A Theoretical Study of WSD

Haozhou Pang
Department of Computing Science
University of Alberta
haozhou@ualberta.ca

Abstract

Previous study (Gale et al., 1992) has shown that in a well-written discourse, a polysemous word appears two or more times are very likely (98% in their experiment) to share the same sense, this is known as One sense per Discourse (OSPD). In this paper, we replicate the experiment on the OntoNotes corpus and our result (80.3%) does not support the OSPD hypothesis as strongly. After some fine-grained senses get grouped to the same homonym class, we show that the One Homonym per Discourse hypothesis holds with 99.7% probability. Also, we proposed another hypothesis, the One Homonym per Translation hypothesis (OHPT), which states that, homonyms represented by the same word have disjoint translation sets. We found this hypothesis holds with extremely high probability (100%) on the OntoNotes corpus for English-Chinese text. We discuss the implications of these results in this paper.

1 Introduction

Homonyms, each of two or more words having the same spelling but different meanings and origins, are very common in English. For example, the word $bank$ can be interpreted as a financial institution or a river bank depending on the context, so being able to disambiguate the sense of homonyms can largely improve the accuracy when we translate an English text into another language. An important observation proved by the previous work (Gale et al., 1992) showed that with extremely strong probability (98%), a polysemous word share sense in the same discourse. In this paper, we replicate their experiment and verify the One Sense per Discourse hypothesis on the OntoNotes 5.0 multilingual corpus, we showed that 80.3% of polysemous words in OntoNotes support the OSPD. Also, we verify the One Homonym per Discourse hypothesis by mapping the fine-grained senses to homonymous senses, and we show that OHPD holds with 99.7% probability.

In this paper, we proposed another hypothesis, the One Homonym per Translation hypothesis (OHPT), which states that, homonyms represented by the same word have disjoint translation sets. Namely, there must be only one homonym class per translation. We test this hypothesis on the OntoNotes corpus, and discover that it holds with 100% probability for English-Chinese parallel corpus.

These hypotheses are important as they can be applied to many other fields, for example, the OSPD hypothesis can be used in the word sense tagging task, if a word only has one sense in a discourse then we only need to disambiguate one occurrence and use that sense to tag the rest of instances. The OHPT hypothesis can guide a Machine Translation system to better translate ambiguous words since a mistranslation can be detected at early stage if we find a violation of OHPT.

2 Related Work

In the One Sense per Discourse paper, (Gale et al., 1992) tested a sample of 9 polysemous English words. 54 pairs of concordance lines for these words were taken from Groliers Encyclopedia, the majority opinion from 5 different judges found that 94% of them shared the same sense. A similar conclusion (96%) was drawn on the Brown corpus as well.

However, another study by (Krovetz, 1998) showed that only 67% of the ambiguous words in SEMCOR corpus has a single sense per discourse, they also deduced that One Sense per Discourse hypothesis is only true for homonymous senses and not for fine-grained senses. Our task is the same as the previous studies, but we used the Ontonotes corpus.

To the best of our knowledge, there is no published work on One Homonym per Translation, in this paper, we make the first effort towards proving this important hypothesis on English-Chinese corpus.

3 Method

The section presents information about how we approached the current task with more detailed description about the data set.

3.1 Data

In this project, we use the OntoNotes 5.0 multilingual corpus (Weischedel, 2013), which contains parallel English and Chinese text from news, broadcast news, broadcast conversation, telephone conversation and web data. Since the English and Chinese sentences are not perfectly aligned in the OntoNotes corpus, e.g. some English Sentences are missing their Chinese translation and vice versa. The cleaning of dataset takes a skilled labor approximately three hours. The detailed statistics of the corpus are summarized in table 1.

num. of aligned sentences	7453
num. of sense-annotated word tokens	16982
num. of sense-annotated word types (with POS distinctions)	1576
num. of sense-annotated word types (without POS distinctions)	1404
num. of distinct WordNet senses	2413

Table 1: Statistics of the corpus used in this project

3.2 Word Alignment

The word alignment was not implemented in the OntoNotes corpus, so an unsupervised word aligner, $fastalign$ (Dyer et al., 2013), was used to perform the word alignment. Since $fastalign$ is a unsupervised model, its performance highly depends on the size of dataset, the 7453 parallel sentences from OntoNotes corpus augmented with another 15k parallel sentences from OPUS (Tiedemann, 2012) were fed as input for $fastalign$, and a word level mapping from English text to Chinese text was generated.

3.3 One sense per Discourse

The One Sense per Discourse hypothesis states that, in a well-written discourse, a polysemous word appears two or more time are very likely to share the same sense. The OntoNotes corpus contains the subtitle text from interviews, documentaries, and TV shows. Different files are either different interviews or different episodes of the documentary. So in this paper, each different file is considered as a individual discourse. Therefore, we extract all the sense annotated words in each file and check how many of them only have single sense in the particular discourse.

3.4 One Homonym per Discourse

The One Homonym per Discourse hypothesis states that, all occurrences of a homonymous word in the same discourse represent the same homonym. As demonstrated by Figure 1, some senses of a English word might be homonymous if they are in the same homonym class. After mapping the fine-grained senses to homonymous senses, we check how many of them represent the same homonym.

