Lexicon of Changes: Towards the Evaluation of Diachronic Semantic Shift in Chinese

Jing Chen, Emmanuele Chersoni, Chu-Ren Huang
The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
Yuk Choi Road 11, Hung Hom, Kowloon, Hong Kong (China)
jing95.chen@connect.polyu.edu.hk, churen.huang@polyu.edu.hk
emmanuele.chersoni@polyu.edu.hk

Abstract

Recent research has brought a wind of using computational approaches to the classic topic of semantic change, aiming to tackle one of the most challenging issues in the evolution of human language. While several methods for detecting semantic change have been proposed, such studies are limited to a few languages, where evaluation datasets are available.

This paper presents the first dataset for evaluating Chinese semantic change in contexts preceding and following the Reform and Opening-up, covering a 50-year period in Modern Chinese. Following the DURel framework, we collected 6,000 human judgments for the dataset. We also reported the performance of alignment-based word embedding models on this evaluation dataset, achieving high and significant correlation scores.

1 Introduction

Lexical semantic change not only satisfies the appetite for linguistic exploration but also reflects the societal and cultural developments (Varian and Choi, 2009; Michel et al., 2011). Recently, this topic has been receiving growing interest from the NLP community, as witnessed a wealth of papers working on this research questions with computational approaches emerged over the past two decades (Kutuzov et al., 2018; Tahmasebi et al., 2019; Schlechtweg et al., 2020). Among these studies, most make use of distributional word representations with temporal information to model diachronic meaning change (Kim et al., 2014; Hamilton et al., 2016a,b; Giulianelli et al., 2020).

Although a variety of computational methods have been proposed for the task of lexical semantic change, evaluation datasets are only available for a limited number of languages, e.g. English, Latin, Italian, Swedish, German, Russian (Schlechtweg et al., 2020; Rodina and Kutuzov, 2020; Basile et al., 2020; Kutuzov and Pivovarova, 2021). Few studies have investigated Chinese in this domain (Tang et al., 2013, 2016) and there is currently no evaluation dataset for detecting Chinese lexical semantic change.

This paper presents the first Chinese evaluation dataset, **ZhShiftEval**, for the detection task. [1] This dataset allows us to evaluate those shifts that occurred to Modern Chinese from 1953 to 2003, over two roughly equal intervals: sub-corpus **C1** (1953-1978) and the sub-corpus **C2** (1979-2003). These two intervals were chosen on the basis of the *Reform and Opening-up*, the most influential milestone in the recent history of China [2]. It is generally assumed that this remarkable social change brought significant changes to the lexicon of Modern Chinese (Diao, 1995).

The remainder of this paper is organized as follows. Section 2 situates our study within previous work. In Section 3, we introduce how the evaluation dataset has been created following the DURel framework. Section 4 qualitatively discusses the dataset itself, and Section 5 presents the preliminary results of static word embeddings on this evaluation dataset.

2 Related Work

Before SemEval 2020, the field lacked shared standard datasets for evaluating lexical semantic change with computational approaches. Most early works were exploratory, testing whether computational models could capture specific established cases of semantic change, but without a quantitative evaluation of the models' performance (Sagi et al., 2009; Kim et al., 2014; Kulkarni et al., 2014).

Some evaluation datasets consisted of a list of target words labeled as 'changed' and 'unchanged'

[1] Researchers interested in the dataset should contact the first author of the study.

[2] Since the decision for the Reform and Opening-up was officially announced by the end of 1978, we set 1979 as the starting point for C2.

Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 113 - 118
May 26-27, 2022 ©2022 Association for Computational Linguistics

with reference to linguistic papers, dictionaries (Tang et al., 2013; Basile et al., 2020), and WordNet (Mitra et al., 2014). However, these datasets are based on a binary judgment on semantic change, ignoring its cumulative nature. In contrast, Gulordava and Baroni (2011) demonstrated a 'gradable' view towards semantic change, asking native speakers to annotate target words with multiple labels for their changing degrees, according to their intuitions.

