Acquiring Translation Equivalences of Multiword Expressions by Normalized Correlation Frequencies

Ming-Hong Bai[1,2] **Jia-Ming You**[1] **Keh-Jiann Chen**[1] **Jason S. Chang**[2]
1 Institute of Information Science, Academia Sinica, Taiwan
2 Department of Computer Science, National Tsing-Hua University, Taiwan
mhbai@sinica.edu.tw, swimming@hp.iis.sinica.edu.tw,
kchen@iis.sinica.edu.tw, jschang@cs.nthu.edu.tw

Abstract

In this paper, we present an algorithm for extracting translations of any given multiword expression from parallel corpora. Given a multiword expression to be translated, the method involves extracting a short list of target candidate words from parallel corpora based on scores of normalized frequency, generating possible translations and filtering out common subsequences, and selecting the top-n possible translations using the Dice coefficient. Experiments show that our approach outperforms the word alignment-based and other naive association-based methods. We also demonstrate that adopting the extracted translations can significantly improve the performance of the Moses machine translation system.

1 Introduction

Translation of multiword expressions (MWEs), such as compound words, phrases, collocations and idioms, is important for many NLP tasks, including the techniques are helpful for dictionary compilation, cross language information retrieval, second language learning, and machine translation. (Smadja et al., 1996; Gao et al., 2002; Wu and Zhou, 2003). However, extracting exact translations of MWEs is still an open problem, possibly because the senses of many MWEs are not compositional (Yamamoto and Matsumoto, 2000), i.e., their translations are not compositions of the translations of individual words. For example, the Chinese idiom 坐視不理 should be translated as "*turn a blind eye,*" which has no direct relation with respect to the translation of each constituent (i.e., *"to sit", "to see" and "to ignore"*) at the word level.

Previous SMT systems (e.g., Brown et al., 1993) used a word-based translation model which assumes that a sentence can be translated into other languages by translating each word into one or more words in the target language. Since many concepts are expressed by idiomatic multiword expressions instead of single words, and different languages may realize the same concept using different numbers of words (Ma et al., 2007; Wu, 1997), word alignment based methods, which are highly dependent on the probability information at the lexical level, are not well suited for this type of translation.

To address the above problem, some methods have been proposed for extending word alignments to phrase alignments. For example, Och et al. (1999) proposed the so-called *grow-diag-final* heuristic method for extending word alignments to phrase alignments. The method is widely used and has achieved good results for phrase-based statistical machine translation. (Och et al., 1999; Koehn et al., 2003; Liang et al., 2006). Instead of using heuristic rules, Ma et al. (2008) showed that syntactic information, e.g., phrase or dependency structures, is useful in extending the word-level alignment. However, the above methods still depend on word-based alignment models, so they are not well suited to extracting the translation equivalences of semantically opaque MWEs due to the lack of word level relations between the translational correspondences. Moreover, the aligned phrases are not precise enough to be used in many NLP applications like dictionary compilation, which require high quality translations.

Association-based methods, e.g., the *Dice coefficient*, are widely used to extract translations of MWEs. (Kupiec, 1993; Smadja et al., 1996; Kitamura and Matsumoto, 1996; Yamamoto and Matsumoto, 2000; Melamed, 2001). The advantage of such methods is that association relations are established at the phrase level instead of the lexical level, so they have the potential to resolve the above-mentioned translation problem. However, when applying association-based methods, we have to consider the following complications. The first complication, which we call the *contextual effect*, causes the extracted translation to contain noisy words. For

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 478–486,
Singapore, 6-7 August 2009.

example, translations of the Chinese idiom *兩全其美* (*best of both worlds*) extracted by a naive association-based method may contain noisy collocation words like *difficult*, *try* and *cannot*, which are not part of the translation of the idiom. They are actually translations of its collocation context, such as *難以*(*difficult*), *嘗試*(*try*), and *不能*(*cannot*). This problem arises because naive association methods do not deal with the effect of strongly collocated contexts carefully. If we can incorporate lexical-level information to discount the noisy collocation words, the contextual effect could be resolved.

