HR-WSD: System Description for All-words Word Sense Disambiguation on a Specific Domain at SemEval-2010

Meng-Hsien Shih
National Taipei University of Technology
Taipei, Taiwan, ROC.
simon.xian@gmail.com

Abstract

The document describes the knowledge-based Domain-WSD system using heuristic rules (knowledge-base). This HR-WSD system delivered the best performance (55.9%) among all Chinese systems in SemEval-2010 Task 17: All-words WSD on a specific domain.

1 Introduction

Word Sense Disambiguation (WSD) is essential for language understanding systems such as information retrieval, summarization, and machine translation systems (Dagan and Itai, 1994; Schutze and Pedersen, 1995; Ng and Zelle, 1997). In particular due to the rapid development of other issues in computational linguistics, WSD has been considered the next important task to be solved. Among various WSD tasks, the lexical sample task can achieve a precision rate more than 70% in Chinese, so can the all-words task in English, but currently no Chinese all-words WSD system is available. This study proposes an all-words WSD system conducted on a specific domain which can achieve a 55.9% precision rate.

This system makes use of certain characteristics of WordNet. First, the sense inventory in Chinese WordNet is ordered by the "prototypicality" of the words. In other words, the first sense of a word with multiple senses will be the prototype meaning of that word. In addition to semantic relations and sense definitions, Chinese WordNet also includes sense axes which indicate the relations between Chinese senses and corresponding English senses.

2 Proposed Approach

Two heuristic rules are devised to characterize domain texts: In a domain text, domain senses are more likely to occur in words if they have one (Heuristic Rule 1); on the other hand, for words with no domain senses, the most generic usages (prototype senses) are more likely to be adopted (Heuristic Rule 2). Therefore, as proposed by Li et al.(1995) for the WordNet-based domain-independent texts WSD task, two heuristic rules (HR) are taken into consideration in the domain WSD test:

```
for all senses $s_k$ of w do
  if w has domain sense
    choose domain sense $s_k$
  else
    choose prototype sense $s_1$
end
```

Figure 1: Heuristic Rules based WSD

Besides, sense definitions from WordNet were also tested with simplified Lesk algorithm (Lesk, 1986; Kilgarriff and Rosenzweig, 2000) in another experiment to examine the effect of considering sense definitions in domain WSD:

```
for all senses $s_k$ of w do
  if w has domain sense
    choose domain sense $s_k$
  elseif $D_k$ overlaps with $C$:
    choose sense $s_k$ with $D_k$
    that overlaps the most
  else:
    choose prototype sense $s_1$
end
```

Figure 2: HR with simplified Lesk Algorithm. D_k is the set of content words occurring in the dictionary definition of sense s_k. C is the set of content words in the context.

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 433–435,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

3 Procedures

Before the test only preprocessing including segmentation and parts of speech tagging will be applied to the target texts, in order to eliminate those senses of the same word form in other parts of speech; the background documents provided by SemEval-2010 are not used for training since this is not a supervised system. According to Wang (2002), with preprocessing of PoS tagging alone, 20% of word sense ambiguity can be distinguished.

Since the current number of semantic relations in Chinese WordNet is still less than that in English WordNet (PWN), to detect domain senses, the sense axes in Chinese WordNet are exploited. By seeding with English words such as "environment" and "ecology," all English words related to these seed words can be captured with the help of the semantic relations in Princeton WordNet. By mapping these environment-related English words to Chinese words with any kind of semantic relations in the sense axes, the corresponding Chinese domain senses can be identified.

Therefore, the HR-WSD system will first consider any domain senses for the words to be disambiguated; if there is no such sense, the prototype sense will be adopted. Another test where sense definitions from WordNet are considered to facilitate HR-based disambiguation was also conducted.

