

# Try to Substitute: An Unsupervised Chinese Word Sense Disambiguation Method Based on HowNet

Bairu Hou<sup>23\*</sup>, Fanchao Qi<sup>12\*</sup>, Yuan Zang<sup>12\*</sup>, Xurui Zhang<sup>12</sup>, Zhiyuan Liu<sup>12†</sup>, Maosong Sun<sup>12</sup>

<sup>1</sup>Department of Computer Science and Technology, Tsinghua University

<sup>2</sup>Institute for Artificial Intelligence, Tsinghua University

Beijing National Research Center for Information Science and Technology

<sup>3</sup>School of Economics and Management, Tsinghua University

houbr.17@sem.tsinghua.edu.cn

{zangy17, qfc17, zhangxr16}@mails.tsinghua.edu.cn

{liuzy, sms}@tsinghua.edu.cn

## Abstract

Word sense disambiguation (WSD) is a fundamental natural language processing task. Unsupervised knowledge-based WSD only relies on a lexical knowledge base as the sense inventory and has wider practical use than supervised WSD that requires a mass of sense-annotated data. HowNet is the most widely used lexical knowledge base in Chinese WSD. Because of its uniqueness, however, most of existing unsupervised WSD methods cannot work for HowNet-based WSD, and the tailor-made methods have not obtained satisfying results. In this paper, we propose a new unsupervised method for HowNet-based Chinese WSD, which exploits the masked language model task of pre-trained language models. In experiments, considering existing evaluation dataset is small and out-of-date, we build a new and larger HowNet-based WSD dataset. Experimental results demonstrate that our model achieves significantly better performance than all the baseline methods. All the code and data of this paper are available at <https://github.com/thunlp/SememeWSD>.

## 1 Introduction

Word sense disambiguation (WSD) is a long-standing natural language processing task which aims to identify the correct sense of a polysemous word in the context (Navigli, 2009). WSD is fundamental to natural language understanding (Navigli, 2018) and has been proven to be beneficial to many other tasks such as machine translation (Vickrey et al., 2005; Pu et al., 2018), information extraction (Bovi et al., 2015) and information retrieval (Zhong and Ng, 2012).

There are two main kinds of WSD, namely supervised disambiguation and unsupervised knowledge-based disambiguation. Supervised WSD requires large amounts of sense-annotated training corpora that are difficult to obtain (Tripodi and Navigli, 2019). In contrast, unsupervised knowledge-based WSD relies on only an external lexical knowledge base (LKB) as the sense inventory and thus has wider practical use.

Existing unsupervised knowledge-based WSD approaches mainly comprise gloss-based and graph-based methods. The gloss-based methods utilize glosses (sense definitions) to conduct disambiguation. Lesk algorithm (Lesk, 1986) is a seminal gloss-based method which disambiguates a word by selecting the sense whose gloss overlaps most with the context. There are many subsequent methods based on Lesk algorithm (Banerjee and Pedersen, 2003; Basile et al., 2014; Wang et al., 2020). The graph-based methods is the other major type of knowledge-based WSD approaches, which exploit the structures of the LKB for disambiguation (Agirre et al., 2014; Moro et al., 2014; Chaplot et al., 2015; Chaplot and Salakhutdinov, 2018). Besides, a recent method uses both glosses and structural information of LKBs in knowledge-based WSD and achieves state-of-the-art performance (Scarlini et al., 2020).

In Chinese WSD, HowNet (Dong and Dong, 2006) is the most widely used LKB (Wu, 2009). Different from other LKBs, HowNet contains neither glosses nor structures of different senses. Instead, HowNet

---

\*Indicates equal contribution.

†Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

defines a sense by a set of predefined *sememes*, the minimum semantic units in linguistics (Bloomfield, 1926). Therefore, both gloss-based and graph-based methods cannot work in HowNet-based WSD.

