and judgement heavily constrain the accuracy. With the development of computational semantics, models like word embedding make it possible to extract semantic relationships from co-occurrence data. Researchers have been using computation approaches to detect semantic changes, perform case studies, and analyse statistical 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 (Hamilton et al. 2016b); sense competition and sense cooperation: similar, related senses tend to cooperate to survive the competition with other more distant ones (Hu et al. 2019). The findings corroborate and complement the laws of semantic evolution previously derived through human analysis and observation.

1.3 Semantic Change in Chinese Language

One challenge in the study of word embeddings and semantic change is the lack of practice in languages other than English (Kutuzov et al. 2018). Although Hamilton et al. used six corpora in four languages (English, German, French, and Chinese) (2016b), the Chinese data had a relatively shorter time span (1950-1999) compared to other languages (over 200 years), and the performance of the chosen segmentation tools was suboptimal. Due to the lack of large-scale diachronic Chinese corpora and standardised benchmark evaluation datasets, the research on Chinese semantic change is relatively preliminary compared to studies in English. Obstacles to applying existing methods to a new language include adjustments of preprocessing steps, as well as the 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 Related Work

2.1 Temporal Word Embeddings

Static word embedding models, represented by the famous Word2Vec (Mikolov et al. 2013b), are designed upon Firth's distributional hypothesis, that "a word is characterized by the company it keeps" (1957). The distributed word representations computed using neural networks can capture semantic relationships, for example, the linear vector calculation vec("Madrid") - vec("Spain") + vec("France") would be approximately vec("Paris") (Mikolov et al. 2013b). Word embeddings have been widely applied in various Natural Language Processing (NLP) tasks, such as part-of-speech tagging, information retrieval, question answering, sentiment analysis, and more.

Static word embeddings assume that the meaning of a word is always fixed, without taking semantic changes into account. Based on a further assumption that changes in a word's meaning and usage are reflected through its collocational patterns (Hilpert 2008), it is intuitive to measure semantic changes using temporal or diachronic word embeddings. Research on temporal word embeddings initially focused on longer time spans divided into larger slices, for instance, Sagi et al. divided English into Early Middle, Late Middle, and Modern English (2012). Smaller time slices can be used to capture more subtle semantic changes related to socio-cultural factors, as demonstrated by Kim et al. (2014) and Liao & Cheng (2016), who divided data into 1-year slices. In addition to treating the corpus as independent time blocks, Rosenfeld & Erk introduced time as a continuous variable, allowing for a more nuanced depiction of the continuity in semantic change (2018).

Due to the stochastic nature of all neural networks' training process, however, word vectors trained on different time slices would fall into different vector spaces and must be fitted into a unified coordinate system for further comparison (Kulkarni et al. 2015). Hamilton et al. apply Orthogonal Procrustes to align their embeddings based on two assumptions: 1) the meanings of most words remain approximately unchanged over time; and 2) if the embedding dimension is large enough (making the optimisation problem less non-convex), the embeddings of these roughly invariant words across different time slices would differ only by a global rotation, and forced alignment of the vector space can be achieved by calculating this rotation (2016b). The method is accepted as a benchmark by some of the peers, but there are also criticisms, that it is hard to "distinguish artefacts of the approximate rotation from a true semantic drift" (Bamler & Mandt 2017). Other approaches include Zhang et al.'s "local anchor", which involves mapping query words to a small neighbour set of "reference points" from the target time slice using linear projections (2016); and Carlo et al.'s"compass", where the pre-trained atemporal hidden embeddings are kept frozen during training, so that word vectors from all time slices lie in a shared coordinate system (2019).

Once the word embeddings from different periods lie within the same vector space, one can explore the degree and direction of semantic change for words. The former is typically measured by the cosine distance between the given word's embeddings at different times, while the latter is examined by observing how the word's neighbourhood shifts over time. Dimensionality reduction can be used to provide an intuitive visualisation of the change path, as illustrated in Figure 1, the word ***gay*** shifts from a neighbourhood of sweet, cheerful to homosexual, lesbian, bisexual; ***broadcast*** changes from scattering, sowing seeds to transmitting information by radio, newspaper or television; and ***awful*** goes from awe-inspiring, majestic to describing something bad and unpleasant.

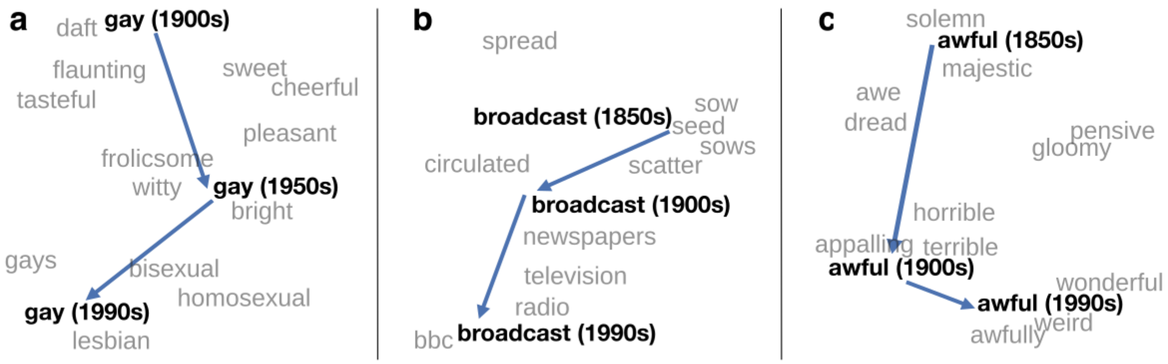


Figure 1: Visualised changes in word meaning (Hamilton et al. 2016b)

2.2 Contextualised Sense Clustering

The single representation models have a common flaw, in that a word can only be represented as one vector for each period, presumably capturing its most dominant sense. This is not precise for words with various senses, where the prominence of one sense may increase while another diminishes. Pre-trained contextualised models (Devlin et al. 2018) are more expressive in representing polysemous words, as they can produce different representations for the same word according to its contexts. Contextualised embeddings of a word are often grouped into sense clusters, which are measured and compared across time to study how new senses emerge while old ones die out. Figure 2 shows how the meaning of ***gay*** shifts from shrinking sense cluster 4, “light-hearted, carefree” to growing ones 2 and 3, adjective “homosexual” and yellow for noun “homosexual person”.

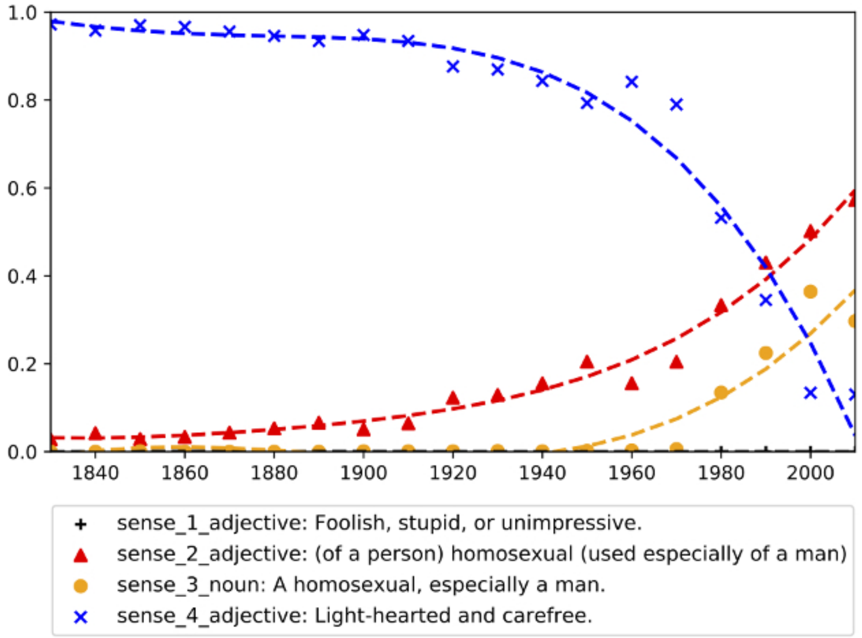


Figure 2: The evolvement of four senses for word ***gay*** (Hu et al. 2019)

There are supervised and unsupervised means of sense clustering. The supervised approach involves calculating target sense representations with example sentences from authoritative dictionaries and using them as cluster centroids, grouping a word's contextualised usages to the target sense with the highest similarity for each occurrence (Hu et al. 2019). The unsupervised approach involves different clustering strategies, including K-Means, DBSCAN, AP (Affinity Propagation), etc.

The supervised method would ideally be more reliable, for it makes use of human knowledge from 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; and 2) an adequate amount of example sentences for each meaning. Hu et al. used the Oxford English dictionary (2019), but such ideal dictionaries are not available in many other languages, including Chinese.

The unsupervised method does not require as much prior knowledge of words' meanings and how they are used, but it also suffers more from difficulties in parameter tuning and the interpretation of results. For K-Means, the selection of number **K** is arbitrary, as we do not know how many sense clusters there should be in advance. There are adaptations to allow more flexibility in the number of clusters, for example, Giulianelli et al. (2020) select a different **K** for each target word between 2 and 10 to maximise the silhouette score (Rousseeuw 1987); the inherent issue of centroid initialisation in the algorithm itself, however, remains unresolved. As an alternative, AP can be employed to let the number of clusters emerge without having to prefix it, but it is sometimes found to produce an unrealistically large number of clusters (i.e., 100) (Periti et al. 2022), which makes it unconvincing to assume each cluster should represent a word sense. This is due to a problem frequently mentioned in publications, the distributional nature of contextualised models to distinguish the contextual variance (word usages) instead of the lexicographic sense (word meanings) (Kutuzov et al. 2022) regardless of the clustering algorithm used. The issue also exists with static word embeddings, but contextualised models are more sensitive to it: a word's usage context may go through noticeable changes despite having a relatively stable meaning (Martinc et al. 2020), thus the detection of semantic shifts in words is more prone to false positives. Finally, the "senses" induced by clusters require additional human effort to map to an external word meanings inventory. The most common approach is to use sentences from the corpus that are closest to the cluster centroid as representative prototypes for generalising a meaning/sense description (Giulianelli et al. 2020).

Apart from clustering methods, there are also variations in how the embeddings are used:

1. The selection of embedding models, such as BERT (Devlin et al. 2018), RoBERTA (Liu et al. 2019), ELMo (Peters et al. 2018), etc. Intuitively, larger models (e.g., bert-base) are mostly believed to guarantee better representation of language features. However, smaller models (bert-small, bert-tiny), have shown competitive performance in detecting semantic change despite having a much smaller number of parameters (Rosin & Radinsky 2022).

2. Type of training, including training from scratch, using a pre-trained model without further tuning, fine-tuning on the entire corpus, and tuning on each separate time slice. Pre-trained models, considering the corpus they are trained on (e.g., Wikipedia), tend to provide a more contemporary representation of words, which can lead to overlooking the temporal aspects. Fine-tuning on the whole diachronic corpus has been proven to improve the quality of word representations for historical texts (Kutuzov & Giulianelli 2020), (Qiu & Xu 2022).

3. Layer(s) from which word embeddings are extracted, e.g., last four, last two, last, etc.
4. How the weights extracted from different layers are aggregated to produce the contextualised word embedding, for example, sum, average, concatenation, etc.

All the above dimensions can be combined in various ways, making the selection, construction, tuning, and comparison of models particularly complex. A comprehensive summary and comparison of different model combinations can be found in the survey by Kutuzov et al. (2022).

2.3 Evaluation and Comparison

For years, the biggest problem in Lexical Semantic Change (LSC) detection was the lack of gold standards. Traditional linguistics research on language change typically consists of a small number of hand-picked examples, which is not sufficient for the evaluation of a system. Moreover, different linguists might have different scales to describe or measure whether or how much a word's meaning changed, thus the ideal testset of a human-annotated list for semantically changed words ranked by the degree of change would be difficult and expensive to acquire.

SemEval-2020 Task 1 (Schlechtweg et al. 2020) is the first shared task for semantic change detection widely acknowledged in the community. It has two subtasks, Binary Change Classification requires the classification of words as stable (without sense(s) loss or gain) or changed (with sense(s) loss or gain), and Grade Change Ranking requires quantification of the extent to which the word lost or gained sense(s). The testsets are constructed using DURel[1] (**D**iachronic **U**sage **R**elatedness) (Schlechtweg et al. 2018), a graph-based evaluation framework that represents the gain and loss of senses for target words with usage graphs, providing a list of high-quality multilingual (English, German, Latin, Swedish) LSC scores based on 100,000 instances of human judgement. Annotators do not rate the degree of change of a word between two periods. Instead, they rate the semantic proximity between usage pairs on a 1-4 relatedness scale, from which the word usage graphs and change scores are automatically generated. The more simple and intuitive judgements significantly reduce the difficulty of annotation and minimise the influence of annotators' subjectivity.

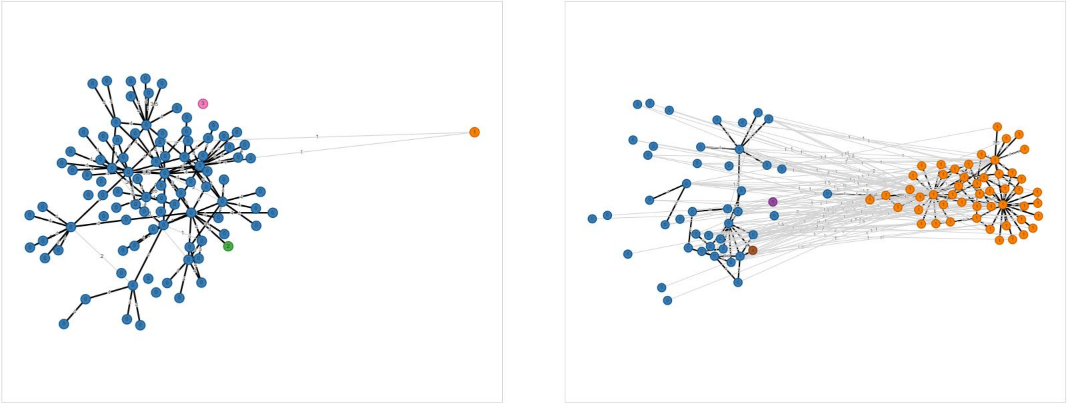


Figure 3: Time-specific word usage subgraphs (Schlechtweg et al. 2018)

[1] `https://www.ims.uni-stuttgart.de/en/research/resources/tools/durel-annotation-tool`

The participating models involve both the static (also form-based/type embeddings) and the contextualised embeddings (also sense-based/token embeddings). The former differ in terms of:

1. Choice of solution to the alignment between time slices, for example, Canonical Correlation Analysis, Orthogonal Procrustes, Temporal Referencing, etc.
2. Choice of function to quantify semantic change, including Cosine Distance, Inverted Cosine Similarity, Euclidean Distance, Local Neighbourhood Distance, etc.

The latter differs in various dimensions as listed in the previous sector (e.g. model training, layer selection, aggregation, clustering algorithm) as well as the choice of the change function, such as Average Pairwise Distance, Jensen-Shannon Divergence, Entropy Difference, etc.

