International Journal of Knowledge and Language Processing
Volume 3, Number 3, **2012**

www.ijklp.org
KLP International ©2012 ISSN 2191-2734
pp. **46-52**

Word Translation Disambiguation Based on Source-Target Dictionary and Target Language Corpus

Liu Pengyuan
Beijing Language and Culture University
No.15, Xueyuan Road, Haidian District, Beijing, China
{ liupengyuan}@blcu.edu.cn

Received June 2012; revised July 2012

ABSTRACT. *This paper introduces a word translation disambiguation method which uses a source-target dictionary and untagged target language corpus. To every ambiguous source language word, first it gets the translations of it. Second, to every translation, instances of that translation can be retrieved easily from an untagged target language corpus. Then these instances which included translations can be seemed as sense-tagged instances of source word. Every source ambiguous word can get many target intances tagged with translations. Finally a classifier can be constructed by using these instances and it can be used for source word translation disambiguation. The experiment result on English lexical sample task of Senseval-2 shows that the performance of the method is good. The recall is 46.7% which outperforms the best unsupervised system of that task.*
Keywords: Word Translation Disambiguation; target language; unsupervised

1. **Introduction.** Word sense disambiguation (WSD) is the task of automatically determining the correct sense for a target word given the context in which it occurs. Word in different source language context may usually have different target language translations. Similarly, the task of determining the correct translation for a target source language word given the context in which it occurs is called word translation disambiguation(WTD). WSD and WTD are some kind of similar task and both are important problems in NLP and the essential preprocessing steps for many applications, including machine translation, question answering and information extraction. However, they both are difficult tasks, and despite the fact that many research have been done over the years, state of-the-art systems are still not good enough for real-task applications. One major factor that makes them difficult is the knowledge acquire bottleneck - lack of manually annotated corpora, which supervised systems heavily rely on.

To address this problem of lack of manually annotated corpora, there has been a significant amount of work on WSD. Among these research lines, one potential research area is to acquire training examples automatically[1-5]. All these research work are focus on the WSD task and the acquisition of example only in source language category. WTD

task is a some kind of different task from WSD especially in the application of machine translation and trans-language information retrieval.

Unlike Wang and Carroll[6], we retrieve translation examples of the source word in target language directly. This work focus on the WTD task which needs to know the distribution information of the context of translation in target language. We try to acquire translation examples automatically and so it is unsupervised. This work is a further research based on our previous work[7].

First, to every ambiguous source language word, the translations of it can be got easily from any source-target dictionary. Second, to every translation, instances of that translation can be retrieved easily from an untagged target language corpus. Then these instances which included translations can be seemed as sense-tagged instances of source word. Every source ambiguous word can get many target instances tagged with translations. Finally a classifier can be constructed by using these instances and it can be used for source word translation disambiguation. We test our method on the English lexical sample task of Senseval-2[8]. The result of experiment shows that the performance of the method is good. The recall is 46.7% which outperforms the best unsupervised system of that task.

This paper is structured as follows. First, Section 2 describes our method and how to retrieve translation examples of the source work from untagged target language corpus by a source-target dictionary. Section 3 and 4 introduces the NBC classifier and the baselines on the English lexical sample task of Senseval-2. Section 5 is the experiment, evaluation and discuss. Finally the conclusions are given in Section 6.

2. **Method.** Generally the source of translations in the task of WTD can be got from source-target dictionary easily. From these translations we can get the examples from search target language corpus or by searching the web. These target language examples can be seemed as tagged instances of which the tags are the target translations of the source word. A classifier can be trained by these tagged instances. When a new instance comes, it can be disambiguate by the classifier.

Taking source ambiguous English word e as a example, we introduce the procedure of how to acquire the corresponding Chinese translation examples.

Algorithm 2.1.

Step 1. Getting the Chinese translations of the English word e and form the translation set $C=\{c_1,c_2\cdots\cdots,c_n\}$. In which n is the total amount of the Chinese translations of e.

