Computational Linguistics and Chinese Language Processing
Vol. 17, No. 3, September 2012, pp. 1-28

Effects of Combining Bilingual and Collocational Information on Translation of English and Chinese Verb-Noun Pairs[1]

Yi-Hsuan Chuang[*], Chao-Lin Liu[*], and Jing-Shin Chang[+]

Abstract

We studied a special case of the translation of English verbs in verb-object pairs. Researchers have studied the effects of the linguistic information of the verbs being translated, and many have reported how considering the objects of the verbs will facilitate the quality of translation. In this study, we took an extreme approach - assuming the availability of the Chinese translation of the English object. In a related exploration, we examined how the availability of the Chinese translation of the English verb influences the translation quality of the English nouns in verb phrases with analogous procedures. We explored the issue with 35 thousand VN pairs that we extracted from the training data obtained from the 2011 NTCIR PatentMT workshop and with 4.8 thousand VN pairs that we extracted from a bilingual version of *Scientific American* magazine. The results indicated that, when the English verbs and objects were known, the additional information about the Chinese translations of the English verbs (or nouns) could improve the translation quality of the English nouns (or verbs) but not significantly. Further experiments were conducted to compare the quality of translation achieved by our programs and by human subjects. Given the same set of information for translation decisions, human subjects did not outperform our programs, reconfirming that good translations depend heavily on contextual information of wider ranges.

[1] This paper was converted from the Master's thesis of the first author, which was partially published in (Chuang *et al*., 2011a) and (Chuang *et al*., 2011b).

[*] National Chengchi University, Taiwan,
E-mail: {98753004, chaolin}@nccu.edu.tw
The author for correspondence is Chao-Lin Liu.
[+] National Chi Nan University, Taiwan
E-mail: jshin@csie.ncnu.edu.tw

Keywords: Machine Translation, Feature Comparison, Near Synonyms in Chinese, E-HowNet, Human Judgments

1. Introduction

In general, the problem we are exploring is an instance of *translation of collocations* (Smadja *et al*., 1996). The collocations consist of the verbs and their direct objects, *i.e*., nouns, in verb phrases. Researchers have extensively studied the translation problems related to individual verbs/nouns (Dorr *et al*., 2002; Lapata & Brew, 2004) and verbs/nouns in phrases (Chuang *et al*., 2005; Koehn *et al*., 2003; Lü & Zhou, 2004). Some techniques have been 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 & Okuyama, 2009), and present an experience in translating common verbs and their direct objects based on bilingual and collocational 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 would improve the quality of verbs' translations. 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 information might contribute to the translation quality. It is understood and expected by many that the Chinese translations of English words might not be directly available for all cases and that a good translator should consider a lot more features to achieve high translation quality. Nevertheless, we thought it would be interesting to know how the availability of such extraordinary information could influence the translation quality within the context that we present in this paper.

The experiments were conducted with the training data available to the participants of the 2011 NTCIR Patent MT task. The original corpus contains one million pairs of a Chinese word and its English translation. We explored four different methods to determine the verb's Chinese translation. These methods utilized the bilingual and contextual information of the English verbs in different ways. Effects of these methods were compared based on experimental evaluation that was conducted with 35 thousand verb-object pairs extracted from the NTCIR corpus. Additional experiments using materials in a bilingual version of *Scientific American*[2] magazine were also conducted. (Since objects are nouns, we will refer to verb-object pairs as verb-noun pairs or VN pairs to simplify the wording.)

[2] http://www.scientificamerican.com/

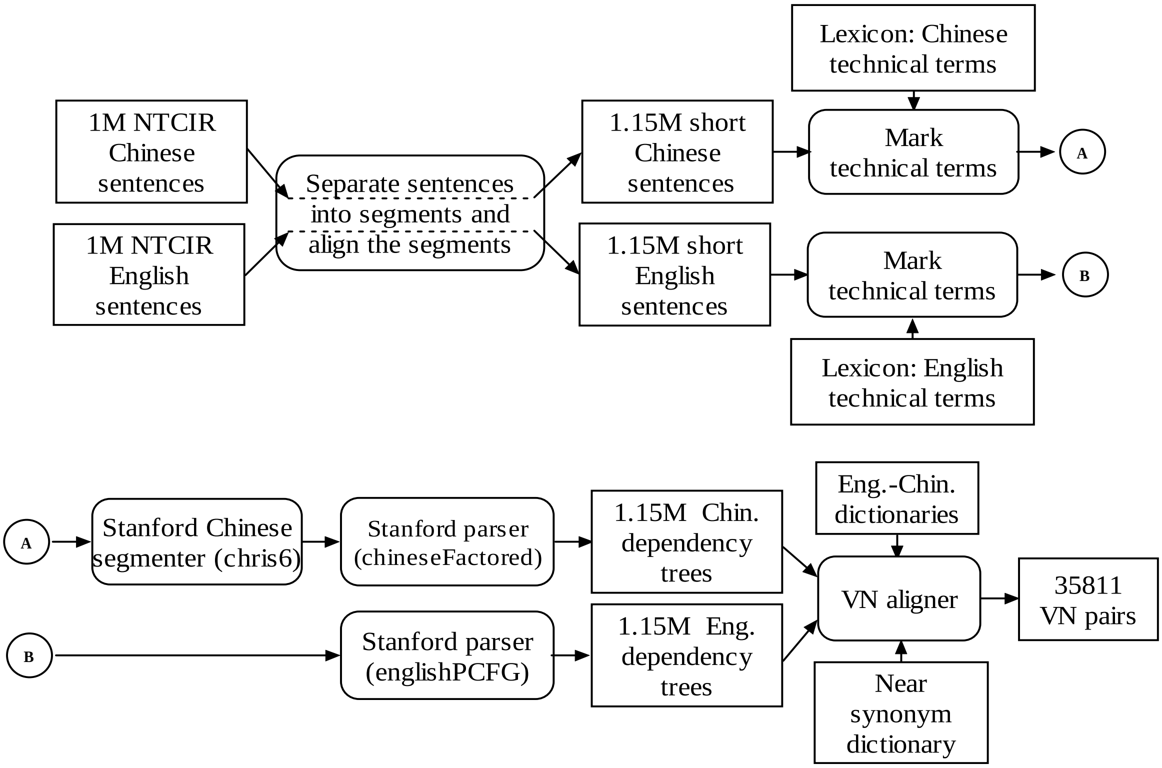
1M NTCIR Chinese sentences
1M NTCIR English sentences
Separate sentences into segments and align the segments
1.15M short Chinese sentences
1.15M short English sentences
Lexicon: Chinese technical terms
Mark technical terms
A
Mark technical terms
B
Lexicon: English technical terms
A
Stanford Chinese segmenter (chris6)
Stanford parser (chineseFactored)
1.15M Chin. dependency trees
B
Stanford parser (englishPCFG)
1.15M Eng. dependency trees
Eng.-Chin. dictionaries
VN aligner
Near synonym dictionary
35811 VN pairs

