Zuo Zhuan Ancient Chinese Dataset for Word Sense Disambiguation

Xiaomeng Pan, Hongfei Wang, Teruaki Oka, Mamoru Komachi
Tokyo Metropolitan University
pan-xiaomeng@ed.tmu.ac.jp, wang-hongfei@ed.tmu.ac.jp
teruaki-oka@tmu.ac.jp, komachi@tmu.ac.jp

Abstract

Word Sense Disambiguation (WSD) is a core task in Natural Language Processing (NLP). Ancient Chinese has rarely been used in WSD tasks, however, as no public dataset for ancient Chinese WSD tasks exists. Creation of an ancient Chinese dataset is considered a significant challenge because determining the most appropriate sense in a context is difficult and time-consuming owing to the different usages in ancient and modern Chinese. Actually, no public dataset for ancient Chinese WSD tasks exists. To solve the problem of ancient Chinese WSD, we annotate part of Pre-Qin (221 BC) text *Zuo Zhuan* using a copyright-free dictionary to create a public sense-tagged dataset. Then, we apply a simple Nearest Neighbors (k-NN) method using a pre-trained language model to the dataset. Our code and dataset will be available on GitHub[1].

1 Introduction

Word sense disambiguation (WSD) is a crucial aspect of NLP, which identifies the sense of polysemous words that best fit the current context. Compared to some languages such as English, a character in Chinese, especially in ancient Chinese, usually has multiple and varying meanings, which greatly increases the difficulty of word sense disambiguation. At the present time, Dang et al. (2002); Li et al. (2005); Hou et al. (2020); Zheng et al. (2021) have made certain advances on modern Chinese WSD tasks. Nevertheless, unlike modern Chinese, ancient Chinese has hardly been explored in WSD tasks for lack of a dataset thus far. The main reason is that the smaller number of Chinese characters in the past led to even greater ambiguity in meaning than in modern Chinese. There are also fundamental differences in usage between ancient and modern Chinese. Figure 1 shows a context

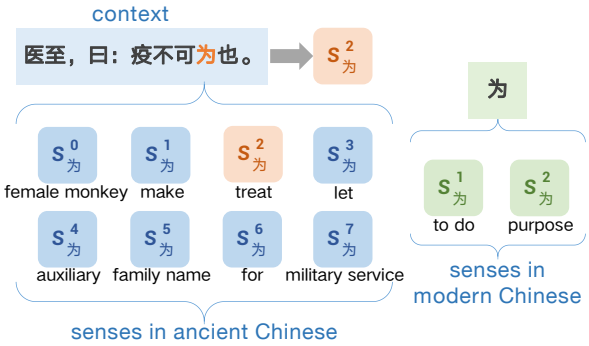


Figure 1: Illustration of choosing the right sense from the given context which means: *The doctor said that the disease could not be treated.* The senses of the target character "为" have different usages in ancient Chinese and modern Chinese. From the eight and two possible senses of the target character "为" in ancient and modern Chinese, the No. 2 sense in ancient Chinese, which denotes *"treat"* best fits the current context.

from *Zuo Zhuan*, a Pre-Qin Chinese book published late in the 4th century BC. The target character "为" has eight senses in ancient Chinese, differing from the two usual senses in modern Chinese. Without WSD, those unfamiliar with ancient Chinese have difficulty determining the correct senses. If WSD can be applied to ancient Chinese, it may contribute to the education of ancient Chinese and also many other tasks such as machine translation for ancient Chinese.

Previous researchers such as Yu et al. (2009); Chang et al. (2013) used few target characters and extracted the contexts to assemble an ancient Chinese lexical sample dataset for their WSD tasks. However, no public dataset for ancient Chinese WSD has yet been established. Consequently, researchers must create their own datasets to test their models for ancient Chinese WSD. Therefore, we choose to self-produce a public dataset for ancient Chinese WSD tasks.

In this study, we selected excerpts from *Zuo Zhuan* that includes approximately 200,000 char-

[1] https://github.com/pxm427/Ancient-Chinese-WSD

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 129 - 135

acters (token). Then we annotated the texts with word senses from an open dictionary *Kangxi* to construct our dataset. In addition, we evaluated a supervised k-NN approach using a pre-trained model (Loureiro and Jorge, 2019) for ancient Chinese WSD tasks.

