Translating Common English and Chinese Verb-Noun Pairs in Technical Documents with Collocational and Bilingual Information

Yi-Hsuan Chuang[†] Chao-Lin Liu[§] Jing-Shin Chang[↑]
[†§] National Chengchi University, Taiwan; [↑]National Chi Nan University, Taiwan
{[†]g9804, [§]chaolin}@cs.nccu.edu.tw, [↑]jshin@csie.ncnu.edu.tw

Abstract. We studied a special case for the translation of English verbs in verb-object pairs. Researchers have studied the effects of the linguistic information about the verbs being translated, and many have reported how considering the objects of the verbs will facilitate the quality of translations. In this study, we took an extreme venue – assuming the availability of the Chinese translations of the English objects. We explored the issue with thousands of samples that we extracted from 2011 NTCIR PatentMT workshop. The results indicated that, when the English verbs and objects were known, the information about the object's Chinese translation could still improve the quality of the verb's translations but not quite significantly.

Keywords: Machine Translation, Feature Comparison

1 Introduction

Researchers have studied extensively the problems related to verbs (e.g., Dorr *et al*., 2002; Lapata and Brew, 2004) and phrases-based translations (e.g., Chuang *et al*., 2005; Koehn *et a*l., 2003). Some techniques were developed for text of special domains (Seneff *et al*., 2006). The techniques are applicable in many real-world problems, including computer-assisted language learning (Chang *et al*., 2008) and cross-language information retrieval (Chen *et al*., 2000).

We work on the processing of patent documents (Lu *et al*., 2010; Yokoama and Okuyama, 2009), and present an experience in translating common verbs and their direct object based on bilingual contextual information. In this study, we took an extreme assumption of the availability of the Chinese translations of the English objects to examine whether the extra information will improve the quality of verbs' translation. The proposed methods are special in that we are crossing the boundary between translation models and language models, by considering information of the target language in the translation task. The purpose of conducting such experiments was to investigate how the availability of such bilingual and collocational information might contribute to the translation quality. It is understood and expected by many that the Chinese translations of English objects might not be available for all cases and that such information should be applied with many other features to achieve high translation quality.

The experiments were conducted with the training data available to the participants of the 2011 NTCIR Patent MT task. The corpus contains one million pairs of Chinese and English pairs. We explored four different methods to determine the verb's Chinese translation. These methods utilized the bilingual and contextual information about the English verb in different ways. Effects of these methods were compared based on experimental evaluation which conducted with 35 thousands of verb-object pairs extracted from the NTCIR corpus. (Since objects are nouns, we will refer to verb-object pairs as verb-noun pairs or VN pairs to simplify the wording.)

We provide a broad outline of our work in Section 2, present our methods for aligning the bilingual VN-pairs in Section 3, explain how we build lexicons with extra information to serve the needs of VN-pair alignment in Section 4, delineate the design of our experiments in Section 5, and discuss the experimental results in Section 6.

2 The Big Picture

Our work consisted of two major stages. We extracted the VN pairs from the original corpus. Then, we applied our translation methods to translate English words to Chinese and vice versa,

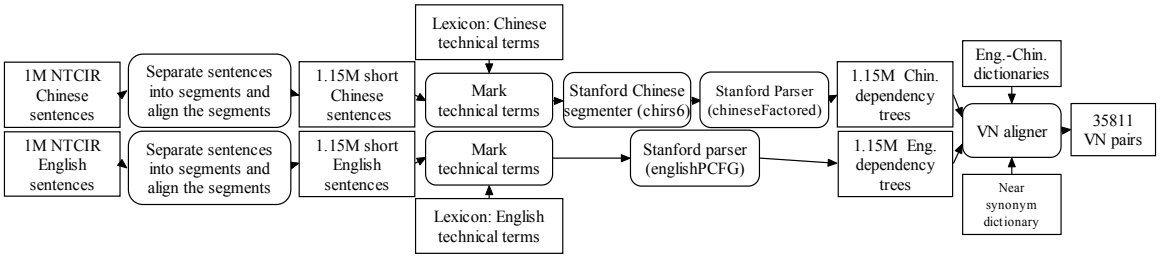

Figure 1: The flow for extracting VN pairs from the original corpora

and compared the translation quality achieved by using different combinations of contextual and bilingual information.