Schlechtweg et al. (2018) later proposed the **D**iachronic **U**sage **R**elatedness (DURel) framework to construct evaluation datasets for the detection task. They asked annotators to compare and grade the semantic relatedness of target words, from unrelated (1) to identical (4), across the context pairs. The ratings, together with target words, formed a small-scale evaluation dataset for German. Following this framework, Rodina and Kutuzov (2020) and Kutuzov and Pivovarova (2021) created a two-period evaluation dataset, *'RuSemShift'* and a three-period evaluation dataset *'RuShiftEval'* for Russian, assessing those meaning shifts that occurred to Russian words from the pre-Soviet period to the Post-Soviet period.

In SemEval 2020, evaluation datasets for English, German, Swedish, and Latin were released as benchmarks for the shared task (Schlechtweg et al., 2020). The datasets were built under the Diachronic Word Usage Graph (DWUG), an extension of the DURel framework, exploiting usage graphs to represent the gain and loss of senses for target words. The usage graph is weighted and undirected. The nodes represent word usages, and the weights are semantic relatedness scores graded by human annotators (Schlechtweg et al., 2021).

3 Dataset Construction

3.1 Corpora

Detecting lexical semantic change over time requires a diachronic corpus having temporal information about texts. The dataset exploited in this study is derived from *People's Daily*, one of the most popular newspapers. This dataset has texts approximately ranging from the 1950s to the early 2000s, which are stored in MD format and in different folders according to the publication year of every newspaper article. To our knowledge, it is by far the largest diachronic Chinese dataset that is publicly accessible to full texts. [3]

The *Reform and Opening up* is assumed as the most influential and significant milestone in the second half of the last century in China. An exploding number of new lexical usages emerged in the process of this pronounced social development, which further introduced significant changes to Modern Chinese (Diao, 1995). Setting the year of the *Reform and Opening-up* as the borderline, we split the dataset into two subcorpora. Thanks to the temporal information of every single text, we obtained two time-specific subcorpora: texts produced from 1953 to 1978 are used to represent the C1 period, before the Reform and opening-up, and those from 1979 to 2003 are set to represent the C2 period, after the Reform and opening-up. The statistics of subcorpora are listed in Table 1.

Periods	Word tokens (million)	Word types (million)
1953 – 1978	262	1.73
1979 – 2003	331	2.54

Table 1: Overview of subcorpora: *C1 and C2*.

3.2 Word List

The word list for annotation includes 20 words, consisting of 10 words that changed their meaning over time and 10 stable words as counterparts. As for the changed words, we first manually picked them from previous literature, such as dictionaries (Guo and Chen, 1999; Shen, 2009) and linguistic books (Diao, 1995) as candidates. We then only included words satisfying the following conditions: 1) have high frequencies in both two corpora; 2) the changes suggested by the linguistic references are reflected in the corpus, either strongly or weakly. This step is conducted by scrutinizing 20 sampled sentences from each subcorpus.

We sampled stable words for each shifted word as counterparts. The changed word and its counterpart must have the same part of speech and the same frequency percentage in both two periods. The diachronic stability of stable words is checked by making use of dictionaries (Diao, 1995; Department of Chinese Lexicography, 2019), as well as with the intuitions of native speakers with linguistic backgrounds.

[3] A reviewer suggested two other diachronic Chinese datasets for consideration. One is the Google Ngram corpus, which contains a Chinese subset, but the access is limited to 5-grams. Another one is the more recent diachronic Chinese corpus (Zinin and Xu, 2020). However, the small scale of the earlier subcorpus (less than 1 million characters) and the fact that it is written in Classic Chinese would make the training process more problematic. These datasets, however, could be useful for future investigations.

3.3 Sampling

In the DURel framework, two metrics are used for quantifying degrees of semantic change (Schlechtweg et al., 2018; Rodina and Kutuzov, 2020): (1) ΔLATER $= Mean_L - Mean_E$, comparing the average score of mean relatedness across the context pairs consisting of two sentences from the LATER group and the context pairs having two sentences from the EARLIER group ; (2) the COMPARE score was obtained by directly calculating the mean relatedness in the COMPARE group comprised of one context in C1 period and the other from the C2 period. According to the design, ΔLATER is specifically robust to detect those monosemous words in the EARLIER period that acquired new senses in the LATER period. However, if a changed word has already finished the process of semantic replacement in the LATER period, probably this metric would not be informative anymore. The COMPARE metric was thus proposed to directly compare words usages from the two time intervals.

