Base Noun Phrase Translation
Using Web Data and the EM Algorithm

Yunbo Cao
Microsoft Research Asia
i-yuncao@microsoft.com

Hang Li
Microsoft Research Asia
hangli@microsoft.com

Abstract

We consider here the problem of Base Noun Phrase translation. We propose a new method to perform the task. For a given Base NP, we first search its translation candidates from *the web*. We next determine the possible translation(s) from among the candidates using one of the two methods that we have developed. In one method, we employ an ensemble of Naïve Bayesian Classifiers constructed with *the EM Algorithm*. In the other method, we use TF-IDF vectors also constructed with *the EM Algorithm*. Experimental results indicate that the coverage and accuracy of our method are *significantly* better than those of the baseline methods relying on existing technologies.

1. Introduction

We address here the problem of Base NP translation, in which for a given Base Noun Phrase in a source language (e.g., 'information age' in English), we are to find out its possible translation(s) in a target language (e.g., '信息时代' in Chinese).

We define a Base NP as a simple and non-recursive noun phrase. In many cases, Base NPs represent holistic and non-divisible concepts, and thus accurate translation of them from one language to another is extremely important in applications like machine translation, cross language information retrieval, and foreign language writing assistance.

In this paper, we propose a new method for Base NP translation, which contains two steps: (1) translation candidate collection, and (2) translation selection. In translation candidate collection, for a given Base NP in the source language, we look for its translation candidates in the target language. To do so, we use a word-to-word translation dictionary and corpus data in the target language on the web. In translation selection, we determine the possible translation(s) from among the candidates. We use non-parallel corpus data in the two languages on the web and employ one of the two methods which we have developed. In the first method, we view the problem as that of classification and employ an ensemble of Naïve Bayesian Classifiers constructed with the EM Algorithm. We will use 'EM-NBC-Ensemble' to denote this method, hereafter. In the second method, we view the problem as that of calculating similarities between context vectors and use TF-IDF vectors also constructed with the EM Algorithm. We will use 'EM-TF-IDF' to denote this method.

Experimental results indicate that our method is very effective, and the coverage and top 3 accuracy of translation at the final stage are 91.4% and 79.8%, respectively. The results are significantly better than those of the baseline methods relying on existing technologies. The higher performance of our method can be attributed to the enormity of the web data used and the employment of the EM Algorithm.

2. Related Work

2.1 Translation with Non-parallel Corpora

A straightforward approach to word or phrase translation is to perform the task by using *parallel* bilingual corpora (e.g., Brown et al, 1993). Parallel corpora are, however, difficult to obtain in practice.

To deal with this difficulty, a number of methods have been proposed, which make use of relatively easily obtainable non-parallel corpora (e.g., Fung and Yee, 1998; Rapp, 1999; Diab and Finch, 2000). Within these methods, it is usually assumed that a number of translation candidates for a word or phrase are given (or can be easily collected) and the problem is focused on translation selection.

All of the proposed methods manage to find out the translation(s) of a given word or phrase, on the basis of the linguistic phenomenon that the contexts of a translation tend to be similar to the contexts of the given word or phrase. Fung and Yee (1998), for example, proposed to represent the contexts of a word or phrase with a real-valued vector (e.g., a TF-IDF vector), in which one element corresponds to one word in the contexts. In translation selection, they select the translation candidates whose context vectors are the closest to that of the given word or phrase.

Since the context vector of the word or phrase to be translated corresponds to words in the source language, while the context vector of a translation candidate corresponds to words in the target language, and further the words in the source language and those in the target language have a many-to-many relationship (i.e., translation ambiguities), it is necessary to accurately transform the context vector in the source language to a context vector in the target language before distance calculation.

The vector-transformation problem was not, however, well-resolved previously. Fung and Yee assumed that in a specific domain there is only one-to-one mapping relationship between words in the two languages. The assumption is reasonable in a specific domain, but is too strict in the general domain, in which we presume to perform translation here. A straightforward extension of Fung and Yee's assumption to the general domain is to restrict the many-to-many relationship to that of many-to-one mapping (or one-to-one mapping). This approach, however, has a drawback of losing information in vector transformation, as will be described.