To the best of our knowledge, only a few studies focus on unsupervised WSD based on HowNet. Yang et al. (2001) propose a representative statistical method which utilizes the co-occurrence of sememes of the target word and context to conduct disambiguation. Tang et al. (2015) learn sememe and sense embeddings and disambiguates a word by choosing the sense that has the closest embedding similarity with the context. These methods work well but far from perfectly.

In this paper, we propose a new unsupervised HowNet-based WSD model with the help of large pre-trained language models. Previous studies have shown that pre-trained language models such as BERT (Devlin et al., 2019) incorporate much sense information (Reif et al., 2019), which can be utilized in WSD. Their pre-training task of masked language model (MLM) is supposed to predict appropriate words for a specified position in the context. In other words, a word with higher MLM prediction score is more suitable for the given context and should have more similar meaning to the original word. Based on this assumption, we design our lexical substitution-based WSD model. For each sense of the target polysemous word, we can find a set of substitution words that involve a sense annotated with the same sememes as the target sense. We calculate the MLM prediction score for each substitution word, and the average of prediction scores of a sense’s all substitution words can reflect the probability that the target word conveys this sense in the context.

The idea of lexical substitution has been applied to WSD in Yuret (2007). Different from our model, it uses a statistical language model to calculate substitution word scores and more importantly, it is not fully unsupervised and requires some sense-annotated corpora in the WSD procedure. Besides, our model resembles the end-to-end BERT-based lexical substitution model in Zhou et al. (2019). However, it is not aimed at WSD and has different calculation methods of substitution word score from us.

In experiments, considering existing HowNet-based WSD dataset is unavailable and based on an outdated version of HowNet, we build a new and larger HowNet-based WSD dataset for evaluation. Experimental results demonstrate that our model significantly outperforms all the baseline methods and achieves state-of-the-art performance.

## 2 Methodology

In this section, we elaborate on our HowNet-based unsupervised WSD model. Before description of the model, we first give a brief introduction to HowNet.

### 2.1 Introduction to HowNet

HowNet (Dong and Dong, 2006) is the most famous sememe knowledge base. It pre-defines a set of about 2,000 sememes and uses them to annotate senses of more than 100,000 Chinese words and phrases. In recent years, HowNet has been successfully applied to diverse natural language processing tasks such as language modeling (Gu et al., 2018), semantic composition (Qi et al., 2019a), sequence modeling (Qin et al., 2020), textual adversarial attack (Zang et al., 2020) and reverse dictionary (Zhang et al., 2020).

Sememe annotations in HowNet are hierarchical, and the sememes of a sense form a tree, as illustrated in Figure 1. But in this paper, following previous work (Yang et al., 2001; Tang et al., 2015), we ignore the hierarchy of sememe annotations and simply regard sememes as discrete semantic labels.

According to the definition of sememe and the philosophy of HowNet, sememes of a sense can convey its meaning, and two senses annotated with the same sememes are supposed to have the same meaning. Therefore, in our model, we select the substitution words whose one sense has the same sememe annotations as the target sense and use their MLM prediction scores to measure the compatibility of the target sense in the context.

### 2.2 Our WSD Model

Suppose we want to disambiguate a target polysemous word  $x_i$  that has  $N_i$  senses in HowNet given the  $L$ -word sentence  $\mathbf{x} = \{x_1, \dots, x_i, \dots, x_L\}$  as the context, for each of its senses  $s_j^i$  we can find a

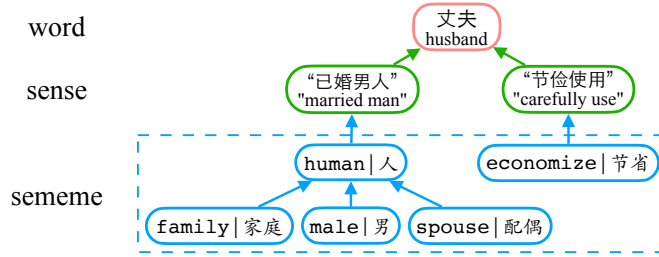


Figure 1: Sememe annotations of the word “husband” in HowNet. It has two senses, namely “married man” (noun) and “carefully use” (verb). The first sense is annotated with 4 sememes while the second is annotated with only 1 sememe.

