


## Article

# Diachronic Semantic Tracking for Chinese Words and Morphemes over Centuries

Yang Chi <sup>1</sup>, Fausto Giunchiglia <sup>1,2,3</sup>  and Hao Xu <sup>1,2,\*</sup>

<sup>1</sup> School of Artificial Intelligence, Jilin University, Changchun 130012, China; yangchi19@mails.jlu.edu.cn (Y.C.); fausto.giunchiglia@unitn.it (F.G.)

<sup>2</sup> College of Computer Science and Technology, Jilin University, Changchun 130012, China

<sup>3</sup> Department of Computer Science and Information Engineering (DISI), University of Trento, 38123 Trento, Italy

\* Correspondence: xuhao@jlu.edu.cn

**Abstract:** Lexical semantic changes spanning centuries can reveal the complicated developing process of language and social culture. In recent years, natural language processing (NLP) methods have been applied in this field to provide insight into the diachronic frequency change for word senses from large-scale historical corpus, for instance, analyzing which senses appear, increase, or decrease at which times. However, there is still a lack of Chinese diachronic corpus and dataset in this field to support supervised learning and text mining, and at the method level, few existing works analyze the Chinese semantic changes at the level of morpheme. This paper constructs a diachronic Chinese dataset for semantic tracking applications spanning 3000 years and extends the existing framework to the level of Chinese characters and morphemes, which contains four main steps of contextual sense representation, sense identification, morpheme sense mining, and diachronic semantic change representation. The experiment shows the effectiveness of our method in each step. Finally, in an interesting statistic, we discover the strong positive correlation of frequency and changing trend between monosyllabic word sense and the corresponding morpheme.

**Keywords:** diachronic semantic tracking; lexical sense; morpheme



**Citation:** Chi, Y.; Giunchiglia, F.; Xu, H. Diachronic Semantic Tracking for Chinese Words and Morphemes over Centuries. *Electronics* **2024**, *13*, 1728. <https://doi.org/10.3390/electronics13091728>

Academic Editors: Arkaitz Zubiaga, Yanping Zhang, Jianjun Yang and Wenlin Han

Received: 16 March 2024

Revised: 20 April 2024

Accepted: 25 April 2024

Published: 30 April 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Lexical meaning constantly changes over time, reflecting the complicated developing process of language and social culture [1]. For example, in the Shang and Zhou dynasties (1000 BC), the original meaning of the Chinese character “师” means “army”. Later, it emerged a new meaning of “an official position specifically responsible for people’s education”, and then it derived the meaning of “teacher”, which is the widest sense at present.

Tracking the lexical semantic changes spanning centuries, such as analyzing which senses are stable and which senses appear, increase, or decrease at which times, contributes to research in the fields of historical linguistics, philology, history, dictionary compilation, and so on. It can also be applied for natural language processing (NLP) and retrieval systems, for instance, automatically discovering semantic change of a word, providing visualization analysis for diachronic lexical semantics, enhancing the retrieval or recommendation of the historical resource by considering the lexical meanings, etc. However, how to extract diachronic lexical semantic knowledge from historical documents is a problem. There are large numbers of historical documents preserved in human history, and there are certain differences between ancient and modern languages. It relies on a large amount of manual labor and expert experience to explore large-scale historical materials and determine the meaning of the target word in each context, and this kind of investigation is limited because it is impossible to cover all of the historical documents through manual work.

In recent years, NLP and a large historical corpus have promoted the development of automated methods in this field. The large-scale pre-trained language model such as BERT [2] can capture complex contextual features and give differentiated representations to the target word in different contexts. Based on this deep contextualization representation, diachronic sense modeling is introduced in recent works, which assigns differentiated embedding to each sense of the target word rather than representing the word as a single vector at each time period, and the model can represent the frequency distribution change of these senses in a fine-grained, smooth, and interpretable way [3].

However, there are still some problems that should be solved for semantic tracking of the Chinese language spanning a long historical period. Most works typically consider words as the smallest semantic unit. Unlike the “word” in Indo-European languages, the Chinese character is the semantic unit naturally formed during the origin of the Chinese language, which can be applied as a mono-syllabic word or a morpheme to compose polysyllabic compound words. Therefore, mining the semantic changes of Chinese characters from the perspective of morphemes is significant for exploring the development of the Chinese language over centuries. In addition, there is currently a lack of authoritative diachronic Chinese datasets for supervised training and semantic tracking. Although the pre-trained language models have learned general contextual semantic information, they are not yet able to identify senses accurately. An effective way is to fine-tune the model further through supervised learning to make similar senses closer in high-dimensional space and distinguish different ones.

Focused on the above problems, we construct a data resource for Chinese semantic tracking. It includes a sense-context dataset, which contains words (including Chinese characters), senses, and contexts with time stage annotation, and a Chinese historical literature corpus for sense tracking. In addition, we expand the semantic tracking framework of work [3], it contains four main steps: contextual sense representation, sense identification, morpheme sense mining, and diachronic semantic change representation. Contextual sense representation and morpheme sense mining are innovative works in this paper. More specifically, at the contextual sense representation step, we train a contextual sense representation model using a sense-context dataset. Then, the model is used for sense identification and morpheme sense mining tasks, which identify the sense for each lexical token in contexts of the Chinese historical literature corpus and discover the character senses that can be applied to morphemes. Finally, the last step provides smooth frequency distribution representations for each sense from the two levels of monosyllabic words and morphemes. The framework can be applied to other languages; however, the morpheme sense mining is only designed for Chinese. For other languages, it cannot analyze the information related to morpheme senses, while other functions are the same as those of Chinese.

The results of the experiment showed that our model performs better for sense identification and morpheme sense mining than the original BERT model. The accuracy of sense identification was improved from 55.08% to 74.19%, and the F1 score for morpheme sense mining was improved from 59.21% to 75.14%. We also give a visualization and analysis case of the semantic change in Chinese characters over centuries. Finally, as an interesting analysis of language development, we give the statistical conclusions observed from semantic tracking of 100 common Chinese characters, discovering the strong positive correlation of diachronic frequency distribution changes between monosyllabic words and corresponding morphemes.

The contributions of this paper include the following:

1. Constructing a resource for Chinese lexical semantic tracking, which includes a sense-context dataset and a Chinese historical literature corpus;
2. Extending the existing lexical semantic tracking framework from the perspective of morphemes;
3. Proposing and training a contextual sense representation model for sense identification and morpheme sense mining tasks;

4. Discovering strong positive correlations of frequency distribution and diachronic change between monosyllabic word sense and corresponding morpheme sense.

This paper consists of the following sections: Section 2 will introduce the related knowledge of Chinese words, characters, and morphemes. Section 3 will introduce the related works. Section 4 will introduce our framework and method, including the contextual sense representation model, sense identification, morpheme sense mining, and diachronic semantic change representation, as well as the datasets we constructed. Section 5 is the evaluation of our method, which will show the experiments and results, and we will also give a semantic tracking visualization case. Section 6 will analyze the correlations of frequency distribution and diachronic change between monosyllabic words and corresponding morphemes. Section 7 is the discussion of the results. Section 8 is the conclusion of the whole work.

## 2. Introduction of Word, Chinese Character, Sense, and Morpheme

We will briefly introduce the concepts of words, Chinese characters, senses, and morphemes that are related to our research.

**Word:** A word is the smallest sentence-making unit composed of morphemes. In this paper, we simplistically classify the Chinese word into two kinds: monosyllabic word and compound word. A monosyllabic word is composed of one Chinese character; for instance, “师” is a monosyllabic word, and one of the meanings is “teacher”. The compound word is composed of more than one Chinese character; for instance, the compound word “教师 (teacher)” is composed of two characters: “教 (educate)” and “师 (teacher)”. According to our statistics on the 256,293 compound words from the Chinese dictionary, 85% (217,967) of them consist of two characters, 6.7% (17,148) of them consist of three characters, and 7.8% (20,016) of them consist of four characters.

**Sense:** A sense is the meaning of a word, and one word can have multiple senses.

**Chinese character:** A Chinese character is the recording symbol in Chinese. From the perspective of expressing meanings, a Chinese character can be seen as a monosyllabic word (e.g., the monosyllabic word “师”). It can also be seen as a morpheme, for example, the compound word “教师 (teacher)” is composed of two morphemes: “教 (educate)” and “师 (teacher)”.

**Morpheme:** A morpheme is the minimal meaningful language unit in a word. For example, the English word “teacher” is composed of morphemes “tech” and “er”. In Chinese, most morphemes are represented by a single Chinese character. In general, the sense of the compound words is similar or related to the sense of their morphemes.

The semantics of Chinese characters are highly correlated to the evolution of the Chinese language and the entire history and culture. In ancient times, the Chinese character (monosyllabic word) was the smallest semantic unit naturally created during the origin of the Chinese language; with the development of society and the increase of cognitive concepts, polysyllabic compound words were created due to the unsustainable expansion of characters, which can be directly composed of existing characters (morphemes). From the above introduction, the expression or usage pattern of the semantics of Chinese characters can be naturally and clearly divided into two types: used as a monosyllabic word or used as a morpheme to compose other compound words. Each sense of Chinese characters may be applied to both two patterns or tend to one of them, and the situation changes over time. For example, in the modern age, many characters can no longer be directly used as monosyllabic words in their certain senses, which does not mean that these senses have been lost—they may still be preserved in morphemes. Therefore, it is necessary to introduce the concept of morphemes into the semantic tracking framework of Chinese characters.

## 3. State of the Art

The most relevant research to our work is lexical semantic change detection (LSCD) in the NLP field. Most scholars currently use parametric distributional models, particularly prediction-based contextual embedding algorithms, to represent the semantic features of

words. By calculating the distance of word embeddings between different times, experts can understand the basics of semantic change. The word representation can be divided into static embeddings and contextualized embeddings [4]. Static models depend on a strong simplification: a single representation is sufficient to model the different meanings of a word. Contextualized approaches can give differentiated representations for one word in different contexts.

