Improving HowNet-Based Chinese WSD with Translations

Xiang Zhang, Bradley Hauer, Grzegorz Kondrak
Alberta Machine Intelligence Institute, Department of Computing Science
University of Alberta, Edmonton, Canada
{xzhang23,bmhauer,gkondrak}@ualberta.ca

Abstract

Word sense disambiguation (WSD) is the task of identifying the intended sense of a word in context. While prior work on unsupervised WSD has leveraged lexical knowledge bases, such as WordNet and BabelNet, these resources have proven to be less effective for Chinese. Instead, the most widely used lexical knowledge base for Chinese is HowNet. Previous HowNet-based WSD methods have not exploited contextual translation information. In this paper, we present the first HowNet-based WSD system which combines monolingual contextual information from a pretrained neural language model with bilingual information obtained via machine translation and sense translation information from HowNet. The results of our evaluation experiment on a test set from prior work demonstrate that our new method achieves a new state of the art for unsupervised Chinese WSD.

1 Introduction

Word sense disambiguation (WSD) is the task of identifying the intended sense of a word in context. It has been identified as a central task in natural language understanding (Navigli, 2018), and has been applied to downstream tasks such as machine translation (Carpuat and Wu, 2007), text summarization (Kouris et al., 2021), and sentiment classification (Hung and Chen, 2016). While supervised WSD systems consistently achieve the best performance, they require large sense-annotated training corpora, which are often difficult to obtain. To avoid this problem, unsupervised WSD systems instead rely on lexical knowledge bases, such as WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2010).

Motivated by the relatively poor coverage of Chinese by BabelNet, a parallel line of research has adopted HowNet (Dong and Dong, 2003) as both a sense inventory and lexical knowledge base for Chinese WSD. Different from WordNet and BabelNet,

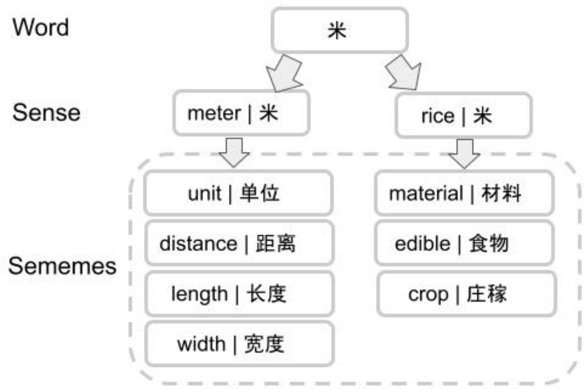

Figure 1: The HowNet representation of two senses of the Chinese word 米. All HowNet senses are associated with a set of sememes, and an English translation.

which define senses by synonymy and describe them using manually-crafted glosses, HowNet associates each sense with a set of sememes, the minimal semantic units in natural language (Qi et al., 2021). In addition, HowNet annotates each sense and sememe with an English translation, as shown in Figure 1. Due to the differences between HowNet and WordNet/BabelNet, particularly their different sense inventories, most WSD systems designed for use with WordNet or a comparable resource cannot be applied to WSD datasets annotated with HowNet senses.

In this paper, we propose a new approach that incorporates both monolingual and multilingual information. For the former, we apply a neural language model pretrained for the masked word prediction task to generate a list of contextually-appropriate substitutions for the target word. For the latter, we exploit the observation that different senses of a word may translate differently (Gale et al., 1992), and leverage a machine translation model to obtain English word translations.

Our experiments show that the proposed method sets a new state of the art for this task on a dataset used in prior work. We observe an improvement of

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4530–4536
December 7-11, 2022 ©2022 Association for Computational Linguistics

2.6% micro-F1 score on the test set, including an improvement of nearly 4% on verbs. We also perform a comparative study which suggests that combining heterogeneous translation methods yields further improvements.

Our contributions are as follows: (1) A novel method for unsupervised Chinese word sense disambiguation with the HowNet sense inventory, which is the first to explicitly leverage translations. (2) An evaluation showing that we set a new state of the art, based on the comparison to four prior methods. (3) A comparative study to evaluate different translation methods, and their combinations.