Figure 1 shows how the VN pairs were extracted from the 1 million parallel sentences, which we obtained from the NTCIR 9 PatentMT task in 2011[1]. The process started from the left of the figure. Most of the original sentences were very long. A sentence has 34 words on average, and the longest sentence has 141 words. Since our goal was to extract VN pairs from the corpus, not doing a full-scale research in machine translation, we choose to segment the sentences into shorter parts at commas and periods. Normally, VN pairs will not cross the punctuations, and, even if some VN pairs did, we afforded to neglect them because we had 1 million pairs of long sentences.

We then re-aligned the short English and Chinese segments with a sentence aligner (Tien *et al*., 2009) that we implemented based on the concept of Champollion (Ma, 2006). We treated the original long sentence pairs as aligned paragraphs, and ran our aligner on the sentences. Like the Champollion, we computed probabilistic scores for the sentence pairs, so we could choose those pairs with higher scores to achieve higher confidence on the aligned pairs. More specifically, we kept only the leading 33% of the short sentence pairs, and obtained 1148632 short sentence pairs.

We employed the Stanford Chinese segmenter[2] to segment the Chinese text. This segmenter allows us to mark the technical terms so that the segmenter will treat the words belonging to technical terms as a unit, preventing them from being segmented again. In addition, currently, our technical terms are nouns, so they are annotated accordingly. When there were more than one possible ways to mark the technical terms in a string, we preferred the longer choices. English texts were tokenized by the Stanford parser[3]. Technical phrases and compound words in English were also marked and would not be treated as individual words either. The special terms came from the glossary that will be explained in Section 4.1.

Based on these short sentence pairs, we aligned VN pairs with the method to be explained in Section 3. This process employed an English-Chinese glossary for technical terms, which we will discuss in Section 4.1, and a bilingual dictionary enhanced with Chinese near synonyms, which we will discuss in Section 4.2. In the end, we accepted 35811 VN pairs for experiments at the second stage.

During the second stage of our work, we split the VN pairs into training and test data. Useful statistics were collected from the training data, and were applied to select Chinese translations for the English words in question.

3 VN Pair Alignment

We employed the Stanford parsers (SPs, henceforth) for English and Chinese to compute the dependency trees for the parallel texts. We extracted the `dobj` relations from the trees and align the VN pairs.

[1] http://ntcir.nii.ac.jp/PatentMT/
[2] http://nlp.stanford.edu/software/segmenter.shtml, version 1.5
[3] http://nlp.stanford.edu/software/lex-parser.shtml, version 1.6.5

3.1 Dependency Trees

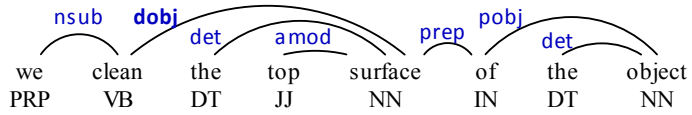

Figure 2: A sample dependency tree with POS tags

Based on the general recommendations on the Stanford site3, we parsed English with the englishPCFG.ser.gz grammar, and parsed Chinese with the chineseFactored.ser.gz grammar.

Figure 2 shows the dependency tree for a simple English sentence "`we clean the top surface of the object`". SPs can provide the part of speeches (POSes) of words and recognize the relationships between the words. POSes are shown below the words, and the relationships are attached to the links between the words. The **`dobj`** link between "`clean`" and "`surface`" indicates that "`surface`" is a direct object of "`clean`", and we could rely on such `dobj` links to identify VN pairs in the corpus.

3.2 VN Pair Alignment

We found 375041 `dobj` links in the 1.15M English sentences and 465866 `dobj` links in the Chinese part. However, not all of the words participating in a `dobj` link were real words. Hence, we looked up the words in our bilingual lexicon, which we will explain in Section 4.2. If the lexicon did not contain the words, we would not use the words in the corresponding `dobj` links as VN pairs. After this step, we had 254091 and 249591 VN pairs in English and Chinese, respectively.

We then tried to align the remaining English and Chinese VN pairs, noting that only those VN pairs that originated from the same pair of long sentence pairs can be aligned.

The alignment method can be quite simple. Let (EV, EN) and (CV, CN) denote an English and a Chinese VN pair, respectively, where EV, EN, CV, and CN denote an English verb, an English noun, a Chinese verb, and a Chinese noun. We just had to check whether CV is a possible translation of EV and whether CN is a possible translation of EN. If both answers are positive, then we aligned the VN pairs.