Most static models, such as skip-gram with negative sampling (SGNS) and continuous bag-of-words (CBOW) [5], are shallow neural language models. Kim et al. [6] first used SGNS to track diachronic semantic changes. They chronologically trained the model by initializing word vectors for subsequent years with the word vectors obtained from previous years. Hamilton et al. [7] independently trained word embeddings in different time intervals and aligned them to quantify semantic changes and reveal statistical laws of semantic evolution. Rosenfeld and Erk [8] built the first diachronic distributional model that represents time as a continuous variable and uses word representations as a function of the time vector. Yin et al. [9] developed a continuous diachronic distributional model based on the global anchor method for quantifying linguistic shifts and domain adaptation. Kaiser et al. [10] point out that the representations are strongly influenced by the size of training corpora, and they use simple pre- and postprocessing techniques to improve the embeddings. However, it is well known that word semantics can be represented with a range of senses, but static models can only assign one embedding representation for each word. Static models can model the coarse-grained semantics of a word from one time to another but cannot represent each sense at a fine-grained level. Some works made extensions of the SGNS model to learn sense-specific embeddings to solve this problem [11,12].

Recently, an increase in large-scale pre-trained language models, e.g., BERT, has attracted considerable attention in the field of NLP, which is fine-tuned with just one additional output layer and achieves state-of-the-art results for a wide range of tasks [2]. These models can ideally capture complex characteristics of word use and how they vary across linguistic contexts to provide contextualized representations for words. Hu et al. [3] first used BERT to represent fine-grained word senses, and they also proposed a framework based on deep contextualization embeddings to deeply detect changes in word senses. Work [13] follows their framework and applies it to Chinese lexical semantic tracking. In particular, they constructed a corpus for word sense disambiguation for Chinese, which is similar to our work. Giulianelli et al. [4] presented the first unsupervised approach for lexical semantic change that makes use of BERT word representations. Kurtyigit et al. [14] demonstrated that both static and contextualized models can successfully be applied to discover new words that are changing meaning. Laicher et al. [15] considerably improved BERT's performance by reducing the influence of orthography on the target word while keeping the rest of the input in its natural form. Teodorescu et al. [16] proposed a novel approach based on framing lexical semantic change detection as a WSD problem, and they utilized the XLM-RoBERTa model [17]. Rosin et al. [18] proposed a temporal contextual language model called TempoBERT, which uses time as an additional context of texts to enhance the performance of semantic change detection. Cassotti et al. [19] proposed a pre-trained bi-encoder model on a large-scale dataset for the word to obtain comparable lexical-based representations.

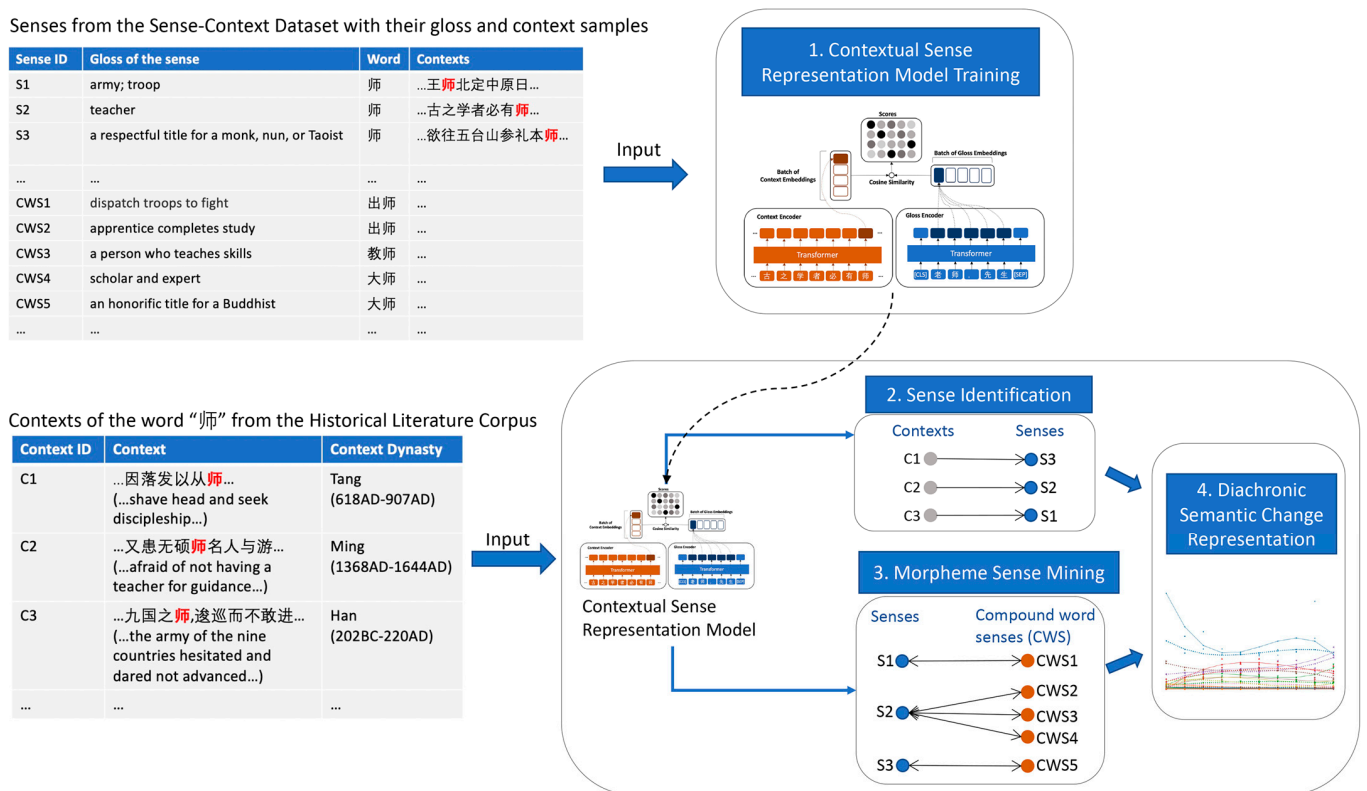
Other related works focus on mining specific lexical changes, such as discovering novel senses, which are usually based on clustering, topic modeling, and network approaches. Cook et al. [20] detected novel senses by comparing a reference corpus and a focus corpus with topic modeling. Mitra et al. [21] identified the sense birth, death, join, and split based on clustering of a co-occurrence graph. Tahmasebi and Risse [22] induced word senses and tracked their changes based on a curvature clustering algorithm. Jana et al. [23] detected the words evolved with a novel sense from a corpus from two different time periods based on network features and a support vector machine (SVM) classifier. The work [24] proposed a scalable method for contextual embeddings clustering that generates interpretable representations. Giulianelli et al. [25] automatically generated nat-

ural language definitions of contextualized word usages as interpretable word and word sense representations.

According to the investigation, the majority of existing works were designed for Indo-European languages, especially English. One of the challenges is to extend existing methods to other languages. One problem is the lack of data resources for this task; the scale of existing resources should be further extended to cover more words and senses. On the other hand, the semantics of Chinese characters and morphemes are also valuable and have not been considered in existing research. At the method level, the accuracy, authority, and interpretability at the sense level are more important for our work, so we extend the framework of paper [3], which constructed a contextualized model to represent each sense, and tracked the fine-grained semantic changes in a smooth process.

### 4. Materials and Methods

We will introduce our framework and method in five parts: dataset construction (Section 4.1), contextual sense representation model (Section 4.2), sense identification (Section 4.3), morpheme sense mining (Section 4.4), and diachronic semantic change representation (Section 4.5). The framework process is shown in Figure 1.



**Figure 1.** Framework of semantic tracking of a Chinese character, taking the Chinese character “师” as an example. It contains four steps: (1) contextual sense representation model training using the sense-context dataset; (2) realizing sense identification based on the model, mapping the target word in the contexts into the corresponding sense label; (3) realizing the morpheme sense mining based on the model, matching the character sense with the corresponding compound word senses; (4) semantic change representation, providing smooth frequency distribution representations for each sense according to the results from step 2 and step 3.

Firstly, we construct a sense-context dataset based on authoritative Chinese dictionaries and train a contextual sense representation model, which can extract contextual features of the word (character), making the embeddings of senses with similar meanings close in space.

Then, given a Chinese character to be analyzed (“师” in Figure 1), we retrieve all the contexts containing the target character in a large-scale historical literature corpus and perform sense identification and morpheme sense mining based on the contextual sense representation model. The purpose of the sense identification task is to select the correct sense for the target character (word) in a context from the set of candidate senses provided by the sense-context dataset. For instance, in Figure 1, the Chinese character “师” has three candidate senses: “army; troop”, “teacher”, and “a respectful title for a monk, nun, or Taoist”, and “army; troop” is the correct sense for the character which is in the context “...九国之师, 逡巡而不敢进...” (“the army of the nine countries hesitated and dared not advanced”).

The morpheme sense mining task aims to answer which senses of the target character (which is used as a morpheme) can be used in compound words and what senses of compound words are composed by it. In this work, we called the senses of the target character, which can be used in a group of compound words to participate in composing their meanings, “morpheme sense”. Therefore, the main process of this task is to match each sense of the target character to the candidate senses of its compound words. For instance, the three morpheme senses of “师” (“army; troop”; “teacher”; “a respectful title for a monk, nun, or Taoist”) in Figure 1 can be, respectively, matched to the senses of compound words of “出师” (dispatch troops to fight), “教师” (a person who teaches skills), and “大师” (an honorific title for a Buddhist).

Finally, for each sense of the target character, after identifying all the corresponding tokens in the contexts of the corpus and finding all related senses of compound words, we can represent and visualize the diachronic frequency distribution of the sense in two perspectives of monosyllabic word and morphemes. The details will be introduced in Sections 4.1–4.5.