English (y)	f_y	$f_{x,y}$	Dice(x,y)
quote **out of context**	22	19	0.56
take **out of context**	17	11	0.35
interpret **out of context**	2	2	0.08
out of context	53	32	0.65

Table 1. The Dice coefficient tends to select a common subsequence of translations. (The frequency of *斷章取義*, f_x, is 46.)

The second complication, which we call the *common subsequence problem*, is that the Dice coefficient tends to select the common subsequences of a set of similar translations instead of the full translations. Consider the translations of *斷章取義* (*quote out of context*) shown in the first three rows of Table 1. The Dice coefficient of each translation is smaller than that of the common subsequence "*out of context*" in the last row. If we can tell common subsequence apart from correct translations, the common subsequence problem could be resolved.

In this paper, we propose an improved precision method for extracting MWE translations from parallel corpora. Our method is similar to that of Smadja et al. (1996), except that we incorporate lexical-level information into the association-based method. The algorithm works effectively for various types of MWEs, such as phrases, single words, rigid word sequences (i.e., no gaps) and gapped word sequences. Our experiment results show that the proposed translation extraction method outperforms word alignment-based methods and association-based methods. We also demonstrate that precise translations derived by our method significantly improve the performance of the Moses machine translation system.

The remainder of this paper is organized as follows. Section 2 describes the methodology for extracting translation equivalences of MWEs. Section 3 describes the experiment and presents the results. In Section 4, we consider the application of our results to machine translation. Section 5 contains some concluding remarks.

2 Extracting Translation Equivalences

Our MWE translation extraction method is similar to the two-phase approach proposed by Smadja et al. (1996). The two phases can be briefly described as follows:

Phase 1: Extract candidate words correlated to the given MWE from parallel text.

Phase 2:

1. Generate possible translations for the MWE by combining the candidate words.
2. Select possible translations by the Dice coefficient.

We propose an association function, called the *normalized correlation frequency*, to extract candidate words in the phase 1. This method incorporates lexical-level information with association measure to overcome the *contextual effect*. In phase 2, we also propose a *weighted frequency* function to filter out false common subsequences from possible translations. The filtering step is applied before the translation selecting step of phase 2.

Before describing our extraction method, we define the following important terms used throughout the paper.

Focused corpus (*FC*): This is the corpus created for each targeted MWE. It is a subset of the original parallel corpora, and is comprised of the selected aligned sentence pairs that contain the source MWE and its translations.

Candidate word list (*CW*): A list of extracted candidate words for the translations of the source MWE.

2.1 Selecting Candidate Words

For a source MWE, we try to extract from the *FC* a set of k candidate words *CW* that are highly correlated to the source MWE. We then assume that the target translation is a combination of some words in *CW*. As noted by Smadja et al. (1996), this two-step approach drastically reduces the search space.

However, translations of collocated context words in the source word sequence create noisy candidate words, which might cause incorrect extraction of target translations by naive statistical correlation measures, such as the Dice coef-

ficient used by Smadja et al. (1996). The need to avoid this context effect motivates us to propose a candidate word selection method that uses the normalized correlation frequency as an association measure.

The rationale behind the proposed method is as follows. When counting the word frequency, each word in the target corpus normally contributes a frequency count of one. However, we are only interested in the word counts correlated to a **MWE**. Therefore, intuitively, we define the normalized count of a target word *e* as the translation probability of *e* given the **MWE**.

We explain the concept of normalizing the correlation count in Section 2.1.1 and the computation of the normalized correlation frequency in Section 2.1.2.

2.1.1 Normalizing Correlation Count

We propose an association measure called the *normalized correlation frequency*, which ranks the association strength of target words with the source MWE. For ease of explanation, we use the following notations: let $\mathbf{f}=f_1,f_2,\ldots,f_m$ and $\mathbf{e}=e_1,e_2,\ldots,e_n$ be a pair of parallel Chinese and English sentences; and let $\mathbf{w}=t_1,t_2,\ldots,t_r$ be the Chinese source MWE. Hence, **w** is a subsequence of **f**.