However, even when an English verb can carry only one sense, there can be multiple ways to translate it into Chinese, and there is no telling whether a dictionary will include all of the possible translations to contain the Chinese translations that were used in the Patent MT corpus. For instance, (improve, quality) can be translated to (改善(gai3 shan4), 品質(pin3 zhi2)) or (改進(gai3 jin4), 品質). If an English-Chinese dictionary only lists "改善" as the translation for "improve" and does not include "改進" as a possible translation, than we could not use that dictionary to align (improve, quality) and (改進, 品質). We need a way to tell that "改進" and "改善" are similar.

Therefore, we expanded the set of possible Chinese translations in a given dictionary with near synonyms, and employed the expanded dictionary to enhance the accuracy of VN pair alignment. The process of constructing such expanded dictionary is provided in Section 4.2.

After completing the VN pair alignment, we obtained 35811 aligned VN pairs, cf. Figure 1.

4 Lexicon Constructions

We explain (1) how we built the glossary of technical terms and (2) how we constructed a bilingual dictionary that contains information about near synonyms in this section.

4.1 Creating a Glossary of Technical Terms

As explained in Section 2, we built a glossary of technical terms so that we can separate technical terms from normal text, thereby achieving higher quality of parsing.

We downloaded 138 different kinds of domain technical term pairs from Taiwan National Academy for Educational Research[4]. The files were stored in excel format, and the total file size is 177MB.

The format of English-Chinese technical term pairs is not always in one-to-one relationship; some English technical terms have more than one translation in Chinese. We converted such pairs into multiple one-to-one pairs, and acquired 804068 English-Chinese technical term one-to-one pairs.

To validate the reliability of the glossary, we conducted a small experiment; that is, to segment patent sentences with the glossary. The results showed that the coverage of these "technical term" pairs was too broad, and almost all of words were considered as technical terms.

We alleviated this problem with E-HowNet[5] (Chen *et al.*, 2005) and WordNet[6]. Treating the words listed in E-HowNet and WordNet as ordinary words, we used them to identify normal words in our technical term pairs. If the original pairs contained any normal Chinese or English words, then the pairs would be removed.

As a result, we removed 14% of the original pairs, and kept 690640 technical term pairs.

4.2 The English-Chinese Dictionary and Near Synonyms

As announced in Section 3.2, we built a bilingual dictionary and enhanced it with information about near synonyms to improve the recall rates of the VN pair alignment.

A good English-Chinese dictionary is the basis for the task of VN pair alignment. We collected and combined the Chinese translations of English words in the Concise Oxford English Dictionary and the Dr.eye online dictionary[7], and acquired 99805 pairs of English words and their translations.

As we explained in Section 3.2, the Chinese translations listed in the dictionaries might not be complete, so we enhanced the merged dictionary with information about near synonyms. We employed two sources of relevant information to obtain near synonyms in this study.

The Web-based service of Word-Focused Extensive Reading System[8] (Cheng, 2004) is maintained by the Institute of Linguistics of the Academia Sinica in Taiwan. The service allows us to submit queries for the near synonyms of Chinese words for free, so we collected the near synonyms from the web site.

Given an entry in our bilingual dictionary, we queried the near synonyms for each of the Chinese translations of an English word, and added the results to the Chinese translations of the English word.

E-HowNet is another source that we could compute and obtain near synonyms. E-HowNet is a lexicon for Chinese. The lexicon contains two levels of detailed semantic information for words: TopLevelDefinition and BottomLevelExpansion. The semantic definitions provided in these two entries come from the E-HowNet Ontology[9], and can be used to compute similarity scores between word senses.

We determine whether two Chinese words are near synonyms with the following procedure. Given a Chinese word, CW, we looked up the E-HowNet for its senses. Let $S_i(CW)$ be one of CW's senses. We combined the semantic definitions listed in the TopLevelDefinition and BottomLevelExpansion of $S_i(CW)$, which might include multiple words. Denote this set by $U_i(CW)$, and let CWW_{ij} be a word in $U_i(CW)$. We looked up the E-HowNet for the senses of CWW_{ij}. Let $S_k(CWW_{ij})$ denote one of the senses of the CWW_{ij}, and $V_{ijk}(CWW_{ij})$ denote the combined semantic definitions listed in the TopLevelDefinition and BottomLevelExpansion of $S_k(CWW_{ij})$. Finally, we computed the union of $U_i(CW)$ and $V_{ijk}(CWW_{ij})$ as a sense vector of CW. Note that

[4] http://terms.nict.gov.tw/
[5] http://ckip.iis.sinica.edu.tw/taxonomy/taxonomy-edoc.htm
[6] http://wordnet.princeton.edu/
[7] http://www.dreye.com/index_en.html
[8] http://elearning.ling.sinica.edu.tw/c_help.html
[9] http://ckip.iis.sinica.edu.tw/taxonomy/taxonomy-edoc.htm

due to lexical ambiguity, a Chinese word might have multiple such vectors. Two Chinese words were considered as near synonyms if the cosine value of any of their sense vectors exceeded 0.7.