Figure 1. The procedure for extracting VN pairs from the original corpora

We provide a broad outline of our work in Section 2, and we present our methods for aligning the bilingual VN-pairs in Section 3. We explain how we build lexicons with information about synonyms to serve the needs of VN-pair alignment in Section 4 and delineate the design of our experiments in Section 5. We discuss the experimental results in Section 6, and we compare the translation quality achieved by human subjects in Section 7. Finally, we wrap up this paper in Section 8.

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 into Chinese and *vice-versa* before comparing the translation quality achieved through different combinations of collocational 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.[3] The process started from the

[3] http://ntcir.nii.ac.jp/PatentMT/

upper left of the figure. Most of the original sentences were very long. A sentence had 34 words on average, and the longest sentence had 141 words. Since our goal was to extract VN pairs from the corpus, not doing a full-scale research project in machine translation, we chose to segment the sentences into shorter parts at commas and periods. Normally, VN pairs will not expand across punctuation; even if some VN pairs did, we could afford 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 we ran our aligner on the sentences that originally belonged to a long sentence. Like the Champollion, we computed 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 1,148,632 short sentence pairs.

We employed the Stanford Chinese segmenter[4] (Chang *et al.*, 2008; Tseng *et al.*, 2005) to segment the Chinese text. This segmenter allows us to mark the technical terms, so 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 multiple ways to mark the technical terms in a string, we preferred the longer choices. English texts were tokenized by the Stanford parser[5] with the PCFG grammar (Klein & Manning, 2003). 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 the VN pairs with the method 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 35,811 VN pairs to be used 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. Details about the design and results of the experiments are provided in Sections 5 and 6.

[4] http://nlp.stanford.edu/software/segmenter.shtml, version 1.5

[5] http://nlp.stanford.edu/software/lex-parser.shtml, with the PCFG grammar, version 1.6.5

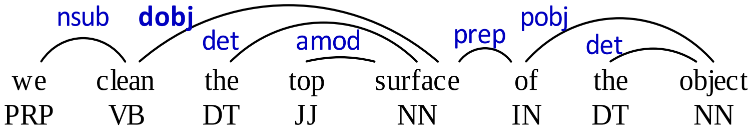

Figure 2. A sample dependency tree with POS tags

Remove the small bar.	移開小塊的木條
root(ROOT-0, Remove-1)	root(ROOT-0, 移開-1)
det(bar-4, the-2)	dep(的-3, 小塊-2)
amod(bar-4, small-3)	assmod(木條-4, 的-3)
dobj(Remove-1, bar-4)	dobj(移開-1, 木條-4)

Figure 3. A pair of aligned sentences and their dependency trees, where the `dobj` relationships can be aligned

3. VN Pair Alignment

We employed the Stanford parsers to compute the dependency trees for the parallel texts for English and Chinese. We extracted the `dobj` relations from the trees and aligned the VN pairs.

3.1 Dependency Trees

Based on the general recommendations on the Stanford site, 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.” Stanford parsers can provide the parts of speech (POSs) of words and recognize the relationships between the words. POSs 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 375,041 `dobj` links in the 1.15M short English sentences and 465,866 `dobj` links in the short Chinese part. Nevertheless, not all of the words participating in a `dobj` link were real words, and the tags for the English and the Chinese short sentences did not always agree. Figure 3 shows the dependency trees of a sample pair of short sentences containing two `dobj` relationships that would be aligned (English on the left; Chinese on the right).

#54098 pair of aligned short sentences

VN pairs in English		VN pairs in Chinese
dobj(round-7, edge-10)		dobj(清除-12, 部分-19)
dobj(remove-15, portion-17)	↔	dobj(使-24, 肩部-27)
		dobj(進-29, 圓滑-31)

Figure 4. Aligning VN pairs within an aligned short sentence

Fill the hole with water.	將洞裝滿水
root(ROOT-0, Fill-1)	nn(洞-2, 將-1)
det(hole-3, the-2)	nsubj(裝滿-3, 洞-2)
dobj(Fill-1, hole-3)	root(Root-0, 裝滿-3)
prep(Fill-1, with-4)	dobj(裝滿-3, 水-4)
pobj(with-4, water-5)	

Figure 5. A pair of aligned sentences that we could not align via the `dobj` relationships

Hence, we took two steps to align the VN pairs. First, we looked up the English and Chinese 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 254,091 and 249,591 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 is not as trivial as it might appear to be. Let (EV, EN) and (CV, CN) denote an English and a Chinese VN pair, respectively; let EV, EN, CV, and CN denote an English verb, an English noun, a Chinese verb, and a Chinese noun, respectively. We 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. An illustration of this basic idea is shown in Figure 4, where the English and the Chinese short sentences contained multiple `dobj` relationships and only one pair could be aligned.

