Mining New Word Translations from Comparable Corpora

Li Shao and **Hwee Tou Ng**
Department of Computer Science
National University of Singapore
3 Science Drive 2, Singapore 117543
{shaoli, nght}@comp.nus.edu.sg

Abstract

New words such as names, technical terms, etc appear frequently. As such, the bilingual lexicon of a machine translation system has to be constantly updated with these new word translations. Comparable corpora such as news documents of the same period from different news agencies are readily available. In this paper, we present a new approach to mining new word translations from comparable corpora, by using context information to complement transliteration information. We evaluated our approach on six months of Chinese and English Gigaword corpora, with encouraging results.

1. Introduction

New words such as person names, organization names, technical terms, etc. appear frequently. In order for a machine translation system to translate these new words correctly, its bilingual lexicon needs to be constantly updated with new word translations.

Much research has been done on using parallel corpora to learn bilingual lexicons (Melamed, 1997; Moore, 2003). But parallel corpora are scarce resources, especially for uncommon language pairs. Comparable corpora refer to texts that are not direct translation but are about the same topic. For example, various news agencies report major world events in different languages, and such news documents form a readily available source of comparable corpora. Being more readily available, comparable corpora are thus more suitable than parallel corpora for the task of acquiring new word translations, although relatively less research has been done in the past on comparable corpora. Previous research efforts on acquiring translations from comparable corpora include (Fung and Yee, 1998; Rapp, 1995; Rapp, 1999).

When translating a word *w*, two sources of information can be used to determine its translation: the word *w* itself and the surrounding words in the neighborhood (i.e., the context) of *w*. Most previous research only considers one of the two sources of information, but not both. For example, the work of (Al-Onaizan and Knight, 2002a; Al-Onaizan and Knight, 2002b; Knight and Graehl, 1998) used the pronunciation of *w* in translation. On the other hand, the work of (Cao and Li, 2002; Fung and Yee, 1998; Koehn and Knight, 2002; Rapp, 1995; Rapp, 1999) used the context of *w* to locate its translation in a second language.

In this paper, we propose a new approach for the task of mining new word translations from comparable corpora, by combining both context and transliteration information. Since both sources of information are complementary, the accuracy of our combined approach is better than the accuracy of using just context or transliteration information alone. We fully implemented our method and tested it on Chinese-English comparable corpora. We translated Chinese words into English. That is, Chinese is the source language and English is the target language. We achieved encouraging results.

While we have only tested our method on Chinese-English comparable corpora, our method is general and applicable to other language pairs.

2. Our approach

The work of (Fung and Yee, 1998; Rapp, 1995; Rapp, 1999) noted that if an English word e is the translation of a Chinese word c, then the contexts of the two words are similar. We could view this as a document retrieval problem. The context (i.e., the surrounding words) of c is

viewed as a query. The context of each candidate translation e' is viewed as a document. Since the context of the correct translation e is similar to the context of c, we are likely to retrieve the context of e when we use the context of c as the query and try to retrieve the most similar document. We employ the language modeling approach (Ng, 2000; Ponte and Croft, 1998) for this retrieval problem. More details are given in Section 3.

On the other hand, when we only look at the word *w* itself, we can rely on the pronunciation of *w* to locate its translation. We use a variant of the machine transliteration method proposed by (Knight and Graehl, 1998). More details are given in Section 4.

Each of the two individual methods provides a ranked list of candidate words, associating with each candidate a score estimated by the particular method. If a word e in English is indeed the translation of a word c in Chinese, then we would expect e to be ranked very high in both lists in general. Specifically, our combination method is as follows: we examine the top M words in both lists and find $e_1, e_2, ..., e_k$ that appear in top M positions in both lists. We then rank these words $e_1, e_2, ..., e_k$ according to the average of their rank positions in the two lists. The candidate e_i that is ranked the highest according to the average rank is taken to be the correct translation and is output. If no words appear within the top *M* positions in both lists, then no translation is output.

Since we are using comparable corpora, it is possible that the translation of a new word does not exist in the target corpus. In particular, our experiment was conducted on comparable corpora that are not very closely related and as such, most of the Chinese words have no translations in the English target corpus.

3. Translation by context

In a typical information retrieval (IR) problem, a query is given and a ranked list of documents most relevant to the query is returned from a document collection.

For our task, the query is $C(c)$, the context (i.e., the surrounding words) of a Chinese word c. Each $C(e)$, the context of an English word e, is considered as a document in IR. If an English word e is the translation of a Chinese word c, they will have similar contexts. So we use the query $C(c)$ to retrieve a document $C(e^*)$ that best matches the query. The English word e^* corresponding to that document $C(e^*)$ is the translation of c.