When counting the word frequency, each word in the target corpus normally contributes a frequency count of one. However, since we are interested in the word counts that correlate to **w**, we adopt the concept of the translation model proposed by Brown et al (1993). Each word *e* in a sentence **e** might be generated by some words, denoted as **r**, in the source sentence **f**. If **r** is non-empty the relation between **r** and **w** should fit one of the following cases:

1) All words in **r** belong to **w**, i.e., $\mathbf{r} \subseteq \mathbf{w}$, so we say that *e* is only generated by **w**.
2) No words in **r** belong to **w**, i.e., $\mathbf{r} \subseteq \mathbf{f} - \mathbf{w}$, so we say that *e* is only generated by context words.
3) Some words in **r** belong to **w**, while others are context words.

Intuitively, In Cases 1 and 2, the correlation count of an instance *e* should be 1 and 0 respectively. In Case 3, the normalized count of *e* is the expected frequency generated by **w** divided by the expected frequency generated by **f**. With that in mind, we define the *weighted correlation count*, *wcc*, as follows:

$$wcc(e;\mathbf{e},\mathbf{f},\mathbf{w}) = \frac{\sum_{\forall f_i \in \mathbf{w}} p(e \mid f_i) + \Delta \mid \mathbf{w} \mid}{\sum_{\forall f_j \in \mathbf{f}} p(e \mid f_j) + \Delta \mid \mathbf{f} \mid},$$

where Δ is a very small smoothing factor in case *e* is not generated by any word in **f**. The probability $p(e \mid f)$ is the word translation probability trained by IBM Model 1 on the whole parallel corpus.

The rationale behind the *weighted correlation count*, *wcc*, is that if *e* is part of the translation of **w,** then its association with **w** should be stronger than other words in the context. Hence its *wcc* should be closer to 1. Otherwise, the association is weaker and the *wcc* should be closer to 0.

2.1.2 Normalized Correlation

Once the weighted correlation counts *wcc* is computed for each word in *FC*, we compute the normalized correlation frequency for each word *e* as the total sum of the $wcc(e;\mathbf{e},\mathbf{f},\mathbf{w})$ of all **w** in bilingual sentences (**e**, **f**) in *FC*. The *normalized correlation frequency (ncf)* is defined as follows:

$$ncf(e;\mathbf{w}) = \sum_{i=1}^{n} wcc(e;\mathbf{e}^{(i)},\mathbf{f}^{(i)},\mathbf{w}).$$

We choose the top-n English words ranked by *ncf* as our candidate words and filter out those whose *ncf* is less than a pre-defined threshold. Table 2 shows the candidate words for the Chinese term 斷章取義 (quote/take/interpret out of context) sorted by their *ncf* values. To illustrate the effectiveness *ncf*, we also display candidate words of the term with their Dice values in Tables 3. As shown in the tables, noise words such as *justify, meaning* and *unfair* are ranked lower using *ncf* than using Dice, while correct candidates, such as *out*, *take* and *remark* are ranked higher. We present more experimental results in Section 3.

2.2 Generation and Ranking of Candidate Translations

After determining the candidate words, candidate translations of **w** can be generated by marking the candidate words in each sentence of *FC*. The word sequences marked in each sentence are deemed possible translations. At the same time, the weakly associated function words,

Candidate words *e*	freq	*ncf*(*e*,**w**)
context	54	31.55
out	58	24.58
quote	26	5.84
take	23	4.81
remark	8	1.84
interpret	3	1.38
piecemeal	1	0.98
deliberate	3	0.98

Table 2. Candidate words for the Chinese term *斷章取義* sorted by their global normalized correlation frequencies.