Nevertheless, 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 and contain the Chinese translations actually 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” but does not include “改進” as a possible translation, then 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 interchangeable.

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

After completing the VN pair alignment, we obtained 35,811 aligned VN pairs, *cf*. Figure 1. Note that, although we started with 1 million pairs of long sentences, we identified less than 36 thousand VN pairs. Many problems contributed to the small number of extracted pairs. We have mentioned that translators might not use the words in dictionaries available to us when they translated. We removed `dobj` relationships that contained words not in our dictionaries. Also, the parser might not parse sentences as one might expect, and we show an example of this in Figure 5. The parser considered the “hole” as the object in the English sentence and considered “water” (水 (shui3)) as the object of the “fill” (裝滿 (zhuang1 man3)) in the Chinese sentence. Hence, the two `dobj` relationships could not be aligned.

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 to distinguish technical terms from normal text, thereby achieving higher quality of parsing.

We downloaded 138 different kinds of domain-dependent dictionaries from Taiwan National Academy for Educational Research.[6] The files contained technical term pairs in the form of (English word(s), Chinese word(s)) that were stored in Excel format. The total file size is 177MB.

The format of English-Chinese technical term pairs is not always a 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 804,068 English-Chinese technical one-to-one term 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 a plethora of ordinary words were considered technical terms.

We alleviated this problem with E-HowNet[7] (Chen *et al*., 2005) and WordNet.[8] Treating

[6] http://terms.nict.gov.tw/

[7] http://ckip.iis.sinica.edu.tw/taxonomy/taxonomy-edoc.htm

the words listed in E-HowNet and WordNet as ordinary words, we used them to identify ordinary words in our technical term pairs. If the English or the Chinese parts of the original pairs were also listed in E-HowNet or WordNet, then the pairs would be removed.

As a result, we removed 14% of the original pairs and kept 690,640 technical term pairs. The English and Chinese parts of the term pairs then were used as two dictionaries of "technical terms," shown in Figure 1.

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[9] to acquire 99,805 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[10] (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 of computing and obtaining near synonyms. E-HowNet is a lexicon for Chinese. Each entry in E-HowNet provides the information about a sense of a Chinese word. If a word can carry multiple senses, the word will have an entry for each of its senses. Among other items, an entry contains two levels of detailed semantic information for a word: TopLevelDefinition and BottomLevelExpansion. The TopLevelDefinition item in a lexical entry records the higher semantic information in the E-HowNet Ontology[11] (Chen *et al*, 2005). In contrast, the BottomLevelExpansion item in a lexical entry records the semantic information at the lowest level in the E-HowNet Ontology. The TopLevelDefinition may not contain any information when the TopLevelDefinition is the same as the same as the

[8] http://wordnet.princeton.edu/

[9] http://www.dreye.com/index_en.html

[10] http://elearning.ling.sinica.edu.tw/c_help.html

[11] http://ckip.iis.sinica.edu.tw/taxonomy/taxonomy-edoc.htm

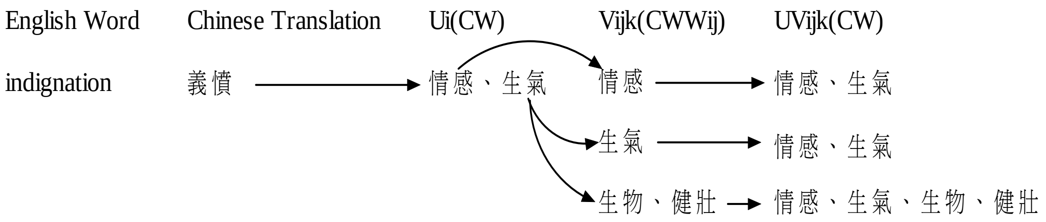

Figure 6. Expanding the Chinese translations of an English word with near synonyms

BottomLevelExpansion. The semantic definitions provided in these two entries can be used to compute similarity scores between word senses.

We determine whether two Chinese words are near synonyms by the following procedure. Given a Chinese word, CW, we looked in 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 Chinese words. Denote this set of Chinese words by U_i(CW), and let CWW_{ij} be a word in U_i(CW). We looked in E-HowNet for the senses of CWW_{ij}. Let $S_k(CWW_{ij})$ denote one of the senses of the CWW_{ij}, and let $V_{ijk}(CWW_{ij})$ denote the set of Chinese words in 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 UV_{ijk} of S_i(CW). Note that, due to lexical ambiguity, a Chinese word might have multiple such vectors.

Figure 6 shows an illustration of the process of finding near synonyms for "義憤" (yi4 fen4), which is a possible translation for "indignation". In this illustration, we assume (1) that there is only one sense for "義憤" and (2) that its semantic information contains two Chinese words: "情感" (qing2 gan3) and "生氣" (sheng1 qi4). Namely, we have CW="義憤", U_1(CW)={"情感", "生氣"}, CWW_{11}="情感", and CWW_{12}= "生氣". There is only one sense for CWW_{11}, and its combined semantic information contains only one Chinese word "情感". Hence, $V_{111}(CWW_{11})$={"情感"}. There are two senses for CWW_{12}. The combined semantic information for $S_1(CWW_{12})$ contains only "生氣," and the combined semantic information for $S_2(CWW_{12})$ contains only "生物" (sheng1 wu4) and "健壯" (jian4 jhuang4). Therefore, $V_{121}(CWW_{12})$={"生氣"} and $V_{122}(CWW_{12})$={"生物", "健壯"}. Finally, we compute the unions of U and V sets to acquire UV_{111}(CW)={"情感" , "生氣"}, UV_{121}(CW)={"情感", "生氣"}, and UV_{122}(CW)={"情感", "生氣", "生物", "健壯"}. Although we have three sets, only two of them are different. Similar to how we compute the sense vectors for "義憤" in Figure 6, we can compute the sense vectors for any Chinese words.

