# Word Sense Disambiguation with Massive Contextual Texts

Ya-fei Liu and Jinmao Wei(✉)

College of Computer Science, Nankai University, Tianjin, China
liuyf@mail.nankai.edu.cn, weijm@nankai.edu.cn

**Abstract.** Word sense disambiguation is crucial in natural language processing. Both unsupervised knowledge-based and supervised methodologies try to disambiguate ambiguous words through context. However, they both suffer from data sparsity, a common problem in natural language. Furthermore, the supervised methods are previously limited in the all-word WSD tasks. This paper attempts to collect all publicly available contexts to enrich the ambiguous word's sense representation and apply these contexts to the simplified Lesk and our M-IMS systems. Evaluations performed on the concatenation of several benchmark fine-grained all-word WSD datasets show that the simplified Lesk improves by 9.4% significantly and our M-IMS has shown some improvement as well.

**Keywords:** WSD · Massive contextual texts · Simplified Lesk · M-IMS

## 1 Introduction

Word sense disambiguation (WSD) is an open problem in natural language processing, which identifies word sense used in a given context. It's considered as the fundamental cornerstone for machine translation, information extraction and retrieval, parsing, and question answer. What's bad is that all methods on WSD highly depend on knowledge sources like corpora of texts which may be unlabeled or annotated with word sense [1]. Ineluctably these knowledge sources all suffer from data sparsity to varying degrees. Apart from the sparsity, a common agreement is that supervised methods are restricted in the all-word tasks as labeled data for the full lexicon is sparse and difficult to obtain [2], while knowledge-based methods only requiring an external knowledge source are more suitable for the all-word tasks [4]. In summary, this paper is chiefly involved with the data sparsity and the adaptability of supervised algorithms. Accordingly, two main contributions are summarized as follows:

- We relieve the data sparsity by assembling almost all publicly available contextual texts from different corpora.
- We modify It Make Sense (IMS) [7] by embedding a knowledge-based method to ensure the latter starts to work in case the former fails.


G. Li et al. (Eds.): DASFAA 2019, LNCS 11448, pp. 430–433, 2019.
https://doi.org/10.1007/978-3-030-18590-9_60

## 2 Methodology

### 2.1 Corpora Sources

The first main point of this paper lies in more corpora with massive instance sentences uniformly annotated by one sense repository. Here are five publicly available corpora annotated with WordNet: WordNet, SemCor [3], OMSTI [6], MASC[1], GMB[2]. WordNet is not only a lexical dictionary as the sense repository here but also a source of example sentences.

### 2.2 M-IMS

**Preprocessing and Feature Extraction.** Preprocessing aims to convert various texts from different corpora into formatted instance sentences. In contrast to IMS, we include two additional procedures: Standardization and Sense Mapping. Standardization intends to unify the formats and preserve texts with POS, annotation and lemma. While Sense Mapping deals with the annotation version problem according to the sense key.

Feature Extraction is conducted on the massive contexts (MC) as how IMS does. A small modification to surrounding words feature here is that the surrounding words can be only in the current sentence, not including the adjacent sentences, because we disambiguate ambiguous words on sentence-level.

**Classification.** Another major contribution of this paper lies in the modification here. The Classification comprises three components: Supervised Classification, Decision Component, and Knowledge-based Classification.

*Supervised Classification and Knowledge-Based Classification.* The supervised classification part is almost the same as the classifier in the IMS. As for the knowledge-based, we select simplified Lesk as the knowledge-based algorithm to make the disambiguation. The overlapping way of simplified Lesk to calculate the similarity between gloss and context conforms to the characteristic of the MC.

*Decision Component.* The rhombus with a question mark inside in Fig. 1 represents the decision component. It determines whether or not the knowledge-based methods are introduced into the disambiguation. Here we recommend two boundary conditions for the decision:

- Strict condition: Only if annotations for a word cover all senses of this word with the same part of speech, can the decision output yes/y, otherwise no/n.
- Loose condition: As long as annotations for a word cover at least one sense of this word with the same part of speech, the decision outputs yes.

This paper adopts relatively loose setting: As long as annotations for a word cover at least two senses of this word with the same part of speech, we consider the trained model is helpful in a way.

[1] http://www.anc.org/data/masc/.
[2] http://gmb.let.rug.nl/.

