Learning Sense Representation from Word Representation for Unsupervised Word Sense Disambiguation (Student Abstract)

Jie Wang,[1,2,◇] **Zhenxin Fu,**[1] **Moxin Li,**[1] **Haisong Zhang,**[3] **Dongyan Zhao,**[1,2] **Rui Yan**[1,2*]
[1]Wangxuan Institute of Computer Technology, Peking University, China
[2]Center for Data Science, Peking University, China [3]AI Lab, Tencent, China
[◇]Email: JaneWangle@gmail.com

Abstract

Unsupervised WSD methods do not rely on annotated training datasets and can use WordNet. Since each ambiguous word in the WSD task exists in WordNet and each sense of the word has a gloss, we propose SGM and MGM to learn sense representations for words in WordNet using the glosses. In the WSD task, we calculate the similarity between each sense of the ambiguous word and its context to select the sense with the highest similarity. We evaluate our method on several benchmark WSD datasets and achieve better performance than the state-of-the-art unsupervised WSD systems.

Introduction

Word Sense Disambiguation (WSD) is a task to identify the correct sense of an ambiguous word in the text. In many natural language processing tasks such as conversation systems (Yan 2018), WSD can play a key role. Although supervised methods achieve better performance in the WSD task, the cost of constructing annotated training datasets is expensive. To this end, we propose an unsupervised WSD method.

Our method is overlap-based which is a popular category of unsupervised WSD methods. The method is simple but effective which selects the sense with the highest similarity with its context for each ambiguous word. To get the similarity between each sense of the ambiguous word and its context, we need to obtain sense representations and context representations. Actually, in the WSD task, the correct sense of each ambiguous word is corresponding to a sense in WordNet (Miller 1995). And each sense in WordNet is corresponding to a gloss which is a sequence of words, so we propose a model which can learn the sense representation from the gloss.

However, it is not that easy to get the sense representation of a sense with the gloss alone without supervision. Since senses can not exist alone without words and there has been effective pre-trained word representations (such as Glove, Word2Vec, and fastText), so we try to use pre-trained word representations as gold answers and build a model to learn the word representation for a word in WordNet based on its glosses. The sense representations are the intermediates of the model. We firstly propose **SGM** which only considers monosemous words and furthermore **MGM** which considers both monosemous words and polysemous words to learn the representations.

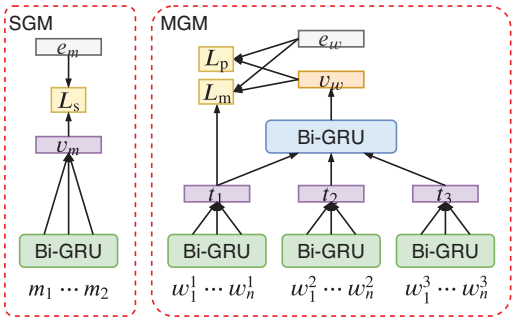

Figure 1: An overview of SGM and MGM.

Sense Representation Learning

Single Gloss Model (SGM) For a monosemous word, its word representation is the same as its sense representation, so it is appropriate to use pre-trained word representation to supervise the learning of sense representation. Given the gloss $\{m^1, m^2, \cdots, m^n\}$ of a monosemous word m where n is the length of the gloss, the model encodes the gloss into a vector v_m through Bi-directional GRU.[1] For a monosemous word, v_m is also the word representation since its sense representation and word representation is equivalent. Inspired by such idea, the loss function is defined as the cosine distance between the learned word representation and the standard pre-trained word representation e_m of word m: $L_s = \cos(v_m, e_m)$.

Multi Gloss Model (MGM) To utilize not only monosemous words but also polysemous words in WordNet, we propose the Multi Gloss Model. Monosemous words can be seen as special polysemous words with single senses. Given word w with multiple senses $\{d_1, d_2, \cdots, d_l\}$, each sense d_i is represented as a gloss g_i which is a sequence of words: $\{w_1^i, w_2^i, \cdots, w_n^i\}$ where n is the length of the gloss. The first layer of the model is to learn the sense representation of each sense d_i of w through $t_i = \text{Bi-GRU}(\{w_1^i, w_2^i, \cdots, w_n^i\})$.

[1]The gloss is a sequence of words and is first transformed into word embedding through looking up word embedding table.

*Corresponding author: Rui Yan (ruiyan@pku.edu.cn)

After getting the sense representation, we need to aggregate them to obtain the word representation v_w. In this paper, we adapt the Bi-GRU to aggregate them: $v_w = \text{Bi-GRU}(\{t_1, \cdots, t_l\})$. The model jointly learns the word representation v_w and the sense representations $t_1, t_2, \cdots, t_l$ of w.