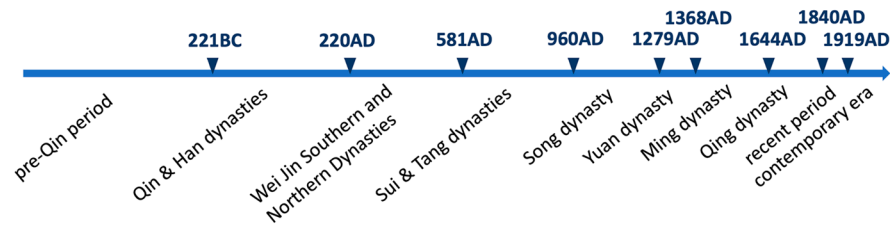
#### 4.1. Dataset Construction

We build a sense-context dataset and a historical literature corpus from an authoritative dictionary and an open historical database, respectively, used for model training and semantic tracking.

##### 4.1.1. Diachronic Chinese Sense-Context Dataset

The diachronic Chinese sense-context dataset has 10 information tags as follows:

1. Context: One context of the target word, usually one sentence;
2. Time stage: The time stage of the context. We divide time intervals by relatively broad Chinese dynasties because dynasty is the most direct and important basis for dividing Chinese historical stages, including politics, life, culture, language, and so on. Although the change nodes in language are not entirely consistent with dynasties, the more fine-grained time division is hard to define, and the exact years of the majority of ancient documents are no longer traceable, which makes it hard to realize automatic annotation. The time ranges in this work divided by dynasties are shown in Figure 2;
3. Author: The author of the context;
4. Title: The name of the literary work from which the context is from;
5. Word position: The position index of the target word in the context;
6. Word ID: The identifier code of the word;
7. Word: The target word in the context. It can be a monosyllabic word (character) or a compound word;
8. Sense ID: The identifier code of the sense for the target word in the context;
9. Sense gloss: The definition of the sense;
10. Sense number: The number of the sense that the target word has.



**Figure 2.** Division of time intervals for Chinese semantic tracking in the whole historical period.

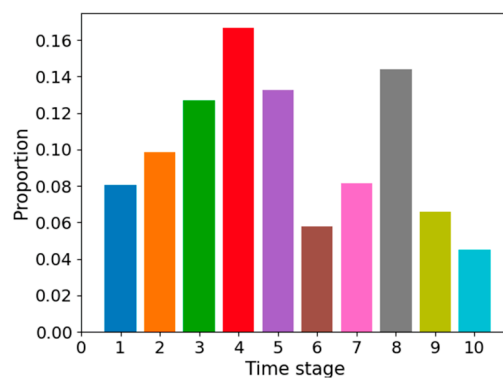
The original data is from open Chinese dictionary resources such as The Great Chinese Dictionary. We constructed the dataset as per the following steps:

1. We extracted the structured information from the dictionary, such as the word (character), gloss, example sentence, author, title, and so on;
2. We simplified the definitions by only retaining the first three sentences and deleting descriptive words unrelated to the meaning. This is because we find that most of the sense definitions in the dictionary are short, within 8 characters and two sentences, and for those long contexts, only the first few sentences are the interpretation of the senses;
3. We aligned the same sense from different resources according to the cosine similarity of definitions. If the similarity score was up to 0.85, we considered them as the same sense;
4. We annotated the dynasty for each document based on their author and title information by automatically searching Baidu Baike [26].

Ultimately, there are 265,289 words, 392,213 senses, and 844,236 contexts in the sense-context dataset. On average, each word (including the monosyllabic word or character) has 1.48 senses, each character has 5.26 senses, and each word sense has 2.15 contexts, and each character sense has 3.91 contexts. The details can be seen in Table 1; we, respectively, give the statistics of the two kinds of words, monosyllabic word (character) and compound word, and their senses and contexts. The proportion of the contexts per time stage is shown in Figure 3.

**Table 1.** Statistics of the Word, Sense, and Context number in the sense-context dataset, the Chinese Word is divided into Monosyllabic word (Character) and Compound word.

|                               | Word    | Sense   | Context |
|-------------------------------|---------|---------|---------|
| Monosyllabic word (Character) | 8996    | 47,341  | 185,071 |
| Compound word                 | 256,293 | 344,872 | 659,165 |



**Figure 3.** Proportion of the contexts per time stage, the time stage numbers (1–10) correspond to the 10 time stages in Figure 2.

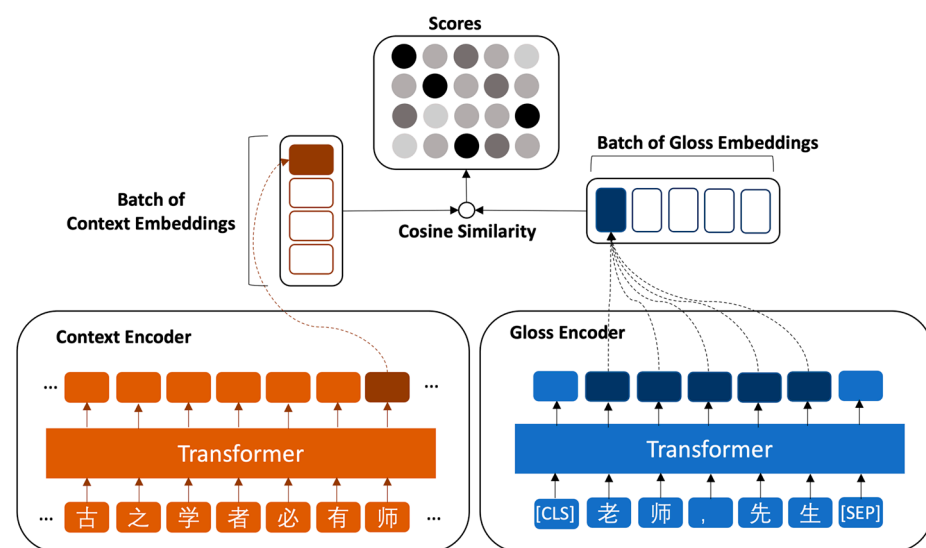
#### 4.1.2. Chinese Historical Literature Corpus

The historical literature corpus includes documents and their historical dynasty. This corpus is extracted from the Daizhige corpus [27], which has a scale of over 2 billion characters and covers documents passed down throughout the Chinese historical period, including various categories such as poetry, novels, essays, dramas, etc. We also annotated the dynasty for each document by automatically searching Baidu Baike [26]. After deleting the duplicate documents and those that failed to find the dynasty, we obtained 4569 documents distributed throughout various dynasties, with a total of 4.13 GB.

#### 4.2. Contextual Senses Representation Model

In this work, we define contextual sense representation as the token embedding for the target character (or compound word) in a certain context. Although the BERT model has already learned several general semantic features from the pre-trained procedure, to give better contextual sense representations to distinguish different senses more accurately, we fine-tune the BERT model under the WSD task using a sense-context dataset. More formally, given a character (or compound word)  $w$  and context  $c$ , a WSD system is a function  $f$  such that  $f(w, c) = s$  subject to  $s \in S_w$ , where  $S_w$  is the set of all possible candidate senses of  $w$ .

This work utilizes a jointly optimized bi-encoder model to import the information in the sense-context dataset to improve the contextual sense representation. The architecture of the model is shown in Figure 4. Here, we trained a single model to cover all dynasties rather than separate models for each period. This is because the pre-trained transformer model, such as BERT, has the ability to distinguish semantic features of contexts from different eras, and the costs of time and space for training multiple models are much more than the single one.



**Figure 4.** Architecture of the bi-encoder contextual senses representation model in this work.

The bi-encoder architecture independently encodes contexts and sense glosses. Each of the two models is initialized with BERT, which has 12 layers, 768 hidden units, and 12 heads. In this work, we use the bert-ancient-Chinese model [28], which was acquired by further training on the large-scale corpus of ancient Chinese literature from Google's bert-base-Chinese [29] model to make it more suitable for processing ancient Chinese. The training of the BERT model is on the Masked Language Model (MLM) and Next Sentence Prediction (NSP), which are the two unsupervised pre-training tasks.

The input of the context encoder is  $c = [\text{CLS}], w_1, \dots, w_i, \dots, w_n, [\text{SEP}]$ , where  $w_1$  to  $w_n$  are  $n$  characters in the context  $c$ ;  $w_i$  is the target character to be predicted; [CLS] and [SEP] are BERT-specific start and end symbols [2]. In our model,  $w_i$  is randomly replaced



with a [MASK] symbol in a certain proportion to conceal the information of the character and enable the model to better predict the sense based on the context. We take the corresponding output of  $w_i$  in the final layer of context encoder as the representation  $r_{w_i}$ . For compound words that are tokenized into multiple characters, we represent them as the average of the embeddings of their characters.

The input of the gloss encoder is  $g = [\text{CLS}], d_1, \dots, d_m, [\text{SEP}]$ , where  $g$  is the gloss (definition) of a candidate sense of  $w_i$ ;  $d_1$  to  $d_m$  are  $m$  characters in the gloss. We take the average of  $m$  embeddings from  $d_1$  to  $d_m$  that are outputted from the final layer of the gloss encoder as the gloss representation  $r_g$ .