For other methods using non-parallel corpora, see also (Tanaka and Iwasaki, 1996; Kikui, 1999, Koehn and Kevin 2000; Sumita 2000; Nakagawa 2001; Gao et al, 2001).

2.2 Translation Using Web Data

Web is an extremely rich source of data for natural language processing, not only in terms of data size but also in terms of data type (e.g., multilingual data, link data). Recently, a new trend arises in natural language processing, which tries to bring some new breakthroughs to the field by effectively using web data (e.g., Brill et al, 2001).

Nagata et al (2001), for example, proposed to collect partial parallel corpus data on the web to create a translation dictionary. They observed that there are many partial parallel corpora between English and Japanese on the web, and most typically English translations of Japanese terms (words or phrases) are parenthesized and inserted immediately after the Japanese terms in documents written in Japanese.

3. Base Noun Phrase Translation

Our method for Base NP translation comprises of two steps: *translation candidate collection* and *translation selection*. In translation candidate collection, we look for translation candidates of a given Base NP. In translation selection, we find out possible translation(s) from the translation candidates.

In this paper, we confine ourselves to translation of noun-noun pairs from English to Chinese; our method, however, can be extended to translations of other types of Base NPs between other language pairs.

3.1 Translation Candidate Collection

We use heuristics for translation candidate collection. Figure 1 illustrates the process of collecting Chinese translation candidates for an English Base NP 'information age' with the heuristics.

1. Input 'information age';
2. Consult English-Chinese word translation dictionary:
   information -> 信息
   age -> 年龄 (how old somebody is)
   时代 (historical era)
   成年 (legal adult hood)
3. *Compositionally* create translation candidates in Chinese:
   信息年龄; 信息时代; 信息成年
4. Search the candidates on web sites in Chinese and obtain the document frequencies of them (i.e., numbers of documents containing them):
   信息时代 10000
   信息年龄 10
   信息成年 0
5. Output candidates having non-zero document frequencies and the document frequencies:
   信息时代 10000
   信息年龄 10

Figure 1. Translation candidate collection

3.2 Translation Selection -- EM-NBC-Ensemble

We view the translation selection problem as that of classification and employ EM-NBC-Ensemble to perform the task. For the ease of explanation, we first describe the algorithm of using only EM-NBC and next extend it to that of using EM-NBC-Ensemble.

Basic Algorithm

Let $\tilde{e}$ denote the Base NP to be translated and $\tilde{C}$ the set of its translation candidates (phrases). Suppose that $|\tilde{C}| = k$. Let $\tilde{c}$ represent a random variable on $\tilde{C}$. Let E denote a set of words in English, and C a set of words in Chinese. Suppose that $|E| = m$ and $|C| = n$. Let e represent a random variable on E and c a random variable on C. Figure 2 describes the algorithm.

Input: $\tilde{e}$, $\tilde{C}$, contexts containing $\tilde{e}$, contexts containing all $\tilde{c} \in \tilde{C}$;

1. create a frequency vector $(f(e_1), f(e_2), \ldots, f(e_m))$, $e_i \in E, (i = 1, \ldots, m)$ using contexts containing $\tilde{e}$; transforming the vector into $(f_E(c_1), f_E(c_2), \ldots, f_E(c_n))$, $c_i \in C, (i = 1, \ldots, n)$, using a translation dictionary and the EM algorithm;

2. **for each** ($\tilde{c} \in \tilde{C}$){
   estimate with Maximum Likelihood Estimation the prior probability $P(\tilde{c})$ using contexts containing all $\tilde{c} \in \tilde{C}$;
   create a frequency vector $(f(c_1), f(c_2), \ldots, f(c_n))$, $c_i \in C, (i = 1, \ldots, n)$ using contexts containing $\tilde{c}$;
   normalize the frequency vector , yielding $(P(c_1 | \tilde{c}), P(c_2 | \tilde{c}), \ldots, P(c_n | \tilde{c})), c_i \in C, (i = 1, \ldots, n)$;
   calculate the posterior probability $P(\tilde{c} | \mathbf{D})$ with EM-NBC (generally EM-NBC-Ensemble), where $\mathbf{D} = (f_E(c_1), f_E(c_2), \ldots, f_E(c_n)), c_i \in C, (i = 1, \ldots, n)$

3. Sort $\tilde{c} \in \tilde{C}$ in descending order of $P(\tilde{c} | \mathbf{D})$;

Output: the top sorted results

Figure 2. Algorithm of EM-NBC-Ensemble

Context Information

As input data, we use 'contexts' in English which contain the phrase to be translated. We also use contexts in Chinese which contain the translation candidates.