Despite the success of recent contextualised paradigms which proves their superiority over static word embeddings in many NLP applications, the performance of these embeddings on the SemEval shared task is outperformed by their static counterparts. Notably, the highest-performing systems are all based on static embeddings and vary only in their approaches to the alignment issue.

A later survey by Kutuzov et al. explores more different models on a wider range of languages and testsets (2022), and found that no single approach can consistently achieve the best performance across all corpora. This suggests that the effectiveness of a model may be language-dependent, so performing well in one language does not necessarily mean it would also be suitable for others.

2.4 Open Challenges

The study of semantic change using distributional word embeddings has a considerable number of open challenges (Kutuzov et al. 2018), (Kutuzov et al. 2022), including:

1. The expansion of existing methods to a wider scope of languages. The majority of publications are still focused on English corpora due to a lack of multilingual datasets. English resources include Google Book Ngrams[2] and COHA (**C**orpus of **H**istorical **A**merican **E**nglish)[3]. However, it is difficult to find an equivalent of comparable size and time span in other languages.
2. Limited gold standard testsets. Due to the cost of annotation, testsets typically involve a few dozen words from two periods, unable to cover a larger vocabulary across longer time.
3. Interpretation. Certain barriers still exist in the interdisciplinary field. Linguists may not have a deep understanding of computational language models, while computer scientists may lack knowledge of linguistic theory. As a result, studies are constrained to describing specific cases of detected semantic changes without more abstract analyses, such as classifying the changes, identifying their causes, and summarising patterns based on existing theories.
4. Application scenarios. Despite the potential use of semantic change detection in real-world scenarios such as historical information retrieval, lexicography, and linguistic research proposed in a range of literature, most works are still focusing on the theoretical side of the problem, with little progress in actually exploring and realising these possibilities.

[2] https://books.google.com/ngrams

[3] http://corpus.byu.edu/coha/

3 Methodology

3.1 Model Selection

For this work we the static embedding over contextualised models because:

1. It is simpler, requiring less computational power and time to train.
2. Its performance on shared tasks like SemEval-2020 is comparable to that of more recent complex and resource-intensive contextualised models.
3. The results are more straightforward to analyse, demanding less human effort or linguistics expertise, making it suitable as a preliminary approach for work on new languages.

Distributional word representations are drawn from co-occurrence relationships. Mikolov et al.'s Word2Vec represents each word w_i with two dense, low-dimensional vectors: the target embedding $\vec{u}_i$ and context embedding $\vec{c}_i$, both can be used to capture the word's meaning. The model's inner structure has two variations, **C**ontinuous **B**ag **O**f **W**ords (CBOW) and **S**kip-**G**ram with **N**egative **S**ampling (SGNS) (2013a). CBOW predicts the target word appearing in a given fixed-size context window, while Skip-Gram predicts the context (i.e., words that occur within the window) given a target word. In both architectures, the training process places the target embedding of each word in its local semantic neighbourhood with the context embeddings of words that occur with it.

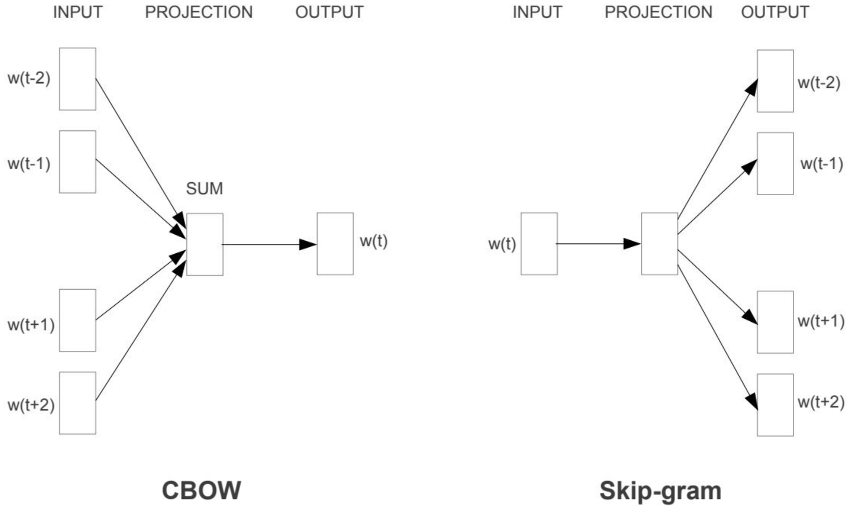


Figure 4: Word2Vec model architectures (Mikolov et al. 2013a)

Most temporal word embedding models struggle with the trade-off between dynamism and staticness in finding a robust alignment strategy. For this matter, we construct temporal Word2Vec-based word embeddings using Hamilton's pairwise orthogonal Procrustes alignment (Hamilton et al. 2016b) and Di Carlo's frozen atemporal compass (Carlo et al. 2019). We run both models on the corpus to evaluate the alignment quality and compare their performance on various tasks.

3.1.1 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 $\mathbf{W}^{(t)} \in \mathbb{R}^{d\times|V|}$ as the embedding matrix at time slice t, where d is the vector size and $|V|$ 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 (Schönemann 1966). 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.

3.1.2 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 $\gamma(w_k) = \left\langle w_{j_1}, \ldots, w_{j_M} \right\rangle$ are the M words in the context of word w_k at time t, the optimisation problem for this single training example is

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

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.

3.2 Diachronic Chinese News Corpus

Relevant works on English corpora have mainly been using time-stamped collections of novels and magazines (e.g., Google Ngrams and COHA) or posts from online communities (e.g., Facebook and Twitter). Yao et al. used the *New York Times* and suggested that news articles are 1) more likely to maintain a consistent language style compared to other text genres and 2) better for studying language evolution from the perspective of social events (2018). Considering these factors, we also chose to implement a Chinese news corpus for this preliminary study.

To construct our diachronic Chinese corpus, we use 4.96 GB of raw news text between the years 1946 and 2023 from the Chinese newspaper 人民日报 *(People's Daily)*. *People's Daily* is one of the most influential newspapers in China with a profound historical legacy. As a newspaper that has been published since before the founding of the People's Republic of China, it has nearly witnessed the entire development of modern Chinese. It covers a wide range of topics, including politics, economics, culture, technology and international affairs, combining authority and breadth, making it a valuable resource for research in linguistics, history, and sociology. Additionally, it stands out as one of the Chinese newspapers with the best digitalised online archive.

3.2.1 Data Collection

The news text collected by GitHub user prnake[4] was merged from various sources including the newspaper's official online achieve, 人民日报图文数据库[5] and non-profit information archive website laoziliao, 老资料网[6] and uploaded to huggingface[7]. We noticed that the data source was missing news from January 1971, and supplemented it with data crawled from laoziliao.

Figure 5 shows a general increase in the number of news articles and the total word count per year. Since the early 1950s when the newspaper was just established, it has grown to nearly five times in size. The number of articles and word count declined in the 1960s and did not recover until the 1980s, which may have been influenced by the Cultural Revolution. The number of news articles per month ranges from 500 to 7,000, with most years averaging between 2,000 and 4,000 articles each month. The average word count per article is between 500 and 1,500. Note that here the "word count" for raw Chinese text without segmentation refers to the number of characters.

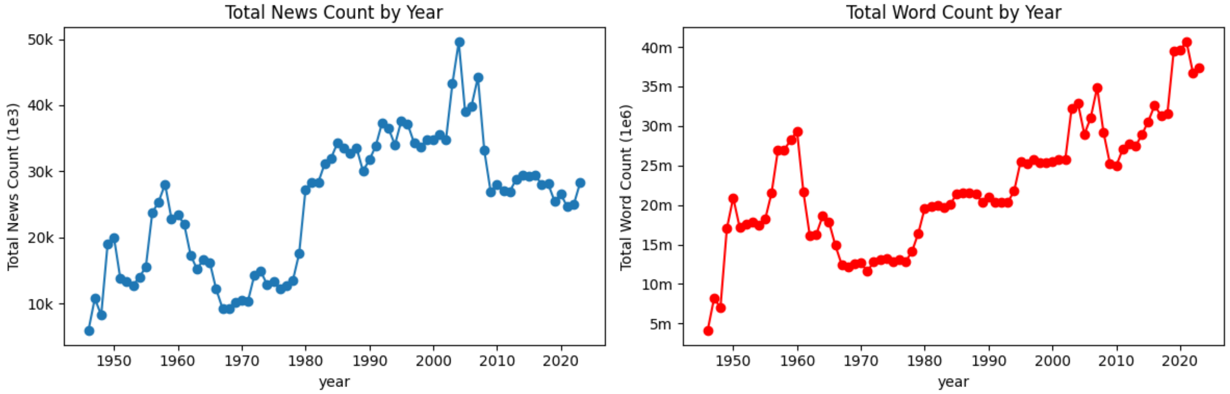


Figure 5: Statistics of raw news data for each year

[4] https://github.com/prnake/CialloCorpus

[5] http://data.people.com.cn/rmrb

[6] https://www.laoziliao.net/rmrb

[7] https://huggingface.co/datasets/Papersnake/people_daily_news

3.2.2 Preprocessing

The LTP (**L**anguage **T**echnology **P**latform)[8] (Che et al. 2021) tool is used for sentence segmentation on news articles. The average length of most sentences is between 40 and 50 characters.

As an isolating language, the quality of Chinese word embeddings is significantly affected by the choice of tokenizer. For example, in the published HistWords[9] models (Hamilton et al. 2016b), 红楼梦(*Dream of the Red Chamber*) was segmented into three separate characters: 红(red), 楼(building), and 梦(dream); 毛主席(Chairman Mao) was segmented into 毛(Mao) and 主席(chairman), etc. Overly fine-grained segmentation resulted in semantic neighbourhoods occupied not by synonyms but by character combinations from common phrases and named entities.

We eventually choose the jieba[10] tokenizer to perform word segmentation. We have also experienced with pkuseg[11] and thulac[12], which claim higher precision and recall scores on certain testsets, but in practice we found despite good performance segmenting Chinese characters, both tokenizers perform poorly on words with numbers and letters, abbreviations, etc. Such words we would ideally wish to keep for later studies, for example when we create a task to query equivalences of ***Coronavirus*** across years, we would like to keep ***H1N1***, ***H5N6*** as potential answers. The segmentation tool provided by LTP seems to be the most effective; however, its runtime is excessively slow (taking ten times as much as many other popular tokenizers), so we eventually did not use it.

Stopwords are removed according to the list published by the Harbin Institute of Technology[13]. The news was then divided in a total of three ways: two slices (before and after the Chinese economic reform in 1978, to match the settings for one of the testsets), 1-year and 5-year.

3.3 Evaluation Design

For evaluation, we have four tasks. The synchronic word similarity and word analogy are used to evaluate the quality of word embeddings within each time slice, or in other words, to demonstrate that the impact of enforced alignment on the embedding quality (its ability to capture time-independent semantic relationships) has been minimised. The diachronic word similarity test assesses the models' ability to measure the degree of semantic change, and the temporal word analogy test measures the alignment between semantic spaces (the ability to capture temporal relationships).

3.3.1 Synchronic Word Similarity

In this task, models are required to compute the semantic similarity/relatedness of given word pairs to test the within-time-period quality of word embeddings. We take the Chinese word similarity datasets wordsim-240 and wordsim-297 (Chen et al. 2015) and compute the Spearman correlation ρ ($\times$100) between human-rated relatedness scores and cosine similarity for each word pair.

[8] https://github.com/HIT-SCIR/ltp

[9] https://nlp.stanford.edu/projects/histwords

[10] https://github.com/fxsjy/jieba

[11] https://github.com/lancopku/pkuseg-python

[12] https://github.com/thunlp/THULAC-Python

[13] https://github.com/goto456/stopwords

3.3.2 Synchronic Word Analogy

The word analogy task consists of relationship questions in the form of “男人(man) : 女人(woman) :: 父亲(father) : ?”. With the word offset technique of simple algebraic operations, we expect to find the word whose embedding vector is closest to vec(女人) - vec(男人) + vec(父亲) should be 母亲(mother). The testset (Chen et al. 2015) has three analogy types:

1. capital-country, 687 pairs: 北京(Beijing) : 中国(China) :: 东京(Tokyo) : 日本(Japan)
2. city-province, 175 pairs: 武汉(Wuhan) : 湖北(Hubei) :: 杭州(Hangzhou) : 浙江(Zhejiang)
3. family (gender), 240 pairs: 父亲(father) : 母亲(mother) :: 丈夫(husband) : 妻子(wife)

We test the embeddings from each time slice on separate categories and also with all three types of analogy combined. Instead of accuracy, that is, how many 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 $\text{rank}[i] = j$ 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 $\text{P@K}[i] = 1$ if the target word is among the top-K results of the query and 0 otherwise.

3.3.3 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 (**C**hinese **W**ord **U**sage **G**raph) (Chen et al. 2023) is built with the DURel (Schlechtweg et al. 2018) framework using over 61,000 human semantic relatedness judgments. The dataset covers 50 years of Modern Chinese from 1953 to 2003, split into two sub-corpus, **C1** (1954-1978) and **C2** (1979-2003). The pivot point 1978 is chosen according to the *Reform and Opening-up*, a remarkable social transformation that is believed to have led to substantial changes in China and the lexicon of Modern Chinese.

To evaluate the embeddings’ ability to decide whether a word’s meaning has changed and measure to what extent it has changed, we compare the cosine distance$(\mathbf{v^{C1}}, \mathbf{v^{C2}}) = 1 - \frac{\mathbf{v^{C1}} \cdot \mathbf{v^{C2}}}{\|\mathbf{v^{C1}}\|\|\mathbf{v^{C2}}\|}$ of a word’s vectors from **C1** and **C2** to the ChiWUG CHANGE score through Spearman correlation. Decades are separated according to ChiWUG’s original setting, 1954-1978 versus 1979-2003.

3.3.4 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" (Szymanski 2017). The tests can be based on 1) publicly recorded knowledge for particular roles and positions with different names listed each year, such as the U.S. president, "Reagan : 1987 :: Clinton : 1997", and 2) more interesting but less clearly-defined examples like emerging technologies, brands and major events, such as "Walkman : 1987 :: iPod : 2007" (Yao et al. 2018). 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, therefore it may not be accurate if we simply translate the English testsets. We selected nine roles or positions, the supreme leader[14] and premier of China, president and secretary of states of the U.S., prime minister of the U.K., president of France, chancellor of Germany, prime minister 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 within a question. Similar to the static word analogy test, we report the MRR, MP@1, MP@5 and MP@10 at various time depths from 5 to 80 years. 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.