We train the model based on the idea of comparative learning [30]. For each target character (or compound word)  $w$ , its candidate sense set  $S_w$  contains all of the senses in the corresponding mini-batch of gloss encoder. We then score each gloss  $g$  of candidate sense  $s \in S_w$  for target character  $w$  by calculating cosine similarity:

$$\text{sim}(w, g) = \frac{r_w r_g}{\|r_w\| \cdot \|r_g\|} \quad (1)$$

For the correct pair of characters (or compound word) and sense gloss  $(w_i, g_k)$ , our goal is to make the representation of  $r_{w_i}$  close to  $r_{g_k}$  and far away from other candidate sense gloss embeddings in the feature space. We use a cross-entropy loss on the scores for the glosses of candidate senses to train the bi-encoder model; the training objective for  $w_i$  is defined as:

$$\ell_{w_i} = -\log \frac{e^{\text{sim}(w_i, g_k)/\tau}}{\sum_{j=1}^{|S_{w_i}|} e^{\text{sim}(w_i, g_j)/\tau}} \quad (2)$$

where  $\tau$  is a temperature hyperparameter, we directly use the default value as 0.5. We fine-tune all the parameters using the contrastive learning objective (Equation (2)). The batch size of the gloss encoder is set as the default value of 64, and the proportion of the [MASK] symbol in the context encoder is 0.2. Here, we experimented with the values in the set [0.1, 0.2, 0.3], and 0.2 generated the best results in the sense identification task. We train the model on the sense-context dataset for 50 epochs until the results of sense identification have not continued to be improved. After training, the output  $r_{wc}$  of the final layer of the context encoder can be directly used as the contextual sense representation for the character (or compound word)  $w$  in the certain context  $c$ .

#### 4.3. Sense Identification

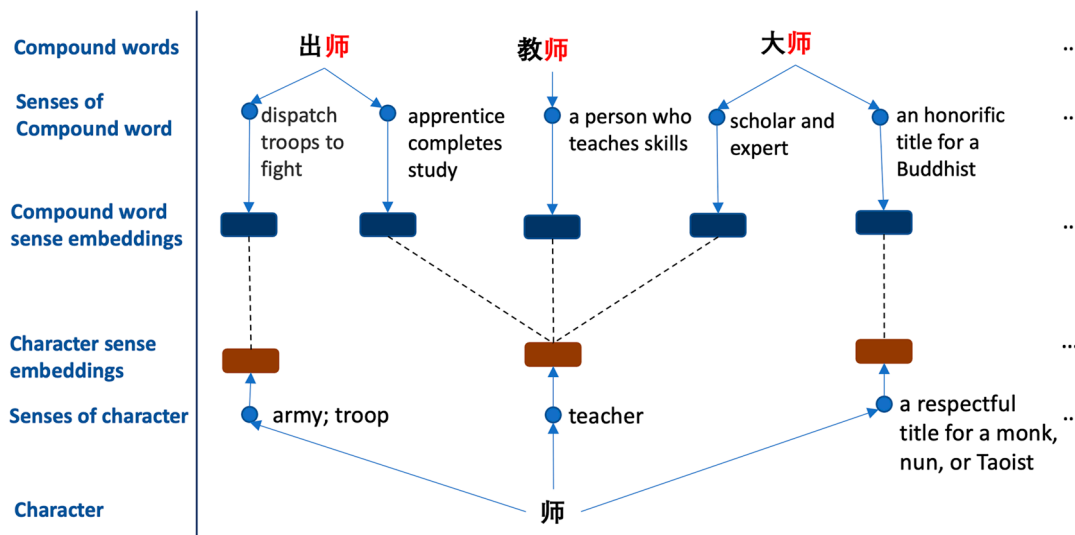
It is convenient to identify the sense of the character (or compound word) in a certain context using our model because it is trained under the task of WSD. Given the target character (or compound word)  $w$  and the certain context  $c$ , we input  $c$  to the context encoder of the trained model and achieve the contextual sense embedding  $r_{wc}$  from the final hidden layer of the model. For each candidate sense  $s$  of  $w$ , we feed its gloss into the gloss encoder and output the gloss embedding  $r_g$ . In this way, we can achieve the contextual sense embedding  $r_{wc}$  and all of the candidate gloss embeddings  $\{r_{g_1}, r_{g_2}, \dots, r_{g_n}\}$  (assuming  $w$  has  $n$  senses). We directly compute the cosine similarities between  $r_{wc}$  and  $n$  gloss embeddings and select the one with the highest similarity score as the predicted sense for  $w$  in context  $c$ .

#### 4.4. Morpheme Sense Mining

The morpheme sense mining task aims to answer which senses can be applied to the morpheme to compose compound words and what senses of compound words can be composed by each morpheme sense. More specifically, for each target character  $w$ , this task finally obtains a morpheme sense set  $MS_w$  for it, as well as the corresponding senses of compound words composed by each morpheme sense in  $MS_w$ .  $MS_w$  is the subset of  $S_w$ , which is the set of senses for the target character. Therefore, the key point is to match each

sense of the target character with the corresponding senses of compound words that are composed by the character.

Figure 5 shows an example of morpheme sense mining for the Chinese character “师”. The method retrieves all the compound words that are composed by the character “师” from the dictionary and then matches their senses to the corresponding sense of “师”. For instance, the first sense of the compound word “出师” means “dispatch troops to fight”, which should be matched to the sense “army; troop” of “师”; the second sense of “apprentice completes study” should be matched to the sense “teacher” of “师”. Details are shown as follows.



**Figure 5.** An example of a morpheme sense mining task (matching three senses of the Chinese character “师” with the senses of compound words of “出师”, “教师”, and “大师”).

For target character  $w$ , we can acquire the contextual sense representations  $\{r_{wc_1}, r_{wc_2}, \dots, r_{wc_m}\}$  of  $w$  from the context encoder of our model, assuming  $w$  has  $m$  contexts in the historical literature corpus. Then, we execute sense identification (Section 4.3) to find the sense of  $w$  in each context and obtain sense representations  $\{r_{s_1}, r_{s_2}, \dots, r_{s_k}\}$  of  $w$ , assuming  $w$  has  $k$  senses. Sense representation  $r_s$  is defined as the average of contextual sense representations that are identified to sense  $s$ .

For each compound word which is composed by  $w$ , we also obtain its sense representations in the same way. For each sense of each compound word, we calculate the cosine similarities between its representation and  $\{r_{s_1}, r_{s_2}, \dots, r_{s_k}\}$ , matching the compound word sense to the character sense with the highest similarity score. If the maximum score is less than 0.01, the sense of the compound word will not be matched to any character sense. We set this parameter by observing the experimental results in a small range of data samples. Our candidate values are [0.01, 0.02], and we found most of the matching is incorrect when the similar score is  $< 0.01$ . After this step, for each sense of  $w$ , it has a corresponding set of compound word senses. When the number of matched compound word senses is greater than a threshold  $\beta$ , we add the sense to the morpheme sense set  $MS_w$ . We set  $\beta = 3$  in this work. It is also set by observing the results in a small range of samples. Our candidate parameters are [1, 2, 3, 4, 5], and we found most of the matchings are incorrect when  $\beta < 3$ .

After this task, for each target character  $w$ , we achieve a morpheme sense set  $MS_w$ , as well as the set of matched senses of compound words for each morpheme sense in  $MS_w$ .

#### 4.5. Diachronic Semantic Change Representation

We follow the work [3] to represent the change of frequency distribution of each sense of Chinese characters (or compound words) over time. For Chinese characters, we extend

the model to represent the semantic change at two levels: monosyllabic words and morphemes. Here is a brief introduction below.

Given the target character  $w$  to be represented and compound words of it, we conveniently identify their tokens in contexts to the corresponding sense based on the method described in Section 4.3. After processing morpheme sense mining based on the method in Section 4.4, we obtain the morpheme sense set  $MS_w$ , as well as the corresponding senses of compound words for each character sense in  $MS_w$ . According to this information, we conduct statistics on the sense labels of context in each time range (Section 4.1) separately to achieve the frequency distribution of each sense of the character  $w$  in the whole historical period. For each character sense  $s_i$ , the diachronic change of frequency distribution is represented by:

$$T_{word}(s_i) = \{P_{t_1}^{s_i}, P_{t_2}^{s_i}, \dots, P_{t_n}^{s_i}\} \quad (3)$$

$$T_{morpheme}(s_i) = \{PM_{t_1}^{s_i}, PM_{t_2}^{s_i}, \dots, PM_{t_n}^{s_i}\} \text{ if } s_i \in MS_w \quad (4)$$

where  $T_{word}(s_i)$  is the representation of a diachronic change of frequency distribution for  $s_i$  when it is applied to a monosyllabic word;  $T_{morpheme}(s_i)$  is the representation for  $s_i$  when character  $w$  is used as a morpheme in compound words. If  $s_i$  is not in  $MS_w$ ,  $T_{morpheme}(s_i)$  will not exist.  $\{t_1, t_2, \dots, t_n\}$  are sequential  $n$  time ranges.  $P_t^{s_i}$  and  $PM_t^{s_i}$  are, respectively, defined as:

$$P_t^{s_i} = \frac{N_t^{s_i}}{\sum_{k=1}^m N_t^{s_k}} \quad (5)$$

$$PM_t^{s_i} = 0.5 \times \frac{NM_t^{s_i}}{\sum_{k=1}^m NM_t^{s_k}} + 0.5 \times \frac{NS_t^{s_i}}{\sum_{k=1}^m NS_t^{s_k}} \quad (6)$$

where  $N_t^{s_i}$  is the number of tokens in contexts identified as sense  $s_i$  at time  $t$ , assuming  $w$  has  $m$  senses.  $NM_t^{s_i}$  is the number of tokens in contexts of compound words that are matched to  $s_i$  at time  $t$  after sense identification and morpheme sense mining processes.  $NS_t^{s_i}$  is the number of senses of compound words that are matched to  $s_i$  at time  $t$ . Finally, we conduct quartic polynomial curve fitting for  $T_{word}(s_i)$  and  $T_{morpheme}(s_i)$ . This method gives a continuous frequency distribution representation for each sense of the target Chinese character (word), clearly monitoring the status of each individual sense, whether it is growing or decreasing.