Following this rationale, we formulated 3 groups of use pairs for each target word, named *C1*, *C2* and *C1C2*, and then randomly sampled 20 use pairs from our subcorpora (see Table 1). In total, each target word would have 60 use pairs, and 1,200 use pairs for all 20 target words.

Each usage pair (see Table 2) is comprised of two sentences containing the target word sampled from relative subcorpora. Enough context information for each sentence is guaranteed by manually checking. The average length of context is around 15 words.

Target word	Context 1	Context 2	Score	Comment
火	极苦的生活和残酷的压迫激起了采煤工人的暴动， 暴动的工人一把火点燃了煤窑	鲁菜卖火了— 山东由农业大省向强省迈进		

Table 2: An example of the use pair in COMPARE group: 火 ‘fire’.

3.4 Annotation

We recruited five native speakers of Mandarin Chinese with linguistics backgrounds as annotators, all of them with a MA degree in Linguistics.

Following Schlechtweg et al. (2018), annotators are asked to give scores to target words by comparing the semantic relatedness across each usage pair (see Table 3). They are also allowed to give a 0 score if they cannot make a decision.

Excluding judgments with 0 grades, 5,968 responses have been collected. The Krippendorff’s alpha was calculated based on five annotators’ ratings. The inter-annotator correlation score is 0.515, comparable to the scores reported for other datasets constructed under the framework of DURel (Schlechtweg et al., 2018, 2020; Kutuzov and Pivovarova, 2021).

	Description
1	Unrelated
2	Distantly related
3	Closely related
4	Identical

Table 3: Four-point scale of relatedness. Taken from Schlechtweg et al. (2018).

4 Dataset Analysis

As described in previous sections, the ΔLATER metric subtracts the mean relatedness of the EARLIER group from the LATER group. Therefore, a positive ΔLATER value is assigned when usages of the annotated word in the C2 group are more similar, whereas negative ΔLATER is assigned to words with less similar usages in the C2 group. Positive and negative ΔLATER values can be considered as two different sub-types of semantic change: innovative meaning change and reductive meaning change, roughly representing the gain or loss of word senses (Schlechtweg et al., 2018). The absolute ΔLATER value assesses the strength of semantic change.

As shown in Figure 1, most annotated words are predicted as stable words, with ΔLATER values around 0. The two topmost words ‘推出’(*push out; launch*), ‘机制’(*machine-made; mechanism*) and the two bottommost words: ‘拐’(*crutch, traffic*), ‘炒’(*to fry, to speculate(in the stock market)*) are predicted as the words with stronger effects of semantic change.

The successful predictions on ‘推出’, ‘机制’, ‘炒’ coincide with documented linguistic publications (Diao, 1995; Shen, 2009), verifying ΔLATER as an effective measure of lexical semantic change. Interestingly, the metric predicted ‘拐’ as a changed word, despite it being originally a stable control word. A closer inspection of all three groups of sampled sentences suggested that ‘拐’ is used more frequently with the ‘crutch’ meaning in the subcorpus C1, but it shows a high prevalence of the ‘trafficking’ meaning in the sub-corpus C2. However, the usage fluctuation detected here has to take into account the corpus bias, as ‘trafficking’ is more likely to occur in a newspaper corpus.

Technically speaking, words such as ‘机制’ and ‘拐’ are homographs with different meanings, i.e. different words with less related or even unrelated meanings. The detected shift actually shows the competition among different meanings with the same surface form, rather than the gain or loss of senses. For example, ‘机制’ in the sampled texts from C1 period dominantly refers to a way of manufacturing as ‘machine-made’ (against ‘handmade’). With the process of industrialization, ‘ machine-made ’ objects became so prevalent in everyday life that the need to mention this feature quickly became obsolete and the usage slipped into obscurity. Meanwhile, the program of Reform and Opening-up was carried out thoroughly, especially concerning the revolution of the Socialist market economy system and mechanism. For this reason, ‘机制’ with the ‘mechanism’ meaning became dominant in the C2 period.