Here, a context containing a phrase is defined as the surrounding words within a window of a predetermined size, which window covers the phrase. We can easily obtain the data by searching for them on the web. Actually, the contexts containing the candidates are obtained at the same time when we conduct translation candidate collection (Step 4 in Figure 1).

EM Algorithm

We define a *relation* between E and C as $R \subseteq E \times C$, which represents the links in a translation dictionary. We further define $\Gamma_c = \{e | (e, c) \in R\}$.

At Step 1, we assume that all the instances in $(f(e_1), f(e_2), .., f(e_m))$ are independently generated according to the distribution defined as:

$$P(e) = \sum_{c \in C} P(c) P(e | c) \qquad (1)$$

We estimate the parameters of the distribution by using the Expectation and Maximization (EM) Algorithm (Dempster et al., 1977).

$$\text{E} - \text{Step} \qquad P(c | e) \leftarrow \frac{P(c) P(e | c)}{\sum_{c \in C} P(c) P(e | c)}$$

$$\text{M} - \text{Step} \qquad P(c) \leftarrow \sum_{e \in E} f(e) P(c | e)$$

$$P(e | c) \leftarrow \frac{f(e) P(c | e)}{\sum_{e \in E} f(e) P(c | e)}$$

Figure 3. EM Algorithm

Initially, we set for all $c \in C$

$$P(c) = \frac{1}{|C|},$$

$$P(e | c) = \begin{cases} \frac{1}{|\Gamma_c|}, & \text{if } e \in \Gamma_c \\ 0, & \text{if } e \notin \Gamma_c \end{cases}$$

Next, we estimate the parameters by iteratively updating them, until they converge (cf., Figure 3). Finally, we calculate $f_E(c)$ for all $c \in C$ as:

$$f_E(c) = P(c) \sum_{e \in E} f(e) \qquad (2)$$

In this way, we can transform the frequency vector in English $(f(e_1), f(e_2), .., f(e_m))$ into a vector in Chinese $\mathbf{D} = (f_E(c_1), f_E(c_2), .., f_E(c_n))$.

Prior Probability Estimation

At Step 2, we *approximately* estimate the prior probability $P(\tilde{c})$ by using the document frequencies of the translation candidates. The data are obtained when we conduct candidate collection (Step 4 in Figure 1).

EM-NBC

At Step 2, we use an EM-based Naïve Bayesian Classifier (EM-NBC) to select the candidates $\tilde{c}$ whose posterior probabilities are the largest:

$$\arg\max_{\tilde{c}\in\tilde{C}} P(\tilde{c}\mid \mathbf{D}) = \arg\max_{\tilde{c}\in\tilde{C}}\left(\log P(\tilde{c}) + \sum_{c\in C} f_E(c)\log P(c\mid\tilde{c})\right) \quad (3)$$

Equation (3) is based on Bayes' rule and the assumption that the data in $\mathbf{D}$ are independently generated from $P(c\mid\tilde{c}), c\in C$.

In our implementation, we use an equivalent

$$\arg\min_{\tilde{c}\in\tilde{C}}\left(-\alpha\log P(\tilde{c}) - \sum_{c\in C} f_E(c)\log P(c\mid\tilde{c})\right) \quad (4)$$

where $\alpha \geq 1$ is an additional parameter used to emphasize the prior information. If we ignore the first term in Equation (4), then the use of one EM-NBC turns out to select the candidate whose frequency vector is the closest to the transformed vector $\mathbf{D}$ in terms of KL divergence (cf., Cover and Tomas 1991).