## 5. Evaluation

We evaluate the method through three indicators around contextual sense representation, sense identification, and morpheme sense mining tasks as follows:

1. Contextual sense representation: given the embedding of the target character outputted by the model in a certain context, we calculate the similarity scores between it and all senses in the dictionary and count the number of synonymous senses with it in the top ranking. The purpose is to evaluate the effect of the model on representing senses in certain contexts, including whether it can distinguish different senses and give similar representations for synonymous senses;
2. Sense identification: given the target character (or compound word) and the corresponding context, the model selects the sense for the target character (word) from the list of candidate senses. We use accuracy as the indicator to evaluate the effect of sense identification;
3. Morpheme sense matching: given the target character sense and the corresponding compound word senses matched to it by the model. We use precision, recall, and f1 as indicators to evaluate whether the morpheme sense matching is correct and comprehensive.

The specific methods and results for each indicator will be introduced in Sections 5.1–5.3, respectively. Finally, as a qualitative analysis, we provide a semantic tracking case of the character “师” in Section 5.4.

### 5.1. Results of Contextual Sense Representation

In this section, we evaluate the contextual sense representation of our fine-tuned model that was trained in the sense-context dataset to verify whether it can distinguish different senses and give similar representations for the synonymous senses. And we compare it with the original BERT model.

In the experiment, for the target character  $w$  and certain context  $c$ , we evaluate the contextual sense representation  $r_{wc}$  by calculating the cosine similarity between it and the representations  $\{r_{s_1}, r_{s_2}, \dots, r_{s_n}\}$  of all senses  $\{s_1, s_2, \dots, s_n\}$  in the dictionary, assuming that there is a total  $n$  senses recorded in the dictionary. We sort the senses in the dictionary according to their similarity scores and use the number of synonymous senses to  $w$  in the Top 5–30 ranking as the indicator for evaluation. The sense representation  $r_s$  is defined as the average of contextual sense representations for all contexts of  $s$  which are recorded in the sense-context dataset.

We randomly select 300 contexts that do not exist in the training dataset distributed in various historical eras and corresponding commonly used target characters as the samples. Before the selection, we had given unique identifications for each context in the sense-context dataset to make sure the testing samples were not in the training set. Three highly educated Chinese native speakers who can understand basic ancient Chinese were asked to judge whether the candidate senses in the top rankings are synonymous with the corresponding meaning of the target character. The candidate sense will be set as the synonymous sense if more than two judges support it. It should be noted that the dictionary data source had provided the corresponding sense label for the target character in context, so the task of the annotators did not include judging the sense of the character in the ancient Chinese context; they just needed to compare the definitions (written by modern Chinese) between the character sense and the similar senses given by machine. Finally, we count the number of captured synonymous terms in Top-5, Top-10, and Top-30 rankings and take the average of 300 samples. We compare the effects between the original bert-ancient-Chinese model (BERT-ancient) and our model (BERT-ancient-trained). The results are shown in Table 2. It can be seen that our trained model captures more synonymous senses in the similarity ranking of the Top 5–30, which means that it can give closer representations of the synonymous senses.

**Table 2.** Number of synonymous terms in the Top 5–30 ranking of similar senses given by our trained model and original BERT model (BERT-ancient is bert-ancient-Chinese model; BERT-ancient-trained is our model trained from bert-ancient-Chinese).

| Model                       | Top-5        | Top-10       | Top-30       |
|-----------------------------|--------------|--------------|--------------|
| BERT-ancient                | 1.120        | 1.613        | 3.193        |
| BERT-ancient-trained (ours) | <b>2.257</b> | <b>3.767</b> | <b>7.410</b> |

Table 3 shows four examples of Top-5 similar senses from the two models, respectively. It can be found that though the original BERT model can generate differentiated representations for characters depending on different contexts, its ability to distinguish senses is limited. For instance, for the second character sense “行” (walk), all of the Top-5 similar senses recommended by the BERT-ancient model belong to the same character “行”, though they are not synonymous. In comparison, the trained model avoids this problem by selecting similar senses from all Chinese characters. Another problem is that BERT tends to find senses that are related in topic, which often exist in similar contexts but are not synonymous. For instance, for the third character sense, “愕” (amazed), the BERT model finds a similar sense, “惶” (confused), which is also a kind of psychological state but is not synonymous with “愕”. By further learning the gloss and contextual differences between different senses, the trained model can capture the features not only about contexts but also about their meaning. So, it achieved better results in this experiment.

**Table 3.** Four cases of Top-5 similar senses given by two models.

| Input Senses and Contexts  | Models                | Top-5 Similar Senses   |
|--|-----------------------|--|
| character: 师<br>sense: n. army<br>context:<br>王师北定中原日，家祭无忘告乃翁。<br>(When the army recaptured the lost land in the Central Plains, don't forget to tell me when you hold a family sacrifice) | BERT-ancient          | 1. 师s5 (the organizational unit of the military); 2. 迹s5 (achievements, deeds); 3. 犹s14 (plan); 4. 国s1 (nation; country); 5. 还s1 (return; go back) |
|  | BERT-ancient- trained | 1. 师s6 (army); 2. 军s1 (army); 3. 贼s6 (People who cause serious harm to the country and society); 4. 国s1 (nation; country); 5. 王s1 (king)           |
| character: 行<br>sense: v. walk<br>context:<br>独行踽踽。岂无他人? (Walking alone. Is there anyone else?)  | BERT-ancient          | 1. 行s32 (righteously); 2. 行s4 (walk); 3. 行s1 (road; path); 4. 行s46 (ancient military system); 5. 行s52 (position in the family hierarchy)           |
|  | BERT-ancient- trained | 1. 行s4 (walk); 2. 徒s1 (walk); 3. 步s1 (walk); 4. 刳s3 (an ancient capacity unit); 5. 蹈s5 (walk)  |
| character: 愕<br>sense: adj. amazed<br>context:<br>天地事物之变，可喜可愕，一寓于书。<br>(The changes of things in the world are both delightful and amazed)   | BERT-ancient          | 1. 愕s1 (amazed); 2. 愕s2 (speak bluntly); 3. 鄂s7 (amazed); 4. 讶s2 (amazed); 5. 惶s2 (confused)   |
|  | BERT-ancient- trained | 1. 愕s1 (amazed); 2. 惊s4 (amazed); 3. 异s9 (amazed); 4. 诧s3 (amazed); 5. 怪s3 (amazed)  |

### 5.2. Results of Sense Identification

In the experiment, given the test context and the target character (word) contained in the context, the model chooses the correct sense for the character from the candidate sense set.

We randomly select 12,000 sentences and target character (word) from the dictionary that do not exist in the training dataset to construct a test dataset for sense identification. We divide the test set into two cases: the target word is a character or a compound word because the number of the candidate sense of compound words (1.56 for average) is significantly less than that of characters (6.12 for average) in the dictionary. We also divide the test set according to the number of sample contexts in the training set into three cases: the target sense containing only one training sample in the training set, containing multiple training samples (>7), and in between, because as usual, the more context samples for training, the better effect will be acquired. Therefore, we divide 12,000 test context samples into six groups of 2000 samples each: character with multiple samples (C-multi); character with less samples (C-less); character with one sample (C-one); compound word with multiple samples (W-multi); compound word with less samples (W-less), and compound word with one sample (W-one).

The experiment is based on two pre-trained BERT models: bert-base-Chinese and bert-ancient-Chinese. The baselines are the long short-term memory (LSTM) model and the original BERT models (BERT-base and BERT-ancient). For the original BERT models, for each sense in the dataset, we use the average embedding of tokens from all example contexts in the training set as the representation of it and choose the sense that has the highest score of cosine similarity with the target word. And for the supervised model, we

follow the method introduced in Section 4.3. We use accuracy to evaluate this task, and the results are shown in Table 4.

**Table 4.** Results of sense identification in test dataset (BERT-base is bert-base-Chinese; BERT-ancient is bert-ancient-Chinese).

| Models                      | C-Multi       | C-Less        | C-One         | W-Multi       | W-Less        | W-One         |
|-----------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
| LSTM                        | 43.45%        | 37.25%        | 16.16%        | 60.26%        | 56.40%        | 35.85%        |
| BERT-base                   | 52.68%        | 45.15%        | 23.61%        | 69.90%        | 69.20%        | 43.65%        |
| BERT-ancient                | 55.08%        | 46.10%        | 25.46%        | 73.15%        | 70.70%        | 42.30%        |
| BERT-base-trained (ours)    | 68.08%        | 63.25%        | 58.18%        | 76.15%        | 76.55%        | 65.65%        |
| BERT-ancient-trained (ours) | <b>74.19%</b> | <b>66.55%</b> | <b>60.03%</b> | <b>78.70%</b> | <b>78.10%</b> | <b>67.10%</b> |

The results show that the effects of unsupervised methods are very limited, especially when there is only one context sample in the training dataset, only 23.61% for BERT-base in C-one. By introducing information from the training dataset and fine-tuning the model, the effect of sense identification can be significantly improved both for character and compound words in all of the groups. However, there is still room for improvement; the effects of character groups are significantly worse than the compound word groups due to the excessive candidate senses and less context samples in the training set.

### 5.3. Results of Morpheme Sense Mining

For each character sense to be evaluated, we use accuracy, recall, and f1 as indicators to determine whether the senses of compound word matched to it are correct and comprehensive. The gold standard is annotated by three Chinese native speakers, and the process is as follows:

We randomly select 200 sense samples of commonly used characters in both ancient and modern times. Given the information, including the character sense to be evaluated, corresponding compound words, and their senses, three annotators were asked to provide all compound word senses that should be matched to the character sense, which means that these compound words are composed by the target character (morpheme) and their meanings are related to the morpheme sense. If more than two annotators thought that one candidate sense of the compound word should be matched, we would add it to the matching list of the target character sense. Finally, we obtained the matching lists of 200 sense samples and compared them with the lists given by BERT-ancient and Bert-ancient-trained.