The main contributions of this paper are as follows:

1. We created a large public ancient Chinese WSD dataset for a lexical sampling task.

2. We applied a supervised k-NN approach using a pre-trained ancient Chinese language model to the ancient Chinese dataset.

2 Related Work

Word sense disambiguation is a task to predict the correct sense using an input word and its context. For example, "bank" has two meanings in English which refers to "a financial institution" and "sloping land". The ambiguity of word can cause noises in downstream tasks. Therefore, it is necessary to uniquely determine the meaning of a word. In Chinese, especially in ancient Chinese, one character usually has multiple and varying meanings. Hence, it adds more difficulties in distinguishing different meanings. Although there is the aforementioned educational aspect, WSD of ancient Chinese can improve machine translation (to modern-Chinese) and full-text search systems.

2.1 Chinese WSD Methods

Modern Chinese. Dang et al. (2002) adopted a maximum entropy method to investigate contextual features for Chinese. Li et al. (2005) used a naïve Bayes model based on local collocation and topical contextual features. Recently, Hou et al. (2020) used an unsupervised method based on HowNet (Dong et al., 2010) and made use of a pre-trained language model. Zheng et al. (2021) proposed FormBERT with word-formation for WSD and created a Chinese lexical sample dataset. All these approaches have performed effectively for Chinese WSD, but their target was modern, not ancient, Chinese.

Ancient Chinese. Yu et al. (2009) applied the CRF (Lafferty et al., 2001) model to tackle ancient Chinese WSD by using contextual words and linguistic features. They tested the model on six target characters with the best average F-score of 83.04% and proved that linguistic features can improve the WSD results for ancient Chinese. Chang et al. (2013) built a knowledge repository of ancient Chinese polysemous words and proposed an unsupervised method for ancient Chinese WSD based on a vector space model. They tested it on ten target characters and obtained an average accuracy of 79.5%. However, both were tested on limited numbers of characters and their datasets were non-public. In our study, we create a public ancient Chinese WSD dataset with 25 target characters, and then apply a k-NN approach using a pre-trained language model to our dataset.

2.2 Resources for Chinese WSD

HowNet is an online common-sense knowledge base including relationships between concepts and attributes with their English equivalents (Dong et al., 2010). It has been used on modern Chinese WSD task (Hou et al., 2020; Zhang et al., 2021), but cannot be applied to ancient Chinese because of the semantic diversity over 2000 years.

Zhang et al. (2012) used *Great Chinese Dictionary* as the knowledge resource and performed WSD of *Zuo Zhuan* by using a semi-supervised machine learning method. Owing to the copyright on the *Great Chinese Dictionary*, the authors have not made the corpus public. Unlike their approach, we used a public dictionary to annotate the word senses and thus can make our corpus publicly available.

Recently, the Pre-Qin Ancient Chinese WordNet (PQAC-WN), which contains 45,498 Pre-Qin basic words and 63,230 semantic classes was constructed by Xu et al. (2020). PQAC-WN organizes information based on semantic relationships and establishes lexical semantic mappings among Pre-Qin ancient Chinese, modern Chinese, and English. Nevertheless, it is not yet public. Therefore, we created an ancient Chinese WSD dataset that can be used freely for research purposes.

3 Construction of the *Zuo Zhuan* Ancient Chinese WSD Dataset

Since there is no public dataset for ancient Chinese WSD, we created the *Zuo Zhuan* Ancient Chinese WSD Dataset for ancient Chinese WSD. In this section, we describe the process of creating the dataset.