substitution word set  $\mathbb{W}_j^i$ , each word of which has a sense that is annotated with the same sememes as  $s_j^i$ . Then we can obtain the probability score that  $x_i$  conveys  $s_j^i$  in the context of  $\mathbf{x}$ :

$$Q(s_j^i|\mathbf{x}) = \frac{1}{|\mathbb{W}_j^i|} \sum_{w \in \mathbb{W}_j^i} P(w|\hat{\mathbf{x}}_i), \quad (1)$$

where  $\hat{\mathbf{x}}_i = \{x_1, \dots, x_{i-1}, [\text{MASK}], x_{i+1}, \dots, x_L\}$  is the masked sentence, and  $P(w|\hat{\mathbf{x}}_i)$  is the MLM prediction score for  $w$  calculated by the pre-trained language model.

Finally we select the sense  $s_j^i$ , that has the highest probability score  $Q(s_j^i|\mathbf{x})$  as the WSD result of  $x_i$  in the context of  $\mathbf{x}$ .

To improve performance, we take part-of-speech into consideration. Specifically, given the part-of-speech of the target word  $x_i$  in the context, only its senses with correct part-of-speech are possible WSD results. And we restrict the substitution words to having a sense whose part-of-speech and sememe annotations are both the same as one possible sense of the target word.

In addition, existing Chinese pre-trained language models are based on characters rather than words. Therefore, in the calculation of  $P(w|\hat{\mathbf{x}}_i)$ , we replace the original word  $x_i$  with the same number of [MASK] tokens as the character number of the substitution word. For example, suppose  $x_i$  is a two-character word and one substitution word of its certain sense is three-character, then the original character sequence is  $\mathbf{c} = \{\dots, [x_{i,1}], [x_{i,2}], \dots\}$  and the masked character sequence is  $\hat{\mathbf{c}} = \{\dots, [\text{MASK}], [\text{MASK}], [\text{MASK}], \dots\}$ . The prediction score of a substitution word is the average of the prediction scores of all its characters.

### 3 Experiments

In this section, we evaluate our model on the newly built HowNet-based WSD dataset by us.

#### 3.1 Construction of the HowNet-based WSD Dataset

To the best of our knowledge, the only HowNet-based Chinese WSD dataset<sup>1</sup> is based on an outdated version of HowNet that cannot be found now, which actually makes the dataset unusable. And besides, it is a little small (containing only 1, 173 instances for 20 target polysemous words). Therefore, we build a new and larger HowNet-based Chinese WSD dataset based on the Chinese Word Sense Annotated Corpus used in SemEval-2007 task 5 (Jin et al., 2007), whose sense inventory is Chinese Semantic Dictionary. This corpus comprises 3, 632 word-segmented and part-of-speech tagged instance sentences for 40 Chinese polysemous words (19 nouns and 21 verbs).

We ask Chinese native speakers to manually annotate the target polysemous words in each instance sentence of the corpus with corresponding senses of HowNet or a special option of “no appropriate sense”, where each instance is annotated by 3 annotators. Among the 40 target polysemous words, 4 words have only one sense in HowNet and thus the other 36 target words’ 3, 328 instances are annotated in total. After annotation, we reject 170 instances whose three annotations are all different (namely

<sup>1</sup><https://web.eecs.umich.edu/~mihalcea/senseval/senseval3/tasks.html#ChineseLS>

1:1:1). For the remaining 1,209 instances of 2:1 and 1,949 instances of 3:0, we obtain the final annotation results by voting. Then we discard 189 instances whose final annotation results are “no appropriate sense”. Finally, we obtain the HowNet-based Chinese WSD dataset, which comprises 2,969 instances for 36 target polysemous words (17 nouns and 19 verbs).

