UROP I
Literature Review
Cross Lingual Word Sense Disambiguation

HENG Low Wee
U096901R

1 Introduction

In any human language, there are some words, or even phrases, that have multiple interpretations depending on the context in which a word is used. These words are called homographs, and have multiple meanings (or *senses*), making language itself ambiguous. Here are some examples:

1. She is interested in the *interest* rates of the bank.
2. He developed an *interest* in art.

Observe that the word *interest* in the two sentences clearly have different meanings. This is the ambiguity we are talking about. Paraphrasing Weaver [14], if we examine some text only *one* word at a time, it is impossible to determine the meaning of it. Only by looking at its surrounding words then can one decide its meaning. It is almost an unconscious act for a human to interpret the word's correct sense. But because human languages are complex, a machine need to run through a series of analytical processes before it can guess the best answer. In general, this process is *Word Sense Disambiguation* (*WSD*). More formally, WSD is the process of deciphering the intended meaning of a homographic word(s) in a given context.

There are four conventional methods to WSD, namely:

1. Dictionary-based & knowledge-based
   This method relies mainly on dictionaries and lexical knowledge bases to disambiguate senses. The Lesk Algorithm [8] is such a method. Satanjeev & Ted adapted the Lesk Algorithm but instead of using a standard dictionary, they used WordNet [6] as a source of words' senses (for details, refer to [1]). After evaluation testings, it was found that their implementation outperformed a more traditional Lesk approach with an accuracy rate of 32%, double of the traditional approach. This demonstrated the fact that an approach that integrates WordNet, or other lexical databases, would be able to achieve a higher rate of accuracy in WSD.
2. Supervised
   Generally, this method relies on manually sense-tagged corpora. Using manually sense-tagged corpora allows supervised methods to achieve a higher accuracy as compared to most unsupervised methods. The downside of using manually sense-tagged corpora, however, is its high costs and labour intensiveness to create.

3. Semi-supervised
   This method makes use of both labeled and unlabeled data. Similarly, it refers to using multiple untagged corpora to provide concurrent information to supplement a tagged corpus.

4. Unsupervised
   This is probably the most challenging among all the approaches to WSD. This method may also be closely related to *Word Sense Induction*, where senses can be induced from analyzing the words in a given text. Schütze [12] described a disambiguation algorithm based on clustering words, and then senses are interpreted as a cluster of similar context of an ambiguous word. To determine the senses, some algorithms may also map the words to a collection of senses. Mihalcea [10] described an unsupervised approach by using the articles on Wikipedia, together with mapping of senses with a WordNet resource. It addressed the issue of *knowledge acquisition bottleneck*, as Wikipedia serves as a dynamic source for senses because it is updated by users all around the world regularly. More importantly, it addresses the case where languages are gaining new words, and that words are gaining new senses, a common thing in today's modern society.

In this review, we will study how we can improve bilingual (English-Chinese) translations by improving its accuracy in disambiguating polysemic words in a given context, so that the original context would not be lost during the translation process. We will also look at the disambiguation process's ability to handle new words and new senses in human languages. Thus, we want to specifically focus on the topic of having Unsupervised *Cross Lingual* WSD. In the following section, we will look at works related to the construction of parallel corpora, and some techniques used in word sense disambiguation.

2 Literature

2.1 Word Sense Tagging using Parallel Corpora

Supervised word sense disambiguation systems rely heavily on manually sense-tagged corpora, and to produce more reliable results they need high quality annotations. However, manual tagging is very labour intensive and costly. It is also very impractical as doubling the training corpora only reduces errors by 3 to 4% [15].

Diab & Resnik described a form of automated annotation and sense-tagging by analyzing an ambiguous word's translations in a second language [5]. Their approach involves two languages, for example English and French. The first step is to identify the target words and their corresponding translations in the source corpus. Word alignment is carried out and the corresponding positions of a target word in both languages are captured. Then, grouping the target words into *target sets*. Next, in each set consider all possible senses for each word and then tag a word with the sense most similar to the other words. Last, we make use of the sense tags in the target sets and project them to the source corpus. (for details, refer to [5])

The advantage of this approach is that it attempts to eliminate the need to carry out manual sense-tagging or annotation. Automation would address the issue about this task being labour-intensive and costly. One disadvantage, however, might be the reduction

of accuracy in the tagging and annotating. A solution may be to require gold-standard sense-tags, as mentioned in [5].