We treated two Chinese words as near synonyms if the cosine value of any of their sense vectors exceeded 0.7.[12] To compute the cosine value of two sense vectors, we first computed the union of the words in two vectors, treated each different word as a different dimension, and converted the word vector into a Boolean vector. Therefore, if a word in a vector did not appear in another vector, a "0" would be used in its place. Assume that we were to compute the cosine of UV_{121}(CW) and UV_{122}(CW) in the preceding paragraph, we would create a 4-dimension space of {"情感," "生氣," "生物," "健壯"}, UV_{121}(CW) would become {1, 1, 0, 0}, and UV_{122}(CW) would become {1, 1, 1, 1}.

Given an entry in our bilingual dictionary, we computed the near synonyms of the Chinese translations of each English word. This was carried out by comparing the sense vectors of Chinese translations in every English-Chinese pair with the sense vectors of 88,074 Chinese words in E-HowNet. The qualified words were added to the Chinese translations of the English words in our dictionary.

Thus, an entry for an English word in our English-Chinese dictionary includes four parts. The first part is the English word itself. The second part is the Chinese translations that we found in our dictionaries (Oxford and Dr.Eye). The third part is the synonyms, obtained from Cheng's (2004) system, for the words in the second part. The fourth part is the near synonyms that we computed with the aforementioned procedure (with E-HowNet).

The purpose of adding information about near synonyms into our bilingual dictionary was to increase the recall rates of VN-pair alignment. Having not-very-good Chinese near synonyms may not hurt our performance, unless the translators of the PatentMT corpus happened to use the same erroneous translations. Nevertheless, more complex methods for identifying synonyms, *e.g.* Bundanitsky and Hirst (2006) and Chang and Chiou (2010), may be instrumental for the study.

5. Design of the Experiments

We conducted experiments to translate from English to Chinese and from Chinese to English. In addition, in separate experiments, we tried to find the best translations of verbs, and tried to find the best translations of objects of the verbs given appropriate contexts. Nevertheless, we present the design of our experiments only with the experiments of translating English verbs to Chinese verbs in this section. Other experiments were conducted with the same procedure.

[12] Given that we did not have the context to do word sense disambiguation at this stage, we have to consider two words synonymous to each other if any of their senses are close enough. This threshold of 0.7 was chosen based on observed results of small-scale experiments and was not chosen scientifically.

5.1 Statistics about the Aligned VN pairs

We calculated the frequencies of the verbs in the 35,811 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 30,376 such aligned VN pairs. The most frequent English verb appeared 4,530 times, as shown in Table 1. The 100th most frequent English verb is “lack,” and it appeared 47 times.

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

Some of the English verbs are 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”.

Table 2. 22 most “challenging” English verbs and their indices

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

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. Intuitively, larger challenging indices imply that it is easier to achieve good translations via the most frequent translations. Table 2 lists the 22 verbs that had the smallest challenging indices.

5.2 Translation Decisions

Given the aligned VN pairs, we could compute conditional probabilities and apply the conditional probabilities to determine the Chinese translation of English words.

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)

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, at the test stage, Equation (1) prefers the most frequent Chinese translation of EV in the training data.

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 as an 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 in Equation (5). C(EV, EN) denotes the frequency that we observed the occurrences of EV and EN in the training data, and C(EV, EN, CVi) denotes the frequency that we observed the occurrence of EV, EN, and CVi in the training data.

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

The remaining equations, (3) and (4), take extreme assumptions. 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, based on the training data.

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

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

We felt that the exploration of using the information about the Chinese translation of the English noun would be interesting. Would the information about CN provide more information, assuming we had information about EV and EN? What 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 the 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 training case, we needed C(EV, EN, CN) in Equation (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 Equation (6) for the test case. When such situations occurred, we chose to allow our system to admit that it was not able to recommend a translation, rather than resorting to smoothing techniques.

6. Experimental Results

Using the formulas listed in Table 3 would allow our systems to recommend only one Chinese translation. In fact, we relaxed this unnecessary constraint by allowing our systems to consider the 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 also 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 24,300 instances of aligned VN pairs for training and 6,076 instances of aligned VN pairs for testing.

We measured four rates as the indication of the performance of using a particular formula in Table 3: rejection rates, inclusion rates, average number of actual recommendations, and average ranks of the answers.

The ***rejection rate*** is the percentage of not being able to respond to the test cases. This is due to our choosing not to smooth the probability distributions, as we explained at the end of Section 5.2.

The rejection rates were 0, 0.201, 0.262, and 0.218 when we applied Equations (1) through (4) in the experiments. It is not surprising that the rejection rates increased as we considered more information in the formulas. As expected, we encountered the highest rejection rate when using Equation (3), when we essentially collected information about four grams at the training stage. Note that using Equation (4) resulted in higher rejection rates than using Equation (2). To have to reject a test instance when we used Equation (2), we must have had no prior experience with the EN in our training data. In contrast, to have to reject a test instance when we used Equation (4), we must have had no prior experience with the CN in our training data. In reality, it was much likely not to have observed a CN for the EN in our training data than not to have observed the EN at all. Hence, it is more likely for Pr(CV_i | EV CN) to be zero than Pr(CV_i | EV EN), and the rejection rates for Equation (4) were higher.

Table 4. Inclusion rates for the top 100 verbs

Inclusion	*k*=1	*k*=3	*k*=5
Eq(1)	0.768	0.953	0.975
Eq(2)	0.786	0.913	0.918
Eq(3)	0.795	0.911	0.916
Eq(4)	0.791	0.910	0.916

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

The comparison between the results for using Equations (3) and (4) and the results of using Equation (2) show that using the bilingual information about CN improved the translation quality when *k*=1, but the changes in the inclusion rates were marginal.