Given an entry in our bilingual dictionary, we compute the near synonyms of the Chinese translations of each English word. This was carried out by comparing the sense vectors of a Chinese translation with 88074 Chinese words in E-HowNet. The qualified words were added to the Chinese translations of the English word in our dictionary.

The purpose of adding information about near synonyms into our bilingual dictionary was to increase the recall rates of VN-pair alignment. More complex methods for identifying synonyms, e.g. (Bundanitsky and Hirst, 2006; Chang and Chiou, 2010) will be instrumental for the study.

5 Design of the Experiments

We have conducted experiments to translate from English to Chinese and from Chinese to English. In addition, we have tried to find the best translations of verbs, and tried to find the best translations of objects of the verbs given their contexts. Nevertheless, we present only the experiments of translating English verbs to Chinese verbs in this paper.

5.1 Statistics about the Aligned VN pairs

We calculated the frequencies of the verbs in the 35811 aligned VN pairs, and ranked the verbs based on the observed frequencies. Table 1 shows the 20 most frequent English verbs and their frequencies. We identified the 100 most frequent English verbs and the corresponding aligned VN pairs in our experiments. In total, there were 30376 such aligned VN pairs. The most frequent English verb appeared 4530 times, as shown in Table 1. The 100th most frequent English verb is "lack", and it appeared 47 times.

Some of the English verbs are relatively easier to translate than others. We can calculate the frequencies of the Chinese translations of verbs to verify the differences. For instance, "add" was translated in five different ways: "增加"(zeng1 jia1) 48 times, "添加"(tian1 jia1) 44 times, "加入"(jia1 ru4) 43 times, "加上"(jia1 shang4) 2 times , and "增添"(zeng1 tian1) 1 time. The distribution, (48, 44, 43, 2, 1), is not very skewed, and the frequencies of the most frequent translation and the second most frequent translation are close. Therefore, we would not achieve very good results, if we should choose to use the most frequent translation for all occurrences of "add".

Based on this observation, we defined the *challenging index* of a word as the ratios of the frequency of their most frequent translation against the frequency of their second most frequent translation. The challenging index of "add" mentioned in the previous paragraph is 1.09.

This challenging index is not a scientifically-proven index for difficulty for translation, but could serve as a heuristic. The larger the index; the easier to achieve good translation by using the most frequent translation to translate. Table 2 lists 22 verbs that had the smallest challenging indexes.

5.2 Translation Decisions

Given the aligned VN pairs, we could compute conditional probabilities, and apply the condi-

Table 1: 20 most frequent English verbs in the aligned VN pairs

Verb	have	provide	use	include	comprise	contain	form	receive	reduce	perform
Freq.	4530	3345	1993	1954	1588	1080	914	863	774	616
Verb	increase	produce	maintain	determine	represent	show	obtain	achieve	improve	allow
Freq.	465	453	397	382	373	352	329	329	322	287

Table 2: 22 most "challenging" English verbs and their indexes

Verb	make	exhibit	add	represent	retain	leave	enhance	reduce	lack	improve	achieve
	1.00	1.09	1.09	1.21	1.21	1.22	1.25	1.26	1.27	1.33	1.39
Verb	employ	reach	create	give	replace	take	apply	adjust	obtain	carry	explain
	1.41	1.43	1.50	1.54	1.69	1.69	1.69	1.72	1.76	1.82	2.00

tional probabilities to determine the Chinese translation of English words.

Table 3 lists four possible ways to choose a Chinese translation for an English verb in a VN pair. Equation (1) is the most simplistic. Let EV denote a specific English verb, and CV_i be one of EV's translations observed in the training data. Given the English verb, the equation chooses the CV_i that maximizes the conditional probability. Namely, it prefers the most frequent Chinese translation of EV in the training data.