4 Results

4.1 Quality of Word Embeddings

A SoTA system VCWE published by Sun et al. (2019) combines character compositionality, scoring **57.81** and **61.29** respectively on the wordsim-240 and wordsim-297 datasets. First, we compare the results across the two largest time slices **C1** (1954-1978) and **C2** (1979-2003), from Table 1 and Table 2 it can be observed that alignment-based and compass-based methods, as well as SGNS and CBOW, exhibit varying performance across different subcorpora and word similarity test sets. All models consistently achieve higher scores on **C2** compared to **C1**. Given that the test set annotations are based on recent judgments, it is reasonable as more recent news better reflects current language usage. The performance of the models on **C2** is comparable to SoTA, thus the Word2Vec models and the news corpus used are sufficient for modelling Chinese semantic relationships.

	alignment		compass	
	SGNS	CBOW	SGNS	CBOW
C1	**51.60**	44.86	49.15	44.49
C2	**56.52**	50.89	56.25	46.10

Table 1: Spearman correlation ρ (×100) on wordsim-240

[14] during different eras in China, this role may or may not be accompanied by the title, 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.

	alignment		compass	
	SGNS	CBOW	SGNS	CBOW
C1	54.88	51.97	53.79	**55.14**
C2	58.47	57.36	**60.28**	56.00

Table 2: Spearman correlation ρ ($\times$100) on wordsim-297

Next, we examine the results with smaller time slices. For five-year slices, the scores for wordsim-240 range from 31.55 to 58.10, while wordsim-297 varies between 44.52 and 61.66. With one-year slices, the scores for wordsim-240 fall between 15.41 and 60.57, whereas wordsim-297 ranges from 25.52 to 66.51. Empirically, the compass-based model is more robust on these smaller time slices, consistent with the claims by Carlo et al. (2019) (see Figure 6 and Figure 7).

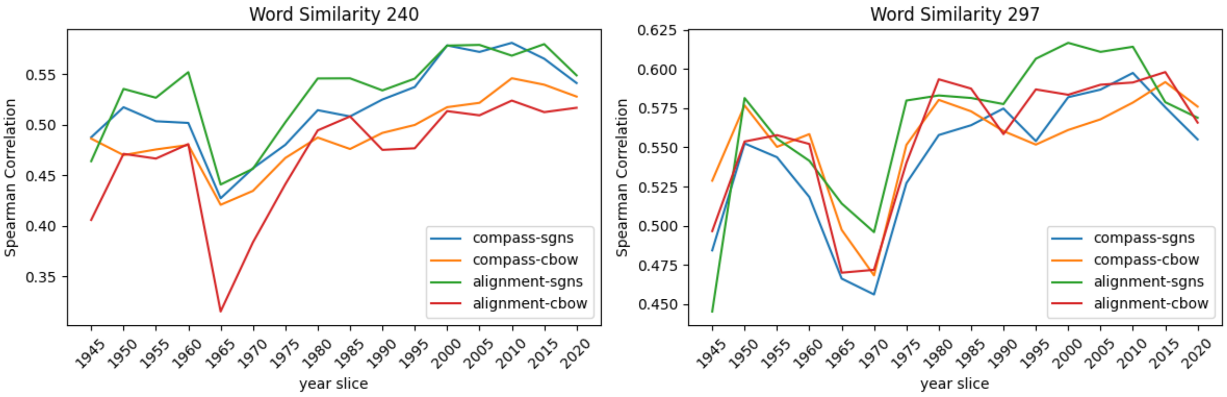


Figure 6: Word similarity tasks on 5-year time slices

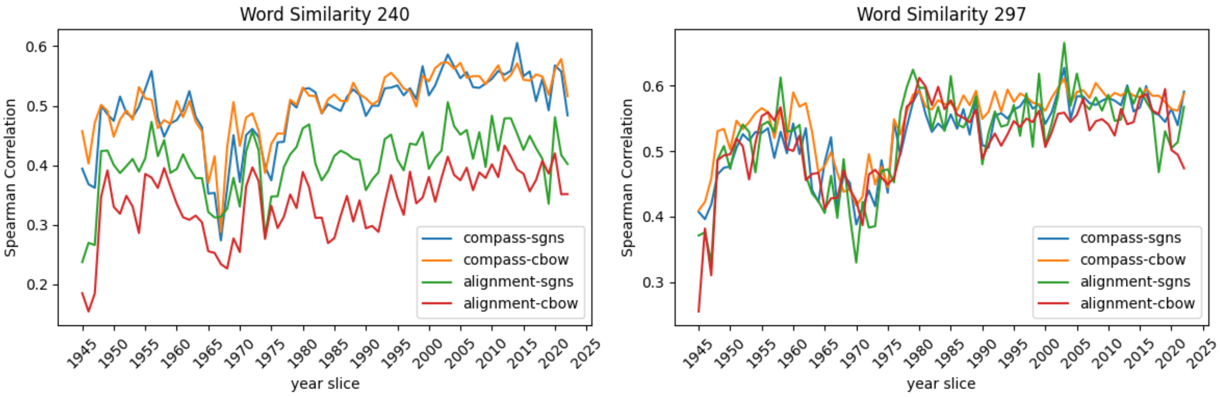


Figure 7: Word similarity tasks on 1-year time slices

However, the impact of text recency and quantity on embedding quality is complicated. Similar to observations on the 25-year time slices, later slices generally score higher than earlier ones. It is important to note that over the years, the annual volume of the newspaper fluctuated with an overall increase, as illustrated previously in Figure 5. The Spearman score ρ ($\times$100) between time (year) and yearly word count is 75.48. Here we present ρ between models' wordsim scores with one-year time slices, the years and the total word count that year (see Table 3, Table 4). The scores from the compass-based model are more strongly correlated with both parameters. It is also noted that wordsim scores from the 1960s and 1970s are even lower than the 1940s. Besides the reduction in the news volume during these years, this may also be related to the distinctive language style prevalent during the Cultural Revolution (Wei 2014) and the limited range of topics and content covered by *People's Daily*, an official party and government media, during that period.

	alignment		compass	
	SGNS	CBOW	SGNS	CBOW
time (year)	45.97	51.43	68.63	78.04
word count	57.74	70.89	75.70	78.78

Table 3: Spearman correlation between year, word count and wordsim-240 score

	alignment		compass	
	SGNS	CBOW	SGNS	CBOW
time (year)	51.43	48.33	76.21	69.82
word count	59.35	60.89	74.35	71.09

Table 4: Spearman correlation between year, word count and wordsim-297 score

For the word analogy task, a SoTA model JWE by Yu et al. (2017) scored **0.9188** (capital), **0.9371** (city), **0.6250** (family) and **0.8505** (total) on accuracy, which can be compared with our MP@1.

		alignment		compass	
		SGNS	CBOW	SGNS	CBOW
capital	MRR	**0.7365**	0.7334	0.6393	0.6239
	MP@1*	**0.6632***	0.6455*	0.5347*	0.5229*
	MP@5	0.8390	**0.8419**	0.7755	0.7637
	MP@10	0.8671	**0.8730**	0.8272	0.8272
city	MRR	**0.8486**	0.7617	0.7993	0.7217
	MP@1*	**0.7829***	0.6914*	0.7200*	0.6457*
	MP@5	**0.9257**	0.8571	0.9143	0.8229
	MP@10	0.9371	0.9029	0.9371	0.8400
family	MRR	0.5250	0.5238	0.4940	**0.5366**
	MP@1*	0.4297*	0.4336*	0.4063*	**0.4609***
	MP@5	0.6484	**0.6680**	0.6211	0.6367
	MP@10	0.7148	**0.7578**	0.6641	0.7031
total	MRR	**0.7053**	0.6895	0.6310	0.6192
	MP@1*	**0.6282***	0.6038*	0.5343*	0.5280
	MP@5	**0.8087***	0.8042	0.7617	0.7437
	MP@10	0.8430	**0.8511**	0.8069	0.8005

Table 5: Word analogy on subcorpus **C1** (MP@1* is equal to accuracy)

		alignment		compass	
		SGNS	CBOW	SGNS	CBOW
capital	MRR	**0.9102**	0.8388	0.8625	0.8209
	MP@1*	**0.8493***	0.7533*	0.7858*	0.7267*
	MP@5	**0.9852**	0.9453	0.9527	0.9394
	MP@10	**0.9956**	0.9705	0.9852	0.9764
city	MRR	0.9730	0.9283	**0.9731**	0.9254
	MP@1*	**0.9543***	0.8971*	**0.9543***	0.8800*
	MP@5	**1.0000**	0.9714	0.9886	0.9829
	MP@10	**1.0000**	0.9829	**1.0000**	0.9829
family	MRR	0.6486	0.6938	0.6488	**0.7288**
	MP@1*	0.5352*	0.5977*	0.5391*	**0.6406***
	MP@5	0.7969	0.8398	0.7891	**0.8477**
	MP@10	0.8672	0.8750	0.8867	**0.9141**
total	MRR	**0.8597**	0.8194	0.8306	0.8161
	MP@1*	**0.7933***	0.7401*	0.7554*	0.7310*
	MP@5	**0.9440**	0.9251	0.9206	0.9251
	MP@10	**0.9666**	0.9504	0.9648	0.9630

Table 6: Word analogy on subcorpus **C2** (MP@1* is equal to accuracy)

It can be seen from Table 5, Table 6, alignment-based models generally outperform compass-based ones, and SGNS architecture outperforms CBOW, with the exception of family (gender) analogies. It is also understandable that embeddings trained on subcorpus **C1** had lower accuracy, for historically the capital of countries/states/provinces have been changing. For example, for analogy questions on capital-country relationships like "北京(Beijing) : 中国(China) :: 莫斯科(Moscow) : ?", word embeddings trained on **C1** would give the technically not "wrong" but outdated answer "苏联(The Soviet Union)". The MP@1 scores on **C2** are also comparable to the SoTA accuracy.

We also test on five-year and one-year time slices, reporting only the MMR and MP@1 across all queries. For more detailed scores across different categories, please refer to the code repository.

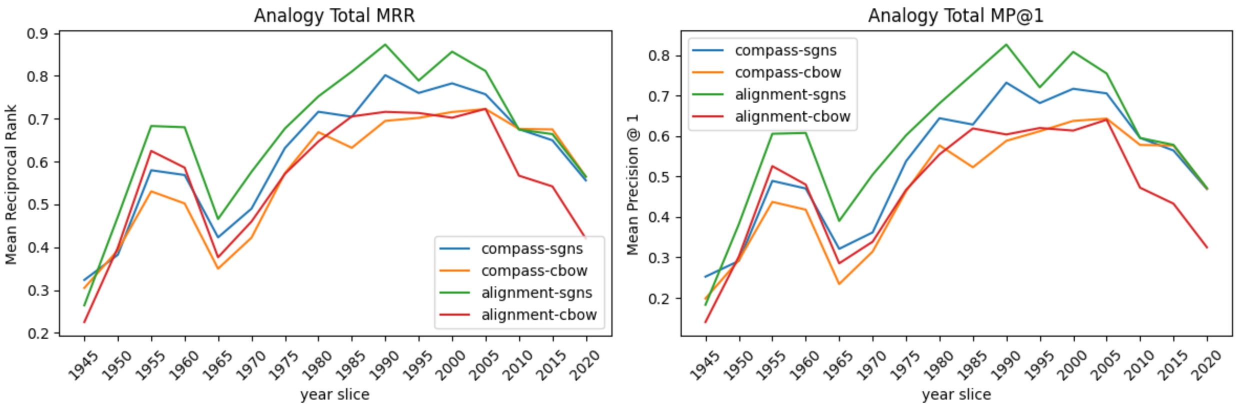


Figure 8: Word analogy tasks on 5-year time slices

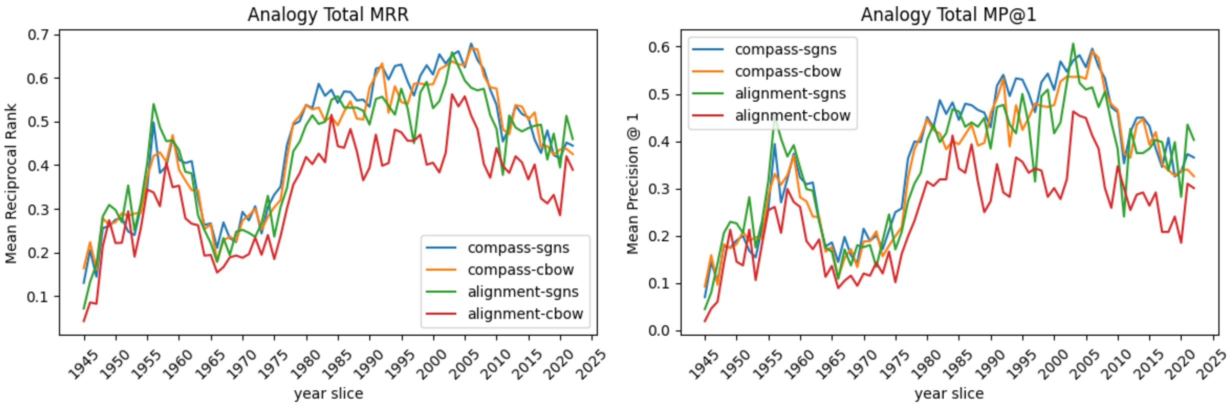


Figure 9: Word analogy tasks on 1-year time slices

For this test, it is more evident that corpus size would affect the word embedding's ability to draw analogical inferences, for embeddings trained on 5-year slices tend to always outscore those trained on 1-year slices at the same point in time. The curve's shape closely mirrors that of wordsim and news/word count, showing an overall upward trend over time, with an anomalous decline around the 1960s, followed by a recovery in the 1980s. Alignment + SGNS perform well on five-year time slices, but its robustness is inferior to that of the compass-based models on one-year slices.

4.2 Degree of Change

The results for the ChiWUG dataset are shown in Table 7. Once again, the alignment-based model outperformed the compass-based, and SGNS worked better than CBOW. Note that the Spearman correlation scores for all models are statistically significant at $p < 0.05$.

alignment		compass	
SGNS	CBOW	SGNS	CBOW
51.97	46.35	43.62	42.34

Table 7: Spearman correlation ρ ($\times 100$) between cosine distance and ChiWUG CHANGE

We also noticed a remarkable difference between the models' performance on single-character and two-character words (Table 8), 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 relatedness annotation, there are cases like 热(hot), a polysemy character that have appeared in:

1. 有过太多的热门话题 (*there were too many trending topics*)
2. 每一个革命者只能做热心家 (*every revolutionary can only be an enthusiastic supporter*)
3. 全厂主要动力的热电站 (*the main power sources of the plant, the thermal power station*)

where the character is clearly not a word unit on its own but part of a compound word. We argue that these one-character "words" are not suitable for this task and should be removed from the testset.

	alignment		compass	
	SGNS	CBOW	SGNS	CBOW
single-character	18.18	23.08	3.50	**32.17**
two-character	**71.26**	66.16	66.68	64.63

Table 8: Spearman correlation ρ ($\times$100) for two-character words

4.3 Temporal Analogy Queries

Recall that we only test temporal word analogy on the compass-based models, for the vocabulary of the alignment-based models only contains words that appeared across all time slices. Consequently, the names of politicians who were not even born 70 years ago, let alone be frequently reported in the news, could not be included. We can see from Figure 10 that the CBOW version of Word2Vec outperformed SGNS, and despite that performance declined as the maximum time depth[15] of the query increased for both models, the CBOW model remained more stable.