It may also be surprising that the inclusion rates for Equations (2) through (4) seem to be saturated when we increase *k* from 3 to 5. This was because our systems actually could not recommend 5 possible translations when they were allowed to. 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 number of VN pairs qualified for training and testing, consequently limiting the actual numbers of available translations to recommend. Table 5 shows the average number of actual recommendations in the tests. Even when we allowed 5 recommendations (*k*=5), using Equations (2) through (4) produced only about 2 recommendations on average. This phenomenon limited the chances to increase the inclusion rates when we increased *k*.

Table 5. Average number of actual recommendations

Recommend	*k*=1	*k*=3	*k*=5
Eq(1)	1.000	2.919	4.614
Eq(2)	1.000	1.923	2.225
Eq(3)	1.000	1.847	2.107
Eq(4)	1.000	1.920	2.244

Table 6. Average ranks of the answers

Ranking	*k*=1	*k*=3	*k*=5
Eq(1)	1.000	1.241	1.310
Eq(2)	1.000	1.166	1.185
Eq(3)	1.000	1.151	1.168
Eq(4)	1.000	1.153	1.173

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 translations in the recommended translations. The first word in the recommendation list is considered Rank 1, the second word is Rank 2, *etc*. Hence, we preferred to have smaller average ranks. The average ranks improved as we considered more information from Equation (1) to Equation (2) and to Equation (3). Using Equation (2) achieved almost the same quality of translations as using Equation (4). Equation (2) achieved better inclusion rates, but Equation (4) offered better average ranks.

6.2 Improving Results for the Top 100 Verbs in Patent Documents

Results reported in the previous subsection indicated that Equation (1) is robust in that it could offer candidate answers all the time. Methods that employed more information could recommend 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 responsiveness while maintaining the translation quality. To this end, we examined all of the combinations of the basic methods listed in Table 3.

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

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

Using Equation (1) 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. Nevertheless, using Equation (1) is relatively less precise (*cf.* Table 6). Hence, we used Equation (2) through Equation (4) before using Equation (1) as a backup. Naturally, in these 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 to recommend *k* candidates.

In Tables 7 and 8, we compare the performance of these combined methods. We copy the inclusion rates of Equation (1) from Table 4 to Table 7 to facilitate the comparison, because Equation (1) 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 Equation (1) and Equation (3) from Table 6 to Table 8. Using Equation (1) and using Equation (3) led to the worst and the best average ranks, respectively, in Table 6. Again, using the combined methods, we improved the average ranks marginally over the results of using Eq. 1.

Table 9. Inclusion rates for the 22 challenging verbs

Inclusion	*k*=1	*k*=3	*k*=5
Eq(1)	0.449	0.865	0.923
Eq(2)	0.561	0.818	0.820
Eq(3)	0.564	0.827	0.829
Eq(4)	0.550	0.827	0.829

Table 10. Average number of recommendations

Recommend	*k*=1	*k*=3	*k*=5
Eq(1)	1.000	2.977	4.756
Eq(2)	1.000	2.090	2.364
Eq(3)	1.000	2.022	2.230
Eq(4)	1.000	2.106	2.411

Statistics in Table 7 suggest that using this machine-assisted approach to translate verbs in common VN pairs in the PatentMT data is feasible. Providing the top five candidates to a human translator to choose will allow the translator to find the recorded answers nearly 98% of the time. Statistics in Table 7 and Table 8 show that the combined methods were able to improve the inclusion rates and the ranks of the correct answers at the same 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 in the translation decisions (*cf.* Table 3) contributed similarly to the translation quality of the English verbs.

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 Equation (1) when *k*=1. Although the inclusion rates reduced noticeably when we used Equation (2), Equation (3), and Equation (4) as well, the drop in the inclusion rate for Equation (1) (when *k*=1) was the most significant. The 22 verbs have small challenging indices (Section 5.1), so providing only one candidate allowed considerably fewer chances to include the correct answers.

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. (Having more possible ways to translate the word made it relatively difficult for computer algorithms to translate correctly.)

Table 11. Average ranks of the answers

Ranking	*k*=1	*k*=3	*k*=5
Eq(1)	1.000	1.607	1.773
Eq(2)	1.000	1.365	1.373
Eq(3)	1.000	1.374	1.383
Eq(4)	1.000	1.394	1.400

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

Corresponding numbers in Table 6 and Table 11 support the claim that translating the 22 challenging words is 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 that recommends the top *k* possible translations for these 22 verbs, the performance would not be as good as what we could achieve by building 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.

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 result in significant differences. When we replaced Equation (2) with Equation (4), or *vice-versa*, in the experiments, we observed very similar results in Tables 12 and 13 most of the time.

6.4 Translating English Nouns

We repeated the experiments that we discussed in Sections 6.1, 6.2, and 6.3 for the top 100 nouns in the PatentMT data. The top 100 nouns appeared in 19,756 VN pairs. The word “method” was the most frequent object in the VN pairs, and it appeared 982 times. For experiments with these nouns, we had 15,804 training instances and 3,952 test instances.

Table 14. Translation decisions for nouns

$\arg\max_{CN_i} \Pr(CN_i \mid EN)$	(8)
$\arg\max_{CN_i} \Pr(CN_i \mid EV, EN)$	(9)
$\arg\max_{CN_i} \Pr(CN_i \mid EV, EN, CV)$	(10)
$\arg\max_{CN_i} \Pr(CN_i \mid EN, CV)$	(11)

Table 15. Average ranks of the answers for translating the nouns

Ranking	*k*=1	*k*=3	*k*=5
Eq(8)	1.000	1.171	1.223
Eq(9)	1.000	1.118	1.138
Eq(10)	1.000	1.104	1.125
Eq(11)	1.000	1.116	1.142

The goal was to find the best Chinese translation of the English objects, given its collocational and bilingual information. The structure of the experiments was analogous to what we have reported for the experiments for finding the best translations of English verbs. 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. Hence, the translation decisions that we listed in Table 3 became those in Table 14.