Table 3: Translation decisions

$\arg\max_{CV_i} \Pr(CV_i \mid EV)$	(1)
$\arg\max_{CV_i} \Pr(CV_i \mid EV, EN)$	(2)
$\arg\max_{CV_i} \Pr(CV_i \mid EV, EN, CN)$	(3)
$\arg\max_{CV_i} \Pr(CV_i \mid EV, CN)$	(4)

We could obtain the conditional probability $\Pr(CV_i|EV)$ by dividing the frequency of observing the VN pair (EV, CV_i) in the training data by the frequency of observing EV in any VN pairs. Using the data for "add" that we mentioned in Section 5.1 for example, we observed 135 occurrences of "add". Therefore, Pr("增加" | "add") = 48/135=0.356 and Pr("加上" | "add") = 2/135=0.015.

Let EN be a specific English noun, Equation (2) considers the object of the verb when choosing the verb's translation. Let $C(\cdot)$ denote the frequency of a given event. The conditional probability in Equation (2) is defined below.

$$\frac{C(EV, EN, CV_i)}{C(EV, EN)} \tag{5}$$

The remaining equations, (3) and (4), take an extreme assumption. We assumed the availability of the Chinese translation of the English object at the time of translation, and used this special information in different ways. Equation (3) considers the words EV, EN, and CN. In a strong contrast, Equation (4) considers only EV and CN to determine the translation of the English verb. The conditional probabilities in Equations (3) and (4) were calculated using Equation (6) and (7), respectively.

$$\frac{C(EV, EN, CV_i, CN)}{C(EV, EN, CN)} \tag{6}$$

$$\frac{C(EV, CN, CV_i)}{C(EV, CN)} \tag{7}$$

We considered the exploration of using the information about the Chinese translation of the English noun is interesting. Would the information about CN provide more information, given we had information about EV and EN? How would we achieve when we had information about only EV and CN but not EN?

In all of the experiments, we used 80% of the available aligned VN pairs as the training data, and the remaining 20% as test data. The training data were randomly sampled from the available data.

As a consequence, it was possible for us to encounter the zero probability problems. Take Equation (6) for example. If, for a test case, we needed C(EV, EN, CN) in (6), but we happened not to have observed any instances of (EV, EN, CN) in the aligned VN pairs in the training data, then we would not be able to compute (6) for the test case. When such cases occurred, we chose to allow our system to admit that it was not able to recommend a translation, rather than resorting to the smoothing technologies.

6 Experimental Results

Using the formula in Table 3 would allow our systems to recommend only one Chinese translation. In fact, we relax this unnecessary constraint by allowing our systems to consider that largest *k* conditional probabilities and to recommend *k* translations.

Although we have been presenting this paper with the 1 million parallel sentences in NTCIR PatentMT data as the example, we have run our experiments with the English-Chinese bilingual

version of Scientific American. Moreover, we ran experiments that aimed at finding the best Chinese translations of English objects. The formulas were defined analogously with those listed in Table 3.

6.1 Basic Results for the Top 100 Verbs in Patent Documents

When we conducted experiments for the top 100 verbs (cf. Section 5.1), we had 24300 instances of aligned VN pairs for training and 6076 instances of aligned VN pairs for test.

We measured four rates as the indication of the performance of using a particular formula in Table 3. The *rejection rate* is the percentage of not being to respond to the test cases. This is due to our choosing not to smooth the probability distributions as we explained in Section 5.2.

Table 4: Inclusion rates for top 100 verbs

inclusion	*k*=1	*k*=3	*k*=5
Eq1	0.768	0.953	0.975
Eq2	0.786	0.913	0.918
Eq3	0.795	0.911	0.916
Eq4	0.791	0.910	0.916

Table 5: Average number of recommendations

recommend	*k*=1	*k*=3	*k*=5
Eq1	1.000	2.919	4.614
Eq2	1.000	1.923	2.225
Eq3	1.000	1.847	2.107
Eq4	1.000	1.920	2.244

Table 6: Average ranks of the answers

ranking	*k*=1	*k*=3	*k*=5
Eq1	1.000	1.241	1.310
Eq2	1.000	1.166	1.185
Eq3	1.000	1.151	1.168
Eq4	1.000	1.153	1.173

It is not surprising that the rejection rates increased as we considered more information in the formulas. The rejections rates were 0, 0.201, 0.262, and 0.218 when we applied Equations (1) through (4) in the experiments. As expected, we encountered the highest rejection rate when using (3). Note that using (4) resulted in a higher rejection rate than using (2). This was because that there were no less than one possible Chinese translations for any English verbs. Hence, the distributions for $\Pr(CV_i \mid EV\ CN)$ would be more sparse than $\Pr(CV_i \mid EV\ EN)$ on average.

