The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

Identify Bilingual Patterns and Phrases from a Bilingual Sentence Pair

Yi-Jyun Chen
Department of Computer Science
National Tsing Hua University
yijyun@nlplab.cc

Hsin-Yun Chung
Department of Mathematics
National Chung Cheng University
maggie100573@gmail.com

Jason S. Chang
Department of Computer Science
National Tsing Hua University
jason@nlplab.cc

Abstract

This paper presents a method for automatically identifying bilingual grammar patterns and extracting bilingual phrase instances from a given English-Chinese sentence pair. In our approach, the English-Chinese sentence pair is parsed to identify English grammar patterns and Chinese counterparts. The method involves generating translations of each English grammar pattern and calculating translation probability of words from a word-aligned parallel corpora. The results allow us to extract the most probable English-Chinese phrase pairs in the sentence pair. We present a prototype system that applies the method to extract grammar patterns and phrases in parallel sentences. An evaluation on randomly selected examples from a dictionary shows that our approach has reasonably good performance. We use human judge to assess the bilingual phrases generated by our approach. The results have potential to assist language learning and machine translation research.

Keywords: Pattern Grammar, Phrase Translation, Word Alignment

1 Introduction

Verb phrases are prominent components of any sentence. If we can correctly extract English-Chinese bilingual grammar patterns and phrases in a bilingual sentence pair, it will be helpful for English and Chinese language learners and machine translation systems. These results can be used to demonstrate the synchronous structure of parallel sentences. However, Chinese sentence parsing technology is still immature, which leads to difficulties in obtaining grammar patterns and phrases in Chinese sentences.

There are still many problems with existing Chinese parsers. To extract Chinese patterns and phrases more accurately, we utilize an English sentence parser which is much more mature than Chinese parsing technology to parse English sentences and extract English patterns and phrases. We also utilize some statistical methods to estimate translation probability of word pairs using bilingual corpora. Then, we extract Chinese patterns and phrases more accurately by finding counterparts of English patterns and phrases with statistical results from bilingual corpora.

We present a system that returns bilingual verb patterns and phrases of a bilingual sentence pair. Our system identifies phrases in sentences by using the pattern table and translation probability model created in advance. The pattern table created by parsing English sentences and calculating counterparts in bilingual parallel corpus with word alignment. The translation probability model is created by the alignment probability of each English and Chinese word pair with consideration of not only the word itself but also related words.

At the runtime, our system starts with an English-Chinese bilingual sentence pair submitted by the user. The system parses the English sentence and extracts English patterns and phrases, and retrieves the counterpart of the pattern from the table created in advance to be Chinese pattern. Then, the system extracts the counterpart of words in the English phrase by the probability model created in advance. Finally, The system combines the Chinese pattern and the counterpart of words in the English phrase to form a Chinese phrase.

The system can assist language learning or be used to generate training data for machine

translation research, especially research related to phrases. The rest of the article is organized as follows. We review the related work in the next section. Then we present our method for automatically identifying bilingual phrases from a bilingual sentence pair in section 3. The experiment and an evaluation based on human judgment are described in Section 4. Finally, we summarize our conclusions in Section 5.

2 Related Work

Machine Translation is a time-honored and yet active research area. Shifting from the rule-based approach toward data-intensive approach after the seminal paper by Brown et al.,1990, an increasing number of bilingual corpora have made statistical machine translation more and more feasible. In our work we address an aspect of machine translation is not a direct focus of Brown et al. (1990). We also consider on more general linguistic class of units in translation where the translation may not be litateral and may need a presentation that reflects the similarity and differences between two languages involved, and the user might be interested in multiple ways (structures) of translating a phrase (e.g., consider the phrase "harmful to the ocean" and its translation).

More recent researches concentrate on learning word translation and extracting bilingual word translation pairs from bilingual corpus, and then calculate the degree of mutual relationship between word pairs in parallel sentences, thereby deriving the precise translation (Catizone et al. (1989); Brown et al. (1990); Gale and Church (1991); Wu and Xia (1994); Fung (1995); Melamed (1995); Moore (2001)).

In our system, we focus on identifying patterns and phrases by a patterns table and a word translation model which are created using statistical methods in bilingual corpora with word alignment.

In the area of phrase alignment, Ko (2006) proposed a method for verb phrase translation. For specific verb fragments (e.g. make a report to police), automatic alignment is applied to calculate the collocation relationship across two language (e.g. when make and report appear together, report often corresponds to "報案"), then word and phrase correspondences are generated (e.g. make a report to police correspond to "向 警察 報案") ,to tally translations and counts. Chen et al. (2020) focus on the translation of noun prepositional collocations. Using statistical methods to extract translations of nouns and prepositions from bilingual parallel corpora with sentence alignment, and then adjust the translations with additional information of Chinese collocations extracted from a Chinese corpus.