Step 2. To every c_i in C do:

Step 2.1. Building a set of Chinese translation examples CT_i, setting it empty;

Step 2.2. Searching and getting the sentences s_i included c_i by set c_i as the target in the untagged Chinese corpus;

Step 2.3. Putting s_i into CT_i;

Step 3. Throwing out the repeat examples in CT_i:

After running algorithm 2.1, to any ambiguous English words, the corresponding Chinese language corpus CT_i in which every instance included its translations has been

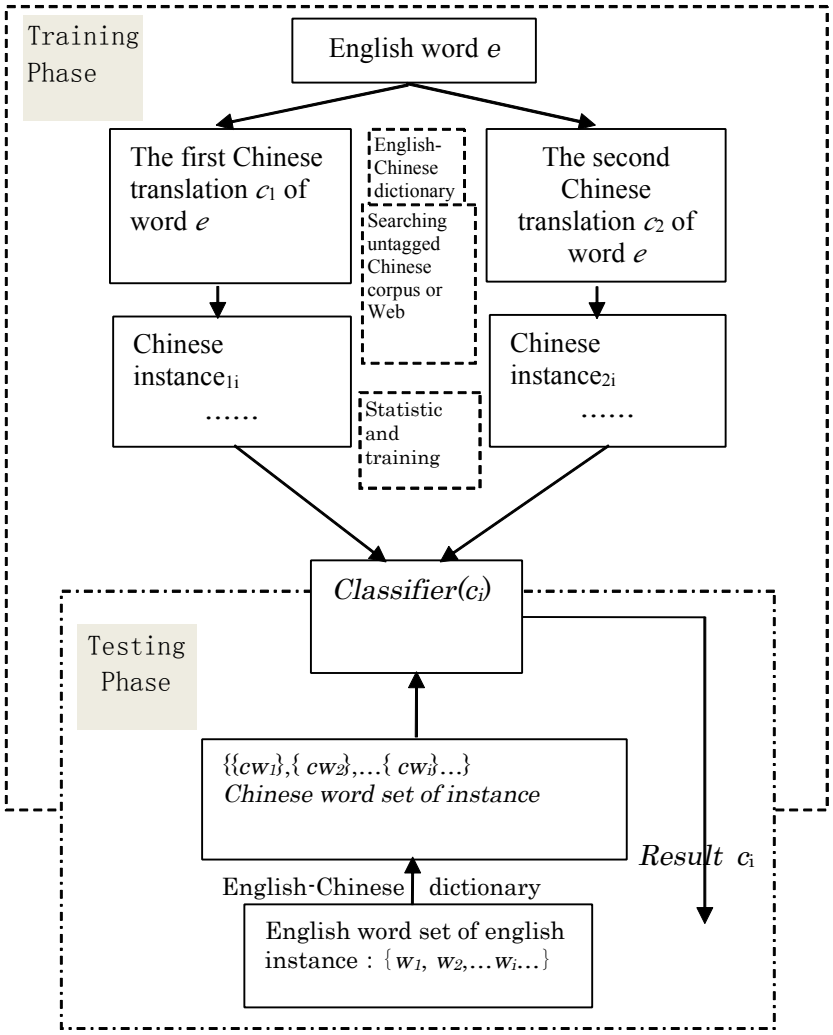

FIGURE 1. The framework of the method.

gotten. For nowadays large-scale untagged monolingual corpus like English, Chinese, German and Russian can be acquired easily. Even if it is not enough, it still can be acquired from mining web[9].

Regarding CT_i as training dataset, we can utilize any available machine learning method to form a classifier for every target English ambiguous word. After that we can use the classifier to disambiguate any Chinese example included the translation whereas in real application we need to determine which is the most appropriate Chinese translation of English word in a given English example. Therefore this classifier can not be used directly. We use an English-Chinese translation dictionary to translate every words w_i(except e) of any english example s included ambiguous word e into Chinese. Then we get the set $\{cw_i\}$. It can be seemed as the corresponding Chinese instance and we can use the classifier to find the most appropriate translation for that Chinese instance. The framework of the method can be described by figure 1.

Based on this framework, any machine learning method can be chosen and then training by the bag of word feature in $\{cw_i\}$.