2 Prior Work

HowNet is one of the most widely used lexical knowledge bases in Chinese NLP research. While BabelNet is a popular resource for multilingual tasks, its coverage of Chinese is relatively poor; indeed, during our experiments, many words covered by HowNet could not be found in BabelNet. Moreover, we found issues with the accuracy of the Chinese information that BabelNet does include. In contrast, HowNet provides much more complete coverage with more than 100,000 Chinese words, each with its own inventory of manually defined senses. In its 20-year history, it has become very well-known in the Chinese NLP community, with frequent updates and a Python API. We have noticed increasing usage of HowNet in recent years for tasks such as sequence modeling (Qin et al., 2019), sememe prediction (Qi et al., 2022), text matching (Lyu et al., 2021), adversarial attacks (Zang et al., 2020), sense embedding (Zhou et al., 2022) and language representation (Liu et al., 2020).

Most HowNet-based WSD systems are unsupervised (Hou et al., 2020) and semi-supervised (Zhou et al., 2020), due to the lack of large corpora annotated with HowNet senses. Yang et al. (2000) perform sense selection based on co-occurrence statistics between the target words and the words in its context. Tang et al. (2015) produce a sense embedding by averaging the embeddings of its sememes, and then compute the cosine similarity between each sense embedding and an embedding of the context of the target word. The method of Ustalov et al. (2018) is designed to use WordNet synset information to create sense embeddings and compute similarity scores between each sense and the context. They adapt this method for use with HowNet by using sememe information to approximate synsets. The current state-of-the-art method of Hou et al. (2020) uses substitution word scores from a masked language model to choose the sense that best fits in the context.

We are not aware of prior methods which leverage translation information from HowNet; however, Luan et al. (2020) exploit translation knowledge from BabelNet. Our HowNet-based WSD task is more challenging, since, unlike BabelNet, HowNet does not include synonyms for each sense, and translations are only available in English, precluding the synergistic use of multiple languages.

3 Methods

The goal of our task is to map a given word x_t (the "target" word) in a given context to the correct sense, as represented in a given sense inventory. Thus, each WSD instance is essentially a multi-class classification problem, with the sense inventory comprising the set of classes. Formally, given a source text $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_t, \ldots, \mathrm{x}_T)$ as input, we try to identify the most suitable sense s_t^i among the N_{x_t} HowNet senses for x_t. We assume that the Chinese input text is tokenized, which is the case in the datasets that we work with. Our approach uses a carefully designed scoring function which combines monolingual contextual information from a pretrained neural language model with bilingual translation information from a machine translation model. In this section, we describe how the language model, HowNet sense representation, and translation knowledge are combined. An overview of our approach is shown in Figure 2.

3.1 Score Function Overview

We define our score function $f\left(\mathrm{s}_t^i; \mathbf{x}, \mathrm{t}\right)$ to measure the fitness of a sense for the target word x_t in context $\mathbf{x}$. It has two parts: (1) A masked language model (MLM) score $f_{LM}\left(\mathrm{s}_t^i; \tilde{\mathbf{x}}\right)$ that leverages HowNet knowledge and a pretrained language model. (2) A translation score $f_{TS}\left(\mathrm{s}_t^i; \mathbf{x}, \mathrm{t}\right)$ that takes advantage of the bilingual information obtained by translating the input sentence from Chinese to English. These score functions are combined as follows:

$$f\left(\mathrm{s}_t^i; \mathbf{x}, \mathrm{t}\right) = f_{LM}\left(\mathrm{s}_t^i; \tilde{\mathbf{x}}\right) \cdot f_{TS}\left(\mathrm{s}_t^i; \mathbf{x}, \mathrm{t}\right)^{\gamma} \quad (1)$$

where γ is a parameter controlling the relative weight of the two scores. The result $f\left(\mathrm{s}_t^i; \mathbf{x}, \mathrm{t}\right)$ is the score for sense s_t^i. We compute the score of

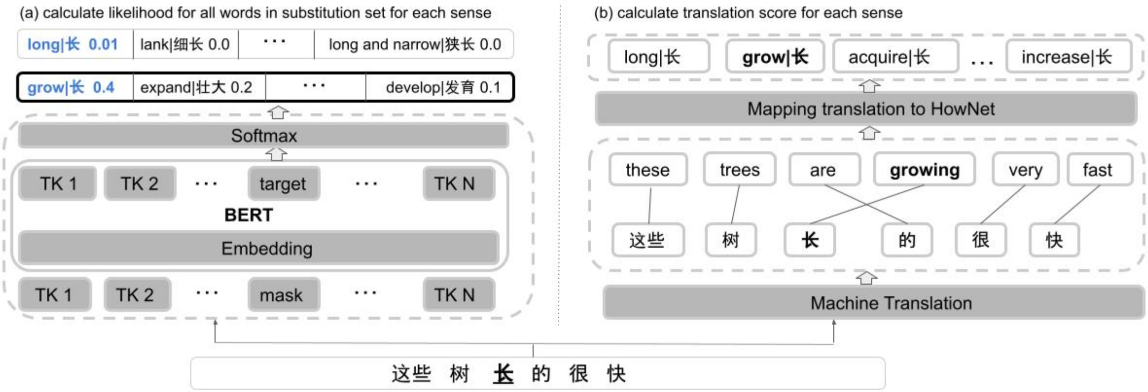

Figure 2: An overview of our combined sense scoring function.

each of the N_{x_t} senses of x_t and return the sense with the highest score.