3 Methodology

We attempt to identify bilingual grammar patterns and phrases in an English-Chinese sentence pair using a lexical translation model and bilingual grammar patterns. Our identification process is shown in Figure 1.

(1) Create Bilingual Grammar Patterns and Phrases Table (section 3.1)
(2) Create Word Translation Probability Model (section 3.2)
(3) System Runtime - Extracting Pattern Grammar and Phrases in Sentence Pair (section 3.3)

Figure 1. Identification process

3.1 Creating Bilingual Grammar Patterns and Phrases Table

In the first stage of the identification process (Step (1) in Figure 1), we extract Chinese grammar patterns for each English grammar pattern. For example, the verb "use" has a grammar pattern "use n to inf", our goal is to extract Chinese counterparts such as "使用 n 來 v" and "使用 n 去 v" for "use n to inf".

The input to this stage is English-Chinese sentence pairs in a word-aligned bilingual parallel corpus. We parse each English sentence into a tree structure to reveal the dependency of words in the sentence and use a recursive approach to extract the grammar pattern and the phrase for each verb in the sentence. Then, we extract Chinese counterparts of the English grammar pattern in the Chinese sentence according to the word alignment, and convert them into Chinese grammar patterns according to the English word in English pattern each Chinese word corresponds to.

For example, for the sentence pair "We use computers to solve the problem." and "我們 使用 電腦 來 解決 問題", we extract the English grammar pattern "use n to inf" and

English phrase “use computer to solve problem” for the verb “use”. Then, we extract the Chinese counterparts “使用 電腦 來 解決 問題” and convert it into Chinese pattern “使用 n 來 v” with converting “電腦” into “n” and converting “解決 問題” into “v” according to their correspondence to English phrase and pattern.

For each English grammar pattern, we compute the frequency of Chinese patterns, and filter patterns with high frequency. Then, we sort the patterns by counts and by pattern length as shown in table 1.

English Grammar Pattern	Chinese Grammar Pattern	Count	Rank
use n to inf	使用 n 來 v	180	1
use n to inf	用 n 來 v	157	2
use n to inf	利用 n 來 v	70	3
use n to inf	用 n 去 v	41	4
use n to inf	使用 n v	287	5
use n to inf	用 n v	186	6

Table 1. Chinese grammar pattern for English pattern “use n to inf”, ranked based on frequency count and pattern lengths. Note that the English are based on a pre-determined templates and the Chinese patterns are automatically derived through word alignment.

The output of this stage is a table which contains sorted Chinese patterns for each English grammar pattern.

3.2 Creating Word Translation Probability

In the second stage of the identification process (Step (2) in Figure 1) , we calculate lexical translation probability for each English word and Chinese word pair.

The input to this stage is English-Chinese sentence pairs in a word-aligned bilingual parallel corpus. We calculate the counts and probability of each Chinese word aligned to each English word in the bilingual corpus. For each English-Chinese word pair in the bilingual corpus, we compute weighted translation probability of the Chinese word to the English word. We also take into consideration the tense, synonym and derivative words of English words. We set weight for these English words according to their degree of relevance to the English word in the pair. Then, we multiply the probability by their weight and sum up these weighted probabilities to generate an adjusted probability for the word pair to represent how likely they are to translate to each other.

For example, we calculate the probability of word pair (討論, discussion) with consideration of the words related to “discussion” such as “discuss” and “talk” and give them weight. Some words we consider for the pair (討論,discussion) are shown in Table 2. After multiplying the probability by their weight and summing up these weighted probabilities, we finally get a adjusted probability 0.62 to represent the probability of “討論” as a translation of “discussion”.

word	probability	Weight
discussing	0.29	1
discussed	0.24	1
talk	0.03	0.5

Table 2. Word forms related the pair (討論,discussion) and weights according to morphology and synonyms

Note that because there are many errors in automatic word alignment, we only consider words with original alignment probability more than 0.01. This approach makes the adjusted probabilities of most word pairs with unrelated meanings will be zero. For example, the word pair (“他”,”you”) has zero probability as shown in Table 4.

Beside creating the translation probability model, for each English word, we also filter some Chinese words with high translation probability to it to be its translations. Translations of some English words are shown in table 3.

Word	Translations
use	應用、利用、採用、運用、使用、用、用途、動用
discuss	討論、商討、探討、論述、談
mate	伴侶、交配、隊友、搭檔、配偶、夥伴、大副

Table 3. Translations of some English words

The output of this stage is a model which gives adjusted estimation of lexical translation probability of English and Chinese words in the bilingual corpus, and translations of each English word. A sample of the translation probability model is shown in table 4.