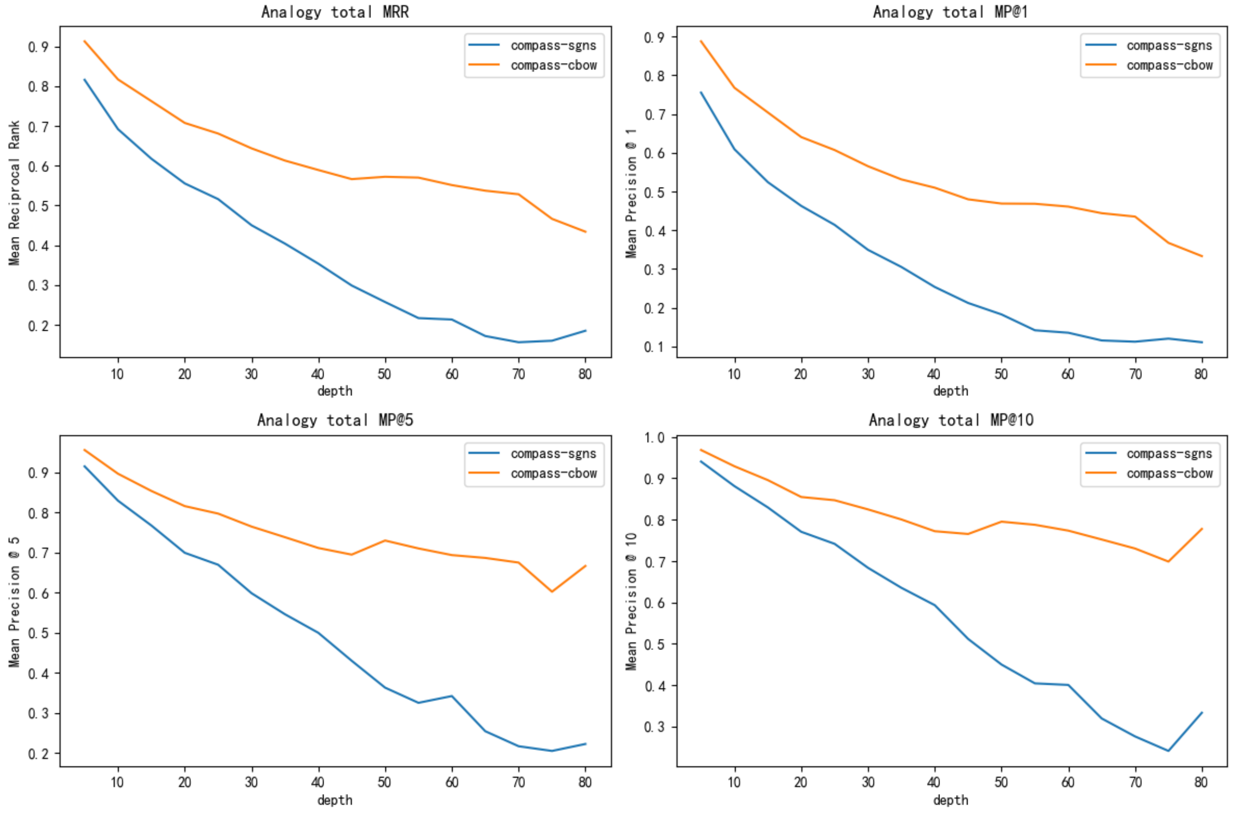


Figure 10: Temporal analogy queries on all politicians

It is also observed that for certain roles including the US president, the UK and the Japanese prime ministers, the model performed very well even at the time depth of over 70 years, with accuracy above 0.80. For others like the German chancellor and the French president, although accuracy

[15] for query "$w_1 : t_1 :: w_2 : t_2$", the time depth is $|t_1 - t_2|$. We slice the time into five-year intervals. For example, 10 on the x coordinate represents the average scores of all queries with a time depth between 5 and 10 years.

dropped with increasing time depth, MP@5 and MP@10 remained relatively high, indicating that the correct answer to the query is still within the approximate semantic neighbourhood. However, with the Korean president, the model performed well within a 20-year time span, but its MP@1, MP@5 and MP@10 all hastily dropped to around zero beyond that range (see Figure 11).

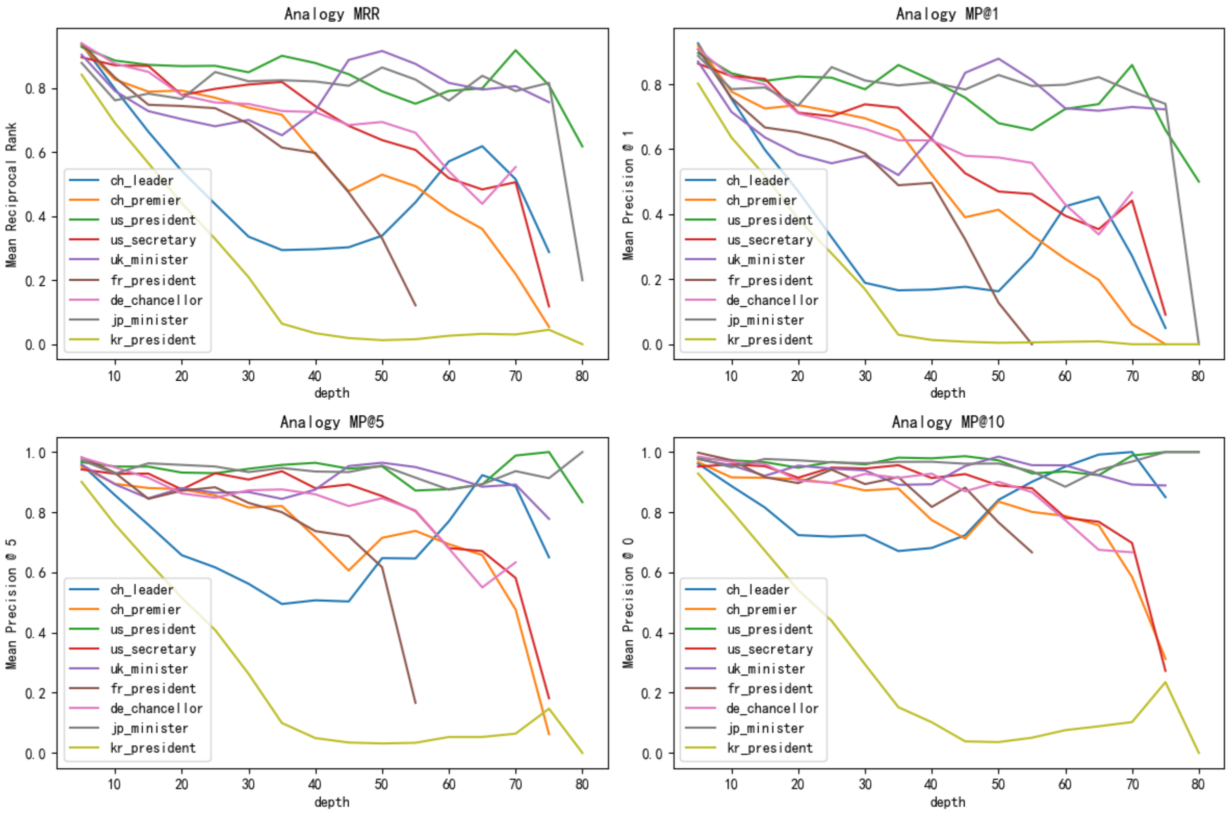


Figure 11: Temporal analogy queries on different politicians

The reasons for this exception remain to be explored. We have ruled out the impact of frequency. As shown in Table 9, the average number of times the Korean president's name appears in the news each year is not the lowest among these politicians; in fact, the UK prime minister appears even less frequently, yet its accuracy remains very high. On the contrary, although the occurrence frequency of the two Chinese leaders is much higher, their scores on the query tasks are only average.

ch leader	ch premier	us president	us secretary	uk minister
6542	2385	1525	539	258
fr president	de premier	jp premier	kr president	
296	231	316	327	

Table 9: Average occurrence per year for 9 national leader positions

We also examined the semantic neighbourhood of potential answers using different years paired with the incumbent South Korean president's name as the query. We observed that over the 30 years since the normalisation of relations and the establishment of diplomatic ties between China and

South Korea in the 1990s, the vast majority of query results have been quite accurate. Before this period, queries were influenced by the political climate and carried strong emotional overtones. For example, in most cases involving state leaders, even when the query returned incorrect answers, the top 5 or 10 results were names of other politicians, with confusion in country, position or term (e.g., a query for the incumbent U.S. president might return the French president, U.S. secretary of state, or a former U.S. president). However, when using Korean president 李承晚(Syngman Rhee) from the Korean War (1950-1953) era as the query, the semantic space included many non-personal terms, such as “美帝国主义” (American imperialism), “走狗” (running dog), and “傀儡政权” (puppet state). As a result, the names returned by the query often did not correspond to the “South Korean president” but to state leaders who played the role of “American imperialism’s puppet” in Chinese official propaganda at different times. For instance, during the Korean War, the semantic space included the name of Filipino president 季里诺(Elpidio Quirino); during the Taiwan Strait crises, the query returned the president of The Republic of China[16] 蒋介石(Chiang Kai-shek), while during the “Khmer Republic” period, this position was occupied by the pro-American Cambodian leader 朗诺(Lon Nol). Other leader names that appeared in this space include 吴庭艳(Ngo Dinh Diem) and 阮文绍(Nguyen Van Thieu) of the “Republic of Vietnam” (South Vietnam).

李承晚(Syngman Rhee)’s appearance in the news further confirms our observation. In the year the Korean War broke out (1950), as China was an ally of North Korea, *People’s Daily* frequently used terms such as “匪帮” (bandit), “帝国主义的走狗” (imperialist lackey), “傀儡” (puppet) and “卖国贼” (traitor) in its reports on the conflict, expressing condemnation of South Korea and support for its ally. These terms often appear close within the context window of 李承晚(Syngman Rhee), leading the model to associate their meanings. This example illustrates the significant impact of the corpus narrative on what is modelled by the embeddings. Temporal queries do not always capture the desired analogical relationships, and the results require manual verification and analysis.

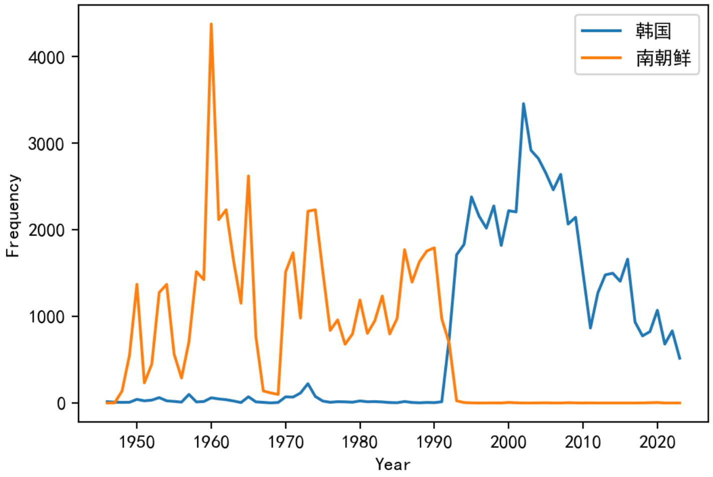


Figure 12: Frequency of “韩国” and “南朝鲜” (1946-2023)

[16] The Republic of China (ROC), commonly referred to as Taiwan, whose political status is contentious, should not be confused with The People’s Republic of China (PRC), also known as Mainland China.

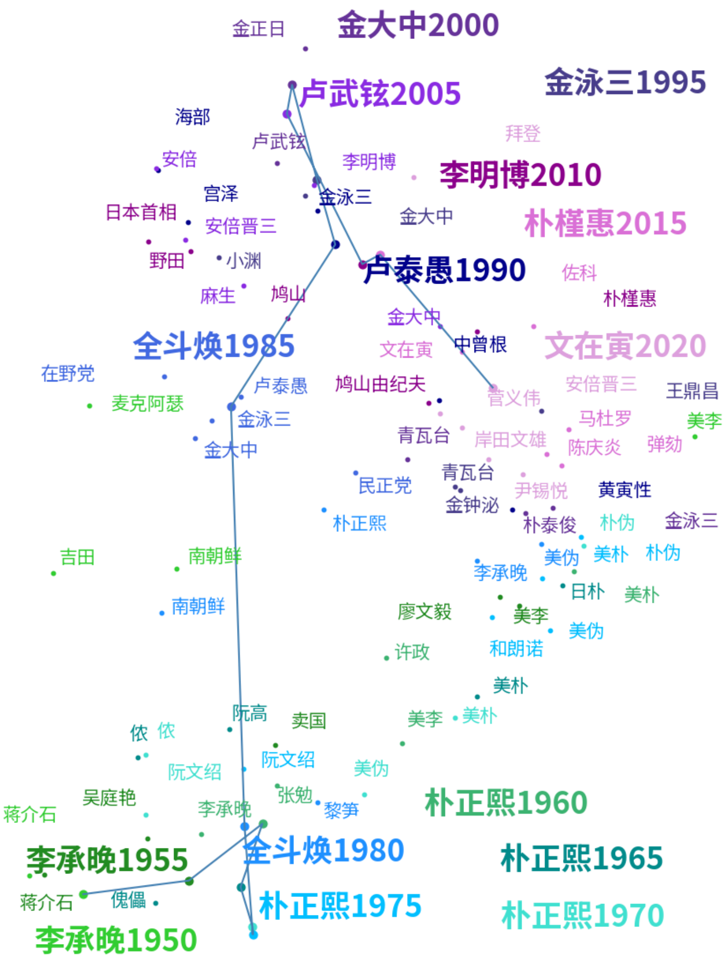


Figure 13: Semantic space of South Korean presidents on 5-year slices

Dividing temporal queries into two periods, the 1960s-1980s and 1990s-2020s, by the normalisation of China-Korean relations, the accuracy within each period remains high. However, a significant shift in the semantic space of "president of South Korea" occurred between these two eras, parallel with changes in China's diplomatic policies and attitudes towards the country. In the earlier period, Chinese official media referred to South Korea as "南朝鲜" (South Chosun)[17], often in a dismissive, accusatory and hostile tone. In 1991, the term "韩国" (Hanguk)[18] was first used, since then the term "南朝鲜" (South Chosun) gradually faded, appearing for the last time in 1995 before disappearing completely. Subsequent reports on South Korea shifted to a neutral political and diplomatic discourse tone. Figure 12 shows the frequency changes of these two terms, and Figure 13 the shifting semantic space of representative Korean president names. Strictly speaking, the changes observed here are not internal to the Chinese language itself, but rather reflect shifts in international political dynamics and social perceptions. Nonetheless, uncovering these extralinguistic changes is also one of the potentials of language models which holds substantial research value.

[17] in Chinese, "朝鲜" (Chosun) refers specifically to 北朝鲜 **North** Chosun (Korea).

[18] officially the Republic of Korea, 大韩民国, commonly written as South Korea.