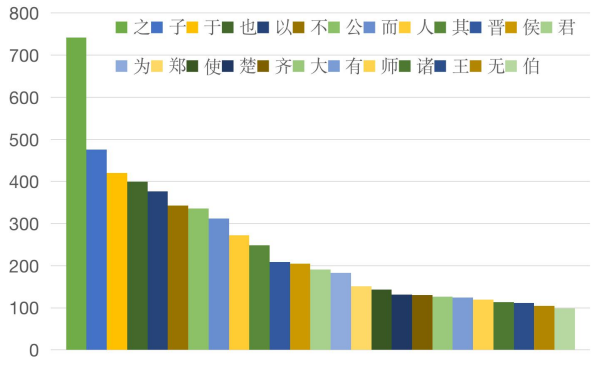


Figure 2: Data statistics. The figure shows the occurrences of 25 high-frequency characters. The vertical axis is the frequency of character, and the horizontal axis is the target characters of this study.

Problems	Examples
No Corresponding Notes	王为中君 → 率领 The king leads his army.
Multiple Similar Notes	晋人许之 → 语助词/他 The Jin promised (him).
Cannot Read	晋人以公为贰于楚 → ?

Figure 3: The three main problems encountered when annotating were termed: "No Corresponding Notes", "Multiple Similar Notes", and "Cannot Read".

3.1 Corpus

3.1.1 Data Selection

We used *Zuo Zhuan* following the previous study (Zhang et al., 2012). As one of the most famous ancient books, *Zuo Zhuan* is free from copyright restrictions, so that we can annotate and make it public. As shown in Figure 2, we selected those with a high-frequency as our target characters (approximately one hundred occurrences for each character) and ranked them from 1 to 25. We selected a total of 2,490 sentences containing the target characters from *Zuo Zhuan*, accounting for 12%. For each target character, we planned to select one hundred sample sentences randomly for annotation. However, the same target character may have appeared several times in the same context. Consequently, there are fewer than one hundred unique sentences for some characters. In such cases, we only chose the first target character to annotate for the context.

3.1.2 Annotation

We discerned the correct meaning of the target character in each context. To be more specific, first, we read every context including the target character and determined all the possible senses. Second, we

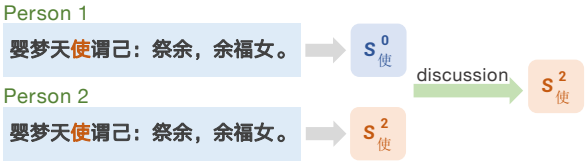


Figure 4: Two researchers annotated the same instance and discussed their readings for accuracy and reliability. Two senses are mentioned in the figure, No. 0 sense: *"make somebody do"* and No. 2 sense: *"Angel"*. This context means: *Zhao Ying dreamed that an angel said to him: Sacrifice me, and I will bless you.*

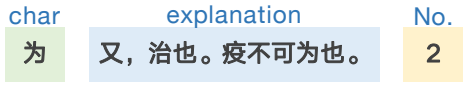
selected the optimal meaning for the target character in the current context.

However, problems may be encountered when annotating. For example, the correct sense could sometimes not be found in the dictionary. As shown in Figure 3, the explanation of the first context is: *The king leads his army.* Here, the correct sense of character "为" is *"lead"* which can not be found in the dictionary. The second context can be translated into: *"The Jin promised."* by choosing the sense which refers to *"auxiliary word"*, or *"The Jin promised him."* by selecting the sense that means *"him"*. It is difficult to determine the most suitable one. In the third context, the correct sense is hard to choose because we could not accurately discern the meaning of the sentence. In such cases, we assigned a special tag *-1* to represent the undetermined sense.

Furthermore, to improve the accuracy and reliability of the annotation, two researchers, native Chinese PhD and master students majoring in NLP, annotated the same target characters separately and discussed them for final confirmation. As shown in Figure 4, occasionally situations arose where different senses were chosen by two researchers for the same instance. This may have been caused by different interpretations of the dictionary and the sentences.

We picked up one character for calculating the inter-annotator agreement of the dataset. For the character "使", the same one hundred sentences have been annotated separately by two researchers with two tags: No.0 and No.1. One researcher annotated 92 sentences with tag No.0 and 8 sentences with tag No.1, and the other researcher annotated 95 sentences with No.0 and 5 sentences with No.1. Using this data, the Cohen's kappa of two independent annotations was 0.75, which indicates moderate agreement (Carletta, 1996).

Fortunately, such consistency problems were



(Treat. For example: The disease can not be treated.)