The results are shown in Table 5. It can be seen that the F1 value of our BERT-ancient-trained model is 75.14%, and the original BERT-ancient model is only 59.21%, which proves that introducing information from our dataset to enhance the ability of contextual sense representation of the model can improve the effect of the morpheme sense mining task.

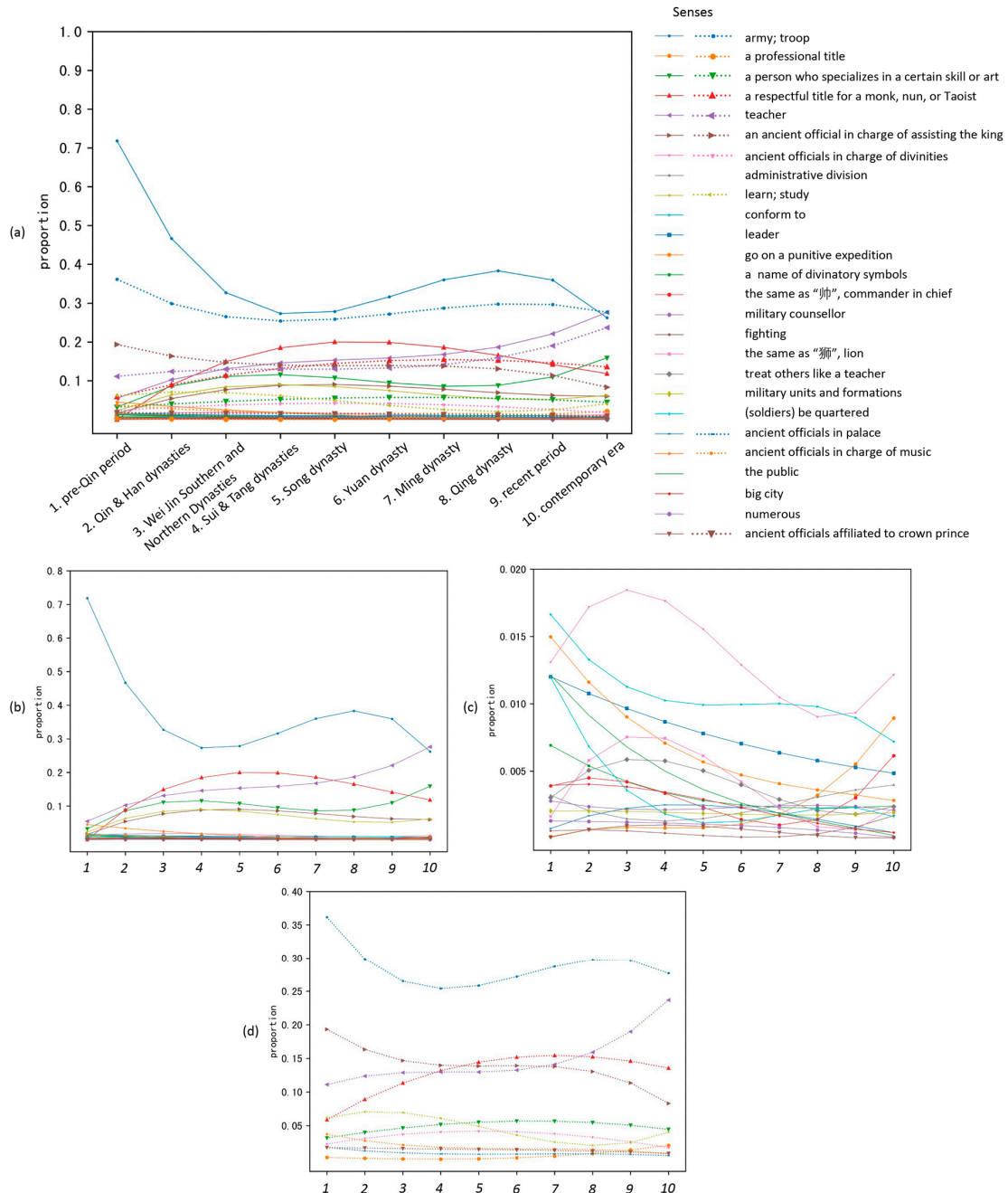
**Table 5.** Results of morpheme sense matching for 200 senses of Chinese characters.

| Models                      | Precision     | Recall        | F1            |
|-----------------------------|---------------|---------------|---------------|
| BERT-ancient                | 67.71%        | 61.23%        | 59.21%        |
| BERT-ancient-trained (ours) | <b>77.25%</b> | <b>79.12%</b> | <b>75.14%</b> |

### 5.4. Visualization Case of Semantic Tracking

Figure 6 shows a visualization case of semantic tracking of the character “师”. The figure shows the change in the frequency distribution of each sense of the character in a continuous and fine-grained way. Different colors and shapes represent different senses. The solid lines indicate the senses that are applied to monosyllabic words, and the dotted line is the senses that are applied to compound words as morphemes (morpheme sense). The method discovered 11 morpheme senses in this case. From Figure 6, it can be concluded that the sense “army; troop” had the absolute highest proportion in the earliest ancient period, indicating that “army; troop” is the original meaning of this character. Although it was in decline continuously in later ages, it is still a commonly used meaning now. These

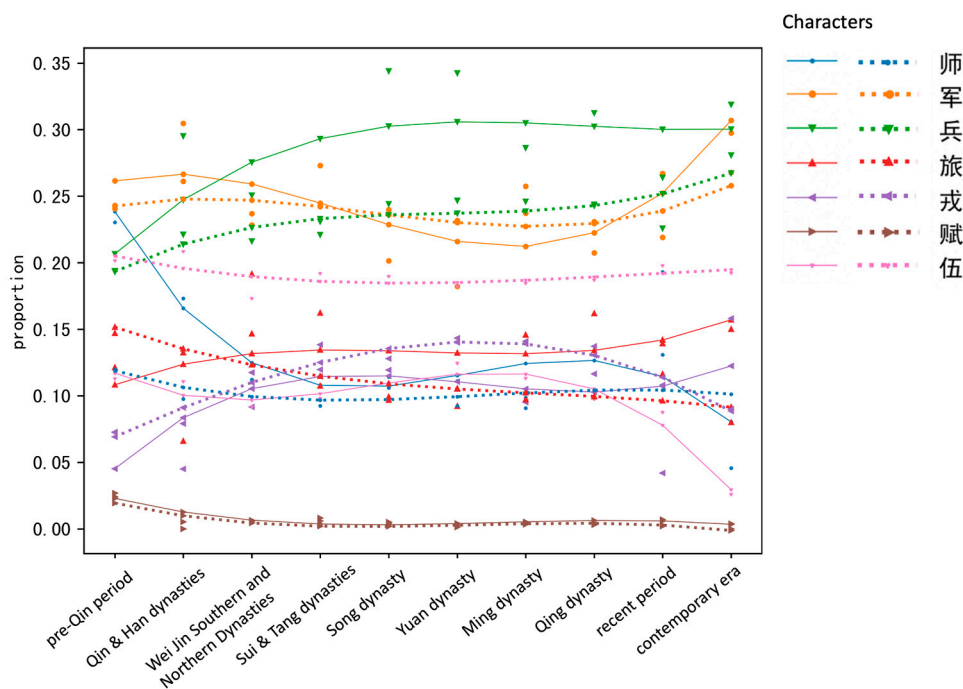
senses, such as “teacher” and “a person who specializes in a certain art”, have gradually increased, especially the sense of “teacher” has replaced “army” as the most commonly used sense in modern times. The sense “a respectful title for a monk, nun, or Taoist” had increased in the middle ancient periods, but there had been a downward trend in recent periods. In comparison, the majority of other senses occupy very small proportions (<2%).



**Figure 6.** The visualization of semantic tracking for the Chinese character “师”, solid lines indicate the sense in monosyllabic words, and dotted lines indicate the sense in morphemes of compound words. Figure (a) is the semantic changes of all of the character senses and morpheme senses; (b) is for all character senses; (c) is for character senses whose proportion < 0.02 in all the time stages; (d) is for all of the morpheme senses.

We also analyze the changes in the wording of a concept. By processing all synonymous characters that can describe the concept, we can draw the frequency distribution of

their corresponding senses in different eras. Figure 7 is an example of the concept “army; troop”, which can be lexicalized by the characters such as “师”, “军”, “兵”, “旅”, “戎”, “赋”, and “伍”. Figure 7 shows that the character “师” was widely used in the earliest ancient times but decreased rapidly; the two most commonly used characters are “军” and “兵” in all historical periods and they present growth trends. Specially, we observe the inconsistencies in the frequency between the usage of monosyllabic words and morphemes for the character “伍”, which is frequently used in a morpheme to compose compound words; however, it is rarely directly used in a monosyllabic word.



**Figure 7.** Visualization of concept tracking for “army”. The “师”, “军”, “兵”, “旅”, “戎”, “赋” and “伍” are seven Chinese characters that have the sense of “army”.