4 Evaluation

The results were evaluated according to three manually tagged documents in SemEval-2010 Task 17: All-words WSD on a Specific domain (Agirre et al., 2010). The most frequent sense baseline (MFS) refers to the first sense in WordNet lexical markup framework (In Chinese WordNet senses are ordered according to annotations in hand-labelled corpora). In these tagged domain texts, only nouns and verbs (two major types of content words) as a single word are disambiguated. Therefore, in this system only these two kinds of words will be tagged with senses. Adjectives, adverbs, or words in multiple forms (e.g., idioms and phrases) are not considered, in order to simplify the test and observe the results more clearly.

5 Results

By observing that the HR-WSD system* (Rank 1) outperformed other systems and was closest to the MFS performance we can infer that Heuristic Rule 2 works. However, since this system performance is still worse than MFS, it may indicate that Heuristic Rule 1 does not work well, or even decreases the system performance, so the mechanism to detect domain senses needs to be refined. Besides, the inclusion of simplified Lesk algorithm** did not perform better than the original HR-WSD system, further investigation such as more fine-grained definition can be expected.

Rank	Precision	Recall
MFS	0.562	0.562
1*	0.559	0.559
2**	0.517	0.517
3	0.342	0.285
4	0.322	0.296
Random	0.32	0.32
5	0.310	0.258

Table 1: Results.

6 Discussion and Future Development

Although PoS tagging may help filter out senses from other parts of speech of the same word form, incorrect PoS tagging will lead to incorrect sense tagging, which did happen in the HR-WSD system, in particular when there is more than one possible PoS tag for the word. For instance, 'nuan-hua' in 'quan-qiu nuan-hua' (global warming) is manually tagged with a verbal sense in the answer key from SemEval-2010, but tagged as a noun in the pre-processing stage of the HR-WSD system. The difference between manual tagged texts and automatic tagged texts should be examined, or consider allowing more than one PoS tag for a word, or even no PoS pre-processing at all.

To disambiguate with the help of gloss definition, gloss words of the polysemous word must have direct overlapping with that of its context word, which does not always occur. To solve this problem, we may expand gloss words to related words such as hyponyms, hypernyms, meronyms, or the gloss definition of the current gloss words.

Apart from nouns and verbs, if function words and other kinds of content words such as adjectives and adverbs are to be disambiguated, the performance of the current WSD system needs to be re-examined.

As mentioned in the beginning, WSD is an essential part in language understanding systems.

With this Chinese WSD program, information retrieval, summarization, or machine translation tasks would be more plausible. The proposed heuristic rules may also work for other languages with similar WordNet resources. Besides, this system was currently tested on three texts from the environment domain only. It can be expected that this Chinese WSD can work on texts of other domains.

References

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*,34:15–48.

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Shu-kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen and Roxanne Segers. 2010. SemEval-2010 Task 17: All-words Word Sense Disambiguation on a Specific Domain. *In Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010), Association for Computational Linguistics, Uppsala, Sweden..*

Hinrich Schutze and Jan O. Pedersen. 1995. Information Retrieval Based on Word Senses. *In Proceedings of the ACM Special Interest Group on Information Retrieval.*

Hui Wang. 2002. A Study on Noun Sense Disambiguation Based on Syntagmatic Features. *International Journal of Computational Linguistics and Chinese Language Processing*,7(2):77–88.

Hwee Tou Ng and John Zelle. 1997. Corpus-Based Approaches to Semantic Interpretation in Natural Language Processing. *AI magazine*,18(4):45–64.

Ido Dagan and Alon Itai. 1994. Word-Sense Disambiguation Using a Second-Language Monolingual Corpus. *Computational Linguistics*,20(4):563–596.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine from a ice cream cone. *In Proceedings of the 5th International Conference on Systems Documentation, Toronto, CA, pp. 24–26..*

Xiaobin Li, Stan Szpakowicz, and Stan Matwin. 1995. A WordNet-based Algorithm for Word Sense Disambiguation. *The 14th International Joint Conference on Artificial Intelligence.*