EM-NBC-Ensemble

To further improve performance, we use an ensemble (i.e., a linear combination) of EM-NBCs (EM-NBC-Ensemble), while the classifiers are constructed on the basis of the data in different contexts with different window sizes. More specifically, we calculate

$$P(\tilde{c}\mid\mathbf{D}) = \frac{1}{s}\sum_{i=1}^{s} P(\tilde{c}\mid\mathbf{D}_i) \quad (5)$$

where $\mathbf{D}_i$, (i = 1, … , s) denotes the data in different contexts.

3.3 Translation Selection -- EM-TF-IDF

We view the translation selection problem as that of calculating similarities between context vectors and use as context vectors TF-IDF vectors constructed with the EM Algorithm. Figure 4 describes the algorithm in which we use the same notations as those in EM-NBC-Ensemble.

The *idf* value of a Chinese word *c* is calculated in advance and as

$$idf(c) = -\log(df(c)/F) \quad (6)$$

where $df(c)$ denotes the document frequency of *c* and *F* the total document frequency.

Input: $\tilde{e}$, $\tilde{C}$, contexts containing $\tilde{e}$, contexts containing all $\tilde{c}\in\tilde{C}$, $idf(c), c\in C$;

1. create a frequency vector $(f(e_1), f(e_2), \ldots, f(e_m)), e_i \in E, (i = 1, \ldots, m)$ using contexts containing $\tilde{e}$;
   transforming the vector into $(f_E(c_1), f_E(c_2), \ldots, f_E(c_n)), c_i \in C, (i = 1, \ldots, n)$, using a translation dictionary and the EM algorithm;
   create a TF-IDF vector
   $\mathbf{A} = (f_E(c_1)idf(c_1), \ldots, f_E(c_n))idf(c_n)), c_i \in C, (i = 1, \ldots, n)$

2. **for each** ($\tilde{c}\in\tilde{C}$){
   create a frequency vector $(f(c_1), f(c_2), \ldots, f(c_n)), c_i \in C, (i = 1, \ldots, n)$ using contexts containing $\tilde{c}$;
   create a TF-IDF vector
   $\mathbf{B} = (f(c_1)idf(c_1), \ldots, f(c_n))idf(c_n)) \; c_i \in C, (i = 1, \ldots, n)$;
   calculate $tfidf(\tilde{c}) = cos(\mathbf{A}, \mathbf{B})$; }

3. Sort $\tilde{c}\in\tilde{C}$ in descending order of $tfidf(\tilde{c})$;

Output: the top sorted results

Figure 4. Algorithm of EM-TF-IDF

3.4 Advantage of Using EM Algorithm

The uses of EM-NBC-Ensemble and EM-TF-IDF can be viewed as extensions of existing methods for word or phrase translation using non-parallel corpora. Particularly, the use of the EM Algorithm can help to accurately transform a frequency vector from one language to another.

Suppose that we are to determine if '信息时代' is a translation of 'information age' (actually it is). The frequency vectors of context words for 'information age' and '信息时代' are given in A and D in Figure 5, respectively. If for each English word we only retain the link connecting to the Chinese translation with the largest frequency (a link represented as a solid line) to establish a many-to-one mapping and transform vector A from English to Chinese, we obtain vector B. It turns out, however, that vector B is quite different from vector D, although they should be similar to each other. We will refer to this method as 'Major Translation' hereafter.

With EM, vector A in Figure 5 is transformed into vector C, which is much closer to vector D, as expected. Specifically, EM can split the frequency of a word in English and distribute them into its translations in Chinese in a theoretically sound way (cf., the distributed frequencies of 'internet'). Note that if we assume a many-to-one (or one-to-one) mapping