3.2 Masked Language Modeling Score

We adapt the score function of Hou et al. (2020), which enables transformer-based language models to capture the sense of a word in given context. Specifically, taking an input sentence $\tilde{\mathbf{x}} = (\mathrm{x}_1, \ldots, \mathrm{x}_{t-1}, [\mathrm{MASK}], \mathrm{x}_{t+1}, \ldots, \mathrm{x}_T)$ with the word at target position t masked, the language model can be applied to predict the likelihood of a word w appearing at the masked position, $p\left(\mathrm{w}|\tilde{\mathbf{x}}\right)$.

For each sense s_t^i, we construct a substitution set $\mathcal{S}_t^i$. The substitution set consists of words that have at least one sense with the same HowNet sememes as s_t^i. The words in a substitution set are not necessarily synonyms, but rather words with senses that have similar semantic properties. For example, the substitution set for a sense of the word "funny" could contain "hilarious", "interesting", "ludicrous", etc.

Finally, averaging the language model's estimated likelihood of all words in $\mathcal{S}_t^i$ gives the language model score for sense s_t^i in the given context:

$$f_{LM}\left(\mathrm{s}_t^i; \tilde{\mathbf{x}}\right) = \frac{1}{|\mathcal{S}_t^i|} \sum_{\mathrm{w} \in \mathcal{S}_t^i} p\left(\mathrm{w}|\tilde{\mathbf{x}}\right) \tag{2}$$

This score therefore combines sememe information with contextual information to estimate the fitness of a sense of the target word.

3.3 Contextual Translation Score

HowNet labels every Chinese word sense with an English translation, which may be a single word or a phrase. Since different senses of a word may have different translations, the translation of the target word in its context can be a valuable source of knowledge for WSD. In this section, we describe our method of leveraging lexical translations to define our translation-based scoring function.

For a given input sentence $\mathbf{x}$ in Chinese, we first use machine translation system to obtain its English translation $\mathbf{e} = (\mathrm{e}_1, \ldots, \mathrm{e}_N)$. A word-level alignment is then performed on this bitext to find the matched set of words from $\mathbf{e}$ for the target word x_t, denoted by the set $\mathcal{E}_{ALIGN}^t$. This set may contain one or multiple words, or may be empty, depending on whether x_t is aligned to one, many, or zero words in the translation.

The intuition behind our $f_{TS}\left(\mathrm{s}_t^i; \mathbf{x}, \mathrm{t}\right)$ function is to assign higher score to the senses whose HowNet translations have more overlap with aligned word set $\mathcal{E}_{ALIGN}^t$ obtained by translating and aligning sentence $\mathbf{x}$. For a given sense s_t^i, let $\mathcal{H}_t^i$ be the set of words in its English translation in HowNet. The scoring function is as follows:

$$f_{TS}\left(\mathrm{s}_t^i; \mathbf{x}, \mathrm{t}\right) = 1 + \frac{|\mathcal{H}_t^i \cap \mathcal{E}_{ALIGN}^t|}{|\mathcal{H}_t^i|} \tag{3}$$

If the translation is a perfect match, $\mathcal{H}_t^i = \mathcal{E}_{ALIGN}^t$, and the score function is maximized at 2. If there is no overlap between the words in the HowNet translation of a sense and the words with which the target word is aligned, the intersection has size 0, and a minimum score of 1 is assigned. This avoids completely eliminating senses, yet still gives priority to those that have better agreement in translations.

Since our translation score is independent from the language modelling score, it can be used with any HowNet-based WSD system.

4 Experiments

In this section, we empirically evaluate our method, and compare its results to methods from prior work.

We also include a comparison of various translation models.