We highlighted this work written by Diab & Resnik [5] precisely because their approach involves the usage of parallel corpora, or in other words, two languages. The general idea of their work is about sense-tagging another corpus using one that has already been sense-tagged. Adapting to this, possibly, we can construct a English-Chinese parallel, which will capture the Chinese translations that are derived (with similar sense or context) from an English word and vice versa. This essentially will improve the quality of bilingual translations.

2.2 Word Sense Disambiguation using Wikipedia

Wikipedia articles, manually created by users, are generally correct in terms of its contents. Some examples of Wikipedia links (on an article), in the *MediaWiki* syntax, are [[**bar(law)**|**bar**]] and [[**bar(counter)**|**bar**]]. The senses of the word *bar*, namely *law* and *counter*, can be derived by extracting them from the annotated links. Mihalcea [10] made use of this information and described an approach to build a sense-tagged corpora using Wikipedia articles. It begins with extracting all paragraphs from Wikipedia that contain the occurrences of a given word. It follows that the senses of each word are extracted, then mapped onto their corresponding WordNet senses. Then, the approach moves on into the disambiguation algorithm [10]. A target text is tokenized, and each token is tagged with its part-of-speech information. Then, collocations are identified. Next, local and topical features are extracted from the context of the ambiguous word. This set of features is similar to the one used by Ng & Lee [11]. (for details, refer to [10])

The advantage of this method is it addresses the *knowledge acquisition bottleneck* issue. The dynamic nature of Wikipedia makes it extensible. Therefore, by using Wikipedia as a corpus, one can be sure that it will always be up-to-date and will contain any new words or new senses. A disadvantage, however, is data inconsistency. Different users might use a different name for the same object or entity. For example, *handphone* and *mobile phone*. This also relates back to the fact that there are new words or senses emerging regularly in today's modern language.

We can observe that this approach would have the flexibility to handle new words and senses. This is because, considering how popular Wikipedia is, any new word or sense used by people, say *tweet* (a post on Twitter), would most likely appear on Wikipedia faster than any formal dictionary databases. But of course, we must also consider the fact that if the number of new words and senses grows as fast as the number of articles, there may be a need to re-construct the sense-tagged corpus. Hence, design considerations for the corpus must include *expandability*. Alternatively, we may use the APIs on MediaWiki to send a query request for a given word, and with some parsing of the output we can extract the needed information. Thus, we can build a real-time WSD system that is highly independent of lexical databases and corpora. Performance issues aside, this could minimize the need to reconstruct a Wikipedia-based corpus on an occasional basis.

2.3 Word Sense Disambiguation using Dependency Knowledge

Similar to Section 2.2, this approach described by [3] begins with the construction of the corpus. In short, ambiguous words are sent to Web search engines to retrieve the relevant pages. These pages are cleaned, segmented, and then parsed with a *dependency parser*, Minipar [9] to retrieve the parsing trees, which are merged to form the *context knowledge base* [3]. Our focus in [3] would be on its problem formulation and its WSD algorithm itself. Formulated into *weighted directed graphs*, it is effective in telling the dependencies between the words in a given text, simply by computing the values of the weighted nodes. The weight assignments and score computations are handled by the *TreeMatching* function [3], which is the score calculator for the weights in the parsing trees. A target sentence is passed into the WSD algorithm together with WordNet sense inventory and the context knowledge base built earlier. *TreeMatching* then assigns weights to the nodes based on rules and dependency relation instances, and returns the score of a WordNet gloss that an ambiguous word was compared with. Subsequently, either the sense with the best score or the first sense will be determined as the correct sense. (for details, refer to [3])

Undoubtably, the algorithm is effective and accurate in matching the dependency relations to determine the correct senses, as shown in the evaluation results in [3]. However, before *TreeMatching* can be done, all the sentences and glosses have to be pre-processed, and parsed into parsing trees. The parsing process, especially, takes a lot of time [3]. Then, there is this concern regarding the dependency parser, Minipar. According to [9], Minipar was able to achieve 89% precision for parsing sentences. Although [3] claimed that the WSD algorithm will minimize those erroneous output, it was not explicitly defined how it actually did it.

3 Conclusion

In this review we touched on the field of *Word Sense Disambiguation* (*WSD*). WSD is a very challenging task because it involves working with the complexity of human languages. First formulated as a distinct computational task during the 1940s, WSD is one of the oldest problems in computational linguistics. Weaver [14] wrote in his memorandum, that if we examine text *one* word at a time, it is impossible to determine the meaning of it. Only by looking at its surrounding words then can one decide its meaning.