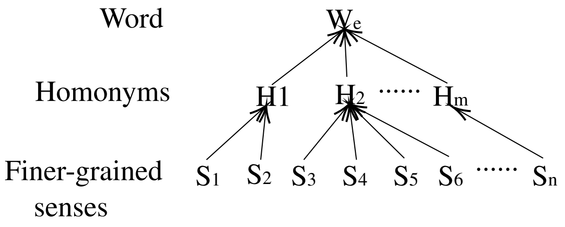

Figure 1: Visualization of homonymous senses

3.5 One Homonym per Translation

For each sense-annotated word in the English text, we use the word alignment to find its corresponding Chinese translation. As shown by Figure.2, suppose an English word W_e has n different senses (s_1, s_2, ... , s_n) in the sense inventory, each sense occurs to be translated to a set of Chinese words ($c_1, c_2, ...$). An English word is considered to support the One Sense per Translation hypothesis if there is no overlap among the Chinese translations of each sense. Otherwise, there must be at least two senses of a word have been translated to the same Chinese word, that indicates a violation of the hypothesis.

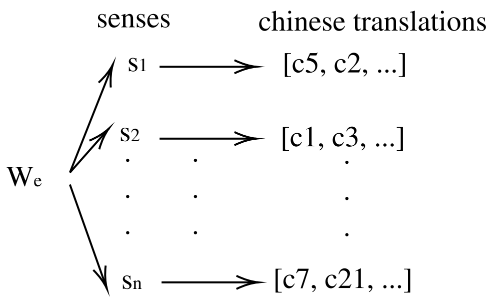

Figure 2: Visualization of mapping from senses to Chinese translations

We merge the Chinese translations of homonymous senses and One Homonym per Translation hypothesis is violated only if there exists overlap among the translations of non-homonymous senses.

3.6 Evaluation

The evaluation was conducted based on the percentage of ambiguous words that support the hypothesis.

4 Result

The section present the results of current task followed by detailed error analysis.

4.1 OSPD

Source Text	Sense-annotated word tokens	Words with single sense per discourse	%
CCTV	1082	938	86.7
CNN	1988	1532	77.1
MSNBC	539	417	77.4
Phoenix	625	514	82.2
Total	4234	3401	**80.3**

Table 2: Percentage of ambiguous words that have single sense per discourse in OntoNotes

Table. 1 shows that 80.3% of ambiguous words in OntoNotes corpus has single sense per discourse, if we differentiate words based on part of speech, the result increases to 83%. Overall, our result coincides with the result (67%) by (Krovetz, 1998) and does not support the One Sense per Discourse hypothesis as strongly as (Gale et al., 1992) study (98%).

4.2 OHPD

After some homonymous senses get grouped together, we find that the One Homonym per Discourse holds with 99.7% probability. The only violation we detect is the word $line$, two non-homonymous senses of $line$ occur in the same discourse.

- line - v2: form a line $(e.g\ line\ up)$

- line - n3: unit of a text, picture or melody $(e.g\ the\ second\ line\ of\ the\ textbook)$

However, if we differentiate words based on part of speech, then it will lead to a 100% probability that OHPD is true. Since the word $line$ in the above example has different pos taginggs, then the line-v and line-n will be considered as different words.

4.3 OHPT

We experiment on 133 homonym words occur in the OntoNotes corpus, and we find that the One Homonym per Translation hypothesis holds with extremely high probability for English-Chinese text (100%), namely, we did not find any violation of OHPT in the dataset we have.

4.4 Error analysis

The quality of the word alignment was checked by taking a random sample of size 100 sentences, 93 of them were aligned correctly and the alignment error rate (AER)[1] is 98.9%.

5 Conclusion

The current task focused on investigating the validity of the One Sense per Discourse, One Homonym per Discourse, and One Homonym per Translation hypotheses. We found more occurrence of multi-senses per discourse than reported in (Gale et al., 1992) (19.7% instead of 2%). After the homonymous senses get grouped together, we show that the OHPD hypothesis is true for 99.7% of the cases. Also, we find that the OHPT hypothesis holds with extremely high probability on En-Ch text by showing no violations out of 133 homonym words in OntoNotes corpus. In future work, we will experiment whether One Homonym per Translation holds for other language pairs.

Acknowledgments

First and far most, I would like to thank prof. Greg Kondrak and Bradley Hauer for their advice and help. Also, many thanks go to my fellow student Issei Luan for helpful discussions.

[1] $AER = 2 * |A \cap S|/(|A| + |S|)$

References

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *In Proc. NAACL*.

William A. Gale, Kenneth Ward Church, and David Yarowsky. 1992. One sense per discourse. In *HLT*.

Robert Krovetz. 1998. More than one sense per discourse. In *NEC Princeton NJ Labs., Research Memorandum*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*.

Ralph et al Weischedel. 2013. Ontonotes release 5.0 ldc2013t19. In *Linguistic Data Consortium*.