Table 4 shows the rates of the correct answers were included in the recommended *k* translations. We did not consider the cases that our systems could not answer in computing the statistics in Table 4. Hence, the data show the average inclusion rates when our systems could answer. As one may have expected, when we increased *k*, the inclusion rates also increased.

It may be surprising that the inclusion rates for Equations (2) through (4) seem to have saturated when we increase *k* from 3 to 5. This was because our systems could not actually recommend 5 possible translations, when they were allowed. Although we had hundreds or thousands of aligned VN pairs for an English verb, cf. Table 1. Including more conditioning information in Equations (2) through (4) still reduced the amount of VN pairs qualified for training and testing, thereby limiting the actual numbers of recommended translations. Table 5 shows the average number of actual recommendations in the tests.

The main advantage of using Equations (2) through (4) is that they were more precise, when they could answer. Table 6 shows the average ranks of the correct translation in the recommended translations. The average ranks improve as we considered more information from Equation (1) to (2) and to (3). Using (2) achieved almost the same average rank with using (4), but using (4) led to slightly better performance.

6.2 Improving Results for the Top 100 Verbs in Patent Documents

Results reported in the previous subsection indicated that Equation (4) is robust in that it could offer candidate answers all the time. Methods that employed more information could choose translations more precisely, but were less likely to respond to test cases. Hence, a natural question is whether we could combine these methods to achieve better performance. To answer this question, we conducted all of the combinations of the basic methods listed in Table 3.

In Tables 7 and 8, we use the notation EqX+EqY to indicate that we used EqX to find as many candidate translations as possible, before we reached a total of *k* recommendations. If applying EqX could not offer sufficient candidate translations, we applied EqY to recommend more candidate translations until we acquired *k* recommendations.

Using Eq1 is sufficiently robust in that the conditional probabilities would not be zero, unless the training data did not contain any instances that included the English verb. Hence, in our experiments, the rejection rates for "Eq2+Eq1", "Eq3+Eq1", and "Eq4+Eq1" became zero. In other words, our systems responded to all test cases when we used these combined methods by allowing all methods to recommend up to *k* candidates.

Table 7: Inclusion rates (combined methods)

inclusion	*k*=1	*k*=3	*k*=5
Eq1	0.768	0.953	0.975
Eq2+Eq1	0.772	0.960	0.979
Eq3+Eq1	0.778	0.960	0.979
Eq4+Eq1	0.776	0.959	0.978

Table 8: Average ranks of the correct answers (combined methods)

ranking	*k*=1	*k*=3	*k*=5
Eq1	1.000	1.241	1.310
Eq3	1.000	1.151	1.168
Eq2+Eq1	1.000	1.240	1.301
Eq3+Eq1	1.000	1.234	1.294
Eq4+Eq1	1.000	1.233	1.296

We compare the performance of these combined methods with the best performing methods in Tables 7 and 8. We copy the inclusion rates of Eq1 from Table 4 to Table 7 to facilitate the comparison, because Eq1 was the best performer, on average, in Table 4. The combined methods improved the inclusion rates, although the improvement was marginal.

Moreover, we copy the average ranks for Eq1 and Eq3 from Table 6 to Table 8. Using Eq1 and using Eq3 led to the worst and the best average ranks in Table 6, respectively. Again, using the combined methods, we improve the average ranks marginally over the results of using Eq1.

Statistics in Table 7 suggest that using machine-assisted approach to translating verbs in common VN pairs in the PatentMT data is feasible. Providing the top five candidates to a human translator to choice will allow the translator to find the recorded answer s nearly 98% of the time.

It is interesting to find that using Equation (2) and Equation (4) did not lead to significantly different results in Tables (4) through (8). The results suggest that using either the English nouns or the Chinese nouns as a condition contributed similarly to the translation quality of the English verbs. More specifically, the translation of the English verb may be conditionally independent of the information about the noun's Chinese translation given the English verb and the English noun. This does not imply that the translation of the English verb is unconditionally independent of the information of the English noun's translation. That Eq4 performed better than Eq1 in Table 6 offered a reasonable support.