Considering that the learned sense representations of polysemous words are not at the same space with the learned word representations, it may exist some bias if calculating the loss based on learned word representations and later using the learned sense representations in our WSD method directly. To reduce the bias, considering that the sense representation and word representation are the same for a monosemous word, we calculate the loss of monosemous words and polysemous words separately. For monosemous words, both the sense level and word level supervision are adapted as in Equation 1 where $\cos(t_1, e_w)$ makes the sense representation and word representation in the same space. For polysemous words, we calculate the loss as in Equation 2.

$$L_m = \alpha\cos(t_1, e_w) + \beta\cos(v_w, e_w) \quad (1)$$

$$L_p = \cos(v_w, e_w) \quad (2)$$

where α and β are adjustable parameters. The objective of the model is to minimize the co-sine distance between the learned word representation and the standard pre-trained word representation.

WSD Method

For an ambiguous word, we can get the sense representation of each sense based on section 2. And we can calculate the context representation which is the average word representations of important words in the context (which is three sentences in this paper). The filtering rule of the important words is based on Pelevina et al. (2016).

For each sense of an ambiguous word, we calculate the cosine similarity of the sense representation and the context representation. Meanwhile, we use a linear combination of the similarity and sense frequency (Agirre, de Lacalle, and Soroa (2018) proved important) as the final score of each sense. The sense with the highest score is the selected sense of the ambiguous word.

Experiments and Results

Dataset of SGM and MGM Each instance is composed of input: glosses of words from WordNet and label: the pre-trained word representation of a word. The size of the dataset of monosemous words is 36,743 where 30,000 instances are used for training and the rest are for validation. The size of the dataset of polysemous words is 22,447 where 20,000 instances are for training and the rest are for validation. We adopt fastText pre-trained word representation in the paper.

WSD Dataset Our WSD task uses the same five evaluation datasets with Agirre, de Lacalle, and Soroa (2018) : Senseval-2 (SE2), Senseval-3 task 1 (SE3), SemEval-07 task 17 (SE7), SemEval-13, and SemEval-15 task 13 (SE15). The total count of the instances in the five datasets is 7,253.

Systems	ALL	SE2	SE3	SE7	SE13	SE15
Basile14	63.7	63.0	63.7	56.7	66.2	64.6
Babelfy	65.5	67.0	63.5	51.6	66.4	70.3
WSD-TM	66.9	69.0	**66.9**	55.6	65.3	69.6
UKBppr_w2w	67.3	68.8	66.1	53.0	**68.8**	70.3
SGM	67.7	68.6	66.2	57.6	67.1	**74.2**
MGM	**68.2**	**69.5**	66.5	**58.2**	67.6	73.6

Table 1: F1 scores.

Results In Table 1, we compare our overall F1 scores with five other knowledge-based unsupervised systems (Basile14 (Basile, Caputo, and Semeraro 2014), Babelfy (Moro, Raganato, and Navigli 2014), UKBppr_w2w (Agirre, de Lacalle, and Soroa 2018), and WSD-TM (Chaplot and Salakhutdinov 2018)). Except for results on five individual evaluation datasets (SE2, SE3, SE7, SE13 and SE15), we also display the results on ALL dataset which is the concatenation of the five datasets.

Most related works focus on the results on ALL dataset. And our proposed method achieves better F1 score of 68.2 in the ALL dataset as compared to the state-of-the-art score of 67.3. It proves that our model is competitive in the WSD task.

Conclusions

In this paper, we propose a model to learn the sense representations and use an overlap-based method to select the correct sense of each ambiguous word. And our model achieves the state-of-the-art in the evaluation datasets.

Acknowledgement

This work was supported by the National Key R&D Program of China (No. 2017YFC0804001), the National Science Foundation of China (NSFC No. 61876196 and NSFC No. 61672058).

References

Agirre, E.; de Lacalle, O. L.; and Soroa, A. 2018. The risk of sub-optimal use of open source nlp software: Ukb is inadvertently state-of-the-art in knowledge-based wsd. *arXiv preprint arXiv:1805.04277*.

Basile, P.; Caputo, A.; and Semeraro, G. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *COLING 2014*, 1591–1600.

Chaplot, D. S., and Salakhutdinov, R. 2018. Knowledge-based word sense disambiguation using topic models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Moro, A.; Raganato, A.; and Navigli, R. 2014. Entity linking meets word sense disambiguation: a unified approach. *TACL* 2:231–244.

Pelevina, M.; Arefiev, N.; Biemann, C.; and Panchenko, A. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, 174–183. Berlin, Germany: ACL.

Yan, R. 2018. " chitty-chitty-chat bot": Deep learning for conversational ai. In *IJCAI*, volume 18, 5520–5526.