4.1 Set Up

Datasets. Since our approach is unsupervised, we do not need a training corpus. We tune our weight parameter on the Senseval3[1] Chinese lexical sample dataset annotated with HowNet senses. It contains 20 different Chinese words with 20 to 100 sense-annotated example sentences each. Senseval3 uses an older version of HowNet than what our method uses; therefore, we first update the annotations and file format in the Senseval3 dataset to correspond to the senses in the more recent version. The annotation involved only a small number of senses, and was done manually by a Chinese speaker. We then evaluate our model on the dataset built by Hou et al. (2020), which contains 2969 instances, representing 36 polysemous target words (17 nouns and 19 verbs), annotated with HowNet senses.

Metrics. Following prior work (Hou et al., 2020), we evaluate our model with micro F1 and macro F1 scores. Micro F1 is the harmonic mean of precision and recall over all instances. Macro F1 is computed by grouping all instances of each word, and computing a precision, recall, and micro F1 for each such group individually; the arithmetic mean of the F1 scores of the groups is then computed to obtain the macro F1. We also report the micro and macro F1 scores on nouns and verbs separately.

Micro F1 is a more reliable metric in the sense that it provides a better estimate of the performance of a WSD system on text corpora, in which the frequencies of words vary greatly. This is because micro F1, being calculated at the level of individual instances rather than words, is influenced by the relative frequencies of words; more frequent words will have more impact on the micro F1. Macro F1, on the other hand, weighs all words equally, regardless of frequency.

Model Configuration. We use Chinese $\mathrm{BERTbase}$ as our language model (Devlin et al., 2019). We adopt the default model settings and tokenizer. We use OpenHowNet (Qi et al., 2019) to acquire sense information, such as sememes and translations, and to build our substitution sets. Lastly, we use the Google Translate API[2] for Chinese-to-English sentence translation, and SimAlign (Jalili Sabet et al., 2020) for word alignment. We use the default parameter settings for Google Translate, and the parameter settings recommended by the authors for SimAlign.

4.2 Results

Table 1 shows the results on the test dataset. We compare our model to the four comparison systems discussed in Section 2. We also include a baseline which chooses a sense of the target word at random. The methods of Yang et al. (2000) and Tang et al. (2015) are unable to outperform the random baseline in some cases. The method of Hou et al. (2020), on the other hand, easily outperforms not only the baseline, but all other previously published systems.

Our approach yields universally better WSD results, across all metrics and subsets of the data, with an overall increase of 2.6% micro-F1 compared to the previous state-of-the-art. The improvement is particularly pronounced on the verb instances, where our method achieves a 3.9% improvement.

4.3 Comparing Translation Models

In order to assess the impact of the Chinese-to-English translation model, we conduct further tests with three commercial MT systems: Google Translate (used in the experiments described above), CaiYun, and YouDao[3]. We also experiment with combining translation systems by averaging the $f_{TS}\left(\mathrm{s}_t^i;\mathbf{x},\mathrm{t}\right)$ scores from each system before computing the final score for each sense. Finally, we implement our own LSTM-based sequence-to-sequence translation model using the FairSeq framework (Ott et al., 2019), and train it on the WMT17 Chinese-English dataset.[4]

Table 2 shows the results. Row 1 does not use translation information, making the method dependent on the language modeling module alone. Rows 2, 3, and 4 contain the results for the individual commercial systems. Rows 5, 6, and 7 show the results with combinations of different translation systems. We observe that combining translation systems yields better performance. We speculate that combining translation information from multiple sources mitigates the impact of translation or alignment errors by providing multiple alternative translations. Finally, in Row 8, our LSTM-based

[1] https://web.eecs.umich.edu/~mihalcea/senseval/senseval3/data.html

[2] https://pypi.org/project/googletrans/

[3] https://pypi.org/project/translators/

[4] https://www.statmt.org/wmt17/translation-task.html

Group	#	Model	Nouns		Verbs		Overall		
			micro-F1	macro-F1	micro-F1	macro-F1	micro-F1	macro-F1	$\Delta F1$
Baseline	1	Random	37.24	34.83	20.54	20.72	26.98	27.38	0.00
Prior Work	2	Yang et al. (2000)†	38.11	26.53	27.29	21.03	31.47	23.63	4.49
	3	Tang et al. (2015) †	47.38	33.13	32.51	27.85	38.25	30.34	11.27
	4	Ustalov et al. (2018) †	52.36	39.06	35.14	33.01	41.79	35.86	14.81
	5	Hou et al. (2020) ‡	53.76	41.71	52.50	48.02	52.98	45.04	26.00
This Paper	6	Ours	**54.19**	**41.97**	**56.40**	**48.08**	**55.55**	**45.19**	**28.57**

Table 1: Results on the test dataset. The highest value in each column is in bold. $\Delta F1$ is the difference of micro-F1 compared to baseline. † represents results quoted from previous paper and ‡ indicates a result we replicated.