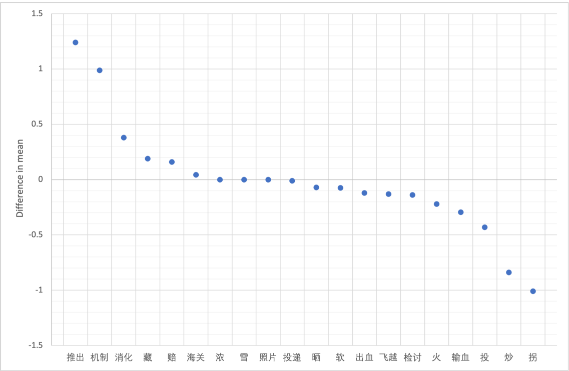

Figure 1: Rank of the target words according to the Δ LATER metric.

The COMPARE metric directly compares the semantic relatedness of a usage pair within the COMPARE group, which consists of sentences from two different periods. Higher COMPARE scores would be assigned to more stable words, like ‘照片(photo),’ ‘雪(snow)’, getting full scores of 4. Lower COMPARE scores are assigned to the shifting ones, e.g. the four changed words predicted by the ΔLATER metric (see Figure 2).

Moreover, this metric captured a shifting word ‘软(soft)’. A closer checking on sampled sentences suggested that ‘软’ is polysemous in the C1 group, but with a dominant usage meaning ‘soft texture of concrete stuff’. In the C2 group, its metaphorical senses even became more diverse, like ‘soft science, soft power’, meanwhile, the ‘soft texture’ sense lost its prevalence based on our observation. The multiple changes made the ΔLATER score not salient *per se*, but they were captured by the COMPARE metric, where usages from two different historical periods are directly compared.

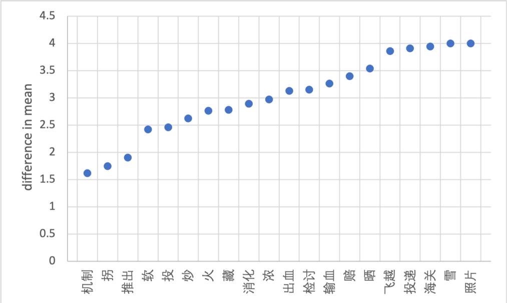

Figure 2: Rank of target words according to the COMPARE metric.

5 Evaluation

The SemEval shared task has indicated that traditional static embeddings may outperform more recent paradigms - e.g., contextualized embeddings (Devlin et al., 2019)- in the task of semantic change detection (Schlechtweg et al., 2020). Therefore, we trained a static word embedding model for this task and evaluated its performance on our newly-created dataset in this study.

We first trained our vectors on each subcorpus using both the Skip-gram model and the Continuous bag of words, which are the two most widely used static word embeddings models (Mikolov et al., 2013a,b). To have an assessment of the quality of the word embeddings trained on our subcorpora, we performed a preliminary evaluation on the Chinese word similarity dataset *COS960*, introduced by Huang et al. (2019).

The results indicated that the quality of the word embedding models was satisfactory (see Table 4). The vectors obtained with the Skip-gram models were better performing, with higher correlation scores for both periods: 0.56 for the C1 period and 0.61 for the C2 period ($p < 0.05$). We thus assumed that Skip-Gram embeddings would provide a better basis for detecting the diachronic semantic change in our study.

We then aligned word representations for the two periods into a shared space with the Orthogonal Procrustes algorithm (Hamilton et al., 2016a,b), projecting word embeddings for the C2 period onto

C1's space and making vectors living in different intervals comparable. The cosine similarity between two vectors for the same word form is calculated as the degree of meaning change. According to the cosine similarity, we ranked those words appearing in both the C1 and C2 periods, where the higher the similarity, the more stable the meaning.

	Skip-gram	CBOW
C1	0.5608	0.4539
C2	0.6144	0.5018

Table 4: Spearman correlation scores between cosine similarities and human ratings for the vectors trained on the subcorpora C1 and C2 (all the correlation scores are significant at $p < 0.05$).

Compared with the scores derived with the COMPARE metric, the Skip-gram model achieved a Spearman correlation score of 0.584. As for the ΔLATER, we took the absolute value indicating the degree of semantic drift for the correlation calculation (the positive and the negative ΔLATER values represent different sub-types of semantic change, but leave this to future investigations). This time, the Skip-gram model achieved a Spearman correlation coefficient of -0.625. Both two correlation scores are statistically significant at $p < 0.05$. As expected, the performance of the Skip-gram model on the detection task is positively correlated with the COMPARE metric and negatively correlated with the ΔLATER metric.