Candidate words *e*	freq	*dice*(*e*,**w**)
context	54	0.0399
quote	26	0.0159
deliberate	3	0.0063
justify	3	0.0034
interpretation	7	0.0032
meaning	3	0.0029
cite	3	0.0025
unfair	4	0.0023

Table 3. Candidate words for the Chinese term *斷章取義* sorted by their Dice coefficient values.

which we fail to select in the candidate word selection stage, should be recovered. The rule is quite simple: if a function word is adjacent to any candidate word, it should be recovered. For example, in the following sentence, the function word *of* would be recovered and added to the marked sequence:

"The financial secretary has been **quoted out** of **context.**
財政司 司長 之 談話 被 **斷章取義.**"

The marked words are shown in boldface.

2.2.1 Generating Possible Translations

Although we have selected a reliable candidate word list, it may still contain some noisy words due to the MWE's collocation context. Consider the following example:

...as **quoted** in the audit report, if **taken out of context...**

In this instance, *quoted* is a false positive; therefore, the marked word sequence **m** "*quoted taken out of context*" is not the correct translation. To avoid such false positives, we include **m** and all its subsequences as possible translations.

quoted taken out of context
quoted taken out of
quoted taken out context
quoted taken of context
quoted out of context
taken out of context
…
quoted out
taken out
quoted
taken
out
context

Table 4. Example subsequences generated of **w** and add them to the candidate translation list.

Table 4 shows the subsequences of **m** in the above example. The generation process is used to increase the coverage of correct translations in the candidate list; otherwise, many correct translations will be lost. However, the process may also trigger the side effect of the common subsequence problem described in Section 1. Since all candidates compete for the best translations by comparing their association strength with **w,** the common subsequences will have an advantage.

2.2.2 Filtering Common Subsequences

To resolve the *common subsequence effect* problem, we evaluate each candidate translation, including its subsequences, by a concept similar to the normalized correlation frequency. As mentioned in Section 1, the Dice coefficient tends to select the common subsequences of some candidates because they have higher frequencies. To avoid this problem, we use the normalized correlation frequency to filter out false common subsequences from the candidate translation list. Here, we also use the weighted correlation count *wcc* to weight the frequency count of a candidate translation. Suppose we have a marked sequence in a sentence, **m,** whose subsequences are generated in the way described in the previous section. If the weighted count of **m** is assigned the score 1, the *weighted count* (*wc*) of a subsequence **t** is then defined as follows:

$$wc(\mathbf{t};\mathbf{e},\mathbf{f},\mathbf{m},\mathbf{w}) = \prod_{\forall e \in \mathbf{m}-\mathbf{t}} (1 - wcc(e;\mathbf{e},\mathbf{f},\mathbf{w})) .$$

The underlying concept of *wc* is that the original marked sequence **m** is supposed to be the most

likely translation of **w** and the weighted count is set to 1. Then, if a subsequence **t** is generated by removing a word *e* from **m**, the weighted count of the subsequence is reduced by multiplying the complement probability of *e* generated by **w**. Note that the weighted correlation count *wcc* is the probability of the word *e* generated by **w**.

After all $wc(\mathbf{t};\mathbf{e},\mathbf{f},\mathbf{m},\mathbf{w})$ in each sentence of the *FC* have been computed, the *weighted frequency* for a sequence **t** can be determined by summing the weighted frequencies over *FC* as follows:

$$wf(\mathbf{t};\mathbf{w}) = \sum_{\forall(\mathbf{e},\mathbf{f})\in FC} wc(\mathbf{t};\mathbf{e},\mathbf{f},\mathbf{m},\mathbf{w}).$$

We compute the *wf* for each candidate translation and then sort the candidate translations by their *wf* values.

Next, we filter out common subsequences based on the following rule: for a sequence **t**, if there is a super-sequence **t'** on the sorted candidate translation list and the *wf* value of **t** is less than that of **t'**, then **t** is assumed be a common subsequence of real translations and removed from the list.

candidate translation list	freq	wf
quote out of context	19	17.55
of context	35	15.45
out of context	32	14.82
quote of context	19	13.32
out	35	11.92
quote	23	11.63
quote out	19	9.42

Table 5. Part of the candidate translation list for the Chinese idiom, *斷章取義*, sorted by the *wf* values.