4.4 Chapter Summary

In this chapter, we demonstrated the quality of word embeddings obtained using Word2Vec on the *People's Daily* corpus and compared the effectiveness of different alignment methods. On larger time slices, alignment-based models proved superior, while compass-based models showed greater robustness for smaller corpora. We also evaluated the model's ability to measure the degree of semantic change: alignment-based + SGNS performed the best. For temporal analogy tasks, only the compass-based method is structurally capable, with the CBOW architecture outperforming SGNS. Considering these factors, we have chosen the compass-based + CBOW approach as a compromise to perform insightful case studies that are not covered by the testsets.

5 Exploration

5.1 Semantic Change Mining

We attempted to automatically pick out words that may have undergone semantic changes. A total of 2,618,233 different words appeared in the *People's Daily* corpus, we filtered out those that appeared more than 100 times in every 5-year slice from beginning to end, leaving 5,983 words. For these words, we calculated their cosine distance between the earliest (1950-1954) and latest (2015-2019) complete 5-year slices (because according to our segmentation method, the periods 1946-1949 and 2020-2023 are incomplete). We then sorted them in descending order according to the distance. As shown in Figure 14, most of the words exhibit small cosine distances between their initial and final states, indicating minimal changes in their meanings. There are 687 words with a cosine distance greater than 0.4, which is about one-ninth. Due to the ambiguity of modelling the semantics of single characters in Chinese, we only consider 546 of them with two or more characters.

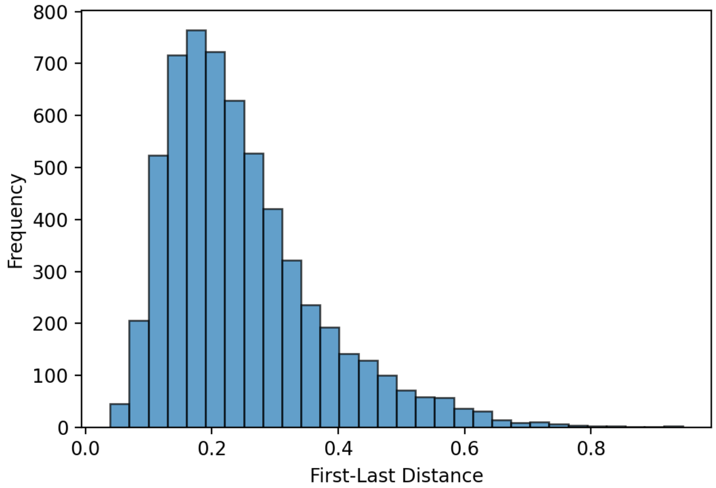


Figure 14: Distribution of first-last distance

We analysed the meaning and usage of these words based on the below three criteria:

1. A word's top 10 neighbours on each time slice;
2. A visualisation of its path in the semantic neighbourhood;
3. Definitions in the 7th edition of 现代汉语词典 (*A Dictionary of Current Chinese*)[19].

The nearest neighbours list, the visualisation and annotation are made available at[20]. The annotation includes the cosine distance before and after, whether there is a semantic change, the word's (old) part of speech, the type of semantic change, a remark on the change direction, the approximate time the change occurred, and the dictionary definitions. Out of all 546 words, there are 90 that we believe have indeed undergone a change in meaning. To assess the reliability of semantic change mining, we computed precision scores at intervals of 50 rows from our dataset (sorted based on the cosine distance). For each interval, precision was calculated by evaluating the proportion of true positives (confirmed changes), using the minimum distance value of that interval as its representation. We then plotted these precision scores against the corresponding distance values as shown in Figure 15. It can be seen that in the first interval (the top 50 words), half of which are confirmed to have changed their meanings. Precision initially decreases as the distance increases, reaching nearly zero at a distance of 0.45, but then shows some fluctuations. Overall, we would suggest prioritising examining words with a distance greater than 0.5 for semantic change detection, as this threshold ensures that at least one in ten words will likely exhibit detectable changes. If time and resources permit, further exploration could be conducted. However, as the distance decreases, the density of words increases, as well as a further decline in precision, which might not be cost-effective. Due to the high cost of manual annotation, we regret that we are unable to calculate recall, i.e., the proportion of words with semantic changes that fall above a certain distance threshold.

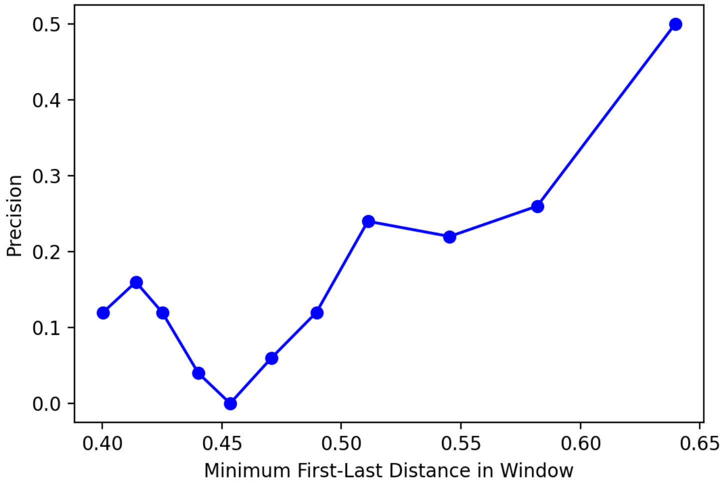


Figure 15: Precision vs. minimum distance in window

[19] ebook `https://archive.org/details/modern-chinese-dictionary_7th-edition`, the MDX version is taken from `https://github.com/bamboo512/ModernChineseDict` for entries extraction.

[20] `https://github.com/RainTreeCrow/Language-Change-CH/tree/main/explore/mining`, see the file neighbours.ipynb for neighbours lists, draw.ipynb for visualised paths, and judgement.csv for annotations.

5.1.1 Statistics

Table 10 shows the part of speech categories for the changed words. Nouns and verbs are the most common, with a few instances of verb phrases, adjectives, adverbs, and idiomatic expressions.

noun	verb	verb phrase	adjective	adverb	idiom
35	38	3	4	6	1

Table 10: Distribution of initial part of speech

Figure 16 shows the temporal distribution of these semantic changes. Two peak periods are observed: one between 1955 and 1960, and another between 1980 and 1995. We believe the former may be attributed to the promotion of Standard Mandarin Chinese in the 1950s, particularly with the publication of the first editions of the 新华字典 (*Xinhua Dictionary*) and the 现代汉语词典 (*A Dictionary of Current Chinese*). Newspaper reporters were among the first to start using Chinese according to the newly published standards. The latter period of language change, as suggested by Chen et al. (2023), is likely a byproduct of the rapid development in politics, economy, science, technology, and culture following the reform and opening-up policy in China.

It has been shown that many of the changes occurring between 1980 and 1995 are related to the economy. For example, the terms 封闭 (literally "close") and 开放 (literally "open") gained new meanings related to the international trade, such as the stagnation and obstruction of imports and exports, as well as the states of openness and circulation. Additionally, 上市 (literally "come into the market") has evolved from the sale of seasonal fresh produce in the market to the listing of stocks, bonds, and funds on stock exchanges upon approval. 供给 (literally "supply") shifted from the distribution of essential living materials under rationing to the supply and demand relationships in a free market. 瞄准 (literally "aim") transitioned from physically aiming at a target (to shoot it) to delving into user psychology and matching market needs. These semantic changes are not unique to Chinese; similar usages can also be found in the corresponding English terms.

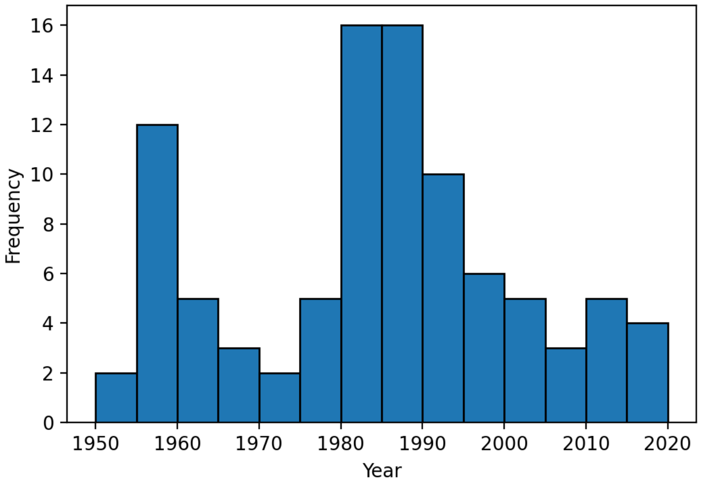


Figure 16: Distribution of the time changes occurred

5.1.2 Types of Changes

To classify the 90 cases of semantic change, we referred to the categories described in the Oxford Research Encyclopedias (Traugott 2017). The six types of changes include:

1. **Metaphorisation**: One concept is understood in terms of another based on similarity.
   tissue: "woven cloth" → "aggregation of cells in animals or plants"

2. **Metonymisation**: One concept represents another based on association or contiguity.
   keel: "the main beam of the vessel" → "the vessel"
   board: "table" → "people sitting around a table" → "governing body"

3. **Pejoration**: A term gains a more negative connotation.
   awful: "awe-inspiring" → "very bad"
   cnafa "boy, young man" → ***knave*** "a dishonest or unscrupulous man"

4. **Amelioration**: A term gains a more positive connotation.
   nice "foolish, stupid" (Middle English) → "good, pleasant"
   cniht "male servant" → ***knight*** "a person honoured by the sovereign for merit"

5. **Narrowing**: A term's meaning becomes more specific.
   deor "animals" → ***deer*** "hoofed ruminant mammal"

6. **Broadening** (generalisation): A term's meaning becomes more inclusive.
   dogge "the Molosser category" → ***dog*** "all domesticated canid"

These categories help us to systematically analyse and understand the nature of semantic changes in the identified cases. Note that they are not entirely distinct; we choose the most representative label for each example. A total of 66 words were classified into these six types, with some words being difficult to categorise, which we will discuss later. Table 11 shows the number of words in each category of semantic change. It is observed that metaphorisation is the most frequent, amelioration is more common than pejoration, and broadening is more common than narrowing.

Metaphorise	Metonymise	Pejorate	Ameliorate	Narrow	Broaden
25	6	0	10	4	19

Table 11: Distribution of semantic change types

For each category of semantic change, we provide a complete list and select representative examples to explain each one. We first look at metaphorisation and metonymisation, which can sometimes be difficult to distinguish. Metaphor focuses more on creating indirect associations based on implicit similarities between things or concepts. For instance, 渠道 (literally "canal"), originally referred to a watercourse dug around rivers, lakes, or reservoirs for irrigation and drainage, extended to describing a way or method for doing things. The two share a commonality in that both involve transferring something from one side (the starting point) to the other (the destination). Another example is 泛滥 (literally "flood"), originally the uncontrolled spread of large volumes of water over a wide area, and extended to the uncontrolled spread of negative phenomena. (Table 12)

Metonymy emphasises the actual relationship between things, where one thing explicitly replaces another. For example, in English, the building "Scotland Yard" is borrowed to refer to the police headquarters in Greater London, and “sweat” is used to symbolises effort. In instances of metonymy we encounter, 两岸 (literally “cross-strait”) referred to the Taiwan Strait, i.e. Mainland China and Taiwan; and 俱乐部 (literally “club”) which is derived from a place for social, cultural, sports and recreational activities, to organisations and groups engaged in these activities. (Table 13)

Word	POS	Old Meaning	New Meaning
攻克	v.	to conquer (e.g. a city)	to overcome (e.g. a difficult challenge)
挑战	n.	a call to compete	an obstacle, a difficult task
用人	vp.	exerting physical labour	appointing talent
基点	n.	pilot programme	focus, starting point
渠道	n.	canal, water channel	way, approach
释放	v.	release (e.g. a prisoner)	unleash, unlock (e.g. the potential)
封闭	v.	to seal (e.g. a crack)	to isolate oneself or be insular
背景	n.	(e.g. stage) backdrop	(e.g. historical/social) background
前沿	n.	(military) frontline	frontier (of research/technology)
打通	v.	to dig through (physically)	to connect (metaphorically)
供给	n.	supply (of food, fuel)	supply (in economics)
泛滥	v.	flooding (of water)	prevalence (e.g. of evils)
缺口	n.	opening (e.g. in the wall)	shortfall (e.g. funding/resource)
侵入	v.	intrude, violate (military)	attack, hack (cyber system)
中心	n.	centre of focus	(e.g. information) centre, hub
进军	v.	advance (military)	go into (e.g. a market/field)
增添	v.	acquire, purchase (items)	add, enhance (e.g. joy)
散发	v.	hand out (e.g. flyers)	give off (e.g. heat, smell)
震撼	v.	physically shaking	emotionally moving, touching
跳出	v.	jump out of (e.g. a pit)	think outside the box
行列	n.	line (of marching people)	the circle (e.g. of developed countries)
向上	adj.	upward (direction)	progressive, striving
堡垒	n.	fortress (military)	barrier (metaphor, e.g. technical)
挖掘	v.	dig (literal, e.g. a hole)	explore, uncover (e.g. potential)
关闭	v.	close (literal, e.g. the door)	shut down (e.g. the factory)

Table 12: Cases of Metaphorisation

Word	POS	Old Meaning	New Meaning
用工	vp.	spend working hours	employ workers
两岸	n.	the two sides (of river/strait)	Mainland China and Taiwan
俱乐部	n.	venue for cultural/sport/ recreational activities	group or organisation engaging in these activities
大陆	n.	a vast landmass/continent	Mainland China
开门	vp.	open the door, start a new business	do sth. publicly, invite broad feedback
内地	n.	inland, landlocked areas	Mainland China

Table 13: Cases of Metonymisation

We did not find examples of pejoration regarding the sentiment carried by words, but there are some cases of amelioration, where words shifted from negative to neutral or positive meanings. For instance, the verb 策划 (literally "to plan") and the idiom 千方百计 (literally "by all means") used to specifically refer to malicious plots and schemes, but now they generally denote neutral plans and efforts. Similarly, 敢于 (literally "dare to) once implied recklessness, impulsiveness or overconfidence, but now it signifies courage in the face of difficulties. (Table 14)

Word	POS	Old (Negative) Connotation	New (Positive/Neutral) Connotation
代理人	n.	a person who serves illegal interests	a person authorised by another party to conduct (e.g. trade, litigation)
集团	n.	gang (e.g of reaction/conspiracy)	economic entity composed of companies of the same type)
十足	adv.	utterly (e.g. foolish, dishonest)	very (e.g. brave, strong)
追随	v.	submit to a(n unjust) force or authority	join or follow sth. (right and progressive)
千方百计	idiom	to scheme and plot (for malicious purposes)	to spare no effort
策划	v.	to scheme (often related to conspiracy)	to plan or devise a strategy (e.g. in the new media industry)
设想	n.	unrealistic speculation/fantasy	practical ideas/plans
支配	v.	dominate/control/enslave	allocate/dispose of (e.g. resource or income)
卷入	v.	actively involved in (e.g. a thriving business)	passively caught up in (unpleasant situations, e.g. war/turmoil)
梦想	n.	fantasy, delusion	dream, aspiration
清算	v.	eliminate (e.g. corruption, crime)	settle (e.g. accounts, transactions)
敢于	v.	dare to (audacity)	dare to (courage)