Related to our topic of interest, it is notable that Lefever & Hoste [7] had demonstrated Cross Lingual WSD using parallel corpora from Europarl[1]. However, the corpora used were mainly in European languages, not applicable when the languages we are focusing on are English and Chinese. There are many papers introducing methods to construct an English-Chinese parallel corpus (see [4, 13, 2], however not discussed in this review for the focus is on word senses). So maybe for our study in Cross Lingual WSD, we could look at these various methods of construct a parallel corpus, and also the method mentioned in Section 2.1.

In general, the characteristics mentioned in this review are *accuracy*, *flexible* (to handle new words & senses) and *unsupervised*. For that, we have studied some approaches

[1] http://www.statmt.org/europarl/

and techniques that may give rise to these characteristics in WSD. Diab & Resnik [5] described a form of automated sense-tagging by analyzing an ambiguous word's translations in a second language. This introduces not only automation to sense-tagging, but also contributes to more *accurate* translations between the languages in the parallel corpora. To be *flexible*, the general idea is to adapt a Web-based resource for sense-extraction. Mihalcea described using Wikipedia for WSD [10]. Advantages include it being publicly accurate in its contents, and it being exposed to new words and senses added by articles' authors regularly. While data inconsistency might be an issue here, it is flexible enough to handle ambiguous words that may not yet appear on formal dictionaries. As for *unsupervised*, [3] introduced an approach that determines the correct sense by on dependency knowledge. Although unsupervised, [3] showed that its precision was pretty close to the supervised methods that were in their evaluation tests. As a whole, I believe that by integrating these techniques, a system would have accurate parallel corpora and word sense disambiguations, be adaptable to new words and senses in languages, and therefore, lead to improved bilingual translations.

Perhaps the next idea to consider is to go real-time for WSD. So far the above mentioned literatures did not touch on this idea, probably because of performance issues, for it can be easily inferred that the construction of large corpora could not possibly be achieved in a matter of seconds. While this poses yet another challenge not exclusive to the field of Word Sense Disambiguation, but as systems' performance are reaching new heights regularly, it should not be impossible in the near future.

References

[1] Banerjee, S., and Pedersen, T. An adapted lesk algorithm for word sense disambiguation using wordnet. In *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing* (London, UK, 2002), Springer-Verlag, pp. 136–145.

[2] Baobao, C. Chinese-English Parallel Corpus Construction and its Application. *Computational Linguistics* (2004), 283–290.

[3] Chen, P., Ding, W., Bowes, C., and Brown, D. A fully unsupervised word sense disambiguation method using dependency knowledge. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09*, June (2009), 28.

[4] Dalianis, H., Xing, H.-c., and Zhang, X. Creating a Reusable English-Chinese Parallel Corpus for Bilingual Dictionary Construction. *Word Journal Of The International Linguistic Association*, 1700–1705.

[5] Diab, M., and Resnik, P. An unsupervised method for word sense tagging using parallel corpora. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Morristown, NJ, USA, 2002), Association for Computational Linguistics, pp. 255–262.

[6] Fellbaum, C., Ed. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May 1998.

[7] Lefever, E., and Hoste, V. Semeval-2010 task 3: cross-lingual word sense disambiguation. In *DEW '09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions* (Morristown, NJ, USA, 2009), Association for Computational Linguistics, pp. 82–87.

[8] Lesk, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation* (New York, NY, USA, 1986), ACM, pp. 24–26.

[9] Lin, D. A dependency-based method for evaluating broad-coverage parsers. *Nat. Lang. Eng. 4*, 2 (1998), 97–114.

[10] Mihalcea, R. Using wikipedia for automatic word sense disambiguation.

[11] Ng, H. T., and Lee, H. B. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics* (Morristown, NJ, USA, 1996), Association for Computational Linguistics, pp. 40–47.

[12] Schütze, H. Automatic word sense discrimination. *Comput. Linguist. 24*, 1 (1998), 97–123.

[13] Sun, L., Xue, S., Qu, W., Wang, X., and Sun, Y. Constructing of a large-scale chinese-english parallel corpus. In *COLING '02: Proceedings of the 3rd workshop on Asian language resources and international standardization* (Morristown, NJ, USA, 2002), Association for Computational Linguistics, pp. 1–8.

[14] Weaver, W. Translation. In *Mimeographed* (1949), MIT Press, pp. 15–23.

[15] Yarowsky, D., and Florian, R. Evaluating sense disambiguation across diverse parameter spaces. *Nat. Lang. Eng. 8*, 4 (2002), 293–310.