6.3 Results for the Most Challenging 22 Verbs in Patent Documents

We repeated the experiments that we conducted for the top 100 verbs for the most challenging 22 verbs (cf. Section 5.1). Tables 9 through 13 correspond to Tables 4 through 8, respectively. The most noticeable difference between Table 9 and Table 4 is the reduction of the inclusion rates achieved by Eq1 when *k*=1. Although the inclusion rates reduced noticeably when we used Eq2, Eq3, and Eq4 as well, the drop in the inclusion rate for Eq1 (when *k*=1) was the most significant.

Table 9: Inclusion rates for 22 challenging verbs

inclusion	*k*=1	*k*=3	*k*=5
Eq1	0.449	0.865	0.923
Eq2	0.561	0.818	0.820
Eq3	0.564	0.827	0.829
Eq4	0.550	0.827	0.829

Table 10: Average number of recommendations

recommend	*k*=1	*k*=3	*k*=5
Eq1	1.000	2.977	4.756
Eq2	1.000	2.090	2.364
Eq3	1.000	2.022	2.230
Eq4	1.000	2.106	2.411

Although we did not define the challenging index of verbs based on their numbers of possible translations, comparing the corresponding numbers in Table 10 and Table 5 suggest that the challenging verbs also have more possible translations in the NTCIR data.

Table 11: Average ranks of the answers

ranking	*k*=1	*k*=3	*k*=5
Eq1	1.000	1.607	1.773
Eq2	1.000	1.365	1.373
Eq3	1.000	1.374	1.383
Eq4	1.000	1.394	1.400

Corresponding numbers in Table 11 and

Table 6 supports the claim that translating the 22 challenging words is relatively more difficult. The average ranks of the answers became worse in Table 11.

Data in Tables 12 and 13 repeat the trends that we observed in Tables 7 and 8. Using the combined methods allowed us to answer all test cases, and improved both the inclusion rates and the average ranks of the answers.

If we built a computer-assisted translation system for these 22 verbs, the performance would not be as good as if we built a system for the top 100 verbs. When the system suggested the leading 3 translations (k=3), the inclusion rates dropped to around 0.90 in Table 12 from 0.96 in Table 7.

Table 12: Inclusion rates (combined methods)

inclusion	*k*=1	*k*=3	*k*=5
Eq1	0.449	0.865	0.923
Eq2+Eq1	0.512	0.896	0.940
Eq3+Eq1	0.503	0.894	0.940
Eq4+Eq1	0.508	0.900	0.942

Table 13: Average ranks of the correct answers (combined methods)

ranking	*k*=1	*k*=3	*k*=5
Eq1	1.000	1.607	1.773
Eq3	1.000	1.374	1.383
Eq2+Eq1	1.000	1.537	1.662
Eq3+Eq1	1.000	1.546	1.677
Eq4+Eq1	1.000	1.547	1.664

Again, using either the English nouns or the Chinese nouns, along with the English verbs, in the conditions of the methods listed in Table 3 did not make significant differences, as suggested by the results in Tables 9 through 13.

6.4 More Experimental Results

We repeated the experiments that we conducted for the top 100 verbs for the top 100 nouns in the PatentMT data. For experiments with the nouns, we had only 3952 test instances. The goals were to find the best Chinese translation of the English objects, given its contextual and bilingual information. More specifically, in addition to the English verbs and the English nouns, we were interested in whether providing the Chinese translations of the English verbs would help us improve the translation quality of the English objects.

Due to the page limits, we could not show all of the tables as we did in Section 6.1. The statistics showed analogous trends that we reported in Section 6.1. Namely, the availability of the Chinese translation of the English verbs did not help, when we already considered the English verbs and objects in the translation decisions.

Scientific American is a magazine for general public. The writing style is more close to ordinary lives. We ran a limited scale of experiment with available text from Scientific American. We had about 1500 training instances and 377 test instances for 25 verbs. The results indicated that using the Chinese translations of the English objects influenced the translation quality of the English verbs. However, the observed differences were not significant. A side observation was that it was relatively harder to find good translation of English verbs in Scientific American than in the PatentMT data. When providing five recommendations, only about 88% of the time the recommendations of our system can include the correct translation. In contrast, we had achieved inclusion rates well above 90% in Tables 7 and 12.