Table 14: Cases of Amelioration

The last pair of change types is narrowing and broadening, which involve restriction and extension of meaning. Narrowing includes cases such as 导师 (literally "mentor"), which has shifted from a general term for a guide or leader (e.g., a spiritual mentor of a revolution) to specifically referring to an advisor or supervisor for research and academic writing. Similarly, 领导人 (literally "leader") narrowed from a someone who leads or is in charge (e.g. of the labour union) to a head of state government. (Table 15) The latter includes examples such as 战略 (literally "strategy") which has broadened from the plans and tactics for guiding warfare to general plans or policies for managing a large-scale situation. 等于 (literally "equals") extended from numerical or quantitative equality to conceptual approximation, similar to "that is to say" or "in other words". (Table 16)

Word	POS	Old Meaning	New Meaning
导师	n.	leader or guide (e.g. in revolution)	academic advisor or supervisor
领导人	n.	director/person in charge	head of state or government
破产	v.	fail (general)	go broke (business)
敬礼	v.	greeting (e.g. in a letter)	(military) salute

Table 15: Cases of Narrowing

Word	POS	Old Meaning	New Meaning
调动	v.	move, transfer (e.g. position)	stimulate, boost (e.g. brainpower)
提升	v.	promote (position)	improve, enhance
产业	n.	the production industry	(any) industry, sector
专车	n.	a private(ly owned) car/ car used for special purpose	a private car hire (in contrast to carpooling
瞄准	v.	aim at (the shooting target)	aim at (e.g. market demand)
创建	v.	to establish (e.g. a company)	to make/turn sth. into (e.g. a civilised city)
开放	v.	open, blossom	lifting restrictions (e.g. for international trade)
流动	v.	tour, travel around (e.g. salesperson, medical team)	flow of resources (e.g. talent, capital)
身边	adv.	(physically) beside, next to	in everyday life
生产者	n.	a person engaged production labour	producer (in finance, aligned with consumer, seller)
主题	n.	the subject matter, theme (e.g. of a literary work)	the theme of an event
人群	n.	a crowd of people (literal)	a group/community (e.g. low income, LGBT)
击败	v.	defeat, conquer (in a war)	win, beat (e.g. in a game/election)
采集	v.	collect, gather, retrieve (e.g. honey, seeds, fruit)	collect (e.g. data)
战略	n.	military strategy	overall strategy or plan
冲击	v.	physically strike (e.g. by gunfire/flood/storm)	disrupt, impact (e.g. on the economy)
布置	v.	arrange or plan (activities)	set up or arrange (items)
等于	v.	equal in number or amount	means (that is to say)
突破	v.	break through defences (military)	overcome (e.g. difficulties), break (e.g. records)

Table 16: Cases of Broadening

For other words that are difficult to categorise into these six types, we distinguish between meaning loss and gain (i.e., the disappearance of old meanings and the emergence of new usages). For example, 打通 (literally "to break through") used to mean "to convince, persuade" (to get the idea through to someone), and 常年 (literally "normal year") once referred to a "typical" year without natural disasters (in agriculture context), both are rarely used in these old senses today. (Table 17) With the advancement of technology and society, some words have also acquired new meanings, such as 清洁 (literally "clean") can now refer to environmentally friendly or non-polluting energy or production methods, and 提起 (literally "to bring up, mention" can now be used in "to file" (a legal lawsuit). (Table 18) It is worth noting that some words also undergo a syntactic change in part of speech as they gain new semantic meanings. For example, 完善 (literally "complete and perfect") emerged from its adjective meaning a verb meaning "to make sth. complete and perfect". The verb 移动 (literally "to move") gained an adjective sense "mobile, portable", and 做工 (literally "to do work") shifted from the verb meaning "to engage in manual labour" (often referring to industrial or handicraft work) to a noun, the technique or quality of craftsmanship. (Table 19)

Word	POS	Lost Sense	Remaining Sense
封闭	v.	seize (illegal business)	close, block
正规	adj.	full-time (education, as opposed to part-time)	compliant with certain rules or standards
打通	v.	persuade	remove obstacles to get through
常年	n.	ordinary year (without disasters)	long-term, year-round
前方	n.	front line (of battlefield)	front

Table 17: Cases of Sense Loss

Word	POS	Old Sense	New Sense
志愿	n.	aspirations	college applications
清洁	adj.	clean, tidy	free from pollution, environmental friendly
自动	adv.	spontaneously	automatically
提起	v.	bring up, mention	file (a lawsuit)
上市	v.	(seasonal fruits and vegetables) come to the market	(product) launch, on sale (company) go public

Table 18: Cases of Sense Gain

Word	Old POS	Old Sense	New POS	New Sense
完善	adj.	complete, perfect	v.	to improve, to perfect
数字	n.	number, figure	adj.	digital
移动	v.	to move, to shift	adj.	mobile, portable
公开	adv.	openly, in public	v.	to announce, to make public
做工	v.	to engage in physical labour (industry and handicrafts)	n.	the quality of a (industrial or handicraft) product
大众	n.	the general public	adj.	popular, mainstream
文明	n.	civilisation	adj.	civilised

Table 19: Cases of Sense Shift with Change in POS

We also noticed some special cases, such as the disambiguation of homophones or similar-looking words (characters) that were once used interchangeably. For instance, 功夫 (gōng fu), skill or accomplishment, and 工夫 (gōng fu), the time and effort spent on a task, as well as 不大 (not big) and 不太 (not very, not much). Although they are still mixed up sometimes in spoken language, they have been standardised and differentiated in written forms, especially in news media. (Table 20) Some terms are associated with newly emerged named entities. For example, 小米 (literally “millet”) has evolved from referring to a type of crop to becoming the name of the technology company Xiaomi. The 一带 (The Belt) and 一路 (The Road) in 一带一路 (Belt and Road) now specifically refer to the *Silk Road Economic Belt* and the *21st Century Maritime Silk Road*. (Table 21)

Finally, there are some terms that are difficult to categorise or describe. 多方 has shifted from “from multiple aspects” or “using various methods” to ”bringing together multiple parties” (e.g., for negotiations). 分化 has evolved from “to incite conflict or disputes between two sides” (e.g., to divide and conquer the enemy) to “increase in differences” (e.g., polarisation). (Table 22)

Word	Meaning	Confusion	Meaning
功夫	skill, expertise	工夫	time and effort
不大	not big	不太	not much, not very

Table 20: Cases of Disambiguation

Word	Old Meaning	New Reference
小米	millet	Xiaomi (electronics company)
一带	region nearby	The Belt (in “Belt and Road”)
一路	along the way	The Road (in “Belt and Road”)

Table 21: Cases of Emerging Name Entities

Word	POS	Old Meaning	New Meaning
国土	n.	territory	land (e.g. resources)
保安	n.	local armed police forces	security guard
多方	adv.	using multiple methods, from multiple perspectives	bring together various parties
分化	v.	to divide, to sow discord	to become diverse, to polarise

Table 22: Other Unclassified Cases

5.1.3 The False Positives

For words with significant distance between vectors but which, based on manual verification, have not actually undergone a semantic change, we have identified the following reasons:

1. **Polysemy and homographs**: For polysemous words, the usage of different meanings may fluctuate over time, such as with “果实” (fruit), which has both a literal and a metaphoric meaning. Consequently, the position of a single representation word embedding may shift in its vector space. For words with multiple parts of speech (homographs), such as “制服” (which can mean both “uniform” as a noun and “to subdue” as a verb), the embedding can be even more unstable and exhibit greater variation. Since we have consistently observed various usages in both the earliest and latest time periods for these words, i.e., the old meanings did not completely disappear, and new meanings were not emerging out of nowhere, fluctuations in frequency among different senses are not classified as semantic change.

2. **Changes in contextual usage**: Although the meaning of the word itself remains unchanged, its common contexts have shifted. For example, the term “全民” (literally “all the people”), which has always meant “the entire population of a country”, has been used in different contexts over time, such as “全民皆兵” (the entire nation as soldiers, during the Chinese Civil War), “全民经济” (people’s economy, during the period of public ownership and state-tun economies), and “全民健身活动” (national fitness campaign, recent, the 21th century). This highlights a major, inherent limitation of most word embedding models: statistical variations in context do not necessarily equate to variations in word meaning.

3. **Reference**: Similar to changes in context, the meanings of these words remain unchanged, but they refer to different things in different periods. For example, “场上” (on the field) was

originally more associated with agricultural production, such as in “打谷场” (threshing field), but now it is more commonly associated with sports, such as “篮球场” (basketball court).

4. **Tokenisation Errors**: Chinese tokenisation often has ambiguities. For example, “建国” is a verb phrase meaning “to establish/build a nation” and also a common name for boys. As a result, many journalists and news figures named “建国” have been incorrectly tokenised, leading to significant semantic confusion about this term.

5.2 More Temporal Analogies

In addition to the publicly recorded change in roles or positions, such as the list of national leaders we previously used for model evaluation, Yao et al. mentions another type of temporal analogy that is less clearly-defined but can also provide interesting insights on technological development or pop culture, such as “Walkman : 1987 :: iPod : 2007” and “yuppie : 1987 :: hipster : 2003” (2018).

5.2.1 Daily Life

The following trends in people’s daily life are observed from temporal queries in Chinese:

1. **Communication Technology**: Mobile phones gradually replaced telephones and fax as a means of communication. Specifically, the analogy words went from 大哥大, old nickname for cellular phone in Cantonese (1992-1998) to 呼机, pager (1996-2002), BP机, beeper (1998-2001), and finally to 手机, mobile phone (1998-). (Table 23)

2. **Home Appliances for Leisure and Entertainment**: A shift in popular household electric device from 收音机 radio and 电视机 television (1949-1979) to 电脑 computer (1980-), 笔记本电脑 laptop (1999-) and most recently, 平板 tablet (2012-). (Table 24)

3. **Music Media**: Early music players like 留声机 gramophone (1949-1964) were replaced by new media and formats, such as 磁带 cassette tape (1978-), CD (1996-), 随身听 Walkman (2000-2007), and MP3 (2000-2014). It is noteworthy that 唱片/黑胶 vinyl record disappeared for a period after 2000, likely due to the rise of more portable CD and MP3 formats. However, they re-emerged at around 2018, possibly because in the days of streaming platforms, music enthusiasts began to revalue the collectible nature of physical records. (Table 25)

4. **Modes of Transportation**: Vehicles such as 自行车 bicycle and 汽车/轿车 car have always been present in analogy terms throughout the years. Others have been out of use, such as 马车 carriage (1947-1985) and 人力车 rickshaw (1949-1950). Some vehicles appeared later, such as 摩托车 motorcycle (1966-) and 电动车 electric bike (2007-). (Table 26)

5. **Means of Communication**: Methods evolved from the earliest 电报 telegraph (1946-1959) and 信 letter (1948-1997) to 电话 telephone (1957-2001), 传真 fax (1993-2002), and then to 电子邮件 email (1995-2014) and 短信 text message (2003-2011). Finally, they transitioned to instant messaging apps QQ (2004-) and 微信 WeChat (2013-). (Table 27)

6. **Sources of News and Information**: People’s way of accessing news and information evolved from reading newspapers 报纸/报刊 (1946-1991) and listening to radio programs 广播电台/栏目 (1946-1999) to checking certain news websites 网站 (1996-2010), such as Sohu

搜狐 (1999-2013), NetEase 网易 (2000-2013), and Sina 新浪网 (2006-2012). Since 2010, people began to access news through social media platforms like Weibo 微博, and from 2014 onwards, the WeChat Official Account 公号 (short for “微信公众号”). (Table 28)

Year	Analogy
1953-2012	电话 telephone
1987-1996	传真 fax
1992-1998	大哥大 cellular phone
1996-2002	呼机 pager
1998-2001	BP机 beeper
1998-	手机 mobile phone

Table 23: Analogy words for “手机” (mobile phone) 2023

Year	Analogy
1949-1978	收音机 radio
1956-1979	电视机 television
1980-	电脑 computer
1999-	笔记本电脑 laptop
2012-	平板 tablet

Table 24: Analogy words for “电脑” (computer) 2023

Year	Analogy
1947-1984	收音机 radio
1949-1964	留声机 gramophone
1949-2001	唱片 record
1978-	磁带 cassette
1996-2020	CD
2000-2007	随身听 Walkman
2000-2014	MP3
2018-	黑胶 vinyl

Table 25: Analogy words for “CD” 2020

Year	Analogy
1946-	自行车 bicycle
1946-	汽车/轿车 car
1947-1985	马车 carriage
1949-1950	人力车 rickshaw
1966-	摩托车 motorcycle
2007-	电动车 electric bike

Table 26: Analogy words for “电动车” (electric bike) 2023

Year	Analogy
1946-1959	电报
1948-1997	信 letter
1957-2001	电话 telephone
1993-2002	传真 fax
1995-2014	电子邮件 email
2003-2011	短信 text message
2004-	QQ
2013-	微信 WeChat

Table 27: Analogy words for “微信” (WeChat) 2023

Year	Analogy
1946-1991	报纸/报刊 newspaper
1946-1982	广播电台 broadcast
1986-1999	栏目 channel
1996-2010	网站 website
1999-2013	搜狐 Sohu.com
2000-2013	网易 NetEase 163.com
2006-2012	新浪网 Sina.com
2010-	微博 Weibo
2014-	公号 WeChat Official Account

Table 28: Analogy words for “微博” (Weibo) 2023

The queries reflect changes in people’s everyday lifestyle in China for the past 78 years, covering various aspects such as communication, transportation and entertainment. The findings demonstrate the potential of temporal word analogies to capture when and how new things emerge and replace old ones, which can reflect the development of human society.

5.2.2 Event Detection

In this section we attempt to study two more subtle cases of event detection, the epidemics of infectious diseases in China, and the major armed conflicts worldwide. In 2022, the most discussed infectious disease keyword in China was 新冠 (COVID-19), while the one of the most influential armed conflict keywords was 俄乌 (Russia-Ukraine). Therefore, we use the two words from 2022 as the baseline vectors for searching. For each year from 1946 to 2023, we search for the top ten most similar words to these vectors and count their occurrences. As we analyse the prevalent diseases and countries/regions involved in conflicts, we merged synonyms and related terms such as 非典 (SARS) and its full name 非典型肺炎 (Severe Acute Respiratory Syndrome), 以巴 (Israel-Palestine) and 巴以 (Palestine-Israel), as well as 美苏 (US-Soviet) and 美俄 (US-Russia). For each task, we selected 16 representative examples (those appearing more than four times) for presentation. We plotted the years in which each word appears, the distribution of points in the figures can be interpreted as the timeline of a disease outbreak in China or an armed conflict in a specific region.