### 3.2 Baseline Methods

There are only a few unsupervised HowNet-based WSD methods, and we choose two most representative ones as the baseline methods. Besides, we compare our model with another three baseline methods that are not specially designed for but can be applied to unsupervised HowNet-based WSD.

- SemCo (Yang et al., 2001). This method utilizes the statistics on the co-occurrence of sememes of the target polysemous word and context to conduct WSD.
- SemEmbed (Tang et al., 2015). This method first learns sememe embeddings and further obtains sense embeddings, and then employs the embedding similarity between senses of the target word and the context for disambiguation.
- Dense (Ustalov et al., 2018). This model is originally designed for WordNet-based WSD, which first obtains sense embeddings from the word embeddings of the corresponding senses’ synonyms and then selects the sense that has the closest embedding similarity with the context. In HowNet-based WSD, we regard the words whose one sense has the same sememes as the target sense as the synonyms.
- Random. This baseline method randomly selects a sense of the target word as the WSD result. A common WSD baseline method is choosing the most frequent sense. But HowNet provides no information about the sense frequency. Therefore, we use Random as an alternative.

### 3.3 Experimental Settings

We use OpenHowNet (Qi et al., 2019b) that provides access to the latest version of HowNet to determine the substitution words for target words. There are 78.7 substitution words for each target polysemous word in the above dataset on average, and the average number of substitution words for each sense is 15.6. In our model, we choose the well-established pre-trained model BERT (Devlin et al., 2019), specifically Chinese BERT<sub>BASE</sub><sup>2</sup>, to calculate the MLM prediction score. For the baseline methods, we use their recommended model settings. We choose micro- and macro-F1 scores as the evaluation metrics.

### 3.4 Experimental Results

Table 1 shows the results of our model as well as all the baseline methods on the newly built HowNet-based WSD dataset. From this table, we can observe that our model consistently and significantly outperforms all the baselines (more than 10 points higher than the best baseline), which demonstrates the effectiveness of our model. We also find that all the baselines perform markedly worse on the verbs than nouns, which is presumably because verbs have more senses than nouns in the dataset (average sense numbers are 5.53 vs. 3.35). In contrast, our model achieves as good performance on verbs as on nouns, which manifests superiority of model.

Model	Nouns		Verbs		All	
	micro-F1	macro-F1	micro-F1	macro-F1	micro-F1	macro-F1
SemCo	38.11	26.53	27.29	21.03	31.47	23.63
SemEmbed	47.38	33.13	32.51	27.85	38.25	30.34
Dense	52.36	39.06	35.14	33.01	41.79	35.86
Random	37.24	34.83	20.54	20.72	26.98	27.38
Ours	<b>53.76</b>	<b>41.71</b>	<b>52.50</b>	<b>48.02</b>	<b>52.98</b>	<b>45.04</b>

Table 1: WSD results (%F1 scores) of all the models.

<sup>2</sup><https://github.com/google-research/bert>

In addition, we make a simple error analysis. Our model performs worst when disambiguating two words including a verb “发” (micro-F1: 8.89, macro-F1: 6.35) and a noun “菜” (micro-F1: 6.85, macro-F1: 7.75). We conjecture that “发” is hard to disambiguate because it has too many senses (18 verbal senses). As for “菜”, although it has only 3 nominal senses, two of them are too similar. One is “a plant that is eaten as food (vegetable)”, and the other is “*part of* a plant that is eaten as food (vegetable)”. Almost half of the wrongly disambiguated instances result from the confusion between the two senses.

## 4 Conclusion and Future Work

In this paper, we propose an unsupervised HowNet-based word sense disambiguation method, which exploits the masked language model task of large pre-trained language models. In addition, we build a new and larger HowNet-based word sense disambiguation dataset. In evaluation, experimental results demonstrate that our model achieves obviously better performance than all the baseline methods.