### 6. Analysis

The sense-tracking framework implemented in this work can be used to analyze the law of language development, for example, to discover the semantic change of lexical meanings in various historical periods. For the first time, we give an analyzed case from the view of the relationship between the two kinds of expressions of Chinese characters: monosyllabic words and morphemes (in compound words). We focus on the following three issues:

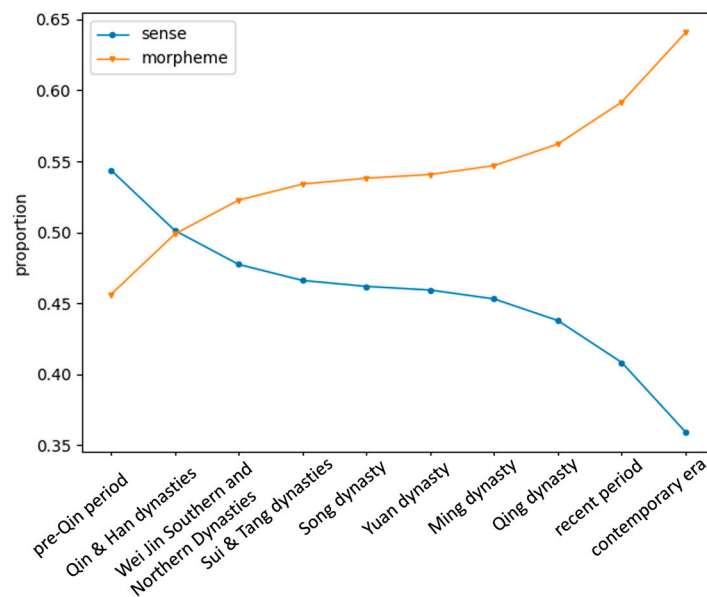
1. The diachronic proportion change of the two kinds of expressions of Chinese characters;
2. The correlation of the frequency between the two kinds of expressions. For instance, if one character sense is always applied to the monosyllabic word, is it frequently or rarely applied to the morpheme?
3. The correlation of the trend between the two kinds of expressions. For instance, if the frequency of one character sense increases on a monosyllabic word, does it have the same or inverse trend on the morpheme?

We selected 100 commonly used Chinese characters in both ancient and modern times as samples for analysis. The average number of senses for each character is 22.19, but the number of commonly used senses is only 3.57 on average. The senses whose proportion is greater than 0.1 during at least one historical period are defined as the commonly used senses. The analysis and results for the three issues are, respectively, shown in Sections 6.1–6.3.



### 6.1. Diachronic Proportion Change between Monosyllabic Word and Morpheme

We calculated the number of contexts in which the target character is applied as a monosyllable and compound word, respectively, in various historical periods and obtained the diachronic proportion change after averaging for all character samples. As shown in Figure 8, the proportion of morphemes (in compound words) continues to increase significantly, while monosyllabic words decrease significantly. The result is in line with the intuition of the development process of Chinese: as human cognitive concepts increased, the infinite development of the number of Chinese characters (monosyllabic words) would greatly increase the memory burden. Therefore, the existing characters were morphemeized, and compound words were formed to express the new concepts. The proportion of compound words especially showed a rapid upward trend during the historical periods with dramatic cognitive changes, such as the recent modern historical period.



**Figure 8.** Diachronic proportion change between monosyllabic words and morpheme expressions.

### 6.2. Correlation of the Using Frequency between Monosyllabic Words and Morphemes

In this section, we will further analyze the correlation of the frequency between the two kinds of sense expressions. Given the two diachronic frequency distribution functions of each sense, which are, respectively, applied to the monosyllabic word and morpheme, we quantified the comprehensive using frequency of the sense for each function and calculated the frequency correlation between the two kinds of expressions in all sense samples.

We used the area enclosed by the function and x-axis to quantify the frequency score of a sense:

$$P(X) = \int f(x)dx \tag{7}$$

The higher the using frequency of the sense, the larger the value of  $P(X)$ ; conversely,  $P(X)$  will tend to be close to 0. We calculated the scores of 1279 senses of 100 common characters in the two kinds of expressions, respectively, and obtained the frequency score sequences in the expression of monosyllabic words:  $Seq_w(P_w(X_1), P_w(X_2), \dots, P_w(X_{1279}))$  and morphemes:  $Seq_m(P_m(X_1), P_m(X_2), \dots, P_m(X_{1279}))$ . Finally, we obtained the Spearman correlation between the two sequences, the result of which is shown in Table 6.

**Table 6.** Spearman correlation of frequency between monosyllabic word and morpheme expressions in samples of 1279 senses.

| Spearman | <i>p</i> |
|----------|----------|
| 0.751    | <0.01    |

The results show that the frequency between monosyllabic words and morpheme expressions has a strong positive correlation. On this basis, to analyze the distribution of the using frequency similarities of 1279 sense pairs of  $(P_w(X_i), P_m(X_i))$ , we further quantified the frequency similarity between the two diachronic frequency distribution functions:

$$fre\_sim(X_1, X_2) = 1 - \frac{|P(X_1) - P(X_2)|}{P(X_1) + P(X_2)} \tag{8}$$

If the using frequency of the two functions is closer, the score is closer to 1; otherwise, it approaches 0. We conducted statistics on the score of 1279 combinations of  $(P_w(X_i), P_m(X_i))$ . For comparison, we also calculated the scores between 1279 random combinations of  $(P_w(X_i), P_m(X_k))$   $i \neq k$ , which means that the two functions are from different senses. The results are shown in Table 7. The median is 0.628, while for the randomized combinations from different senses is only 0.354. We used the Wilcoxon signed-rank test and proved that the scores of matching combinations are significantly higher than randomized ones.

**Table 7.** Distribution of the frequency similarities of 1279 matching and random function combinations.

| Groups                    | Median (P25, P75)    | Wilcoxon Signed-Rank Test |          |
|---------------------------|----------------------|---------------------------|----------|
|                           |                      | <i>z</i>                  | <i>p</i> |
| <i>fre_sim</i> (matching) | 0.628 (0.340, 0.851) | 17.185                    | <0.01    |
| <i>fre_sim</i> (random)   | 0.354 (0.145, 0.657) |                           |          |

Based on the above analysis, we think that the frequency of a sense between monosyllabic words and morpheme expressions has a strong positive correlation; that is, in general, if one character sense is always applied to the monosyllabic word, it is also frequently applied to the morpheme. However, 280 senses ( $sim < 0.35$ ) that do not conform to this rule are observed in 1279 samples, most of which (261) are only applied to morphemes and rarely used to monosyllabic words.

### 6.3. Correlation of the Trend between Monosyllabic Word and Morpheme

In this section, we will analyze the correlation of the trend between the two kinds of sense expressions. Given the two diachronic frequency distribution functions of each sense that are, respectively, applied to the monosyllabic word and morpheme, we interpolated with 0.1 as the interval, and concatenated the values of the two kinds of functions, respectively, of all senses and finally obtained two discrete sequences  $Seq_w$  and  $Seq_m$ :  $Seq_w = Seq_{w_1} \oplus Seq_{w_2}, \dots, \oplus Seq_{w_{1279}}$ ;  $Seq_m = Seq_{m_1} \oplus Seq_{m_2}, \dots, \oplus Seq_{m_{1279}}$ , where  $Seq_{w_i}$  and  $Seq_{m_i}$  are the sequences of the interpolation values of the functions of monosyllabic word and morpheme, respectively, from the  $i$ -th sense sample.

We conducted a Spearman correlation on  $Seq_w$  and  $Seq_m$  to obtain the correlation between the overall diachronic trends of the two kinds of expressions. The result is shown in Table 8.

**Table 8.** Spearman correlation of trend between monosyllabic word and morpheme expressions in samples of 1279 senses.

| Spearman | <i>p</i> |
|----------|----------|
| 0.731    | <0.01    |

The correlation score is 0.731 and  $p < 0.01$ , which shows that the changing trend of sense between monosyllabic words and morphemes has a strong positive correlation. In order to give the distribution of the trend similarity scores of 1279 pairs of  $(Seq_{w_i}, Seq_{m_i})$ , we directly quantified the trend similarity between the two diachronic distribution functions as follows:

$$r_{sim}(X_1, X_2) = \text{Spearman}(Seq_1, Seq_2) \quad (9)$$

We calculated the Spearman correlation scores between  $Seq_{w_i}$  and  $Seq_{m_i}$  for all 1279 senses. For comparison, we also gave the scores between 1279 random combinations of  $(Seq_{w_i}, Seq_{m_k})$   $i \neq k$ , the two sequences of which are from different senses. The results are shown in Table 9.

**Table 9.** Distribution of trend similarities of 1279 matching and random function combinations.

| Groups               | Median (P25, P75)      | Wilcoxon Signed-Rank Test |          |
|----------------------|------------------------|---------------------------|----------|
|                      |                        | <i>z</i>                  | <i>p</i> |
| $tr\_sim$ (matching) | 0.555 (0.051, 0.862)   | 17.308                    | <0.01    |
| $tr\_sim$ (random)   | −0.082 (−0.587, 0.510) |                           |          |

The result shows that the median of the correlation is 0.555, while for the randomized combinations from different senses is −0.082. Wilcoxon signed-rank test shows that the scores of matching pairs are significantly higher than those of randomized pairs.

Based on the above analysis, we think that there is a strong positive correlation of the diachronic trend between the two expressions of one sense. That is, if the frequency of a sense increases or decreases in a certain era when expressed as a monosyllabic word, it often has the same change trend when expressed as a morpheme. While there are 299 senses ( $sim \leq 0$ ) that do not conform to this rule, they are observed in 1279 samples.

## 7. Discussion

### 7.1. Theoretical and Practical Implications

At the theoretical level, this work provides a framework, method, and dataset for lexical semantic tracking of Chinese over a long historical period. According to our investigation, this is the first work that focuses on Chinese characters and morphemes in this field. Our framework inherits the existing basic process of semantic tracking [3,16], which is to identify word senses on a large-scale corpus and generate the diachronic frequency distribution for each sense. However, different from their work, we do not consider “word” as the smallest unit of semantics but rather consider the relationship between morphemes, characters, and words in Chinese for the first time in diachronic semantic modeling and add the morpheme sense mining process. It enables our approach to provide a richer representation of diachronic changes at these three levels, which is crucial for analyzing the evolution of Chinese over a long historical period and can provide theoretical guidance for the diachronic semantic modeling of Chinese. We also give a diachronic analysis case firstly from the perspective of monosyllabic words and morphemes that can provide a reference for future works of automatically exploring the laws of language evolution. At the method level, we do not directly use the pre-trained model as the work [3] but introduce the information of sense definitions to further train a contextual sense representation model in the Word-sense disambiguation task. We proved that by capturing the features of definitions, the model can enhance the contextual sense representation and achieve better

performance in both sense identification and morpheme sense mining tasks. Our model and the open dataset can both serve Chinese NLP research.

