HIT-WSD: Using Search Engine for Multilingual Chinese-English Lexical Sample Task

PengYuan Liu, TieJun Zhao, MuYun Yang
MOE-MS Key Laboratory of NLP & Speech, HIT, School of Computer Science and
Technology, Harbin Institute of Technology, Harbin, Heilongjiang, China
{pyliu,tjzhao,ymy}@mtlab.hit.edu.cn

Abstract

We have participated in the Multilingual Chinese-English Lexical Sample Task of SemEval-2007. Our system disambiguates senses of Chinese words and finds the correct translation in English by using the web as WSD knowledge source. Since all the statistic data is obtained from search engine, the method is considered to be unsupervised and does not require any sense-tagged corpus.

1 Introduction

Due to the lack of sense tagged corpora (and the difficulty of manually creating them), the unsupervised method tries to avoid, or at least to reduce, the knowledge acquisition problem, which the supervised methods have to deal with. In order to tackle the problem of the knowledge acquisition bottleneck, we adopted an unsupervised approach based on search engine, which does not require any sense tagged corpus.

The majority of methods using the Web often try to automatically generate sense tagged corpora (Agirre and Martinez 2000;Agirre and Martinez 2004;Gonzalo et al. 2003; Mihalcea and Moldovan 1999;Santamaria et al. 2003). In this paper, we experiment with our initial attempt on another research trend that uses the Web not for extracting training samples but helping disambiguate directly during the translation selection process.

The approach we present here is inspired by (Mihalcea and Moldovan 1999;Brill 2003; Rosso et al. 2005; Dagan et al. 2006; McCarthy 2002). Suppose that source ambiguous words are apt to appear with its target translation on bilingual web pages either parallel or non-parallel. Instead of searching the source language or target language respectively on web, we try to let the search engine think in a bilingual style. First, our system gets the co-occurrence information of Chinese context and its corresponding English context. Then it computes association measurements of Chinese context and English context in 4 kinds of way. Finally, it selects the correct English translation by computing the association measurements.

In view that this is the first international standard evaluation to predict the correct English translation for ambiguous Chinese word, we built HIT-WSD system as our first attempt on disambiguation by using bilingual web search and just want to testify validity of our method.

2 HIT-WSD System

2.1 Disambiguation Process

HIT-WSD system disambiguates senses of Chinese target ambiguous word and finds the correct translation in English by searching bilingual information on the web. Figure 1 gives the flowchart of our proposed approach. Given an ambiguous word with a Chinese sentence, we easily create its Chinese context. English context can be acquired from a Chinese-English dictionary and the translation mapping set(offered by the Multilingual Chinese-English Lexical Sample Task). System puts Chinese context and English context as queries on search engine individually and collectively. After this step, frequency and co-occurrence frequency of Chinese context and English

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 169–172,
Prague, June 2007. ©2007 Association for Computational Linguistics

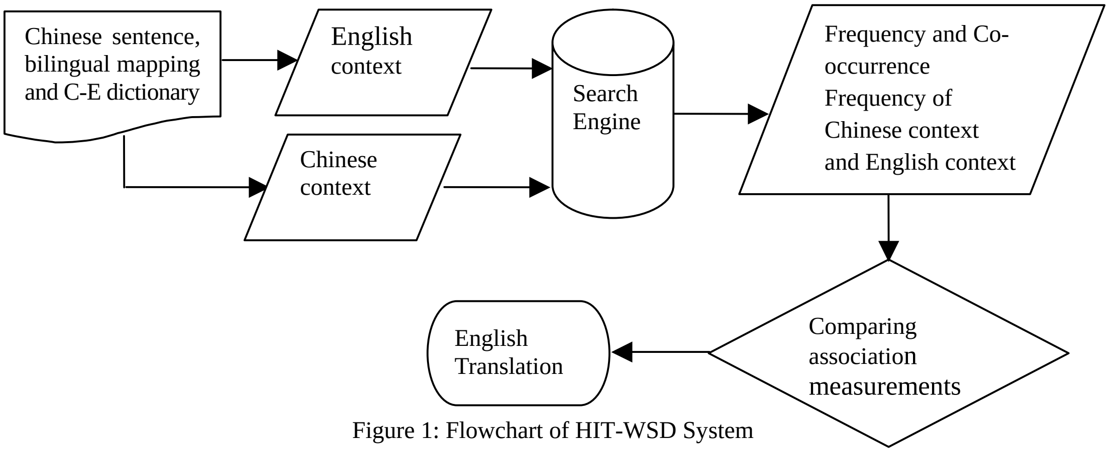
Chinese sentence, bilingual mapping and C-E dictionary
English context
Chinese context
Search Engine
Frequency and Co-occurrence Frequency of Chinese context and English context
Comparing association measurements
English Translation

Figure 1: Flowchart of HIT-WSD System

context will be found. Finally, our system selects the most probable English translation by computing association measurements.