Within IR, there is a new approach to document retrieval called the language modeling approach (Ponte & Croft, 98). In this approach, a language model is derived from each document D. Then the probability of generating the query Q according to that language model, $P(Q \mid D)$, is estimated. The document with the highest $P(Q \mid D)$ is the one that best matches the query. The language modeling approach to IR has been shown to give superior retrieval performance (Ponte & Croft, 98; Ng, 2000), compared with traditional vector space model, and we adopt this approach in our current work.

To estimate $P(Q \mid D)$, we use the approach of (Ng, 2000). We view the document D as a multinomial distribution of terms and assume that query Q is generated by this model:

$$P(Q \mid D) = \frac{n!}{\prod_t c_t!} \prod_t P(t \mid D)^{c_t}$$

where t is a term in the corpus, c_t is the number of times term t occurs in the query Q, $n = \sum_t c_t$ is the total number of terms in query Q.

For ranking purpose, the first fraction $n! / \prod_t c_t!$ can be omitted as this part depends on the query only and thus is the same for all the documents.

In our translation problem, $C(c)$ is viewed as the query and $C(e)$ is viewed as a document. So our task is to compute $P(C(c) \mid C(e))$ for each English word e and find the e that gives the highest $P(C(c) \mid C(e))$, estimated as:

$$\prod_{t_c \in C(c)} P(t_c \mid T_c(C(e)))^{q(t_c)}$$

Term t_c is a Chinese word. $q(t_c)$ is the number of occurrences of t_c in $C(c)$. $T_c(C(e))$ is the

bag of Chinese words obtained by translating the English words in $C(e)$, as determined by a bilingual dictionary. If an English word is ambiguous and has *K* translated Chinese words listed in the bilingual dictionary, then each of the *K* translated Chinese words is counted as occurring *1/K* times in $T_c(C(e))$ for the purpose of probability estimation.

We use backoff and linear interpolation for probability estimation:

$$P(t_c \mid T_c(C(e))) = \alpha \cdot P_{ml}(t_c \mid T_c(C(e))) + (1-\alpha) \cdot P_{ml}(t_c)$$

$$P_{ml}(t_c \mid T_c(C(e))) = \frac{d_{T_c(C(e))}(t_c)}{\sum_{t \in T_c(C(e))} d_{T_c(C(e))}(t)}$$

where $P_{ml}(\bullet)$ are the maximum likelihood estimates, $d_{T_c(C(e))}(t_c)$ is the number of occurrences of the term t_c in $T_c(C(e))$, and $P_{ml}(t_c)$ is estimated similarly by counting the occurrences of t_c in the Chinese translation of the whole English corpus. α is set to 0.6 in our experiments.

4. Translation by transliteration

For the transliteration model, we use a modified model of (Knight and Graehl, 1998) and (Al-Onaizan and Knight, 2002b).

Knight and Graehl (1998) proposed a probabilistic model for machine transliteration. In this model, a word in the target language (i.e., English in our task) is written and pronounced. This pronunciation is converted to source language pronunciation and then to source language word (i.e., Chinese in our task). Al-Onaizan and Knight (2002b) suggested that pronunciation can be skipped and the target language letters can be mapped directly to source language letters.

Pinyin is the standard Romanization system of Chinese characters. It is phonetic-based. For transliteration, we estimate $P(e \mid c)$ as follows:

$$P(e \mid c) = P(e \mid pinyin)$$
$$= \sum_a P(e, a \mid pinyin)$$
$$= \sum_a \prod_i P(l_i^a \mid p_i)$$

First, each Chinese character in a Chinese word *c* is converted to *pinyin* form. Then we sum over all the alignments that this *pinyin* form of *c* can map to an English word *e*. For each possible alignment, we calculate the probability by taking the product of each mapping. p_i is the *i*th syllable of *pinyin*, l_i^a is the English letter sequence that the *i*th *pinyin* syllable maps to in the particular alignment *a*.

Since most Chinese characters have only one pronunciation and hence one pinyin form, we assume that Chinese character-to-pinyin mapping is one-to-one to simplify the problem. We use the expectation maximization (EM) algorithm to generate mapping probabilities from pinyin syllables to English letter sequences. To reduce the search space, we limit the number of English letters that each pinyin syllable can map to as 0, 1, or 2. Also we do not allow cross mappings. That is, if an English letter sequence e_1 precedes another English letter sequence e_2 in an English word, then the pinyin syllable mapped to e_1 must precede the pinyin syllable mapped to e_2.