7 Concluding Remarks

We designed a procedure to extract and align VN pairs in a bilingual corpus of patent documents. The corpus contains 1 million pairs of English and Chinese sentences, and we aligned 35811 VN pairs. We employed the VN pairs to investigate whether the availability of the Chinese translations for nouns in English VN pairs would improve the translation quality of the English verbs. Experimental results suggest that the Chinese translation of the English noun is marginally helpful when both the English verbs and English nouns are already available. Choosing the best Chinese translation of the English verb based on the constraint of its English object or based on the information about the object's Chinese translation achieved similarly in the experiments.

Additional and analogous experiments were conducted with the PatentMT data. In these new experiments, we aimed at the translations of the nouns in the English VN pairs, given different combinations of the bilingual and contextual information. Again, we observed that, after putting the English verb and English noun in the conditions in the attachment decision formulas (analogous to but slightly different with those in Table 3), the Chinese translations of the English verbs did not offer extra help. In addition, using (1) the English verb and the English noun or (2) the Chinese verb and the English noun achieved similar experimental results.

Acknowledgments

The work was supported in part by the funding from the National Science Council in Taiwan under the contracts NSC-99-2221-E-004-007 and NSC-100-2221-E-004-014. The authors are obliged to the anonymous reviewers for their valuable comments because we could not respond satisfactorily to all comments due to the page limits on this paper.

References

Bundanitsky, A. and G. Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):14–47.

Carpuat, M., P. Fung and G. Ngai. 2006. Aligning word senses using bilingual corpora. *ACM Trans. on Asian Language Information Processing*, 5(2):89–120.

Chang, J.-S. and S.-J. Chiou. 2010. An EM algorithm for context-based searching and disambiguation with application to synonym term alignment. *Proc. of the 23rd Pacific Asia Conf. on Language, Information and Computation*, 2:630–637.

Chang, Y. C., J. S. Chang, H. J. Chen and H. C. Liou. 2008. An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3):283–299.

Cheng, C. C. 2004. Word-focused extensive reading with guidance. *Selected Papers from the 13th International Symposium and Book Fair on English Teaching*, 24-32. http://elearning.ling.sinica.edu.tw/WordFocused%20Extensive%20Reading%20with%20Guidance.pdf

Chuang, T. C., J.-Y. Jian, Y.-C. Chang and J. S. Chang. 2005. Collocational translation memory extraction based on statistical and linguistic information. *Int'l J. of Computational Linguistics and Chinese Language Processing*, 10(3):329–346.

Chen, A., H. Jiang and F. Gey. 2000. Combining multiple sources for short query translation in Chinese-English cross-language information retrieval. *Proc. of the 5th Int'l Workshop on Information Retrieval with Asian Languages*, 17–23.

Chen, K.-J., S.-L. Huang, Y.-Y. Shih and Y.-J. Chen. 2005. Extended-HowNet: A representational framework for concepts. *Proc. of the 2005 IJCNLP Workshop on Ontologies and Lexical Resources*, 1–6.

Dorr, B. J., G.-A. Levow and D. Lin. 2002. Construction of a Chinese-English verb lexicon for machine translation and embedded multilingual applications. *Machine Translation*, 17:99–137.

Koehn, P., F. J. Och and D. Marcu. 2003. Statistical phrase-based translation. *Proc. of the 2003 Conf. of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 48-54.

Lapata, M. and C. Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–73.

Lu, B., B. K. Tsou, T. Jiang, O. Y. Kwong and J. Zhu. 2010. Mining large-scale parallel corpora from multilingual patents: An English-Chinese example and its application to SMT. *Proc. of the 1st CIPS-SIGHAN Joint Conf. on Chinese Language Processing*.

Ma, X. 2006. Champollion: A robust parallel text sentence aligner. *Proc. of the 5th Int'l Conf. of the Language Resources and Evaluation*, 489–492.

Seneff, S., C. Wang and J. Lee. 2006. Combining linguistic and statistical methods for bi-directional English Chinese translation in the flight domain. *Proc. of the 7th Conf. of the Association for Machine Translation in the Americas*, 213–222.

Tien, K.-W., Y.-H. Tseng and C.-L. Liu. 2009. Sentence alignment of English and Chinese patent documents, *Proc. of the 21st Conf. on Computational Linguistics and Speech Processing*, 85–99.

Yokoama, S. and M. Okuyama. 2009. Translation disambiguation of patent sentences using case frames. *Proc. of the 3rd Workshop on Patent Translation*, in Machine Translation Summit XII, 33–36.