In the future, we will try to extend our model to other unsupervised knowledge-based word sense disambiguation tasks, e.g., WordNet-based English word sense disambiguation. Besides, we will consider making explorations into supervised word sense disambiguation based on our method.

## Acknowledgements

This work is funded by the NSFC/DFG Collaborative Research Centre SFB/TRR169 “Cross-modal Learning” II and Beijing Academy of Artificial Intelligence (BAAI). Hou and Zang are supported by Tsinghua University Initiative Scientific Research Program, DCST Student Academic Training Program and National Training Program of Innovation and Entrepreneurship for Undergraduates. We also thank the anonymous reviewers for their constructive comments.

## References

- Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.
- Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of IJCAI*.
- Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING*.
- Leonard Bloomfield. 1926. A set of postulates for the science of language. *Language*, 2(3).
- Claudio Delli Bovi, Luis Espinosa Anke, and Roberto Navigli. 2015. Knowledge base unification via sense embeddings and disambiguation. In *Proceedings of EMNLP*.
- Devendra Singh Chaplot and Ruslan Salakhutdinov. 2018. Knowledge-based word sense disambiguation using topic models. In *Proceedings of AAAI*.
- Devendra Singh Chaplot, Pushpak Bhattacharyya, and Ashwin Paranjape. 2015. Unsupervised word sense disambiguation using markov random field and dependency parser. In *Proceedings of AAAI*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.
- Zhendong Dong and Qiang Dong. 2006. *HowNet and the computation of meaning*. World Scientific.
- Yihong Gu, Jun Yan, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. 2018. Language modeling with sparse product of sememe experts. In *Proceedings of EMNLP*.
- Peng Jin, Yunfang Wu, and Shiwen Yu. 2007. SemEval-2007 task 05: multilingual Chinese-English lexical sample. In *Proceedings SemEval*.
- Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*.

- Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *TACL*, 2:231–244.
- Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.
- Roberto Navigli. 2018. Natural language understanding: instructions for (present and future) use. In *Proceedings of IJCAI*.
- Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018. Integrating weakly supervised word sense disambiguation into neural machine translation. *TACL*, 6:635–649.
- Fanchao Qi, Junjie Huang, Chenghao Yang, Zhiyuan Liu, Xiao Chen, Qun Liu, and Maosong Sun. 2019a. Modeling semantic compositionality with sememe knowledge. In *Proceedings of ACL*.
- Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Qiang Dong, Maosong Sun, and Zhendong Dong. 2019b. Open-HowNet: 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. 2020. Improving sequence modeling ability of recurrent neural networks via sememes. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. In *Proceedings of NeurIPS*.
- Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of AAAI*.
- 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*, 29(6):23–29.
- Rocco Tripodi and Roberto Navigli. 2019. Game theory meets embeddings: a unified framework for word sense disambiguation. In *Proceedings of EMNLP-IJCNLP*.
- 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 LREC*.
- David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of HLT-EMNLP*.
- Jie Wang, Zhenxin Fu, Moxin Li, Haisong Zhang, Dongyan Zhao, and Rui Yan. 2020. Learning sense representation from word representation for unsupervised word sense disambiguation (student abstract). In *Proceedings of AAAI*.
- Yunfang Wu. 2009. A survey of Chinese word sense disambiguation: Resources, methods and evaluation. *Contemporary Linguistics*, 11(2):113–123.
- Erhong Yang, Guoqing Zhang, and Yongkui Zhang. 2001. A Chinese word sense disambiguation method based on primitive co-occurrence data. *Journal of Computer Research & Development*, 38(7):833–838.
- Deniz Yuret. 2007. KU: Word sense disambiguation by substitution. In *Proceedings of SemEval*.
- 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 ACL*.
- Lei Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Multi-channel reverse dictionary model. In *Proceedings of AAAI*.
- Zhi Zhong and Hwee Tou Ng. 2012. Word sense disambiguation improves information retrieval. In *Proceedings of ACL*.
- Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of ACL*.