3. **Naïve Bayesian classifier.** For a naïve Bayesian classifier, the joint probability of observing a certain combination of context features with a particular sense is expressed as:

$$p(F_1, F_2, ...F_n, S) = p(S)\prod_{i=1}^{n} P(F_i \mid S) \quad (1)$$

In (1), (F_1, F_2, … , F_n) is feature variables, S is classification variable and $p(S)$ is the prior probability of classification variable. Any parameter that has a value of zero indicates that the associated word never occurs with the specified sense value. These zero values are smoothed by additive smoothing method as expressed below:

$$P(F_i \mid S_k) = \frac{C(F_i, S_k) + \lambda}{C(S_k) + N \cdot \lambda}, \quad \lambda \in (0,1) \quad (2)$$

In (2), λ is the smoothness variable. $C(S_k)$ is the times of instances with S_k label. $C(F_i, S_k)$ is the concurrences times of F_i and S_k. N is the times of total words in the corpus.

4. **The baseline system and the dataset.** First we want to compare our method to the three other works which also focus on acquiring sense-tagged example automatically.

1)Agirre and Martinez. It explores the large-scale acquisition of sense-tagged examples for Word Sense Disambiguation (WSD). They have applied the "WordNet monosemous relatives" method to construct automatically a web corpus that they have used to train disambiguation systems.[5]

2)Wang and Carroll. It present a novel almost-unsupervised approach to the task of Word Sense Disambiguation (WSD). We build sense examples automatically, using large quantities of Chinese text, and English-Chinese and Chinese-English bilingual dictionaries, taking advantage of the observation that mappings between words and meanings are often different in typologically distant languages.[6]

3)our previous work. It describes an unsupervised translation disambiguation method based on the Equivalent Pseudo Translation (EPT). EPT is constructed by using non-ambiguous words of target language, which is semantically equivalent to the source ambiguous words. Sense-tagged examples are automatically extracted from a large scale Chinese corpus, by which a semantic classifier of EPT is formed. In order to apply the EPT classifier, English examples are mapped into a set of Chinese words by Hownet.[7]

All these three methods utilized the non-ambiguous words to replace the original words first. Our method doesn't. All these three methods were evaluated on the noun of English lexical sample task of Senseval-2 therefore we also use this dataset to evaluate our method. We also choose the best competition system UNED[10] of that task as a baseline 4th.

Our method also needs Chinese corpus. We use the People's Daily Corpus and Nist MT 2004 training data(1.4Gbytes). The People's Daily Corpus we used are listed in table 1.

TABLE 1. Details of People's Daily Corpus

Time(Year)	File Num.	MBytes
1993	8	45.9
1994	4	41.6
1995	12	48.5
1996	12	44.5
1998(1-6months)	6	54.6
2000	1	45.2
Total	43	280.3

5. **Experiment and discuss.** We use the same translation set as baseline 2nd and retrieve Chinese examples from People's Daily Corpus first. If it is not enough then we use Nist MT corpus. All other settings of experiment of our method is similar to baseline 2nd except we use Hownet to get the translation set $\{cw_i\}$ from test examples whereas baseline 2nd use a translation system to translate all the Chinese examples it acquired.

Before we use the translation set $\{cw_i\}$, we set average weight for every translation based on its translation number in the synset of Hownet.

The result of our experiment is listed in table 2. It shows that our method is good and it outperforms the best system UNED. From the recall of each word we can compare method and [6]. The knowledge and information acquired from Chinese examples is useful and at the same time it is different from that of acquiring from monosemous relatives examples.

Our method forms the classifier directly after it acquired the target language examples so that the classifier does not include the noise such like [5] which include many errors through translating target language examples to source language.

Comparing to our previous work[7], it shows that although acquiring examples through the non-ambiguous target word can lessen the noise data, the examples included direct translation are slightly better.