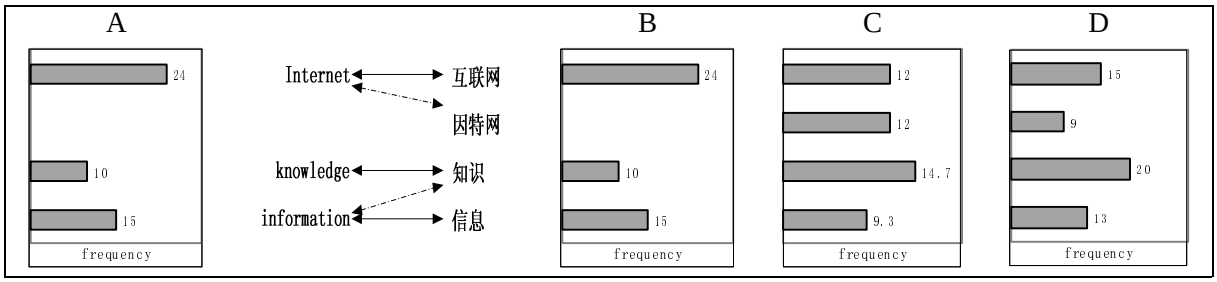

Figure 5. Example of frequency vector transformation

relationship, then the use of EM turns out to be equivalent to that of Major Translation.

3.5 Combination

In order to further boost the performance of translation, we propose to also use the translation method proposed in Nagata et al. Specifically, we combine our method with that of Nagata et al by using a back-off strategy.

1. Input 'information asymmetry';
2. Search the English Base NP on web sites in *Chinese* and obtain documents as follows (i.e., using partial parallel corpora):

> 公司的控制者（所有者）和管理者（我们称为内部人）通常掌握着许多外部投资者所不了解的信息，即在内部人与外部人之间存在信息不对称（information asymmetry）。

3. Find the most frequently occurring Chinese phrases immediately before the brackets containing the English Base NP, using a suffix tree;
4. Output the Chinese phrases and their document frequencies:

信息不对称 5
信息失衡 5

Figure 6. Nagata et al's method

Figure 6 illustrates the process of collecting Chinese translation candidates for an English Base NP 'information asymmetry' with Nagata et al's method.

In the combination of the two methods, we first use Nagata et al's method to perform translation; if we cannot find translations, we next use our method. We will denote this strategy 'Back-off'.

4. Experimental Results

We conducted experiments on translation of the Base NPs from English to Chinese.

We extracted Base NPs (noun-noun pairs) from the Encarta [1] English corpus using the tool developed by Xun et al (2000). There were about 3000 Base NPs extracted. In the experiments, we used the HIT English-Chinese word translation dictionary[2]. The dictionary contains about 76000 Chinese words, 60000 English words, and 118000 translation links. As a web search engine, we used Google (http://www.google.com).

Five translation experts evaluated the translation results by judging whether or not they were acceptable. The evaluations reported below are all based on their judgements.

4.1 Basic Experiment

In the experiment, we randomly selected 1000 Base NPs from the 3000 Base NPs. We next used our method to perform translation on the 1000 phrases. In translation selection, we employed EM-NBC-Ensemble and EM-TF-IDF.

Table 1. Best translation result for each method

	Accuracy (%)		Coverage (%)
	Top 1	Top 3	
EM-NBC-Ensemble	**61.7**	**80.3**	**89.9**
Prior	57.6	77.6	
MT-NBC-Ensemble	59.9	78.1	
EM-KL-Ensemble	45.9	72.3	
EM-NBC	60.8	78.9	
EM-TF-IDF	**61.9**	**80.8**	
MT-TF-IDF	58.2	77.6	
EM-TF	55.8	77.8	

Table 1 shows the results in terms of coverage and top *n* accuracy. Here, coverage is defined as the percentage of phrases which have translations selected, while top *n* accuracy is defined as the percentage of phrases whose selected top *n* translations include correct translations.

For EM-NBC-Ensemble, we set the α in (4) to be 5 on the basis of our preliminary experimental results. For EM-TF-IDF, we used the non-web data described in Section 4.4 to estimate *idf* values of words. We used contexts with window sizes of ±1, ±3, ±5, ±7, ±9, ±11.

[1] http://encarta.msn.com/Default.asp

[2] The dictionary is created by the Harbin Institute of Technology.