Chinese Word	English Word	Adjsuted Probabilty
玩	play	0.43

吸引	attract	0.51
產品	product	0.79
推銷	product	0.02
他	you	0

Table 4. A sample of the translation probability model

3.3 System Runtime - Extracting Pattern Grammar and Phrases in Sentence Pair

Once the bilingual grammar pattern table and weighted word translation probability model are created, our system then evaluates a given sentence pair using the procedure in Figure 2.

(1) Extract English Pattern Grammar and Phrase in English Sentence
(2) Find Suitable Counterparts for Words in English Phrase
(3) Select Chinese Grammar Pattern and Compose Chinese Phrase

Figure 2. Runtime evaluation procedure

The input to the system is an English-Chinese sentence pair such as "Peacocks use their beautiful tails to attract mates" versus "蝴蝶 用 美麗 的 尾巴 來 吸引 配偶".

In the first step (Step (1) in Figure 2), we parse the English sentence and extract the grammar pattern and phrase as described in section 3.1. For example, the pattern grammar "use n to inf" and the phrase "use tail to attract mate" will be extracted from "Peacocks use their beautiful tails to attract mates".

In the second step (Step (2) in Figure 2), for each word in the English phrase except those in the grammar pattern, we find its suitable counterpart in the Chinese sentence using the weighted word translation probability model described in Section 3.2. For each English word w in English phrase, if there are some words in Chinese sentence that have non-zero translation probability to w, we choose the one with the highest probability to be the counterpart of w.

For example, in the sentence pair "Peacocks use their beautiful tails to attract mates" versus "蝴蝶 用 美麗 的 尾巴 來 吸引 配偶", for word "tail" in phrase "use tail to attract mates", we consider the weighted translation probabilities of word pairs ("蝴蝶", "tail"), ("用", "tail"), ("美麗", "tail"), ..., ("配偶", "tail"). Because the word pair ("尾巴," tail") has highest probability, we select "尾巴" to be the counterpart of "tail".

If there are not any Chinese words that have non-zero translation probability to w, we consider their similarity to the translations of w selected in advance by using word embedding. For each word c in the Chinese sentence, we multiply its similarity to each translation c_pre of w by the weighted translation probability of c_pre to w and sum up to be the new probability of c. Then, we choose the one with the highest new probability to be the counterpart of w.

In the final step (Step (3) in Figure 2), we consider the sorted Chinese grammar patterns of the English grammar pattern extracted in Step (1) according to the table created in advance (described in Section 3.1) and the counterparts of words in the English phrase selected in Step (2). We check each Chinese pattern in order whether it is contained in the Chinese sentence and whether the position of the counterpart of each word in English phrase is reasonable. If so, we select the grammar pattern to be the counterpart of the English grammar pattern and combine it with the counterpart of each word in English phrase to form a Chinese phrase.

For example, the Chinese pattern "用 n 來 v" is contained in sentence "蝴蝶 用 美麗 的 尾巴 來 吸引 配偶" and the most suitable counterpart of "tail", "attract" and "mate" in the phrase "use tail to attract mate" are "尾巴", "吸引" and "配偶" and the position of counterparts is reasonable. We combine pattern "用 n 來 v" and words "尾巴", "吸引" and "配偶" to form the phrase "用 尾巴 來 吸引 配偶".

For a verb in the pattern such as the "v" in the "用 n 來 v", we also consider its own bilingual pattern to find its counterparts instead of only by translation probabilities of its words. For example, in the sentence pair "she want to send her son to the school" versus "她 想 把 兒子 送到 這所 學校", there is English pattern "want to inf" with Chinese pattern "想 v" for verb "want", and English pattern "send n1 to n2" with Chinese pattern "把 n1 送到 n2" for verb "send". By replacing the "inf" in "want to inf" by "send n1 to n2" and replacing the "v" in "想 v" by "把 n1 送到 n2", the bilingual pattern pair "want to send n1 to n2" versus "想 把 n1 送到 n2" is generated

and then the bilingual phrase pair "want to send son to school" versus "想 把 兒子 送到 學校" can be extracted from the sentence pair.

The output of the system is bilingual grammar patterns and phrases extracted from the bilingual sentence pair. For example, the grammar patterns "use n to inf" versus "用 n 來 v" and the phrases "use tail to attract mates" versus "用 尾巴 來 吸引 配偶" are the output of the input sentence pair "Peacocks use their beautiful tails to attract mates" versus "蝴蝶 用 美麗 的 尾巴 來 吸引 配偶".

4 Evaluation and Discussion

The purpose of our system is to allow users to retrieve the bilingual patterns and phrases from a bilingual sentences pair. Therefore, in this section, we report the results of preliminary evaluations on the extraction of bilingual patterns and phrases. The evaluation process was conducted on a set of bilingual sentence pairs along with their patterns and phrases extracted.