Table 5 shows an example of the rule's application. The candidate translation list is sorted by the translations' *wf* values. Then, candidates 2-7 are removed because they are subsequences of the first candidate and their *wf* values are smaller than that of the first candidate.

2.3 Selection of Candidate Translations

Having removed the common subsequences of real translations from the candidate translation list of **w**, we can select the best translations by comparing their association strength with **w** for the remaining candidates. The Dice coefficient is a good measure for assessing the association strength and selecting translations from the candidate list. For a candidate translation **t,** the Dice coefficient is defined as follows:

$$Dice(\mathbf{t},\mathbf{w}) = \frac{2p(\mathbf{t},\mathbf{w})}{p(\mathbf{t})+p(\mathbf{w})}.$$

Where $p(\mathbf{t},\mathbf{w})$, $p(\mathbf{t})$, $p(\mathbf{w})$ are probabilities of (**t**,**w**), **t**, **w** derived from the training corpus.

After obtaining the Dice coefficients of the candidate translations, we select the top-n candidate translations as possible translations of **w**.

3 Experiments

In our experiments, we use the Hong Kong Hansard and the Hong Kong News parallel corpora as training data. The training data was preprocessed by Chinese word segmentation to identify words and parsed by Chinese parser to extract MWEs. To evaluate the proposed approach, we randomly extract 309 Chinese MWEs from training data, including dependent word pairs and rigid idioms. We then randomly select 103 of those MWEs as the development set and use the other 206 as the test set. The reference translations of each Chinese MWE are manually extracted from the parallel corpora.

3.1 Evaluation of Word Candidates

To evaluate the method for selecting candidate words, we use the coverage rate, which is defined as follows:

$$coverage = \frac{1}{n}\sum_{\forall \mathbf{w}} \frac{|A_{\mathbf{w}} \cap C_{\mathbf{w}}|}{|A_{\mathbf{w}}|},$$

where *n* is the number of MWEs in the test set, $A_{\mathbf{w}}$ denotes the word set of the reference translations of **w**, and $C_{\mathbf{w}}$ denotes a candidate word list extracted by the system.

Table 6 shows the coverage of our method, NCF, compared with the coverage of the IBM model 1 and the association-based methods MI, Chi-square, and Dice. As we can see, the top-10 candidate words of NCF cover almost 90% of the words in the reference translations. Whereas, the coverage of the association-based methods and IBM model 1 is much lower than 90%. The result implies that the candidate extraction method can extract a more precise candidate set than other methods.

Method	Top10	Top20	Top30
MI	0.514	0.684	0.760
Chi-square	0.638	0.765	0.828
Dice	0.572	0.735	0.803
IBM 1	0.822	0.900	0.948
NCF	0.899	0.962	0.973

Table 6. The coverage rates of the candidate words extracted by the compared methods

Figure 1 shows the curve diagram of the coverage rate of each method. As the figure shows, when the size of the candidate list is increased, the coverage rate of using NCF rises rapidly as *n* increases but levels off after *n*=20. Whereas, the coverage rates of other measures grow much slowly.

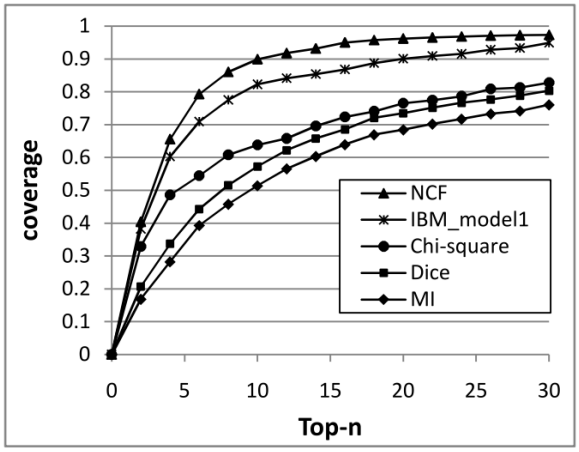

Figure 1. The curve diagram of the coverage of the candidate word list compiled by each method.