At the practical level, most of the existing works discover what words or senses changed based on models of clusters, topics, or networks. However, the conclusions are coarse-grained and fuzzy, which require extensive human explanations. Different from their approaches, the advantage of our framework is interpretability for specific senses, which can provide insights into the nature of semantic change in the morpheme, character, and word levels by identifying specific senses that appear or disappear in texts over time. Therefore, it is suitable for generating fine-grained analysis for experts and amateurs in the fields of historical linguistics, history, and so on. This work can be applied to the Chinese diachronic semantic analysis and information retrieval system of historical literature resources. For instance, generating visual images of diachronic semantic changes, recommending contexts of historical documents for each sense and time range, and automatically discovering statistical laws of language evolution, etc.

### 7.2. Limitations

Limitations of the dataset: One problem is that the number of context samples per sense in the dataset is small for supervised learning. In the sense-context dataset, there are only 2.08 context samples belonging to each sense on average, and only 3.61 belong to the sense of characters. For commonly used senses with more than seven context samples, the accuracy of identification can reach 74.19%, and for these senses with only one sample, the accuracy is only 60.03%. Another problem is that the granularity of senses is fine. As a result, there are often quite a lot of senses for one character, and several of them are similar and overlap in their meanings, for instance: “green bamboo” and “younger bamboo”; “woman” and “elderly women”. It affects the performance of sense identification and morpheme sense mining. For sense identification, many wrong senses predicted by the model are highly similar to the correct ones, which is why there is no need to differentiate; among the 1000 incorrect samples surveyed, we observed 245 that belong to this problem.

Limitation of method: The limitations of the method in this work include the following: (1) This method cannot discover senses that are not recorded in the dictionary: to ensure accuracy and interpretability, our model is implemented based on dictionary information and supervised learning. However, some senses from ancient times may not have been verified and included in the dictionary, and new senses have existed in recent years. (2) This method cannot further judge the more complex semantic relationships between senses, such as widening, narrowing, and semantic extension, which are also important for the study of the evolution of diachronic lexical meanings. (3) For sense identification, we observe that the model tends to classify the tokens to the candidate senses with more context samples in the training set, which is an inevitable problem of supervised learning. It will further lead to the deviation of semantic tracking results: the predicted proportion of the senses with more samples in the dataset will be higher than the actual situation, while the senses with less samples will be the opposite.

In the future, we will expand and improve the existing framework and methods, including (1) expanding the dataset and further improving the effect of contextual sense representation, sense identification, and morpheme sense mining tasks; (2) developing a method to automatically discover new senses that are not recorded in the dictionary; (3) further introduce the method to judge the more complex semantic relationships between lexical senses, such as widening, narrowing, and semantic extension, serving the researchers of diachronic evolution of Chinese; (4) the Large Language Model and Generative Pre-trained Transformer develop fast, and we plan to explore methods based on them. The ultimate goal is to enable the machine to answer questions about language changes using natural languages.

## 8. Conclusions

This paper proposed a semantic tracking framework and method for Chinese characters in a long historical period from two views of monosyllabic words and morphemes, including constructing a contextual sense representation model, realizing sense identification and morpheme sense mining based on the model, so as to process the large-scale historical corpus and obtain the diachronic semantic change representation of each sense of character. Another contribution of this work is to construct datasets from authoritative dictionaries and historical corpus, which includes a sense-context dataset for model training and a Chinese historical literature corpus for sense tracking. In the experiment, we evaluated the method through quantitative and qualitative ways to show the performance. We demonstrated that our model introducing information from the dataset shows better performance for sense representation compared to the original BERT model: our model gave more synonymous terms in the Top 5–30 similar sense rankings, the average number of synonyms from the Top-5 ranking was 2.2 while for original BERT was 1.1. And our model also achieves better results in both sense identification and morpheme sense matching tasks. The accuracy of sense identification was improved from 55.08% to 74.19%, and the F1 score for morpheme sense mining was improved from 59.21% to 75.14%. The visualization and the qualitative analysis cases show that the method can carry out a smooth and continuous representation of the frequency distribution of each sense of characters in historical periods. Finally, for the first time, we gave an analyzed case through 100 commonly used characters from the view of the relationship between monosyllabic words and morphemes. As a result, we proved that there is a strong positive correlation between the frequency and change trend of the two kinds of sense expressions of a monosyllabic word and the corresponding morpheme. The Spearman correlation of frequency between monosyllabic words and morpheme expressions in samples of 1279 senses was 0.751,  $p < 0.01$ . This work can serve the fields of historical linguistics, history, dictionary compilation, NLP, and information processing of historical resources, which is of significance in capturing the laws of Chinese language evolution and exploring the process of social and cultural development.

**Author Contributions:** Conceptualization, Y.C., F.G., and H.X.; methodology, Y.C.; software, Y.C.; validation, Y.C.; formal analysis, Y.C.; investigation, Y.C.; resources, H.X.; data curation, Y.C.; writing—original draft preparation, Y.C.; writing—review and editing, F.G. and H.X.; visualization, Y.C.; supervision, Y.C.; project administration, H.X.; funding acquisition, F.G. and H.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Paleography and Chinese Civilization Inheritance and Development Program Collaborative Innovation Platform (No. G3829).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available here: <https://github.com/YangChijLU/diachronic-Chinese-tracking> (accessed on 24 April 2024).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Kutuzov, A.; Øvrelid, L.; Szymanski, T.; Velldal, E. Diachronic word embeddings and semantic shifts: A survey. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1384–1397.
2. Devlin, J.; Chang, M.-W.; Lee, K.; Google, K.T.; Language, A.I. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2 June 2019; pp. 4171–4186.
3. Hu, R.; Li, S.; Liang, S. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 3899–3908.
4. Giulianelli, M.; Del Tredici, M.; Fernández, R. Analysing lexical semantic change with contextualised word representations. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 3960–3973.

5. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. In Proceedings of the 1st International Conference on Learning Representations, Scottsdale, AZ, USA, 2–4 May 2013; pp. 1–12.
6. Kim, Y.; Chiu, Y.I.; Hanaki, K.; Hegde, D.; Petrov, S. Temporal Analysis of Language through Neural Language Models. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; pp. 61–65.
7. Hamilton, W.L.; Leskovec, J.; Jurafsky, D. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 2116–2121.
8. Rosenfeld, A.; Erk, K. Deep neural models of semantic shift. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 474–484.
9. Yin, Z.; Sachidananda, V.; Prabhakar, B. The global anchor method for quantifying linguistic shifts and domain adaptation. In Proceedings of the 32nd Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 9412–9423.
10. Kaiser, J.; Kurtyigit, S.; Kotchourko, S.; Schlechtweg, D. Effects of pre- and post-processing on type-based embeddings in lexical semantic change detection. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Online, 19–23 April 2021; pp. 125–137.
11. Qiu, L.; Tu, K.; Yu, Y. Context-dependent sense embedding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 183–191.
12. Lee, G.H.; Chen, Y.N. MUSE: Modularizing unsupervised sense embeddings. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 327–337.
13. Shu, L.; Guo, Y.; Wang, H.; Zhang, X.; Hu, R. The Construction and Application of Ancient Chinese Corpus with Word Sense Annotation. *J. Chin. Inf. Process.* **2022**, *36*, 21–30.
14. Kurtyigit, S.; Park, M.; Schlechtweg, D. Lexical Semantic Change Discovery. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 6985–6998.
15. Laicher, S.; Kurtyigit, S.; Schlechtweg, D.; Kuhn, J.; Walde, S.S. Explaining and improving BERT performance on lexical semantic change detection. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Online, 19–23 April 2021; pp. 192–202.
16. Teodorescu, D.; vonder Ohe, S.; Kondrak, G. UAlberta at LSCDiscovery: Lexical Semantic Change Detection via Word Sense Disambiguation. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Dublin, Ireland, 26–27 May 2022; pp. 180–186.
17. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451.
18. Rosin, G.D.; Guy, I.; Radinsky, K. Time masking for temporal language models. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event/Tempe, AZ, USA, 21–25 February 2022; pp. 833–841.
19. Cassotti, P.; Siciliani, L.; DeGemmis, M.; Semeraro, G.; Basile, P. XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 1577–1585.
20. Mccarthy, D.; Baldwin, T. Novel word-sense identification. In Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 1624–1635.
21. Mitra, S.; Mitra, R.; Riedl, M.; Biemann, C.; Mukherjee, A.; Goyal, P. That’s sick dude!: Automatic identification of word sense change across different timescales. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; pp. 1020–1029.
22. Tahmasebi, N.; Risse, T. Finding individual word sense changes and their delay in appearance. In Proceedings of the Recent Advances in Natural Language Processing, Varna, Bulgaria, 2–8 September 2017; pp. 741–749.
23. Jana, A.; Mukherjee, A.; Goyal, P. Network measures: A new paradigm towards reliable novel word sense detection. *Inf. Process. Manag.* **2020**, *57*, 102173. [[CrossRef](#)]
24. Montariol, S.; Martinc, M.; Pivovarova, L. Scalable and Interpretable Semantic Change Detection. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 4642–4652.
25. Giulianelli, M.; Luden, I.; Fernandez, R.; Kutuzov, A. Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 3130–3148.
26. Baidubaike. Available online: <https://baike.baidu.com/> (accessed on 15 March 2024).
27. Daizhige. Available online: <https://github.com/garychowcmu/daizhigev20> (accessed on 15 March 2024).
28. Bert-Ancient-Chinese. Available online: <https://huggingface.co/Jihuai/bert-ancient-chinese> (accessed on 15 March 2024).

- 
29. Bert-Base-Chinese. Available online: <https://huggingface.co/bert-base-chinese> (accessed on 15 March 2024).
  30. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Virtual Event, 7–11 November 2021; pp. 6894–6910.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.