# HIT-CIR: An Unsupervised WSD System Based on Domain Most Frequent Sense Estimation

**Yuhang Guo, Wanxiang Che, Wei He, Ting Liu, Sheng Li**
Harbin Institute of Technolgy
Harbin, Heilongjiang, PRC
yhguo@ir.hit.edu.cn

## Abstract

This paper presents an unsupervised system for all-word domain specific word sense disambiguation task. This system tags target word with the most frequent sense which is estimated using a thesaurus and the word distribution information in the domain. The thesaurus is automatically constructed from bilingual parallel corpus using paraphrase technique. The recall of this system is 43.5% on SemEval-2 task 17 English data set.

## 1 Introduction

Tagging polysemous word with its most frequent sense (MFS) is a popular back-off heuristic in word sense disambiguation (WSD) systems when the training data is inadequate. In past evaluations, MFS from WordNet performed even better than most of the unsupervised systems (Snyder and Palmer, 2004; Navigli et al., 2007).

MFS is usually obtained from a large scale sense tagged corpus, such as SemCor (Miller et al., 1994). However, some polysemous words have different MFS in different domains. For example, in the Koeling et al. (2005) corpus, target word *coach* means *"manager"* mostly in the SPORTS domain but means *"bus"* mostly in the FINANCE domain. So when the MFS is applied to specific domains, it needs to be re-estimated.

McCarthy et al. (2007) proposed an unsupervised predominant word sense acquisition method which obtains domain specific MFS without sense tagged corpus. In their method, a thesaurus, in which words are connected with their distributional similarity, is constructed from the domain raw text. Word senses are ranked by their prevalence score which is calculated using the thesaurus and the sense inventory.

In this paper, we propose another way to construct the thesaurus. We use statistical machine translation (SMT) techniques to extract paraphrase pairs from bilingual parallel text. In this way, we avoid calculating similarities between every pair of words and could find semantic similar words or compounds which have dissimilar distributions.

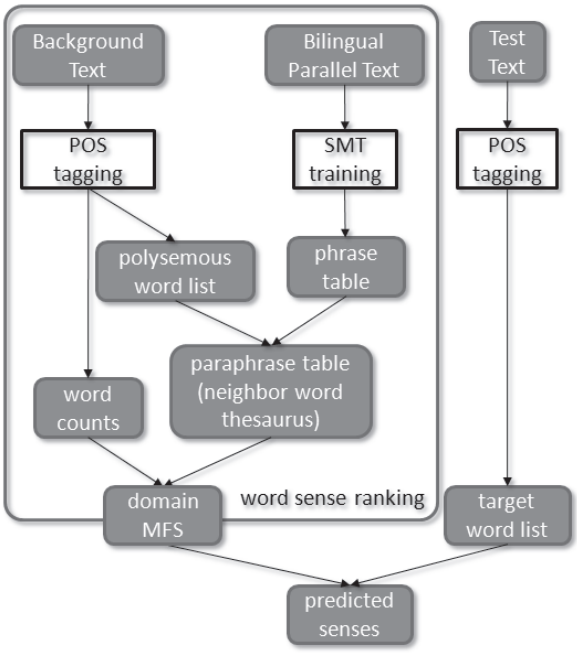


Figure 1: The architecture of HIT-CIR

Our system is comprised of two parts: the word sense ranking part and the word sense tagging part. Senses are ranked according to their prevalence score in the target domain, and the predominant sense is used to tag the occurrences of the target word in the test data. The architecture of this system is shown in Figure 1.

The word sense ranking part includes following steps.

1. Tag the POS of the background text, count the word frequency in each POS, and get the polysemous word list of the POS.

2. Using SMT techniques to extract phrase table



*Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 407–410,
Uppsala, Sweden, 15-16 July 2010. 