Our method differs from (Knight and Graehl, 1998) and (Al-Onaizan and Knight, 2002b) in that our method does not generate candidates but only estimates $P(e \mid c)$ for candidates *e* appearing in the English corpus. Another difference is that our method estimates $P(e \mid c)$ directly, instead of $P(c \mid e)$ and $P(e)$.

5. Experiment

5.1 Resources

For the Chinese corpus, we used the Linguistic Data Consortium (LDC) Chinese Gigaword Corpus from Jan 1995 to Dec 1995. The corpus of the period Jul to Dec 1995 was used to come up with new Chinese words *c* for translation into English. The corpus of the period Jan to Jun 1995 was just used to determine if a Chinese word *c* from Jul to Dec 1995 was new, i.e., not occurring from Jan to Jun 1995. Chinese Gigaword corpus consists of news from two agencies: Xinhua News Agency and Central News Agency.

As for English corpus, we used the LDC English Gigaword Corpus from Jul to Dec 1995. The

English Gigaword corpus consists of news from four newswire services: Agence France Press English Service, Associated Press Worldstream English Service, New York Times Newswire Service, and Xinhua News Agency English Service. To avoid accidentally using parallel texts, we did not use the texts of Xinhua News Agency English Service.

The size of the English corpus from Jul to Dec 1995 was about 730M bytes, and the size of the Chinese corpus from Jul to Dec 1995 was about 120M bytes.

We used a Chinese-English dictionary which contained about 10,000 entries for translating the words in the context. For the training of transliteration probability, we required a Chinese-English name list. We used a list of 1,580 Chinese-English name pairs as training data for the EM algorithm.

5.2 Preprocessing

Unlike English, Chinese text is composed of Chinese characters with no demarcation for words. So we first segmented Chinese text with a Chinese word segmenter that was based on maximum entropy modeling (Ng and Low, 2004).

We then divided the Chinese corpus from Jul to Dec 1995 into 12 periods, each containing text from a half-month period. Then we determined the new Chinese words in each half-month period *p*. By new Chinese words, we refer to those words that appeared in this period *p* but not from Jan to Jun 1995 or any other periods that preceded *p*. Among all these new words, we selected those occurring at least 5 times. These words made up our test set. We call these words Chinese source words. They were the words that we were supposed to find translations from the English corpus.

For the English corpus, we performed sentence segmentation and converted each word to its morphological root form and to lower case.

We also divided the English corpus into 12 periods, each containing text from a half-month period. For each period, we selected those English words occurring at least 10 times and were not present in the 10,000-word Chinese-English dictionary we used and were not stop words. We considered these English words as potential translations of the Chinese source words. We call them English translation candidate words. For a Chinese source word occurring within a half-month period *p*, we looked for its English translation candidate words occurring in news documents in the same period *p*.

5.3 Translation candidates

The context $C(c)$ of a Chinese word *c* was collected as follows: For each occurrence of *c*, we set a window of size 50 characters centered at *c*. We discarded all the Chinese words in the context that were not in the dictionary we used. The contexts of all occurrences of a word *c* were then concatenated together to form $C(c)$. The context of an English translation candidate word *e*, $C(e)$, was similarly collected. The window size of English context was 100 words.

After all the counts were collected, we estimated $P(C(c) \mid C(e))$ as described in Section 3, for each pair of Chinese source word and English translation candidate word. For each Chinese source word, we ranked all its English translation candidate words according to the estimated $P(C(c) \mid C(e))$.

For each Chinese source word c and an English translation candidate word e, we also calculated the probability $P(e \mid c)$ (as described in Section 4), which was used to rank the English candidate words based on transliteration.

Finally, the English candidate word with the smallest average rank position and that appears within the top *M* positions of both ranked lists is the chosen English translation (as described in Section 2). If no words appear within the top *M* positions in both ranked lists, then no translation is output.

Note that for many Chinese words, only one English word e appeared within the top M positions for both lists. And among those cases where more than one English words appeared within the top M positions for both lists, many were multiple translations of a Chinese word. This happened for example when a Chinese word was a non-English person name. The name could have multiple translations in English. For example, 米洛西娜 was a Russian name. Mirochina and Miroshina both appeared in top 10 positions of both lists. Both were correct.

5.4 Evaluation

We evaluated our method on each of the 12 half-month periods. The results when we set M = 10 are shown in Table 1.