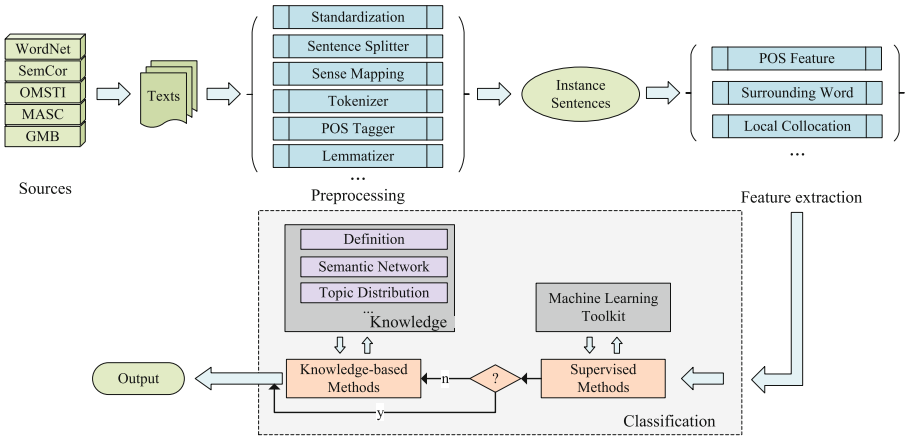


**Fig. 1.** M-IMS system architecture

## 3 Experiments and Results

The first experiment aims at showing the ability of massive contextual texts to relieve the data sparsity. The second makes a comparison among M-IMS, IMS, simplified Lesk etc. And we choose the concatenation of the five standardized datasets (Sem-Union) from [5] as the test dataset.

### 3.1 Results

In Table 1, we have found that contextual texts, like instance sentences, extremely suit for word matching pattern contemporarily. Furthermore, the increment by our MC offers more possibility for previously annotation-lacking senses and relieve the data sparsity to a certain degree.

**Table 1.** The overlap rates and annotation coverages of several {sources}-context pairs.

| Coverage type | {Sources}-context pairs | Rate (%) | Accuracy (%) |
|---|---|---|---|
| Overlap rate | Gloss | 16.1 | 53.7 |
| | WordNet | **82.1** | 57.6 |
| | SemCor | 66.0 | 63.9 |
| | MC | 72.5 | **67.0** |
| Annotation coverage | SemCor | 69.0 | – |
| | OMSTI | 71.5 | |
| | MC | **76.7** | |

In Table 2, it's remarkable that simplified Lesk with MC obtains a much better performance and pushes the overlap-based algorithms to a new high.

What's more, M-IMS uniformly performs better than IMS both on SemCor and MC, but not with a significant margin implying that the performance of knowledge-based algorithms is required to be promoted in the future.

**Table 2.** Comparison of IMS, M-IMS and SL with different sources on Sem-Union.

| Systems | Sources | Sem-Union (%) |
|---|---|---|
| SL | WordNet | 57.6 |
| | MC | **67.0** |
| IMS | SemCor | 67.1 |
| | MC | 67.5 |
| M-IMS | SemCor | 67.4 |
| | MC | **67.7** |

## 4 Conclusion

This paper mainly deals with the data sparsity in WSD with massive contexts and the adaptability of supervised methods. Note that this work is still in progress and we shall release MC in our later research work along with relevant API to enable various applications with detail documentations.

**Acknowledgements.** This work was partially supported by the National Natural Science Foundation of China (61772288), and the Natural Science Foundation of Tianjin City (18JCZDJC30900).

## References

1. Borah, P.P., Talukdar, G., Baruah, A.: Approaches for word sense disambiguation-a survey. IJRTE **3**(1), 35–38 (2014)
2. Chaplot, D.S., Salakhutdinov, R.: Knowledge-based word sense disambiguation using topic models. arXiv preprint arXiv:1801.01900 (2018)
3. Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T.: A semantic concordance. In: Proceedings of the Workshop on Human Language Technology, pp. 303–308. ACL (1993)
4. Miller, T., Biemann, C., Zesch, T., Gurevych, I.: Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In: Proceedings of the 24th COLING, pp. 1781–1796 (2012)
5. Raganato, A., Camacho-Collados, J., Navigli, R.: Word sense disambiguation: a unified evaluation framework and empirical comparison. In: Proceedings of the 15th Conference of ECACL, vol. 1, pp. 99–110 (2017)
6. Taghipour, K., Ng, H.T.: One million sense-tagged instances for word sense disambiguation and induction. In: Proceedings of the 19th CoNLL, pp. 338–344 (2015)
7. Zhong, Z., Ng, H.T.: It makes sense: a wide-coverage word sense disambiguation system for free text. In: Proceedings of the ACL 2010 System Demonstrations, pp. 78–83. ACL (2010)