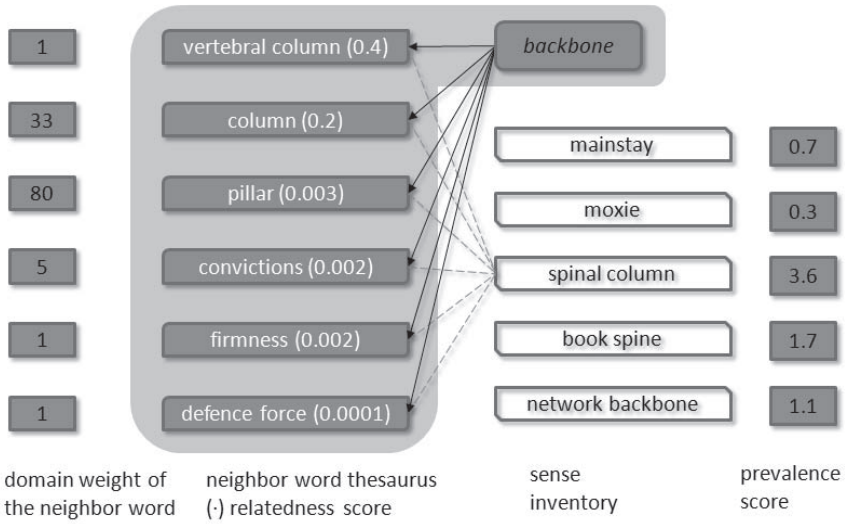


Figure 2: Word sense ranking for the noun *backbone*

from the bilingual corpus. Extract the paraphrases (called as neighbor words) with the phrase table for each word in the polysemous word list.

3. Calculate the prevalence score of each sense of the target words, rank the senses with the score and obtain the predominant sense.

We applied our system on the English data set of SemEval-2 specific domain WSD task. This task is an all word WSD task in the environmental domain. We employed the domain background raw text provided by the task organizer as well as the English WordNet 3.0 (Fellbaum, 1998) and the English-Spanish parallel corpus from Europarl (Koehn, 2005).

This paper is organized as follows. Section 2 introduces how to rank word senses. Section 3 presents how to obtain the most related words of the target words. We describe the system settings in Section 4 and offer some discussions in Section 5.

## 2 Word Sense Ranking

In our method, word senses are ranked according to their prevalence score in the specific domain. According to the assumption of McCarthy et al. (2007), the prevalence score is affected by the following two factors: (1) The relatedness score between a given sense of the target word and the target word's neighbor word. (2) The similarity between the target word and its neighbor word. In addition, we add another factor, (3) the importance of the neighbor word in the specific domain. In this paper, "neighbor words" means the words which are most semantically similar to the target word.

Figure 2 illustrates the word sense ranking process of noun *backbone*. The contribution of a neighbor word to a given word sense is measured by the similarity between them and weighted by the importance of the neighbor word in the target domain and the relatedness between the neighbor word and the target word. Sum up the contributions of each neighbor words, and we get the prevalence score of the word sense.

Formally, the prevalence score of sense $s_i$ of a target word $w$ is assigned as follows:

$$ps(w, s_i) = \sum_{n_j \in N_w} rs(w, n_j) \times ns(s_i, n_j) \times dw(n_j) \quad (1)$$

where

$$ns(s_i, n_j) = \frac{sss(s_i, n_j)}{\sum_{s_{i'} \in senses(w)} sss(s_{i'}, n_j)}, \quad (2)$$

$$sss(s_i, n_j) = max_{s_x \in senses(n_j)} sss'(s_i, s_x). \quad (3)$$

$rs(w, n_j)$ is the relatedness score between $w$ and a neighbor word $n_j$. $N_w = \{n_1, n_2, \ldots, n_k\}$ is the top $k$ relatedness score neighbor word set. $ns(s_i, n_j)$ is the normalized form of the sense similarity score between sense $s_i$ and the neighbor word $n_j$ (i.e. $sss(s_i, n_j)$). We define this score with the maximum WordNet similarity score between $s_i$ and the senses of $n_j$ (i.e. $sss'(s_i, n_j)$). In our system, lesk algorithm is used to measure the sense similarity score between word senses.

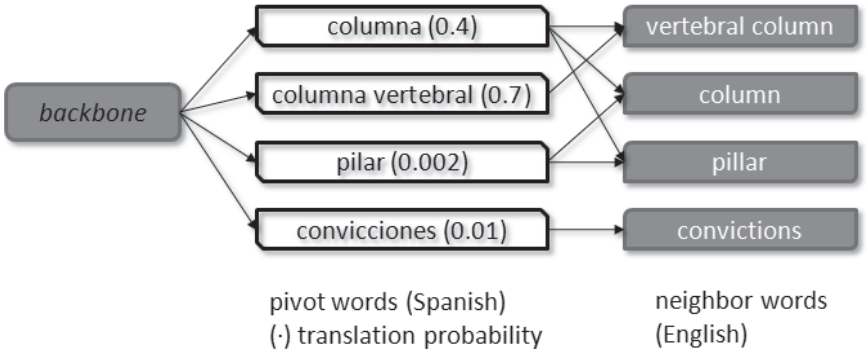


Figure 3: Finding the neighbor words of noun *backbone*

The similarity of this algorithm is the count of the number of overlap words in the gloss or the definition of the senses (Banerjee and Pedersen, 2002). The domain importance weight $dw(n_j)$ is assigned with the count of $n_j$ in the domain background corpus. For the neighbor word that does not occur in the domain background text, we use the *add-one* strategy. We will describe how to obtain $n_j$ and $rs$ in Section 3.