Figure 5: The structure of sense No. 2 for the "为" character from the dictionary. It consists of three parts: the target character, explanation, and sense number. The translation of the explanation part is shown below.

able to resolve after discussion. The consistency of annotation of the whole data (2,490 sentences) was 88% before discussion, but it finally increased to 100% after discussion and confirmation[2].

3.2 Dictionary

We chose *Kangxi* dictionary[3] compiled in 1716, which contains explanations for almost all the characters of the dynasties before the Qing Dynasty and is free of copyright. As shown in Figure 5, for the character "为", the explanation consists of three parts. The first part is the target character, the second part is the explanation of the particular sense, and the last part is the number of the sense.

4 k-Nearest Neighbors Method using a pre-trained language model for WSD

For this study, we applied the k-Nearest Neighbour classification (k-NN) using a pre-trained language model by following the approach from Loureiro and Jorge (2019). Specifically, we used *GuwenBert*, a pre-trained language model for ancient Chinese, to generate the embedding for WSD.

4.1 *GuwenBert*

GuwenBert-base is a RoBERTa (Liu et al., 2019) model pre-trained on ancient Chinese, which consists of 12 layers with 768 hidden units. The training data is from the daizhige dataset (殆知阁古代文献) that consists of 15,694 books in Classical Chinese, approximately 76% of which are punctuated. The total number of characters is 1.7B (1,743,337,673). All the traditional characters are converted to simplified characters.

It has been proved that *GuwenBert* was more effective than Chinese RoBERTa in Named Entity Recognition (NER) task on ancient Chinese[4], but it has not been used in any WSD tasks on ancient Chinese.

4.2 1-Nearest Neighbor

We applied 1-Nearest Neighbor classification by following the method from Loureiro and Jorge (2019). As shown in Figure 6, we combined sentence embedding E_s with gloss embedding E_g.

$$E = Combination(E_s, E_g) \tag{1}$$

Here, sense embedding E is the combination of sentence embedding and gloss embedding using concatenate or average. We compute the sentence embedding as follows:

$$E_s = \frac{1}{|D^{(t,s)}|} \sum_{c \in D^{(t,s)}} v^{(c,t)} \tag{2}$$

$$v^{(c,t)} = Embed(c)_t \tag{3}$$

where $v^{(c,t)}$ represents the embedding of the target character in the context from the dataset, $D^{(t,s)}$ is the set of contexts where target character t is associated with the sense s in the training data, respectively. Here, c and t are *context* from dataset and *target character*. $Embed(\cdot)_t$ returns the contextualized word embedding of the target character. Likewise, we calculate the gloss embedding as follows:

$$E_g = v^{(g,t)} = Embed(g)_t \tag{4}$$

where g means *gloss*, and $v^{(g,t)}$ represents the embedding of the target character in the gloss.

Finally, the similarity between combined sense embedding E and the target character embedding $v^{(c,t)}$ from test data[5] is calculated. We predicted the sense s as the one with the highest cosine similarity.

$$\hat{s} = \arg\max_{s} sim_{\cos}(v^{(c,t)}, E) \tag{5}$$

5 Experiments

5.1 Experimental Settings

Dataset. We first acquired contexts with same sense number for each sense. Then we split them into training data and test data in an 8:2 ratio. The statistics of the data are shown in Table 1.

[2] As the size of the dataset grows in the future, we are discussing and making a manual together so that the consistency will be as high as possible.

[3] https://www.kangxizidian.com

[4] https://github.com/ethan-yt/guwenbert

[5] When using concatenation to obtain E, the dimension of the combined embedding becomes twice as the sense embedding. Therefore, if we want to calculate the similarity, we need to concatenate the sense embedding $v^{(c,t)}$ itself from test data as well, so that the dimensions of both embeddings are identical.

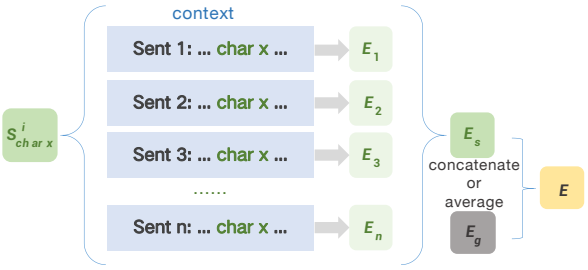


Figure 6: The process of obtaining the combined embedding.