The statistics showed analogous trends that we discussed in the previous sections. Namely, the availability of the Chinese translations of the English verbs was useful but did not help significantly when we already considered the English verbs and objects in the translation decisions, so we do not show all of the tables for the results in this paper. The rejection rates observed when we used Equations (8) through (11) were 0, 0.126, 0.184, and 0.128, respectively. The average ranks of the correct answers for the English nouns are listed in Table 15.

6.5 Experiments using Aligned Sentences in *Scientific American*

Scientific American is a magazine for introducing scientific findings to the general public. The writing style is close to ordinary life. We ran our sentence aligner (Tien *et al*., 2009) to extract aligned sentences from 1,745 articles that were published between 2002 and 2009 in the bilingual version of *Scientific American*[13]. We extracted 63,256 pairs of sentence pairs and ran the procedure depicted in Figure 1 over this set of sentence pairs to obtain 4,814 VN pairs. This scale of experiment is smaller than with the PatentMT corpus.

Since we had only 4,814 VN pairs, we chose only the 25 most frequent verbs in the experiments. This selection further reduced available VN pairs to only 1,885 pairs. With an 8:2 split for training and test data, we had only 1,508 training instances and 377 test instances. The procedure for the experiments was the same as reported in Sections 6.1 and 6.2. Again, the observed statistics indicated that using the Chinese translations of the English objects helped the translation quality of the English verbs, but the improvement was not significant. An incidental observation was that it was harder to find good translations of English verbs in *Scientific American* than in the PatentMT corpus. When providing five recommendations (k=5), only about 88% of the time the recommendations of our system could include the correct translations. In contrast, we had achieved inclusion rates well above 90% in Tables 7 and 12 in the experiments that used PatentMT corpus.

7. A Comparison with Human Performance

Using equations listed in Table 3 and Table 14 to make translation decisions posed a serious constraint on the available information for achieving good translations. A good translator would check a larger context to select the best translations. What would ordinary people achieve if they were provided the same limited information that our systems were provided?

To explore this interesting question, we recruited 52 human subjects who were Computer Science majors at the time of testing. Some of them were undergraduates, and some were graduate students. We placed them into three groups for three different tests: 17, 19, and 16 subjects in Test 1, Test 2, and Test 3, respectively. No human subject participated in different tests because the test questions were similar.

We chose 10 instances of verb translations from our *Scientific American* corpus, and converted each of them into three different formats for different tests. These 10 verbs were among the 25 most frequent verbs in the aligned VN pairs in our *Scientific American* corpus.

[13] http://sa.ylib.com/

Table 16. A sample question for Test 1 and Test 2

English sentence	Investigators are, of course, also exploring additional avenues for **improving** efficiency; as far as we know, though, those other approaches generally extend existing methods.
Chinese sentence	當然，研究人員也在尋找其他可______效率的方法，但就我們目前所知，其他方法一般只是延伸現有的途徑罷了。
Available choices	(1) 增進 (2) 提高 (3) 改進 (4) 改善
Possible translations and their frequencies for “improve” in *Scientific American*	improve={利用=1, 增加=1, 改良=1, 運用=1, 使=2, 加強=3, 提高=4, 改進=4, 增進=11, 改善=22}

Table 17. A sample question for Test 3

Test question	**improve** efficiency: ______ 效率
Available choices	(1) 增進 (2) 提高 (3) 改進 (4) 改善

The formats varied in the information available to the translators. Table 16 shows a test instance for Test 1. In this test, the human translators were provided 10 test instances. In each test instance, there was (1) a complete English sentence with a highlighted verb; (2) a partially translated Chinese sentence for the English sentence, with the translation for the highlighted English verb removed; and (3) four candidate Chinese verbs to be used to translate the highlighted English verb. The candidate Chinese verbs, listed in the row of “Available choices,” were selected from the translations of the highlighted English verbs in our corpus. The very last row shows the complete list of the translations for “improve” in our corpus, but this list was not provided to the human subjects.

In Test 2, the human subjects had to respond to 10 test instances. The format was the same as that for Test 1, except that the candidate Chinese verbs were not provided. The human subjects had to fill in the blanks in the Chinese sentences in Test 2.

Table 17 shows a test question for Test 3. In Test 3, the human subjects would also have to respond to 10 test questions, and they only saw the English verb, the English object, and the Chinese translation of the English object. The subjects had to choose the best translation from the list of candidate translations.

The human subjects could take their time to respond to 10 questions in the tests. There were no time limits. They usually turned in their responses within a short time, but they did not always respond to all questions. Correctness of their responses was judged based on the actual translations in *Scientific American*, even when other alternatives were also reasonable for the test questions. The sample question shown in Table 17 is an obvious example. In this example, all four translations are reasonable Chinese verbs to go with the Chinese noun. That was because there was no contextual information in Table 17 to distinguish the subtle differences between the candidate translations. Nevertheless, the original sentence pairs, shown in Table 16, were translated in exactly one way among the alternatives. Therefore, only one of the choices was considered correct.

Table 18. Average correct rates of human subjects and Equation (3)

	Human Subjects	Equation (3)
Test 1	0.524	0.600
Test 2	0.342	0.600
Test 3	0.395	0.600

We applied Equation (3), *k*=1, in Table 3 in this experiment. The average correct rates achieved by the human subjects and our programs in three tests are collected in Table 18. The correct rate is the portion of test questions with correct responses. More specifically, questions that were not answered were considered incorrect responses, and this principle applied to both human translators and our programs. Our programs made decisions only based on the English verbs, the English nouns, and the Chinese nouns in all tests. Hence, its performance was 0.6 and remained the same in all of the tests. In contrast, the average correct rates achieved by the human subjects varied with the difficulty of the tests. The human subjects performed best in Test 1, partially because they were offered more information to make decisions. Test 2 was the most difficult one, because the subjects had to provide Chinese translations themselves on the fly. The difficulty of the test questions in Test 3 was similar to those in Test 2, but the human subjects were provided with candidate translations, so the average correct rate was higher.