From the evaluation of candidate word selection, we find that the *ncf* method, which incorporates lexical-level information into association-based measure, can effectively filter out noisy words and generates a highly reliable list of candidate words for a given MWE.

3.2 Evaluating Extracted Translations

To evaluate the quality of MWE translations extracted automatically, we use the following three criteria:

1) **Translation accuracy:**
   This criterion is used to evaluate the top-n translations of the system. It treats each translation produced as a string and compares the whole string with the given reference translations. If any one of the top-n hypothesis translations is included in the reference translations, it is deemed correct.
2) **WER** (word error rate):
   This criterion compares the top-1 hypothesis translation with the reference translations by computing the edit distance (i.e., the minimum number of substitutions, insertions, and deletions) between the hypothesis translation and the given reference translations.
3) **PER** (position-independent word error rate):
   This criterion ignores the word order and computes the edit distance between the top-1 hypothesis translation and the given reference translations.

We also use the MT task to evaluated our method with other systems. For that, we use the GIZA++ toolkit (Och et al., 2000) to align the Hong Kong Hansard and Hong Kong News parallel corpora. Then, we extract the translations of the given source sequences from the aligned corpus as the baseline. We use the following two methods to extract translations from the aligned results.

1) **Uni-directional alignment**
   We mark all English words that were linked to any constituent of **w** in the parallel Chinese-English aligned corpora. Then, we extract the marked sequences from the corpora and compute the frequency of each sequence. The top-n high frequency sequences are returned as the possible translations of **w**.
2) **Bi-directional alignments**
   We use the *grow-diag-final* heuristic (Och et al., 1999) to combine the Chinese-English and English-Chinese alignments, and then extract the top-n high frequency sequences as described in method 1.

To determine the effect of the *common subsequence filtering method*, FCS, we divide the evaluation of our system into two phases:

1) **NCF+Dice**:
   This system uses the normalized correlation frequency, NCF, to select candidate words as described in Section 2.1. It then extracts candidate translations (described in Section 2.2), but FCS is not used.
2) **NCF+FCS+Dice**:
   This is similar to system 1, but it uses FCS to filter out common subsequences (described in subsection 2.2.2).

Method	WER(%)	PER(%)
Uni-directional	4.84	4.02
Bi-directional	5.84	5.12
NCF+Dice	3.55	3.24
NCF+FCS+Dice	**2.45**	**2.23**

Table 7. Translation error rates of the systems.

Method	Top1	Top2	Top3
Uni-directional	67.5	79.6	83.0
Bi-directional	65.5	77.7	81.1
NCF+Dice	72.8	85.9	88.3
NCF+FCS+Dice	**78.2**	**89.3**	**91.7**

Table 8. Translation accuracy rates of the systems. (%)

Table 7 shows the word error rates for the above systems. As shown in the first and second rows, the translations extracted from *uni-directional* alignments are better than those extracted from *bi-directional* alignments. This means that the *grow-diag-final* heuristic reduces the accuracy rate when extracting MWE translations.

The results in the third row show that the NCF+Dice system outperforms the methods based on GIZA++. In other words, the NCF method can effectively resolve the difficulties of extracting MWE translations discussed in Section 1.

In addition, the fourth row shows that the NCF+FCS+Dice system also outperforms the NCF+Dice system. Thus, the FCS method can resolve the common subsequence problem effectively.

Table 8 shows the translation accuracy rates of each system. The NCF+FCS+Dice system achieves the best translation accuracy. Moreover, it significantly improves the performance of finding MWE translation equivalences.

4 Applying MWE Translations to MT

To demonstrate the usefulness of extracted MWE translations to existing statistical machine translation systems, we use the XML markup scheme provided by the Moses decoder, which allows the specification of translations for parts of a sentence. The procedure for this experiment consists of three steps: (1) the extracted MWE translations are added to the test set with the XML markup scheme, (2) after which the data is input to the Moses decoder to complete the translation task, (3) the results are evaluated using the BLEU metric (Papineni et al., 2002).