## 3 Thesaurus Construction

The neighbor words of the target word as well as the relatedness score are obtained by extracting paraphrases from bilingual parallel texts. When a word is translated from source language to target language and then translated back to the source language, the final translation may have the same meaning to the original word but with different expressions (e.g. different word or compound). The translation in the same language could be viewed as a paraphrase term or, at least, related term of the original word.

For example, in Figure 3, English noun *backbone* can be translated to *columna*, *columna vertebral*, *pilar* and *convicciones* etc. in Spanish, and these words also have other relevant translations in English, such as *vertebral column*, *column*, *pillar* and *convictions* etc., which are semantically related to the target word *backbone*.

We use a statistical machine translation system to calculate the translation probability from English to another language (called as pivot language) as well as the translation probability from that language to English. By multiplying these two probabilities, we get a paraphrase probability. This method was defined in (Bannard and Callison-Burch, 2005).

In our system, we choose the top $k$ paraphrases as the neighbor words of the target word, which have the highest paraphrase probability. Note that there are two directions of the paraphrase, from target word to its neighbor word and from the neighbor word to the target word. We choose the paraphrase score of the former direction as the relatedness score ($rs$). Because the higher of the score in this direction, the target word is more likely paraphrased to that neighbor word, and hence the prevalence of the relevant target word sense will be higher than other senses. Formally, the relatedness score is given by

$$rs(w, n_j) = \sum_f p(f|w)p(n_j|f), \qquad (4)$$

where $f$ is the pivot language word.

We use the English-Spanish parallel text from Europarl (Koehn, 2005). We choose Spanish as the pivot language because in the both directions the BLEU score of the translation between English and Spanish is relatively higher than other English and other languages (Koehn, 2005).

## 4 Data set and System Settings

The organizers of the SemEval-2 specific domain WSD task provide no training data but raw background data in the environmental domain. The English background data is obtained from the official web site of World Wide Fund (WWF), European Centre for Nature Conservation (ECNC), European Commission and the United Nations Economic Commission for Europe (UNECE). The size of the raw text is around 15.5MB after simple text cleaning. The test data is from WWF and ECNC, and contains 1398 occurrence of 436 target words.

For the implementation, we used bpos (Shen et al., 2007) for the POS tagging. The maximum

number of the neighbor word of each target word $k$ was set to 50. We employed Giza++[1] and Moses[2] to get the phrase table from the bilingual parallel corpus. The WordNet::Similarity package[3] was applied for the implement of the lesk word sense similarity algorithm.

For the target word that is not in the polysemous word list, we use the MFS from WordNet as the back-off method.

## 5 Discussion and Future Work

The recall of our system is 43.5%, which is lower than that of the MFS baseline, 50.5% (Agirre et al., 2010). The baseline uses the most frequent sense from the SemCor corpus (i.e. the MFS of WordNet). This means that for some target words, the MFS from SemCor is better than the domain MFS we estimated in the environmental domain. In the future, we will analysis errors in detail to find the effects of the domain on the MFS.

For the domain specific task, it is better to use parallel text in the domain of the test data in our method. However, we didn't find any available parallel text in the environmental domain yet. In the future, we will try some parallel corpus acquisition techniques to obtain relevant corpus for environmental domain for our method.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) via grant 60803093, 60975055, the "863" National High-Tech Research and Development of China via grant 2008AA01Z144, and Natural Scientific Research Innovation Foundation in Harbin Institute of Technology (HIT.NSRIF.2009069).

## References

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.*

Satanjeev Banerjee and Ted Pedersen. 2002. 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*, pages 136–145, London, UK. Springer-Verlag.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604, Morristown, NJ, USA. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit*, Phuket, Thailand.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 419–426, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590, December.

G. A. Miller, C. Leacock, R. Tengi, and R. Bunker. 1994. A semantic concordance. In *Proc. ARPA Human Language Technology Workshop '93*, pages 303–308, Princeton, NJ, March. distributed as *Human Language Technology* by San Mateo, CA: Morgan Kaufmann Publishers.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35, Prague, Czech Republic, June. Association for Computational Linguistics.

Libin Shen, Giorgio Satta, and Aravind Joshi. 2007. Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 760–767, Prague, Czech Republic, June. Association for Computational Linguistics.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

[1] http://www.fjoch.com/GIZA++.html

[2] http://www.statmt.org/moses/

[3] http://wn-similarity.sourceforge.net/