Figure 2 gives an example of how the proposed approach selects English translations of the Chinese ambiguous word “动摇/dongyao” given the sentence and its translation mapping set. This instance comes from the training data of Multilingual Chinese-English Lexical Sample Task of Semeval2007. According to the translation mapping set, Chinese target word “动摇/dongyao” has two English Translations: shake and vacillate.

English Context Candidates set is the translations set of the Chinese context. System uses translation mapping set to translate Chinese target ambiguous word and uses an Chinese-English dictionary to translate other words in Chinese context. English Context Candidates set could be any combination of translations and each combination could be selected as the English context.

After getting the Chinese context and English context, we put them as queries to search engine and extract page counts (which can be considered as frequency) which search engine returned. We not only search Chinese context and English context individually, but also put them together to search engine.

Association measurements: the Dice coefficient, point-wise mutual information, Log Likelihood score and χ^2 score are computed in the third phase while we got all kinds of statistic results from search engine. Finally, we determine the translation by simply computing the association measurements

Instance: 事实证明了邓小平同志对形势发展的判断，证明了坚持基本**路线不**<head>***动摇***</head>**是实现**中国现代化的根本保证。
Chinese Ambiguous Word: 动摇
Translation Mapping Set: 动摇-shake/动摇-vacillate
Translations of Chinese context in Chinese-English dictionary:不/not,是/is,路线/line,实现/ actualize

Chinese Context(CC): 路线不动摇是实现
English Context Candidates set:
Shake, shake is, not shake, line shake…/vacillate, not vacillate, vacillate is, line vacillate…
English Context(EC): shake/vacillate
Putting on Search Engine and getting counts:
$c(shake) = 1880000, c(vacillate) = 5450$

$c(CC) = 113000, c(CC, shake) = 77, c(CC, vacillate) = 12$

Computing association measurements:

$$Dice(CC, shake) = \frac{2 \times c(CC, shake)}{((c(CC, shake) + c(shake)) \times (c(CC, shake) + c(CC)}$$

$$= \frac{2 \times 77}{(77 + 1880000) \times (77 + 113000))} = 7.24e - 10$$

$$Dice(CC, vacillate) = \frac{2 \times c(CC, vacillate)}{((c(CC, vacillate) + c(vacillate)) \times (c(CC, vacillate) + c(CC))}$$

$$= \frac{2 \times 12}{(12 + 5450) \times (12 + 11300)} = 3.89e - 8$$

Compare and Determine a Translation:
3.89e-8>7.24e-10, So the answer is **vacillate**.

Figure 2: Example of the Chinese ambiguous word “动摇/dongyao” selection process

2.2 Experiment Settings

Although the Chinese context can be represented with local features, topic features, parts of speech and so on, we use sentence segment as Chinese context in our experiment system. The sentence segment is a window size ± n segment of the sentence including the ambiguous words.

English Context Candidates set could be any combination of the translation of words appearing in Chinese context. In our experiment system, we just choose the translation of the Chinese target ambiguous words in the translation mapping set as English context.

We choose google[1] and baidu[2] as our search engine, for they are both most widely used for English and Chinese language respectively.

Putting Chinese context and English context as queries to the search engine, we will get corresponding page counts it returned as figure 2 shows.

Four statistical measurements were used in order to measure the degree of association of Chinese Context (CC) and English Context (EC). CC and EC can be seen as two random events occuring in the web pages:

1. Point-wise mutual information:

$$MI(CC, EC) = \log_2 \frac{n \times a}{(a+b)\times(a+c)} \quad (1)$$

2. DICE coefficient:

$$DICE(CC, EC) = \frac{2 \times a}{(a+b)\times(a+c)} \quad (2)$$

3. χ^2 score:

$$X^2(CC, EC) = \frac{n \times (a \times d - b \times c)^2}{(a+b)\times(a+c)\times(b+d)\times(c+d)} \quad (3)$$

4. Log Likelihood score:

$$LL(CC, EC) = 2 \times (a \times \log \frac{n \times a}{(a+b)\times(a+c)} + b \times \log \frac{n \times b}{(a+b)\times(b+d)} + c \times \log \frac{n \times c}{(c+d)\times(a+c)} + d \times \log \frac{n \times d}{(c+d)\times(b+d)}) \quad (4)$$

Here is the meaning of *a*, *b*, *c*, *d* and *n*.

a: all counts of the web pages which include Both CC and EC.

b: all counts of the web pages which include CC, do not include EC.

c: all counts of the web pages which include EC, do not include CC.

d: all counts of the web pages which include neither CC and EC.