According to the official website of *Chinese Centre for Disease Control and Prevention*[21], China effectively controlled typhoid fever in the early 1980s. Organised smallpox vaccination began in 1952, and by the 1980s, smallpox had been nearly eradicated. For a period after, there were still related news mentions, primarily focused on promoting vaccination. Respiratory, blood-borne, and intestinal infectious diseases, such as diphtheria, whooping cough, dysentery, malaria, and plague, also largely disappeared by the 1980s due to economic development, improved living conditions, and better health education. In addition to these disappearing diseases, new ones have emerged. For example, AIDS, which arrived in the 1980s, and avian influenza outbreaks in the 21st century. Hepatitis B, despite having exist in China for a long time, saw a significant increase in cases in the 1980s due to widespread unregulated blood collection, leading to greater awareness. Figure 17 also highlights the two recent major outbreaks of pneumonia in China: SARS and COVID-19.

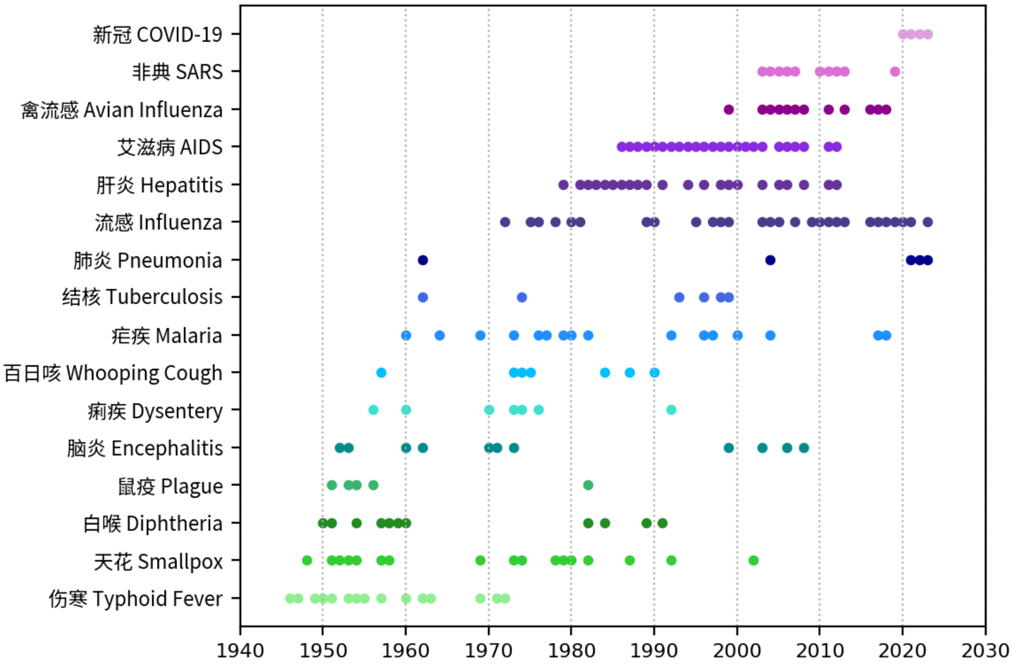


Figure 17: Analogy words for “新冠” (Coronavirus) 2022

Analogy	Years
H1N1	2009, 2010, 2011, 2013, 2017
H7N9	2013, 2016, 2017
MERS	2015
H5N6	2016
猴痘 monkeypox	2022
支原体 mycoplasma	2023

Table 29: Analogy words for “新冠” (Coronavirus) 2022

[21] https://www.chinacdc.cn

We also examined some less frequently occurred diseases. As shown in Table 29, the model correctly identified the sequential outbreaks of various influenza virus subtypes such as H1N1, H5N6, and H7N9 between 2009 and 2020, the 2015 MERS outbreak in South Korea (which received a lot of media attention in China due to South Korea being a neighboring country), as well as the recent emergence of monkeypox and mycoplasma viruses in 2022 and 2023.

Next, we examine the analogy words for 俄乌 “Russia-Ukraine” from 2022. The model successfully captured many significant international events, including the Korean War, the Taiwan Strait crises and the Vietnam War starting from the 1950s, the Lebanese Civil War from 1975 to 1990, the Iran-Iraq War and the Central American crisis in the 1980s, as well as the Syrian Civil War since 2011. (Figure 18) For some conflicts with very short duration (less than a year, sometimes only a few months), the model was also able to take note of them. As shown in Table 30, these include the Iran-Azerbaijan Crisis of 1946, the Sino-Indian War of 1962, the Turkish invasion of Cyprus in 1974, the Falklands War in 1982, and the Russo-Georgian War of 2008.

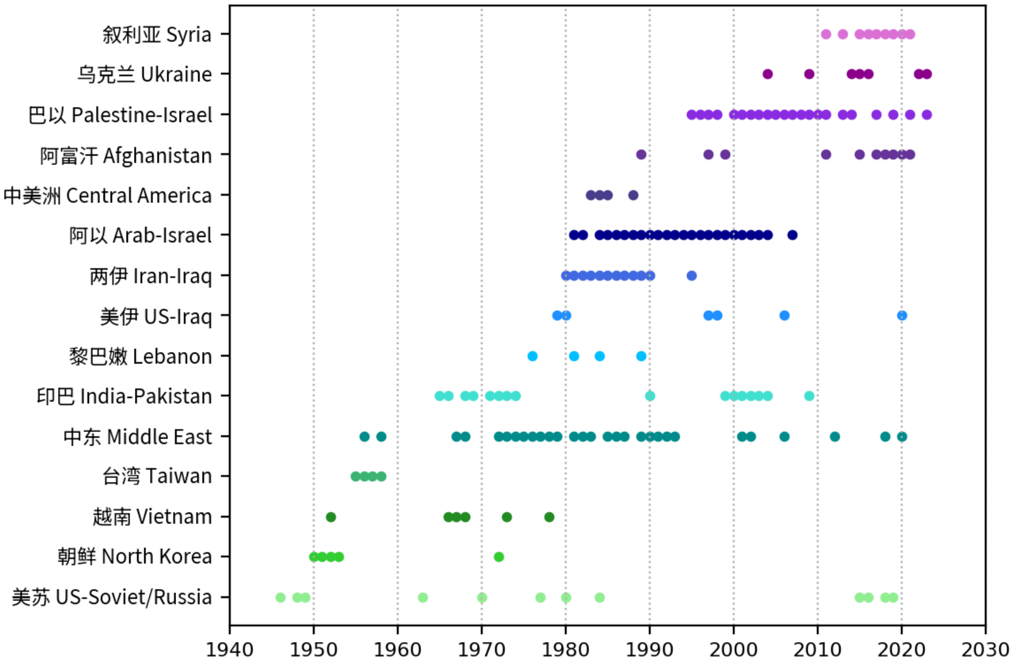


Figure 18: Analogy words for “俄乌” (Russia-Ukraine) 2022

Analogy	Years
阿塞拜疆 Azerbaijan	1946
印中 India-China	1962, 1963
塞浦路斯 Cyprus	1975
马岛 Falkland Islands	1982
格鲁吉亚 Georgia	2008

Table 30: Analogy words for “俄乌” (Russia-Ukraine) 2022

6 Conclusion

Our work employed a Word2Vec-based temporal word embedding model, designed both static and temporal tests to train and evaluate two alignment methods on *People's Daily* news data from 1946 to 2023. We selected a model with balanced performance across tasks for further heuristic analysis and exploration, filtered a list of word with potential semantic changes for manual annotation and classification, and used temporal analogy queries to capture real-world events, demonstrating a potential application of temporal word embeddings in sociology and current affairs.

7 Future Work

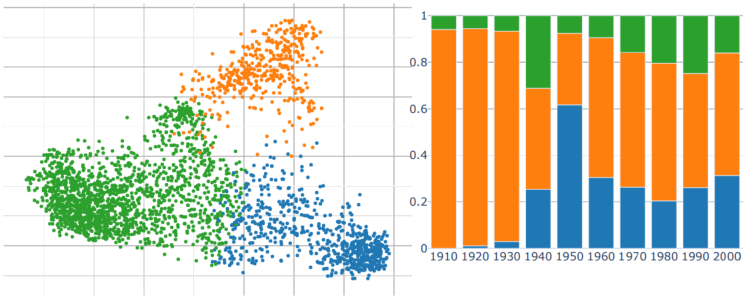
7.1 Expending the Diachronic Chinese Corpora

Our study indicates that the static quality of word embeddings is highly related to the size of the corpus; for example, the quality embeddings on 1-year time slices is lower than on 5-year slices. To improve the quality of word embeddings and model the semantic relationships more accurately on smaller time scales, it is essential to increase the amount of data. We suggest prioritising the inclusion of more news text, which has the advantage of a more consistent language style and ample information on current affairs. However, they are limited in style and can only represent official, written discourse, unable to capture semantic changes in colloquial everyday language. Nonetheless, in preliminary stages, we believe the advantages of news corpora outweigh the disadvantages. For instance, in temporal analogy tasks involving the names of national leaders, official media can ensure standardised and consistent name translations, whereas other media and spoken language may have multiple variants in transliteration, such as "特朗普" and "川普" for Trump, or "卡梅伦" and "卡梅隆" for Cameron. Introducing other types of text would require considerable manual effort for disambiguation. Failure to do so may result in lower scores on such test sets.

Once work on sufficiently large and news corpora has reached its full potential, more diverse corpora sources can be considered, such as building a specialised corpus for literary works or academic papers. In addition, despite that online post are more fragmented and may contain a large number of irregular abbreviations and omissions, linguists and sociologists are highly interested in the study of Internet slang and buzzwords, for they are more reflective of everyday life and subtle shifts in the public psyche than formal, written language. Thus, building a diachronic Chinese corpus of online texts is very meaningful, though the cost would be predictably high.

7.2 Contextualised Models and Polysemy

Due to the high complexity and low interpretability of using contextualised word embeddings + clustering methods for semantic change detection, we chose the static word embeddings to begin with. However, representing words from different contexts as a single vector can lead to semantic ambiguity, making it challenging to accurately model how a word's meanings shift over time. For example, the English word ***mouse*** has two vastly different meanings: "a small rodent" and "a computer device". Forcing a single vector representation for both would inevitably lead to confusion. In the semantic space, the vector for "a rat" may be gradually "pulled" by the meaning of "a computer mouse", shifting away from the domain of animals toward computer accessories, but this trajectory is vague and hard to quantify. We also encounter similar situations in the semantic change mining stages, where the embedding of a polysemous word oscillates between its meanings in the vector space, and we are unable to quantitatively analyse the extent to which it deviates from or leans towards one of the meanings. Context-based modelling are able to address the issue of polysemy more effectively, which is a potential advantages in measuring semantic change. If a model can distinguish when ***mouse*** refers to the animal and when it refers to the computer device, it can precisely identify when and where the meaning of a "computer's mouse" first emerged and how its usage frequency has changed over time. Unfortunately, in reality, the multiple meanings of most words are not as clear and easily distinguishable as those of ***mouse***. As a result, sense clustering often produces results that are difficult to interpret. Currently, the most intriguing use of these models is to create intuitive visualisations, such as the sense cluster distribution maps from Giulianelli et al. (2020) and the semantic dynamic graphs from Ma et al. (2024).



(a) PCA visualisation of the usage representations.

(b) Probability-based usage type distributions along time.

Figure 19: Usage representations and distributions of word ***atom*** (Giulianelli et al. 2020)

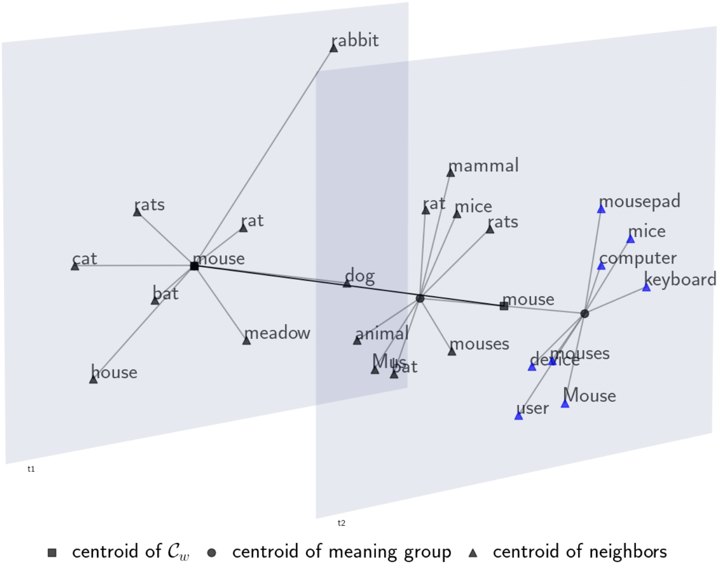


Figure 20: Representation of the temporal dynamic graph for ***mouse*** (Ma et al. 2024)

In the future, we also hope to extend such methods to Chinese to better analyse polysemous words and provide interesting visual results to explain the change of word senses. As mentioned earlier, the main challenges include the cost of training and fine-tuning large models, selecting clustering algorithms and parameters, as well as the manual analysis of clustering results.

7.3 Linguistics Study

Due to the authors' limited linguistic background, the observed semantic change phenomena are only described and categorised in a very simple and basic manner, with few linguistic theoretical tools applied, and the accuracy of the analysis remains to be verified. In the future, we look forward to collaboration with linguists for more solid and in-depth study on semantic change.

7.4 Visualisation and Query API

Considering the reason mentioned above, we would like to build a query website to share the results with 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; based on the analogy task, input a query (word, year) pair and return the results for each year as a list, etc. Compared to this study, which focuses on rough batch analyses of a number of words carried out within a short period of time, the query interface will allow users to take a closer and purposeful look at a small number of words of interest, bringing their prior knowledge to the table.

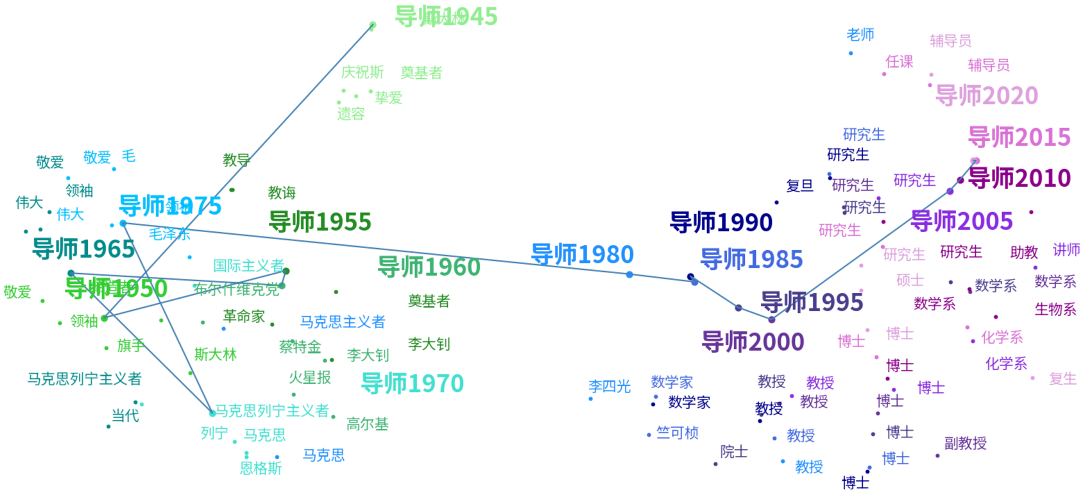


Figure 21: Transition path of “导师” (mentor) in its neighbourhoods

Figure 21 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 22 shows the path of “提升” (promote) shifting from “to raise someone to a higher position” in the neighbourhood of synonym verbs “提拔” (promote), “升

值” (raise) and position nouns “工程师” (engineer), “车间主任” (workshop director), to “improve, encourage” in the neighbourhood of “提高”, “增强” (enhance, increase).