6 Conclusion

This paper presented the first human-annotated evaluation dataset for the task of Chinese lexical semantic change detection. This dataset was built following the DURel framework, which allows us to evaluate the usage drift that occurred in coincidence with the *Reform and Opening-up* in recent Chinese history. Our data further suggested that interpretation of the ΔLATER metric could be extended to the competition among different usages of the same surface form, in order to accommodate historical changes involving homographs. We finally examined the performance of the Skip-gram model on our evaluation dataset and found that it achieves a relatively high correlation coefficient with the two metrics.

This paper served as a first, exploratory study on modeling lexical semantic change in Chinese, on the basis of a limited number of words.

In the near future, our goal is to scale up the dataset and to examine the performance of more models for Chinese, including the more recent contextualized embeddings (Devlin et al., 2019). Moreover, using finer-grained intervals for diachronic meaning change detection and exploring the diatopic variation between different Chinese dialects are also possible directions of our future work (Wang et al., 2022; Zampieri et al., 2019).

References

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In *Proceedings of EVALITA*.

Chinese Academy of Social Science Department of Chinese Lexicography, Institute of Linguistics. 2019. *Contemporary Chinese Dictionary (Xiandai Hanyu Cidian)*, the 7th edition. Commercial Press, Peking.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Yanbin Diao. 1995. *The Development and Reform of Mainland Chinese in the New Era*. Hung Yeh Publishing, Taibei.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of ACL*.

Kristina Gulordava and Marco Baroni. 2011. A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*.

Dasong Guo and Haihong Chen. 1999. *Chinese Neologisms for a Fifty-year (1949-1999)*. Shandong Education Press, Jinan.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of EMNLP*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of ACL*.

Junjie Huang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, and Maosong Sun. 2019. COS960: A Chinese Word Similarity Dataset of 960 Word Pairs. *arXiv preprint arXiv:1906.00247*.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. *arXiv preprint arXiv:1405.3515*.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Statistically Significant Detection of Linguistic Change.

Andrey Kutuzov and Lidia Pivovarova. 2021. Three-part Diachronic Semantic Change Dataset for Russian. In *Proceedings of the ACL International Workshop on Computational Approaches to Historical Language Change*.

Andrey Kutuzov, Lilja vrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic Word Embeddings and Semantic Shifts: A Survey. In *Proceedings of COLING*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, and Peter Norvig. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL-HLT*.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's Sick Dude!: Automatic Identification of Word Sense Change across Different Timescales. In *Proceedings of ACL*.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: A Dataset of Historical Lexical Semantic Change in Russian. In *Proceedings of COLING*.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. *Proceedings of the EACL Workshop on GEMS: Geometrical Models of Natural Language Semantics*.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of SemEval*.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of NAACL-HLT*.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A Large Resource of Diachronic Word Usage Graphs in Four Languages. *arXiv preprint arXiv:2104.08540*.

Mengying Shen. 2009. *New Words and New Expressions in Chinese New Era (1949-299)*. Sichuan Lexicographical Press, Chengdu.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2019. Survey of Computational Approaches to Lexical Semantic Change. *arXiv preprint arXiv:1811.06278*.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2013. Semantic Change Computation: A Successive Approach. In *Behavior and Social Computing*, pages 68–81, Cham. Springer International Publishing.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2016. Semantic Change Computation: A Successive Approach. *World Wide Web*, 19.

Hal Varian and Hyunyoung Choi. 2009. Predicting the Present with Google Trends. *Economic Record*, 88.

Shan Wang, Ruhan Liu, and Chu-Ren Huang. 2022. Social Changes through the Lens of Language: A Big Data Study of Chinese Modal Verbs. *PLOS ONE*, 17:1–31.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A Report on the Third VarDial Evaluation Campaign. In *Proceedings of the NAACL Workshop on NLP for Similar Languages, Varieties and Dialects*.

Sergey Zinin and Yang Xu. 2020. Corpus of Chinese Dynastic Histories: Gender Analysis over Two Millennia. In *Proceedings of LREC*.