Period	#c	#e	#o	#Cor	Prec. (%)
1	420	15505	7	5	71.4
2	419	15863	15	9	60.0
3	417	16434	25	21	84.0
4	382	17237	11	8	72.7
5	301	16106	8	5	62.5
6	295	15905	10	9	90.0
7	513	15315	13	8	61.5
8	465	17121	17	14	82.4
9	392	16075	13	11	84.6
10	361	15970	10	9	90.0
11	329	15924	9	8	88.9
12	205	15066	9	8	88.9
Total	4499	192521	147	115	78.2

Table 1. Accuracy of our system in each period (M = 10)

In Table 1, period 1 is Jul 01 – Jul 15, period 2 is Jul 16 – Jul 31, …, period 12 is Dec 16 – Dec 31. #c is the total number of new Chinese source words in the period. #e is the total number of English translation candidates in the period. #o is the total number of output English translations. #Cor is the number of correct English translations output. Prec. is the precision. The correctness of the English translations was manually checked.

Recall is somewhat difficult to estimate because we do not know whether the English translation of a Chinese word appears in the English part of the corpus. We attempted to estimate recall by manually finding the English translations for all the Chinese source words for the two periods Dec 01 – Dec 15 and Dec 16 – Dec 31 in the English part of the corpus. During the whole December period, we only managed to find English translations which were present in the English side of the comparable corpora for 43 Chinese words. So we estimate that English translations are present in the English part of the corpus for $43/(329+205)\times 4499 = 362$ words in all 12 periods. And our program finds correct translations for 115 words. So we estimate that recall (for M = 10) is approximately $115/362 = 31.8\%$.

We also investigated the effect of varying M. The results are shown in Table 2.

M	Number of output	Precision (%)	Recall (%)
30	378	38.1	39.8
20	246	53.3	36.2
10	147	78.2	31.8
5	93	93.5	24.0
3	77	93.5	19.9
1	35	94.3	9.1

Table 2. Precision and recall for different values of M

The past research of (Fung and Yee, 1998; Rapp, 1995; Rapp, 1999) utilized context information alone and was evaluated on different corpora from ours, so it is difficult to directly compare our current results with theirs. Similarly, Al-Onaizan and Knight (2002a; 2002b) only made use of transliteration information alone and so was not directly comparable.

To investigate the effect of the two individual sources of information (context and transliteration), we checked how many translations could be found using only one source of information (i.e., context alone or transliteration alone), on those Chinese words that have translations in the English part of the comparable corpus. As mentioned earlier, for the month of Dec 1995, there are altogether 43 Chinese words that have their translations in the English part of the corpus. This list of 43 words is shown in Table 3. 8 of the 43 words are translated to English multi-word phrases (denoted as "phrase" in Table 3). Since our method currently only considers unigram English words, we are not able to find translations for these words. But it is not difficult to extend our method to handle this problem. We can first use a named entity recognizer and noun phrase chunker to extract English names and noun phrases.

The translations of 6 of the 43 words are words in the dictionary (denoted as "comm." in Table 3) and 4 of the 43 words appear less than 10 times in the English part of the corpus (denoted as "insuff"). Our method is not able to find these translations. But this is due to search space pruning. If we are willing to spend more time on searching, then in principle we can find these translations.

Chinese	English	Cont. rank	Trans. rank
鲍克	Bork	1	1
达布瓦利镇	Dabwali	1	1
卡斯布拉托夫	Khasbulatov	1	1
纳萨尔	Nazal	1	1
奥斯兰德	Ousland	1	1
杜亚拉	Douala	1	2
艾巴肯	Erbakan	1	2
叶玛斯	Yilmaz	1	120
巴佐亚	Bazelya	1	NA
坩埚	crucible	1	NA
法塔赫	Fatah	2	1
卡达诺夫	Kardanov	2	1
米洛西娜	Mirochina	3	2
马特欧利	Matteoli	4	2
杜卡姆	Tulkarm	8	7
普利法	Preval	8	NA
苏活	Soho	9	1
拉马苏尔	Lamassoure	9	3
卡敏斯基	Kaminski	10	1
莫伦	Muallem	19	52
柴卡斯基	Cherkassky	46	2
艾巴甘	Erbakan	49	2
雷蒂嫩	Laitinen	317	2
库利埃	Courier	328	21
豹式	leopard	1157	NA
纳乌莫夫	Naumov	insuff	
商州市	Shangzhou	insuff	
沃勒尔	Voeller	insuff	
瓦森纳尔	Wassenaar	insuff	
秃发	bald	comm	
碱基	base	comm	
耶诞季	Christmas	comm	
损减	decrease	comm	
恤金	pension	comm	
沙乌地人	Saudi	comm	
赫尔采格–波斯尼亚	Bosnia-Hercegovina	phrase	
圣诞卡	Christmas Card	phrase	
展售馆	exhibition hall	phrase	
孵蛋	hatch egg	phrase	
川崎制铁	Kawasaki Steel Co.	phrase	
圣荷西山	Mount San Jose	phrase	
家邦党	Our Home Be Russia	phrase	
联选	Union Election	phrase	