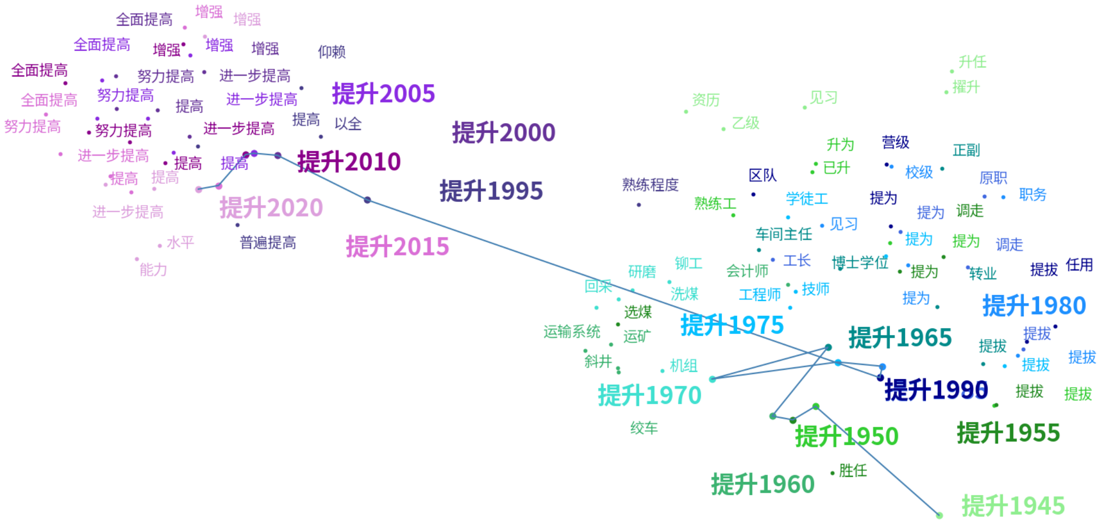


Figure 22: Transition path of “提升” (promote) in its neighbourhoods

Of course, there is a certain amount of randomness in getting neat and beautiful pictures like this one: when we use TSNE or PCA to reduce high-dimensional vectors (e.g., 100) to a two-dimensional plane, sometimes the distances between word vectors are distorted, i.e., the two points that appear to be closer together in a picture don’t necessarily have smaller cosine distances, and vice versa. Especially with TSNE, depending on the random seed settings, the resulting images can vary significantly. Sometimes the visualisation is clear, with points well-dispersed, while other times the points may be either too sparse or too dense, making it difficult to interpret. We often had to try several random seeds to obtain a more aesthetically pleasing image. Another issue is the time required for loading the model, performing nearest neighbour searches for each time slice, dimensionality reduction, plotting and adjusting text layout to avoid overlap. Locally, generating a single image often takes between 10 seconds to a minute. If we wish to provide query API to other users, we must optimise the storage and search of word vectors to speed up the operation.

8 Project Management

8.1 Code and Data

Executable code is provided at [22]. We have included a README file that explains the programme structure and usage, as well as how to download and store the raw corpus data for preprocessing. A requirements file is also provided for Python virtual environment configuration[23].

[22] https://github.com/RainTreeCrow/Language-Change-CH

[23] 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.

8.2 Timeline

Part-time working on dissertation until June:

- February - March 2024 (term 2): background research, literature review, data collection and preprocessing, experiment with Word2Vec alignment strategies.
- April (Easter break): model tuning, rough qualitative analysis of preliminary results.
- May - June (exams): quantitative evaluation on publicly available testsets.

Full-time working from June onwards:

- Week 1 of July: implement the Chinese temporal analogy task.
- Week 2 of July: switch to new data source for the previous one was found to be incomplete (missing half the news data between 1991-1992), rerun the model on new data.
- Week 3-4 of July: experimenting with BERT-based models and clustering, gave up because models take too much time to tune and does not seem to give much new, exciting insights.
- Week 1 of August: analysing the results for Chinese temporal word analogies, attempt to find out the cause for exceptionally poor performance with South Korean presidents.
- Week 2 of August: annotation and classification of potential semantic change cases.
- Week 3 of August: more temporal analogies, event detection, report writing.
- Week 4 of August: report writing.

8.3 Disclaimer

The political descriptions and analyses presented in this paper are derived from the corpus data used in the study. These descriptions are an honest reflection of the inherent biases present within the data and should not be interpreted as the personal views or positions of the authors.

List of Figures

List of Tables

References

BAMLER, ROBERT, und STEPHAN MANDT. 2017. Dynamic word embeddings. *Proceedings of the 34th international conference on machine learning*, *Proceedings of Machine Learning Research*, Vol. 70, 380–389. PMLR. URL `https://proceedings.mlr.press/v70/bamler17a.html`.

BLOOMFIELD, LEONARD. 1923. *Language*. George Allen & Unwin Ltd, london. URL `https://archive.org/details/in.ernet.dli.2015.1477125`.

CARLO, VALERIO DI; FEDERICO BIANCHI; und MATTEO PALMONARI. 2019. Training temporal word embeddings with a compass. *Aaai conference on artificial intelligence*. URL `https://api.semanticscholar.org/CorpusID:174803673`.

CHE, WANXIANG; YUNLONG FENG; LIBO QIN; und TING LIU. 2021. N-LTP: An open-source neural language technology platform for Chinese. *Proceedings of the 2021 conference on empirical methods in natural language processing: System demonstrations*, 42–49. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. URL `https://aclanthology.org/2021.emnlp-demo.6`.

CHEN, JING; EMMANUELE CHERSONI; DOMINIK SCHLECHTWEG; JELENA PROKIC; und CHU-REN HUANG. 2023. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. *Proceedings of the 4th workshop on computational approaches to historical language change*, 93–99. Singapore: Association for Computational Linguistics. URL `https://aclanthology.org/2023.lchange-1.10`.

CHEN, XINXIONG; LEI XU; ZHIYUAN LIU; MAOSONG SUN; und HUANBO LUAN. 2015. Joint learning of character and word embeddings. *Proceedings of the 24th international conference on artificial intelligence*, IJCAI'15, 1236–1242. AAAI Press.

DEVLIN, JACOB; MING-WEI CHANG; KENTON LEE; und KRISTINA TOUTANOVA. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

FIRTH, J. R. 1957. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)* 1952-59.1–32.

GIULIANELLI, MARIO; MARCO DEL TREDICI; und RAQUEL FERNÁNDEZ. 2020. Analysing lexical semantic change with contextualised word representations. *Proceedings of the 58th annual meeting of the association for computational linguistics*, 3960–3973. Online: Association for Computational Linguistics. URL `https://aclanthology.org/2020.acl-main.365`.

HAMILTON, WILLIAM L.; JURE LESKOVEC; und DAN JURAFSKY. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. URL `https://arxiv.org/abs/1606.02821`.

HAMILTON, WILLIAM L.; JURE LESKOVEC; und DAN JURAFSKY. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th annual meeting of the association for computational linguistics*, 1489–1501. Berlin, Germany: Association for Computational Linguistics. URL `https://aclanthology.org/P16-1141`.

HILPERT, MARTIN. 2008. *Germanic future constructions: A usage-based approach to language change*. John Benjamins, Amsterdam, Netherlands. URL https://www.jbe-platform.com/content/books/9789027291035.

HU, RENFEN; SHEN LI; und SHICHEN LIANG. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. *Proceedings of the 57th annual meeting of the association for computational linguistics*, 3899–3908. Florence, Italy: Association for Computational Linguistics. URL https://aclanthology.org/P19-1379.

KIM, YOON; YI-I CHIU; KENTARO HANAKI; DARSHAN HEGDE; und SLAV PETROV. 2014. Temporal analysis of language through neural language models. *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 61–65. Baltimore, MD, USA: Association for Computational Linguistics. URL https://aclanthology.org/W14-2517.

KULKARNI, VIVEK; RAMI AL-RFOU; BRYAN PEROZZI; und STEVEN SKIENA. 2015. Statistically significant detection of linguistic change. *Proceedings of the 24th international conference on world wide web*, WWW '15, 625–635. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

KUTUZOV, ANDREY, und MARIO GIULIANELLI. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. *Proceedings of the fourteenth workshop on semantic evaluation*, 126–134. Barcelona (online): International Committee for Computational Linguistics. URL https://aclanthology.org/2020.semeval-1.14.

KUTUZOV, ANDREY; LILJA ØVRELID; TERRENCE SZYMANSKI; und ERIK VELLDAL. 2018. Diachronic word embeddings and semantic shifts: a survey. *Proceedings of the 27th international conference on computational linguistics*, 1384–1397. Santa Fe, New Mexico, USA: Association for Computational Linguistics. URL https://aclanthology.org/C18-1117.

KUTUZOV, ANDREY; ERIK VELLDAL; und LILJA ØVRELID. 2022. Contextualized embeddings for semantic change detection: Lessons learned. *Northern european journal of language technology, volume 8*, ed. by Leon Derczynski. Copenhagen, Denmark: Northern European Association of Language Technology. URL https://aclanthology.org/2022.nejlt-1.9.

LIAO, XUANYI, und GUANG CHENG. 2016. Analysing the semantic change based on word embedding. *Natural language understanding and intelligent applications*, 213–223. Cham: Springer International Publishing.

LIU, YINHAN; MYLE OTT; NAMAN GOYAL; JINGFEI DU; MANDAR JOSHI; DANQI CHEN; OMER LEVY; MIKE LEWIS; LUKE ZETTLEMOYER; und VESELIN STOYANOV. 2019. Roberta: A robustly optimized bert pretraining approach. URL https://arxiv.org/abs/1907.11692.

MA, XIANGHE; MICHAEL STRUBE; und WEI ZHAO. 2024. Graph-based clustering for detecting semantic change across time and languages. *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 1: Long papers)*, ed. by Yvette Graham und Matthew Purver, 1542–1561. St. Julian's, Malta: Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.93.

MARTINC, MATEJ; SYRIELLE MONTARIOL; ELAINE ZOSA; und LIDIA PIVOVAROVA. 2020. Capturing evolution in word usage: Just add more clusters? *Companion proceedings of the web conference 2020*, WWW '20, 343–349. New York, NY, USA: Association for Computing Machinery. URL https://doi.org/10.1145/3366424.3382186.

MIKOLOV, TOMAS; KAI CHEN; GREG CORRADO; und JEFFREY DEAN. 2013a. Efficient estimation of word representations in vector space. URL https://arxiv.org/abs/1301.3781.

MIKOLOV, TOMAS; ILYA SUTSKEVER; KAI CHEN; GREG CORRADO; und JEFFREY DEAN. 2013b. Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th international conference on neural information processing systems*, NIPS'13, 3111–3119. Red Hook, NY, USA: Curran Associates Inc.

PERITI, FRANCESCO; ALFIO FERRARA; STEFANO MONTANELLI; und MARTIN RUSKOV. 2022. What is done is done: an incremental approach to semantic shift detection. *Proceedings of the 3rd workshop on computational approaches to historical language change*, ed. by Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, und Lars Borin, 33–43. Dublin, Ireland: Association for Computational Linguistics. URL https://aclanthology.org/2022.lchange-1.4.

PETERS, MATTHEW E.; MARK NEUMANN; MOHIT IYYER; MATT GARDNER; CHRISTOPHER CLARK; KENTON LEE; und LUKE ZETTLEMOYER. 2018. Deep contextualized word representations. URL https://arxiv.org/abs/1802.05365.

QIU, WEN, und YANG XU. 2022. Histbert: A pre-trained language model for diachronic lexical semantic analysis. *ArXiv* abs/2202.03612. URL https://api.semanticscholar.org/CorpusID:246652613.

ROSENFELD, ALEX, und KATRIN ERK. 2018. Deep neural models of semantic shift. *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)*, ed. by Marilyn Walker, Heng Ji, und Amanda Stent, 474–484. New Orleans, Louisiana: Association for Computational Linguistics. URL https://aclanthology.org/N18-1044.

ROSIN, GUY D., und KIRA RADINSKY. 2022. Temporal attention for language models. URL https://arxiv.org/abs/2202.02093.

ROUSSEEUW, PETER J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20.53–65. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

SAGI, EYAL; STEFAN KAUFMANN; und BRADY CLARK. 2012. Tracing semantic change with latent semantic analysis. *Current Methods in Historical Semantics*, 161–183. URL https://doi.org/10.1515/9783110252903.161.

SCHLECHTWEG, DOMINIK; BARBARA MCGILLIVRAY; SIMON HENGCHEN; HAIM DUBOSSARSKY; und NINA TAHMASEBI. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. *Proceedings of the fourteenth workshop on semantic evaluation*, ed. by Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, und Ekaterina Shutova, 1–23. Barcelona (online): International Committee for Computational Linguistics. URL https://aclanthology.org/2020.semeval-1.1.

Schlechtweg, Dominik; Sabine Schulte im Walde; und Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)*, ed. by Marilyn Walker, Heng Ji, und Amanda Stent, 169–174. New Orleans, Louisiana: Association for Computational Linguistics. URL https://aclanthology.org/N18-2027.

Schönemann, Peter H. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31.1–10. URL https://api.semanticscholar.org/CorpusID:121676935.

Sun, Chi; Xipeng Qiu; und Xuanjing Huang. 2019. VCWE: Visual character-enhanced word embeddings. *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)*, 2710–2719. Minneapolis, Minnesota: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1277.

Szymanski, Terrence. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 448–453. Vancouver, Canada: Association for Computational Linguistics. URL https://aclanthology.org/P17-2071.

Traugott, Elizabeth Closs. 2017. Semantic change. URL https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-323.

Wei, Li. 2014. 刁晏斌(diao yanbin). 2011.《“文革”期间语言研究》(studies of wenge language). *Chinese Language and Discourse* 5.281–284. URL https://www.jbe-platform.com/content/journals/10.1075/cld.5.2.07wei.

Yao, Zijun; Yifan Sun; Weicong Ding; Nikhil Rao; und Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. *Proceedings of the eleventh acm international conference on web search and data mining*. ACM.

Yu, Jinxing; Xun Jian; Hao Xin; und Yangqiu Song. 2017. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. *Proceedings of the 2017 conference on empirical methods in natural language processing*, ed. by Martha Palmer, Rebecca Hwa, und Sebastian Riedel, 286–291. Copenhagen, Denmark: Association for Computational Linguistics. URL https://aclanthology.org/D17-1027.

Zhang, Yating; Adam Jatowt; Sourav S. Bhowmick; und Katsumi Tanaka. 2016. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering* 28.2793–2807.