$n = a + b + c + d$

Association Measurements	Precision(Micro-average)		
	Context Window Size		
	-1,+1	-1,+2	-2,+2
MI(Baidu)	**0.349**	**0.349**	0.339
XX(Baidu)	0.338	0.344	0.314
LL(Baidu)	0.315	0.320	0.293
DICE(Baidu)	0.285	0.295	0.295
MI(google)	0.334	0.334	0.339
XX(google)	0.322	0.316	0.316
LL(google)	0.295	0.306	0.299
DICE(google)	0.281	0.278	0.272

Table 1:Training data results of Multilingual Chinese-English Lexical Sample Task

	Micro-average	Macro-average
Our result	0.336898	0.395993
Baseline (MFS)	0.4053	0.4618

Table 2:Official results: Multilingual Chinese-English Lexical Sample Task

We applied our method to the training data of Multilingual Chinese-English Lexical Sample Task. The results are as showed in Table 1.

Since only one test result can be uploaded for one system, our system selects the settings of one of the best results. The final settings of our system is: window size is [-1, +2], the search engine is baidu and the association measurement is Point-wise mutual information.

3 Official Results

In multilingual Chinese-English lexical sample task of SemEval-2007, there are 2686 instances in training data for 40 Chinese ambiguous words. All these ambiguous words are either nouns or verbs. Test data consist of 935 untagged instances of the same target words.

The official result of our system in multilingual Chinese-English lexical sample task is reported as in Table 2.

[1] www.google.com.

[2] www.baidu.com.

4 Conclusions

In SemEval-2007, we participated in Multilingual Chinese-English Lexical Sample Task with a fully unsupervised system based on bilingual web search. Our initial experiment result shows that our system fails to reach MFS (Most Familiar Sense) baseline due to our method is too simple where search queries are formed (just uses simple context window and English target translation). Our approach is the first attempt so far as we know on using bilingual web search for translation selection directly. The system is very simple but seemed to achieve a not bad performance when considered the performance of fully unsupervised systems in SENSEVAL-2, SENSEVAL -3 English tasks.

For future research, we will investigate the dependency of bilingual documents, optimize the search queries, filter out potential noises and combine the different results in order to devise an improved method that can utilize bilingual web search better.

References

Agirre, E.and Martinez, D. 2000. *Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web.* Proc. of the COLING-2000.

Agirre, E.and Martinez, D. 2004. *Unsupervised word sense disambiguation based on automatically retrieved examples: The important of bias.* Proc. of the EMNLP 2004(Barcelona, Spain, July 2004).

Brill, E. 2003. *Processing Natural Language Processing without Natural Language Processing.* Lecture Notes in Computer Science, Vol. 2588. Springer-Verlag (2003) 360–369.

Dagan, I., Glickman, O., Gliozzo, A., Marmorshtein, E. and Strapparava, C. 2006. *Direct Word Sense Matching for lexical substitution*. Proceedings of ACL/COLING 2006.

Gonzalo, J., Verdejo, F. and Chugar, I. 2003. *The Web as a Resource for WSD*. 1st MEANING Workshop, Spain.

McCarthy, D. 2002. *Lexical Substitution as a Task for WSD Evaluation.* In Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia, USA.

Mihalcea, R. and Moldovan, D.I. 1999. *An Automatic Method for Generating Sense Tagged Corpora.* Proc. of the 16th National Conf. on Artificial Intelligence. AAAI Press.

Rosso, P., Montes, M., Buscaldi, D., Pancardo, A., and Villase, A., 2005. *Two Web-based Approaches for Noun Sense Disambiguation.* Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2005, Springer Verlag, LNCS (3406), Mexico D.F., Mexico, pp. 261-273

Santamaria, C., Gonzalo, J. and Verdejo, F. 2003. *Automatic Association of WWW Directories to Word Senses.* Computational Linguistics (2003), Vol. 3, Issue 3 – Special Issue on the Web as Corpus, 485–502.