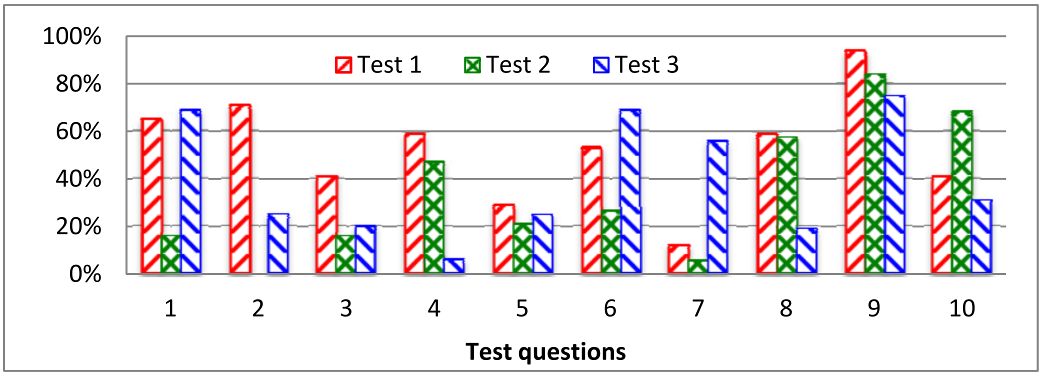

Figure 7. Average correct rates of the human translators

Figure 7 shows the average correct rates for individual questions in the three tests. The averages were computed based on the responses of the human subjects who participated in the tests. Although the average correct rates listed in Table 18 corresponded approximately to the average difficulty levels of the test formats, the performance of human subjects varied with the individual test questions. In Table 18, the average correct rate for Test 1 is the highest. In Figure 7, we can see that the correct rates for questions used in Test 1 did not always exceed those for the corresponding questions used in Test 2 and Test 3.

We do not mean to interpret results of these simple tests as a competition between human beings and computers. The results, however, suggest that translating English verbs based on partial information, *i.e.*, the English verb, the English noun, and the Chinese noun can be difficult for human subjects. The average correct rates can be seriously impacted when we insisted that there was exactly one correct answer for a test question, where the answer was defined based on the original corpus.

A previous reviewer of our work contended that we should treat all of the candidate Chinese translations in Table 16 as correct answers. Although that is a reasonable consideration, when we evaluate a system with a considerable number of test questions, doing so would require a non-negligible amount of human intervention. One possible approach might be to create an evaluation system that considers “acceptable answers” while comparing the outputs of a decoder and the expected translations.

8. More Discussion

We discuss some issues raised by anonymous reviewers in this section.

One reviewer questioned the use of the Stanford parser for both English and Chinese material, and wondered whether we should have used the CKIP parser[14] for Chinese. The point was brought up because the CKIP parser may be more reliable than the Stanford parser for Chinese.

While we agree with the reviewer about the reliability of the CKIP parser, we chose to employ the Stanford parser for both languages for two reasons at the time of our implementation. The first reason was that we needed the parsers to provide not just parse trees but also dependency relationships between words, *i.e.*, the `dobj` relationship. Using the same parser for both languages made our processing more efficient. The second reason was that the Stanford parser is an open system, so we can download the parser and parse our text on our computers. In contrast, we have to submit text material to the CKIP server for services. For copyrighted material, we were not sure that it was appropriate to rely on the CKIP services.

A concern was about how we deal with the forms of English words, *e.g.*, the tenses of verbs, in the translation of the VN pairs. The tenses of English verbs carry information about when the actions were taken, so are crucial for quality translation. Nevertheless, when we generated the VN pairs from the NTCIR corpus (Figure 1), we lemmatized the English words.

[14] http://godel.iis.sinica.edu.tw/CKIP/parser.htm

Hence, the current work, as the reviewers have noticed, did not aim at choosing the correct morphological forms for the English verbs. Similarly, we did not attempt to choose the singular and plural forms for nouns either. This issue should be tackled in further studies.

Table 19. Frequencies of 22 most "challenging" English verbs

Verb	make	exhibit	add	represent	retain	leave	enhance	reduce	lack	improve	achieve
	114	103	138	373	131	61	178	774	47	322	329
Verb	employ	reach	create	give	replace	take	apply	adjust	obtain	carry	explain
	135	119	201	70	53	210	50	69	329	241	54

Another question was about how the selection of verbs (or nouns) influences the general implication of our experimental results. Namely, how general are our results? Table 19 shows the frequencies of the 22 most challenging verbs. Evidently, the sample sizes of these verbs were not as large as those of the 20 most frequent English verbs in our dataset (*cf*. Table 1). Nevertheless, most of them were frequent enough for conducting experiments.

The resulting differences between choosing the most frequent verbs and the most challenging verbs were discussed in Section 6.3. When using the most challenging ones, the most noticeable changes were that it became more difficult to recommend the best translations of the verbs with the same number, *i.e.*, *k*, of recommendations. The inclusion rates dropped, *cf*. Table 4 and Table 9, especially when we recommended only one candidate translation. The ranks of the true answers worsened as well, *cf*. Table 6 and Table 11.

We believe that the changes observed in the experimental results are general because of the definition of degrees of challenging index (*cf*. Section 5.1). A word is more challenging if its most frequent translation is not significantly more frequent than its second frequent translation. Hence, using the challenging words made it more difficult to achieve good translations, given the same contextual information and the same number of recommended translations.