	Moses	MWE +Moses
NIST06-sub	23.12	23.49
NIST06	21.57	21.79

Table 9. BLEU scores of the translation results.

4.1 Experimental Settings

To train a translation model for Moses, we use the Hong Kong Hansard and the Hong Kong News parallel corpora as training data (2,222,570 sentence pairs). We also use the same parallel corpora to extract translations of MWEs. The NIST 2008 evaluation data (1,357 sentences, 4 references) is used as development set and NIST 2006 evaluation data (1,664 sentences, 4 references) is used as test set.

4.2 Selection of MWEs

Due to the limitation of the XML markup scheme, we only consider two types of MWEs: continuous bigrams and idioms. Since the goal of this experiment is not focus on extraction of MWEs, simple methods are applied to extract MWEs from the training data: (1) we collect all continuous bigrams from Chinese sentences in the training data and then simply filter out the bigrams by mutual information (MI) with a threshold[1], (2) we also extract all idioms from Chinese sentences of the training data by collecting all 4-syllables words from the training data and filtering out obvious non-idioms, such as determinative-measure words and temporal words by their part-of-speeches, because most Chinese idioms are 4-syllables words.

In total, 33,767 Chinese bigram types and 20,997 Chinese idiom types were extracted from training data; and the top-5 translations of each MWE were extracted by the method described in Section 2. Meanwhile 1,171 Chinese MWEs were added to the translations in the test set. The Chinese words covered by the MWEs in test data set were 2,081 (5.3%).

4.3 Extra Information

When adding the translations to the test data, two extra types of information are required by the Moses decoder. The first type comprises the function words between the translation and its context. For example, if *經貿 合作/economic cooperation* is added to the test data, possible

[1] We set the threshold at 5.

source sentence	... 進入 <MWE>五光十色</MWE> 的社會 ...
Moses	... entered **blinded by the colourful** community ...
MWE+Moses	... entered **the colourful** community ...
reference	... entered **the colorful** society ...
source sentence	... 不希望看到 <MWE>進一步 升級</MWE> 危機 ...
Moses	... do not want to see **an escalation** of crisis ...
MWE+Moses	... do not want to see **a further escalation** of crisis ...
reference	... don 't want to see **the further escalation** of the crisis ...
source sentence	... 廣大 人民 的 <MWE>根本 利益</MWE> ...
Moses	... the people 's **interests** ...
MWE+Moses	... the people of the **fundamental interests** ...
reference	... the **fundamental interests** of the masses ...

Table 10. Examples of improved translation quality with the MWE translation equivalences.

function words, such as '*in*' or '*with*', should be provided for the translation. Because the Moses decoder does not generate function words that are context dependent, it treats a function word as a part of the translation. Therefore, we collect possible function words for each translation from the corpora when the conditional probability is larger than a threshold[2].

The second type of information is the *phrase translation probability* and *lexical weighting*. Computing the phrase translation probability is trivial in the training corpora, but lexical weighting (Koehn et al., 2003) needs lexical-level alignment. For convenience, we assume that each word in an MWE links to each word in the translations. Under this assumption, the lexical weighting is simplified as follows:

$$p_w(\mathbf{f} \mid \mathbf{e}, a) = \prod_{i=1}^{n} \frac{1}{|\{j \mid (i,j) \in a\}|} \sum_{\forall (i,j) \in a} p(f_i \mid e_j)$$
$$\cong \prod_{i=1}^{n} \frac{1}{|\mathbf{e}|} \sum_{\forall e_j \in \mathbf{e}} p(f_i \mid e_j).$$

Then, it is trivial to compute the simplified lexical weighting of each MWE correspondence when the word translation probability table is provided. Here, we use the IBM model 1 to learn the table from the training data.

4.4 Evaluation Results

We trained a model using Moses toolkit (Koehn et al., 2007) on the training data as our baseline system.