4.1 Experimental setting

The bilingual parallel corpora we used are the Minutes of Legislative Council of the Hong Kong Special Administrative from the legislative council of Hong Kong with 1,640,007 bilingual sentence pairs, and the UM-corpus (Liang, 2014) from university of Macau with 1,827,014 bilingual sentence pairs. We used CKIP (Ma and Chen, 2003) which is a Chinese knowledge and information processing system developed by academic sinica to process Chinese word segmentation and used fast-align (Dyer et al., 2013) to process word alignment of bilingual parallel sentences.

We used Spacy (Honnibal and Montani, 2017) to parse English sentences and extract patterns with their counterparts to create a bilingual patterns table as we described in section 3.1. Then, we calculate and create a probability model for bilingual word pairs as we describe in section 3.2. Finally, we get the result by using our system to evaluate given sentence pairs as we describe in section 3.3.

4.2 Evaluation Metrics

The output of our method are bilingual grammar patterns and phrases of all verbs in sentence pairs. To evaluate our approach, we randomly selected 50 valid sentences (66 verb patterns) from examples in Cambridge dictionary and Macmillan dictionary. The grammar patterns and phrases in each sentence are evaluated by a linguist. Note that the verb "be" and the verb "have" are excluded from our evaluation since they often don't have a direct translation in the Chinese sentence. In some English sentences, there are not any verb grammar patterns with length greater than 1. We treat such sentence pairs as invalid and they would not be included in the 50 sentence pairs. There are totally 66 verb patterns and phrases with length greater than 1 in the 50 English sentences. We evaluated the correctness of Chinese patterns and phrases of these 66 verb patterns. Some patterns and phrases successfully identified are shown in table 5. The overall accuracy is 79% and the overall recall is 29%.

4.3 Discussion

The results of evaluation has high precision rate and low recall rate. It shows that most of the Chinese patterns and phrases identified by our method are correct, but there are many phrases that have not been successfully identified. Because of the limit of the amount of data of parallel corpora, many correct and common patterns are not successfully extracted to put into the pattern table, and then cannot be identified from the sentence pairs submitted at the runtime. It may be the main reason that causes the low recall.

English Pattern	English Phrase	Chinese Pattern	Chinese Phrase
fall on n	fell on floor	掉 在 n	掉 在 地 上
word for n	work for company	為 n 工作	為 公司 工作
agree to inf	agree to form league	同意 v	同意 結成 聯盟
provide n	provide evidence	提供 n	提供 證據
commit n	committed crime	犯 n	犯 罪

Table 5. Some patterns and phrases successfully identified

Counterparts of words in an English phrase are extracted by a probability model which is a weighted probability model adjusted from word alignment probability in the parallel corpora. That means, only words which appear and have

sufficient frequency in the parallel corpora are in the model. Although we also design methods by word embedding to process words which are not in the model, they still brought a relatively high error rate.

5 Conclusion and Future Work

Many avenues exist for future research and improvement of our system. As we mentioned in section 4.3, lack of bilingual patterns in our table created in advance causes that many phrases cannot be identified. One such avenue is to design methods to expand the amount of the patterns. For example, we can consider collocations calculated in the Chinese monolingual corpus which contains a larger amount of sentences, or consider the synonyms of the words in the pattern by using word embedding or dictionaries like WordNet, to generate more bilingual patterns.

In summary, we have introduced a method for identifying patterns and phrases that allow users to submit an English-Chinese sentence pair and get bilingual patterns and phrases in the sentence pair. The method involves parsing English sentences and extracting counterparts of English patterns and phrases by using a bilingual pattern table and word translation probability model created in advance. The result of the evaluations show that our method is highly accurate.

References

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Roberta Catizone, Graham Russell, and Susan Warwick. 1989. Deriving translation data from bilingual texts. In *Proceedings of the First International Lexical Acquisition Workshop*, pages 1–7.

Yi-Jyun Chen, Ching-Yu Helen Yang, and Jason S. Chang. 2020. Improve word alignment for extraction phrasal translations. *International Journal of Computational Linguistics Chinese Language Processing*, 25(2):37–54

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pages 644–648

Pascale Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *arXiv preprint cmp-lg/9505016*.

William A Gale and Kenneth Church. 1991. Identifying word correspondences in parallel texts. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

M. H. Ko. 2006. Alignment of Multi-word Expressions in Parallel Corpora. Master's thesis, National Tsing Hua University, Taiwan.

Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to ckip chinese word segmentation system for the first international chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 168–171.

I Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. *arXiv preprint cmp-lg/9505044*.

Robert C Moore. 2001. Towards a simple and accurate statistical approach to learning translation relationships among words. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Shuo Li, Yiming Wang, Yi Lu, "UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation". In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC'14), Reykjavik, Iceland, 2014.

Dekai Wu and Xuanyin Xia. 1994. Learning an english-chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*