Table 3. Rank of correct translation for period Dec 01 – Dec 15 and Dec 16 – Dec 31. 'Cont. rank' is the context rank, 'Trans. Rank' is the transliteration rank. 'NA' means the word cannot be transliterated. 'insuff' means the correct translation appears less than 10 times in the English part of the comparable corpus. 'comm' means the correct translation is a word appearing in the dictionary we used or is a stop word. 'phrase' means the correct translation contains multiple English words.

As shown in Table 3, using just context information alone, 10 Chinese words (the first 10) have their correct English translations at rank one position. And using just transliteration information alone, 9 Chinese words have their correct English translations at rank one position.

On the other hand, using our method of combining both sources of information and setting $M = \infty$, 19 Chinese words (i.e., the first 22 Chinese words in Table 3 except 巴佐**亚**,坩**埚**,普利法) have their correct English translations at rank one position. If $M = 10$, 15 Chinese words (i.e., the first 19 Chinese words in Table 3 except 叶**玛**斯,巴佐**亚**,坩**埚**,普利法) have their correct English translations at rank one position. Hence, our method of using both sources of information outperforms using either information source alone.

6. Related work

As pointed out earlier, most previous research only considers either transliteration or context information in determining the translation of a source language word *w*, but not both sources of information. For example, the work of (Al-Onaizan and Knight, 2002a; Al-Onaizan and Knight, 2002b; Knight and Graehl, 1998) used only the pronunciation or spelling of *w* in translation. On the other hand, the work of (Cao and Li, 2002; Fung and Yee, 1998; Rapp, 1995; Rapp, 1999) used only the context of *w* to locate its translation in a second language. In contrast, our current work attempts to combine both complementary sources of information, yielding higher

accuracy than using either source of information alone.

Koehn and Knight (2002) attempted to combine multiple clues, including similar context and spelling. But their similar spelling clue uses the longest common subsequence ratio and works only for cognates (words with a very similar spelling).

The work that is most similar to ours is the recent research of (Huang et al., 2004). They attempted to improve named entity translation by combining phonetic and semantic information. Their contextual semantic similarity model is different from our language modeling approach to measuring context similarity. It also made use of part-of-speech tag information, whereas our method is simpler and does not require part-of-speech tagging. They combined the two sources of information by weighting the two individual scores, whereas we made use of the average rank for combination.

7. Conclusion

In this paper, we proposed a new method to mine new word translations from comparable corpora, by combining context and transliteration information, which are complementary sources of information. We evaluated our approach on six months of Chinese and English Gigaword corpora, with encouraging results.

Acknowledgements

We thank Jia Li for implementing the EM algorithm to train transliteration probabilities. This research is partially supported by a research grant R252-000-125-112 from National University of Singapore Academic Research Fund.

References

Y. Al-Onaizan and K. Knight. 2002a. Translating named entities using monolingual and bilingual resources. In *Proc. of ACL*.

Y. Al-Onaizan and K. Knight. 2002b. Machine transliteration of names in Arabic text. In *Proc. of the ACL Workshop on Computational Approaches to Semitic Languages*.

Y. Cao and H. Li. 2002. Base noun phrase translation using web data and the EM algorithm. In *Proc. of COLING.*

P. Fung and L. Y. Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of COLING-ACL*.

F. Huang, S. Vogel and A. Waibel. 2004. Improving named entity translation combining phonetic and semantic similarities. In *Proc. of HLT-NAACL.*

K. Knight and J. Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4): 599-612.

P. Koehn and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proc. of the ACL Workshop on Unsupervised Lexical Acquisition.*

I. D. Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proc. of EMNLP*.

R. C. Moore. 2003. Learning translations of named-entity phrases from parallel corpora. In *Proc. of EACL.*

H. T. Ng and J. K. Low. 2004. Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based? To appear in *Proc of EMNLP*.

K. Ng. 2000. A maximum likelihood ratio information retrieval model. *In Proc. of TREC-8*.

J. M. Ponte and W. B. Croft. 1998. A language modeling approach to information retrieval. In *Proc. of SIGIR*.

R. Rapp. 1995. Identifying word translations in non-parallel texts. In *Proc. of ACL (student session).*

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