The presentation of the human performance triggered some questions. The first one was about the answers to the tests. The test item in Table 17 shows a confusing example, in which some distractors are acceptable to native speakers. Hence, a natural question is about how a "correct" answer was defined.

We touched upon this question at the end of Section 7. Apparently, some distractors are acceptable to native speakers, and some of them should have been considered correct. Nevertheless, when we evaluated a computer program, we normally had one correct answer in the test data. Even though the computer program "knew" a lot of acceptable synonyms of the correct answer, it still has to find "the" answer to be considered "correct" in the evaluation. The example shown in Table 17 is such an example. To make the computers and human subjects be evaluated on the same basis, we allowed only one answer from the available choices. The available choices came from the training data, and the answer for a test item was based on the original English and Chinese sentence pair.

When the human subjects were given contextual information in Test 1 in Section 7, they did not perform very well on average. One obvious reason was because of the multiple attractive candidates, which we discussed in the last paragraph. One may also challenge the language ability of the human subjects. Indeed, we chose the human subjects from engineering majors at the levels of undergraduates and graduate students, but we did not test their language ability before the experiments. If we were to investigate machine translation problems in this line of concern, we would probably have to ask whether all available bilingual textual material were produced by qualified linguistic and domain-dependent experts. This line of work should be important for the research community.

A reviewer stated that the human subjects did not always perform better in “easier” tasks in Test 1. For instance, in the seventh question in Figure 7, the human subjects performed much better in Test 3 than in Test 1. This may be possible for a variety of reasons. For instance, without context, the “correct” answer happened to be the most frequently collocating words, and, with context, the human subjects were distracted by confusing information in the context. As a consequence, it became easier to guess the correct answer without context.

Although we believe it is informative to compare the performance of our methods and the performance of human subjects, we did not intend to design a waterproof psycholinguistic experiment in Section 7. Hence, we chose the test instances arbitrarily from the dataset, and we compared the average performance of just 52 human subjects. A more carefully-designed psycholinguistic investigation may reveal more serious details about human performance in language translation.

9. Concluding Remarks

We designed a procedure to extract and align VN pairs in bilingual corpora. The PatentMT corpus contains 1 million pairs of English and Chinese sentences, and we aligned 35,811 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 information about 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 similar results in the experiments.

Additional and analogous experiments were conducted with the PatentMT data. In these new experiments, we aimed at the translating 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 the English noun in the conditions in the formulas for translation decisions (partially shown in Table 14), the Chinese translations of the English verbs did not offer much extra help.

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, and are grateful to Ms. Chia-Chi Tsai and Mr. Jui-Ping Wang for their help in some important steps of data processing. We also thank Mr. Sun for his help in English.

References

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

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

Chang, J.-S. & Chiou, S.-J. (2010). An EM algorithm for context-based searching and disambiguation with application to synonym term alignment. *Proceedings of the Twenty Third Pacific Asia Conference on Language, Information and Computation*, 2, 630-637.

Chang, P.-C., Galley, M., & Manning, C. D. (2008). Optimizing Chinese word segmentation for machine translation Performance. *Proceedings of the ACL Third Workshop on Statistical Machine Translation*, 224-232.

Chang, Y. C., Chang, J. S., Chen, H. J., & Liou, H. C. (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 Thirteenth 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., Jian, J.-Y., Chang, Y.-C. & Chang, J. S. (2005). Collocational translation memory extraction based on statistical and linguistic information. *International Journal of Computational Linguistics and Chinese Language Processing*, 10(3), 329-346.

Chuang, Y.-H., Liu, C.-L., & Chang, J.-S. (2011a). Translating common English and Chinese verb-noun pairs in technical documents with collocational and bilingual information.

Proceedings of the Twenty Fifth Pacific Asia Conference on Language, Information and Computation, 493-502.

Chuang, Y.-H., Wang, J.-P., Tsai, C.-C., & Liu, C.-L. (2011b). Collocational influences on the Chinese translation of non-technical English verbs and their objects in technical documents. *Proceedings of the Twenty Third Conference on Computational Linguistics and Speech Processing*, 94-108. (in Chinese)

Chen, A., Jiang, H., & Gey, F. (2000). Combining multiple sources for short query translation in Chinese-English cross-language information retrieval. *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, 17-23.

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

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

Klein, D. & Manning, C. D. (2003). Accurate unlexicalized parsing. *Proceedings of the Forty First Meeting of the Association for Computational Linguistics*, 423-430.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 48-54.

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

Lu, B., Tsou, B. K., Jiang, T., Kwong, O. Y., & Zhu, J. (2010). Mining large-scale parallel corpora from multilingual patents: An English-Chinese example and its application to SMT. *Proceedings of the First CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 79-86.

Lü, Y. & Zhou, M. (2004). Collocation translation acquisition using monolingual corpora. *Proceedings of the Forty Second Annual Meeting on Association for Computational Linguistics*, 167-174.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. *Proceedings of the Fifth International Conference of the Language Resources and Evaluation*, 489-492.

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1), 1-38.

Seneff, S., Wang, C., & Lee, J. (2006). Combining linguistic and statistical methods for bi-directional English Chinese translation in the flight domain. *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas*, 213-222.

Tien, K.-W., Tseng, Y.-H., & Liu, C.-L. (2009). Sentence alignment of English and Chinese patent documents. *Proceedings of the Twenty First Conference on Computational Linguistics and Speech Processing*, 85-99. (in Chinese)

Tseng, H., Chang, P.-C., Andrew, G., Jurafsky, D., & Manning, C. D. (2005). A conditional random field word segmenter. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 168-171.

Yokoama, S. & Okuyama, M. (2009). Translation disambiguation of patent sentences using case frames. *Proceedings of the Third Workshop on Patent Translation*, in Machine Translation Summit XII, 33-36.