Split	Characters	Sentences
Train	34,971	1,970
Test	9,648	520

Table 1: Statistics of training data and test data.

Baseline. The most frequent sense (MFS) baseline aims to find the sense which occurs most often in the annotated corpus. We selected the sense which appears most frequently in the training corpus for each character and calculated the accuracy.

k-NN As mentioned in Subsection 4.1, we chose *GuwenBert-base* as our model to obtain contextualized character embeddings in 1-NN classification. We only used it for obtaining the embeddings, so that no fine-tuning was required.

5.2 Results & Analysis

Table 2 shows the accuracy of 25 target characters on *Zuo Zhuan* Ancient Chinese WSD Dataset across MFS and 1-NN.

Dataset. As mentioned in 3.1.2, sometimes we cannot assign a definite sense number for a target character in certain contexts when annotating. Such cases account for 12% of the dataset. The cases for "No Corresponding Notes", "Multiple Similar Notes" and "Cannot Read" respectively account for 71%, 17%, 12% of these sentences. It is reasonable to assume that these cases arise mainly from missing explanations in the dictionary, uncertainties of the sentences themselves, and rare ancient usages.

We also find that the discussion improves the reliability of the dataset. The consistency increases from 84% to 100% after discussion and agreement between two researchers. So it is presumed that the dataset gains accuracy and credibility when annotated by more people.

Char	No.	QTY	MFS	Concat	Avg
之	3/8	11	**0.32**	0.23	0.27
子	-1	18	**0.41**	0.18	**0.41**
于	2	8	**1.00**	**1.00**	**1.00**
也	0	5	**1.00**	**1.00**	**1.00**
以	-1	6	**0.62**	0.24	0.29
不	0	12	**1.00**	**1.00**	**1.00**
公	3	17	**0.81**	0.00	0.10
而	6	10	**0.67**	0.14	0.24
人	0	8	0.77	0.00	**0.90**
其	0	9	0.68	**1.00**	**1.00**
晋	5	10	**1.00**	**1.00**	**1.00**
侯	0	11	0.95	**1.00**	**1.00**
君	0	18	0.73	0.00	**0.77**
为	1	8	0.52	0.52	**0.57**
郑	0	4	**1.00**	**1.00**	**1.00**
使	0	4	**0.90**	**0.90**	0.86
楚	6	14	**0.95**	0.00	**0.95**
齐	9	26	**0.95**	**0.95**	**0.95**
大	0	17	0.38	0.38	**0.48**
有	1	8	**0.86**	**0.86**	**0.86**
师	2	13	**0.86**	**0.86**	**0.86**
诸	6	22	**0.62**	**0.62**	**0.62**
王	0	16	**0.86**	**0.86**	0.82
无	0	15	**1.00**	0.90	0.90
伯	2	12	0.85	0.85	**0.90**
			0.79	0.62	0.75

Table 2: Accuracy results based on our dataset. **No.** means the sense number of the target character in the dictionary. **MFS** is the frequency of most frequent sense for each character from test data. **QTY** means the quantity of senses for each character in the dictionary. **concat** and **avg** mean the accuracy calculated by concatenate approach and average approach. Best results and median are shown in **bold** and underline. The last row of data is the average of the columns.

MFS baseline & 1-NN The MFS baseline assumes a sense annotated corpus from which the frequencies of individual senses are learned. Although this is a fairly naïve baseline without exploiting any contextual information, it has proven difficult to beat.

As shown in Table 2, the characters with low MFS accuracy also tend to be low in 1-NN. This may be related to the occurrence of the most frequently annotated senses. For example, the most frequently annotated sense of "于" appears in every context with an accuracy of 1. Therefore, it is more likely to have a higher 1-NN accuracy. In contrast,

"之" with an MFS accuracy of 0.32 can also be inferred to have a low 1-NN accuracy. Furthermore, we also observe that the sense distribution of the character with lower accuracy is more even. For example, in Table 2, "之" has the two most frequent senses with low accuracy.