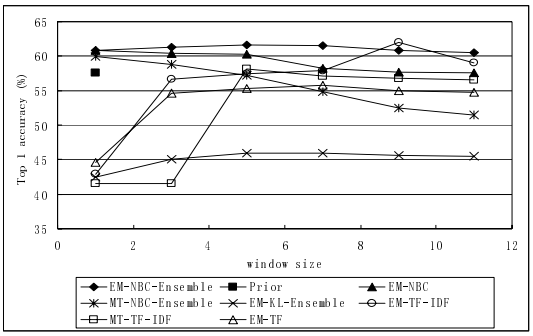

Figure 7. Translation results

Figure 7 shows the results of EM-NBC-Ensemble and EM-TF-IDF, in which for EM-NBC-Ensemble 'window size' denotes that of the largest within an ensemble. Table 1 summarizes the best results for each of them.

'Prior' and 'MT-TF-IDF' are actually baseline methods relying on the existing technologies. In Prior, we select candidates whose prior probabilities are the largest, equivalently, document frequencies obtained in translation candidate collection are the largest. In MT-TF-IDF, we use TF-IDF vectors transformed with Major Translation.

Our experimental results indicate that both EM-NBC-Ensemble and EM-TF-IDF *significantly* outperform Prior and MT-TF-IDF, when appropriate window sizes are chosen. The p-values of the sign tests are 0.00056 and 0.00133 for EM-NBC-Ensemble, 0.00002 and 0.00901 for EM-TF-IDF, respectively.

We next removed each of the key components of EM-NBC-Ensemble and used the remaining components as a variant of it to perform translation selection. The key components are (1) distance calculation by KL divergence (2) EM, (3) prior probability, and (4) ensemble. The variants, thus, respectively make use of (1) the baseline method 'Prior', (2) an ensemble of Naïve Bayesian Classifiers based on Major Translation (MT-NBC-Ensemble), (3) an ensemble of EM-based KL divergence calculations (EM-KL-Ensemble), and (4) EM-NBC. Figure 7 and Table 1 show the results. We see that EM-NBC-Ensemble outperforms all of the variants, indicating that all the components within EM-NBC-Ensemble play positive roles.

We removed each of the key components of EM-TF-IDF and used the remaining components as a variant of it to perform translation selection. The key components are (1) *idf* value and (2) EM. The variants, thus, respectively make use of (1) EM-based frequency vectors (EM-TF), (2) the baseline method MT-TF-IDF. Figure 7 and Table 1 show the results. We see that EM-TF-IDF outperforms both variants, indicating that all of the components within EM-TF-IDF are needed.

Comparing the results between MT-NBC-Ensemble and EM-NBC-Ensemble and the results between MT-TF-IDF and EM-TF-IDF, we see that the uses of the EM Algorithm can indeed help to improve translation accuracies.

Table 2. Sample of translation outputs

Base NP	Translation
calcium ion	钙离子
adventure tale	冒险故事 奇遇故事 冒险传说
lung cancer	肺癌
aircraft carrier	* 飞机承运人
adult literacy	* 成人识字 * 成年识字

Table 2 shows translations of five Base NPs as output by EM-NBC-Ensemble, in which the translations marked with * were judged incorrect by human experts. We analyzed the reasons for incorrect translations and found that the incorrect translations were due to: (1) no existence of dictionary entry (19%), (2) non-compositional translation (13%), (3) ranking error (68%).

4.2 Our Method vs. Nagata et al's Method

Table 3. Translation results

	Accuracy (%)		Coverage (%)
	Top 1	Top 3	
Our Method	61.7	**80.3**	**89.9**
Nagata et al's	**72.0**	76.0	10.5

We next used Nagata et al's method to perform translation. From Table 3, we can see that the accuracy of Nagata et al's method is higher than that of our method, but the coverage of it is lower. The results indicate that our proposed Back-off strategy for translation is justifiable.

4.3 Combination

Table 4. Translation results

	Accuracy （%）		Coverage （%）
	Top 1	Top 3	
Back-off (Ensemble)	62.9	79.7	91.4
Back-off (TF-IDF)	62.2	79.8	

In the experiment, we tested the Back-off strategy, Table 4 shows the results. The Back-off strategy

helps to further improve the results whether EM-NBC-Ensemble or EM-TF-IDF is used.