TABLE 2 Recall(%) on Senseval-2 ELS

Word	[7]	[5]	[6]	[10]	OUR
art	37.8	40.8	45.6	50.1	41.8
authority	19.6	19.6	40.0	34.8	21.7
bar	13.2	37.7	26.4	27.8	35.8
bum	46.7	44.4	57.5	11.1	40.0
chair	84.1	72.5	69.4	81.2	49.3
channel	31.5	31.5	30.9	17.8	27.4
child	40.6	34.4	34.7	43.8	45.3
church	75.0	51.6	49.7	62.5	59.4
circuit	55.3	63.5	49.1	55.3	38.8
day	9.7	15.2	12.5	20.0	13.1
detention	90.6	71.9	87.5	78.1	96.9
dyke	89.3	82.1	80.4	35.7	78.6
facility	31.0	22.4	22.0	25.9	43.1
fatigue	74.4	55.8	75.0	86.0	69.8

feeling	9.8	19.6	42.5	60.8	56.9
grip	58.8	33.3	28.2	21.6	26.5
hearth	75.0	46.9	60.4	65.5	40.6
holiday	48.4	61.3	72.2	54.8	58.1
lady	67.9	41.5	23.9	58.5	54.7
material	37.7	47.8	52.3	53.6	34.8
mouth	10.0	48.3	46.5	48.3	40.0
nation	35.1	29.7	80.6	70.3	62.2
nature	32.6	34.8	34.1	23.9	23.9
post	41.8	38.0	47.4	41.8	50.6
restraint	24.4	26.7	31.4	17.8	22.2
sense	32.1	41.5	41.9	30.2	54.7
spade	66.7	54.5	85.5	54.5	45.5
stress	15.4	17.9	27.6	20.5	41.0
yew	82.1	82.1	77.8	71.4	82.1
Average	46.1	43.7	43.2	46.4	**46.7**

6. **Conclusions and future work.** This paper introduces a word translation disambiguation method which uses a source-target dictionary and untagged target language corpus. The experiment result on English lexical sample task of Senseval-2 shows that the performance of the method is good. The recall is 46.7% which outperforms the best unsupervised system of that task. The result shows that the disambiguation knowledge can be acquired by mining target language corpus. We will use machine translation system on Google/baidu/Bing to translate source language test examples into target language. We hope it will works because the machine translation system have already resolve sense-ambiguous partially.

Acknowledgment. This work is supported by the project of National Natural Science Foundation of China (No.60903063) and the Fundamental Research Funds for the Central Universities. The author also gratefully acknowledges the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] C. Leacock, M. Chodorow, G.A. Miller, Using Corpus Statistics and WordNet Relations for Sense Identification. Computational Linguistics, 1998, 24(1):147~166.

[2] R. Mihalcea and I. Moldovan. A Method for Word Sense Disambiguation of Unrestricted Text. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics, ACL, 1999:152~158.

[3] R. Mihalcea and I. Moldovan. An Automatic Method for Generating Sense Tagged Corpora. In Proceedings of the Conference of the American Association for Articial Intelligence, AAAI, 1999:461~466.

[4] R. Mihalcea. Bootstrapping Large Sense Tagged Corpora. In Proceedings of the International

Conference on Languages Resources and Evaluation, LREC, 2002:1407~1411.

[5] E. Agirre and D. Mart＇inez. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In Proceedings of the International Conference on Empirical Methods in Natural Language Processing, EMNLP, 2004:25~32.

[6] Xinglong Wang and John Carroll. Word Sense Disambiguation Using Sense Examples Automatically Acquired from a Second Language. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, October 2005:547~554.

[7] Liu Peng-yuan, Zhao Tie-Jun, Yang Mu-yun and Li Zhuang. Unsupervised Translation Disambiguation Based on Equivalent PseudoTranslation Model. Journal of Electronics & Information Technology. 2008. Vol30(7):1690~1694.

[8] Adam Kilgarriff. English lexical sample task description. In Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2). Toulouse, France. 2001.

[9] A. Kilgarriff. ‘Web as Corpus’, in Proc. Corpus Linguistic Conf. UCREL-Lancaster Univ, UK. 2001:342~344.

[10] Edmonds, Philip and Adam Kilgarriff, editors. Special Issue on Evaluating Word Sense Disambiguation Systems. Natural Language Engineering. Cambridge University Press, vol. 8(4), 2002.