The size and diversity of the dataset also affect the study. Since our dataset is relatively small, the distribution of senses is limited, and a larger and more comprehensive dataset would considerably improve the accuracy of the 1-NN model that can take advantage of contextualized word embeddings.

Combination strategy. Compared with the concatenate approach, the accuracy of the average approach is generally increased by about 13 points. The reason why the average method outperforms the concatenate method is likely because when using the concatenate approach, it is biased toward the training corpus since we copied the sense embedding from the test data, resulting in a smaller role for the dictionary. Conversely, the average method is more capable of combining the role of the training corpus and the dictionary. Table 2 shows that the accuracy is generally high when the known senses of characters appear in the sentence. In contrast, the appearance of unknown senses (a special tag *-1*) that do not exist in the dictionary cannot be predicted, consequently, resulting in a low accuracy.

Hard characters. It can be observed that the accuracy for the target characters which have unseen senses such as "以" is low in Table 2. The performance for the target characters with diverse senses such as "之" and "大" is also not high. Additionally, characters such as "公" and "而" are considered hard compared to the MFS. We leave improving the performance of these characters for future work.

6 Conclusion and Future Work

In this paper, we created the *Zuo Zhuan* Ancient Chinese WSD Dataset, and then evaluated a 1-NN approach using a pre-trained model *GuwenBert* on our dataset.

In future, we plan to increase the coverage of our dataset, explore whether this approach can detect unknown senses and improve the performance by adapting the pre-trained model to our dataset.

In addition, ancient Chinese and modern Chinese have changed greatly in word meanings and vocabulary. Among these, we would like to make a comparison of the two models for ancient Chinese and modern Chinese to address following questions: "How well do models optimized for modern language model perform in our dataset?" and "How well does the model for our ancient Chinese perform in the modern Chinese dataset?"

References

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254.

E Chang, Changxiu Zhang, Hanqing Hou, and Fuping Hui. 2013. Automatic word sense disambiguation of ancient Chinese based on vector space model. *Library and Information Service*, 057(002):114–118.

Hoa Trang Dang, Ching-yi Chia, Martha Palmer, and Fu-Dong Chiou. 2002. Simple features for Chinese word sense disambiguation. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, Taipei, Taiwan. International Committee on Computational Linguistics.

Zhendong Dong, Qiang Dong, and Changling Hao. 2010. HowNet and its computation of meaning. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 53–56. International Committee on Computational Linguistics.

Bairu Hou, Fanchao Qi, Yuan Zang, Xurui Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Try to substitute: An unsupervised Chinese word sense disambiguation method based on HowNet. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1752–1757, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John Lafferty, Andrew Mccallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, Williamstown, MA, USA.

Wanyin Li, Qin Lu, and Wenjie Li. 2005. Integrating collocation features in Chinese word sense disambiguation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 87–94, Jeju Island, Korea. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Huidan Xu, Siyu Chen, Jingjing Cai, Lin Cao, and Bin Li. 2020. The construction and statistical analysis of Pre-Qin ancient Chinese WordNet. *International Journal of Knowledge and Language Processing*, 11(3):48–61.

Lili Yu, Dexin Ding, Weiguang Qu, Xiaohe Chen, and Hui Li. 2009. Ancient Chinese word sense disambiguation based on CRF. *Microelectronics and Computers*, 26(10):4.

Minghao Zhang, Dongyu Zhang, and Hongfei Lin. 2021. Unsupervised Chinese verb metaphor recognition method based on HowNet. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 258–268, Huhhot, China. Chinese Information Processing Society of China.

Yingjie Zhang, Bin Li, Jiajun Chen, and Xiaohe Chen. 2012. A study in dictionary-based all-word word sense disambiguation for pre-Qin Chinese. *Journal of Chinese Information Processing*, 26(03):65–71.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, and Yang Liu. 2021. Leveraging word-formation knowledge for Chinese word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 918–923, Punta Cana, Dominican Republic. Association for Computational Linguistics.