4.4 Web Data vs. Non-web Data

To test the effectiveness of the use of web data, we conducted another experiment in which we performed translation by using non-web data. The data comprised of the Wall Street Journal corpus in English (1987-1992, 500MB) and the People's Daily corpus in Chinese (1982-1998, 700MB). We followed the Back-off strategy as in Section 4.3 to translate the 1000 Base NPs.

Table 5. Translation results

Data	Accuracy （%）		Coverage （%）
	Top 1	Top 3	
Web (EM-NBC-Ensemble)	**62.9**	**79.7**	**91.4**
Non-web (EM-NBC-Ensemble)	56.9	74.7	79.3
Web (EM-IF-IDF)	**62.2**	**79.8**	**91.4**
Non-web (EM-TF-IDF)	51.5	71.4	78.5

The results in Table 5 show that the use of web data can yield better results than non-use of it, although the sizes of the non-web data we used were considerably large in practice. For Nagata et al's method, we found that it was almost impossible to find partial-parallel corpora in the non-web data.

5. Conclusions

This paper has proposed a new and effective method for Base NP translation by using web data and the EM Algorithm. Experimental results show that it outperforms the baseline methods based on existing techniques, mainly due to the employment of EM. Experimental results also show that the use of web data is more effective than non-use of it.

Future work includes further applying the proposed method to the translation of other types of Base NPs and between other language pairs.

Acknowledgements

We thank Ming Zhou, Chang-Ning Huang, Jianfeng Gao, and Ashley Chang for many helpful discussions on this research project. We also acknowledge Shenjie Li for help with program coding.

References

Brill E., Lin J., Banko M., Dumais S. and Ng A. (2001) *Data-Intensive Question Answering.* In Proc. of TREC '2001.

Brown P.F., Della Pietra, S.A., Della Pietra V.J., and Mercer, R.L. (1993) *The mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics 19(2), pp.263--11.

Cover T. and Thomas J. (1991) *Elements of Information Theory*, Wiley.

Dempster A. P, Laird N. M. and Rubin D. B. (1977) *Maximum likelihood from incomplete data via the EM algorithm*. J. Roy. Stat. Soc. B 39:1--38.

Diab M. and Finch S. (2000) *A statistical word-level translation model for comparable corpora*. In Proc. of RIAO.

Fung P. and Yee L.Y. (1998) *An IR approach for translation new words from nonparallel, comparable texts.* In Proc. of COLING-ACL '1998, pp 414--20.

Gao J. F., Nie J. Y., Xun E. D., Zhang J., Zhou M. and Huang C. N. (2001) *Improving Query Translation for Cross-Language Information Retrieval Using Statistical Models*. In Proc. of SIGIR '2001.

Kikui G. (1999) *Resolving translation ambiguity using non-parallel bilingual corpora*. In Proc. of ACL '1999 Workshop, Unsupervised Learning in NLP.

Koehn P. and Knight K.(2000) *Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm.* In Proc. of AAAI '2000.

Nagata M., Saito T., and Suzuki K. (2001) *Using the Web as a bilingual dictionary*. In Proc. of ACL'2001 DD-MT Workshop.

Nakagawa H. (2001) *Disambiguation of single noun translations extracted from bilingual comparable corpora*. In Terminology 7:1.

Pederson T.(2000) *A Simple Approach to Building Ensembles of Naïve Bayesian Classifiers for Word Sense Disambiguation.* In Proc. of NAACL '2000.

Rapp R. (1999) *Automatic identification of word translations from unrelated English and German corpora.* In Proc. of ACL'1999.

Sumita E.(2000) *Lexical transfer using a vector-space model*. In Proc. of ACL '2000.

Tanaka K. and Iwasaki H. (1996) *Extraction of Lexical Translation from non-aligned corpora.* In Proc. of COLING '1996

Xun E.D., Huang C.N. and Zhou M. (2000) *A Unified Statistical Model for the Identification of English BaseNP*. In Proc. of ACL '2000.