Table 9 shows the influence of adding the MWE translations to the test data. In the first row (NIST06-sub), we only consider sentences containing MWE translations for BLEU score evaluation (726 sentences). In the second row, we took the whole NIST 2006 evaluation set into consideration (1,664 sentences). The Chinese words covered by the MWEs in NIST06-sub and NIST06 were 9.9% and 5.3% respectively.

Adding MWE translations to the test data statistically significantly lead to better results than those of the baseline. Significance was tested using a paired bootstrap (Koehn, 2004) with 1000 samples ($p<0.02$). Although the improvement in BLEU score seems small, it is actually reasonably good given that the MWEs account for only 5% of the NIST06 test set. Examples of improved translations are shown in Table 10. There is still room for improvement of the proposed MWE extraction method in order to provide more MWE translation pairs or design a feasible way to incorporate discontinuous bilingual MWEs to the decoder.

5 Conclusions and Future Work

We have proposed a high precision algorithm for extracting translations of multiword expressions from parallel corpora. The algorithm can be used to translate any language pair and any type of word sequence, including rigid sequences and discontinuous sequences. Our evaluation results show that the algorithm can cope with the difficulties caused by *indirect association* and the *common subsequence effects*, leading to significant improvement over the word alignment-based extraction methods used by the state of the art systems and other association-based extraction methods. We also demonstrate that extracted translations significantly improve the

[2] We set the threshold at 0.1.

performance of the Moses machine translation system.

In future work, it would be interesting to develop a machine translation model that can be integrated with the translation acquisition algorithm in a more effective way. Using the normalized-frequency score to help phrase alignment tasks, as the *grow-diag-final* heuristic, would also be interesting direction to explore.

Acknowledgement

This research was supported in part by the National Science Council of Taiwan under the NSC Grants: NSC 96-2221-E-001-023-MY3.

References

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311.

Gao, Jianfeng, Jian-Yun Nie, Hongzhao He, Weijun Chen, Ming Zhou. 2002. Resolving Query Translation Ambiguity using a Decaying Co-occurrence Model and Syntactic Dependence Relations. In *Proc. of SIGIR'02*. pp. 183 -190.

Kitamura, Mihoko and Yuji Matsumoto. 1996. Automatic Extraction of Word Sequence Correspondences in Parallel Corpora. In *Proc. of the 4th Annual Workshop on Very Large Corpora*. pp. 79-87.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of HLT/NAACL'03*. pp. 127-133.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP'04*. pp. 388-395.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL'07, demonstration session*.

Kupiec, Julian. 1993. An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In *Proc. of ACL'93* . pp. 17-22.

Liang, Percy, Ben Taskar, Dan Klein. 2006. Alignment by Agreement. In *Proc. of HLT/NAACL'06*. pp. 104-111.

Ma, Yanjun, Nicolas Stroppa, Andy Way. 2007. Bootstrapping Word Alignment via Word Packing. In *Proc. of ACL'07*. pp. 304-311.

Ma, Yanjun, Sylwia Ozdowska, Yanli Sun, and Andy Way. 2008. Improving Word Alignment Using Syntactic Dependencies. In *Proc. of ACL/HLT'08 Second Workshop on Syntax and Structure in Statistical Translation.* pp. 69-77.

Melamed, Ilya Dan. 2001. Empirical Methods for Exploiting parallel Texts. *MIT press*.

Och, Franz Josef and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proc. of ACL'00*. pp. 440-447.

Och, Franz Josef, Christoph Tillmann, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In *Proc. of EMNLP/VLC'99*. pp. 20-28.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL'02*. pp. 311-318.

Smadja, Frank, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1):1-38.

Wu, Dekai. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3):377-403.

Wu, Hua, Ming Zhou. 2003. Synonymous Collocation Extraction Using Translation Information. In *Proc. of ACL'03*. pp. 120-127.

Yamamoto, Kaoru, Yuji Matsumoto. 2000. Acquisition of Phrase-level Bilingual Correspondence using Dependency Structure. In *Proc. of COLING'00*. pp. 933-939.