#	Translation System	Overall	
		micro-F1	macro-F1
1	w/o translation	52.98	45.04
2	CaiYun	54.57	45.66
3	YouDao	54.77	44.52
4	Google Translate (GT)	55.55	45.19
5	YouDao+CaiYun	55.31	44.53
6	GT+YouDao	**56.02**	45.07
7	GT+Caiyun	55.48	**45.69**
8	LSTM	53.41	44.70

Table 2: Results with different translation systems.

translation system performs comparably to the language model baseline, but not as well as the commercial systems. This is likely due to lower translation accuracy, which may be caused by inaccurate translations of rare words.

5 Conclusion

We have demonstrated that translation information can improve Chinese word sense disambiguation using HowNet. Our application of both monolingual and multilingual information requires translations in only a single language, and does not require information about synonymy or other semantic relations, which are required by many unsupervised WSD methods. Our methods set a new state-of-the-art for unsupervised WSD in Chinese. In the future, we plan to further investigate combining multiple machine translation systems to further improve our results.

Limitations

Our method depends on both HowNet and a machine translation model (plus alignment). Therefore, our method is only applicable to languages and domains for which comparable resources are available. In particular, it would be difficult to apply our method to low-resource languages.

While our results indicate a new state-of-the-art for our task, our test dataset contains only a relatively small lexical sample of target words, each of which is either a noun or a verb. The ability to generalize to all words or all parts of speech is therefore untested, as no other more extensive datasets are available.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhendong Dong and Qiang Dong. 2003. Hownet - a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 820–824.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, volume 112. Citeseer.

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.

Chihli Hung and Shiuan-Jeng Chen. 2016. Word sense disambiguation based sentiment lexicons for sentiment classification. *Knowledge-Based Systems*, 110:224–232.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. 2021. Abstractive text summarization: Enhancing sequence-to-sequence models using word sense disambiguation and semantic content generalization. *Computational Linguistics*, 47(4):813–859.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2901–2908.

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065, Online. Association for Computational Linguistics.

Boer Lyu, Lu Chen, Su Zhu, and Kai Yu. 2021. Let: Linguistic knowledge enhanced graph transformer for chinese short text matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13498–13506.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11).

Roberto Navigli. 2018. Natural language understanding: Instructions for (present and future) use. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5697–5702. International Joint Conferences on Artificial Intelligence Organization.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Fanchao Qi, Chuancheng Lv, Zhiyuan Liu, Xiaojun Meng, Maosong Sun, and Hai-Tao Zheng. 2022. Sememe prediction for BabelNet synsets using multilingual and multimodal information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Fanchao Qi, Ruobing Xie, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. Sememe knowledge computation: a review of recent advances in application and expansion of sememe knowledge bases. *Frontiers of Computer Science*, 15(5):155327.

Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Qiang Dong, Maosong Sun, and Zhendong Dong. 2019. Openhownet: An open sememe-based lexical knowledge base. *arXiv preprint arXiv:1901.09957*.

Yujia Qin, Fanchao Qi, Sicong Ouyang, Zhiyuan Liu, Cheng Yang, Yasheng Wang, Qun Liu, and Maosong Sun. 2019. Enhancing recurrent neural networks with sememes. *CoRR*, abs/1910.08910.

Gongbo Tang, Dong Yu, and Endong Xun. 2015. An unsupervised word sense disambiguation method based on sememe vector in hownet. *Journal of Chinese Information Processing*, 6:23–29.

Dmitry Ustalov, Denis Teslenko, Alexander Panchenko, Mikhail Chernoskutov, Chris Biemann, and Simone Paolo Ponzetto. 2018. An unsupervised word sense disambiguation system for under-resourced languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Erhong Yang, Guoqing Zhang, and Yongkui Zhang. 2000. The research of word sense disambiguation method based on co-occurrence frequency of hownet. In *Second Chinese Language Processing Workshop*, pages 60–65, Hong Kong, China. Association for Computational Linguistics.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics.

Qiaoli Zhou, Gu Yue, and Yuguang Meng. 2020. *Incorporating HowNet-Based Semantic Relatedness Into Chinese Word Sense Disambiguation*, pages 359–370.

Yangxi Zhou, Junping Du, Zhe Xue, Ang Li, and Zeli Guan. 2022. Chinese word sense embedding with sememewsd and synonym set.