Alignment and Extraction of Bilingual Legal Terminology from Context Profiles

Oi Yee Kwong, Benjamin K. Tsou, Tom B.Y. Lai, Robert W.P. Luk†, Lawrence Y.L. Cheung and **Francis C.Y. Chik**

Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
{rlolivia,rlbtsou,cttomlai,rlylc,rlfchik}@cityu.edu.hk

†Department of Computing
Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
csrluk@comp.polyu.edu.hk

Abstract

In this study, we propose a knowledge-independent method for aligning terms and thus extracting translations from a small, domain-specific corpus consisting of parallel English and Chinese court judgments from Hong Kong. With a sentence-aligned corpus, translation equivalences are suggested by analysing the frequency profiles of parallel concordances. The method overcomes the limitations of conventional statistical methods which require large corpora to be effective, and lexical approaches which depend on existing bilingual dictionaries. Pilot testing on a parallel corpus of about 113K Chinese words and 120K English words gives an encouraging 85% precision and 45% recall. Future work includes fine-tuning the algorithm upon the analysis of the errors, and acquiring a translation lexicon for legal terminology by filtering out general terms.

1 Introduction

Machine translation, parallel text alignment, and translation lexicons are the vertices in their tightly bound triangular relation. The mutual relation between translation lexicons and parallel text alignment is especially close as they together provide foundational resources for the research of machine translation.

Conventional statistical methods for bilingual word alignment and extraction of translation lexicon require large corpora to be effective. Lexical approaches, on the other hand, depend on existing bilingual dictionaries which often only cover general terms. In any case, both methods will probably fall short with a small and domain-specific corpus, as the one in our study.

The parallel corpus in this study consists of bilingual Hong Kong court judgments, in English and Chinese. It is a small and domain-specific corpus. Thus on the one hand, we can imagine that existing bilingual dictionaries would be of little help in the alignment process, as general lexicons do not often cover legal terminologies and their translations. On the other hand, the effectiveness of statistical or probabilistic approaches might not be realised, given the limited corpus size. Hence, neither approach fits the data ideally. Moreover, English and Chinese are from different language families. Also, legal terms are not limited to single words.

Hence, we propose a method for aligning terms in this small corpus of legal texts and extracting a translation lexicon therefrom, based on the consistency observed in legal texts. The method only requires a sentence-aligned corpus and suggests translation equivalences from the frequency profiles of parallel concordances. Pilot testing shows that the method is effective for fulfilling the two purposes, i.e. term alignment and lexicon extraction, simultaneously.

In the following, we will first review related studies in Section 2. Then we will describe the properties of our corpus in Section 3 and our method in Section 4. A pilot experiment with the proposed approach will be reported in Section 5 with results discussed in Section 6, followed by a conclusion in Section 7.

2 Related Work

Conventional bilingual sentence alignment is often based statistically on sentence length (e.g. Gale and Church, 1991), or lexically on cognates (e.g. Simard *et al.*, 1992) and correspondence of word position (e.g. Kay and Roescheisen, 1993; Piperidis *et al.*, 1997). Such criteria, however, are mostly applicable to Indo-European language pairs. Although Wu (1994) found the length criterion applied surprisingly well be-

tween English and Chinese, he supplemented the statistical, length-based method with lexical criteria. Fixed words or phrases with consistent translations, like month names, were identified first, and he observed an improvement of the results.

The derivation of bilingual dictionaries often follows text alignment (possibly at the word level) based on some frequency criterion. Nevertheless, in practice sentence alignment is not always distinctly separated from word alignment, and neither is word alignment and the derivation of translation lexicons thereby. In fact, apart from Gale and Church's length-based method, most other methods worked at the lexical level to some extent.

Word alignment can be done statistically by learning the translation association or token co-occurrences between the source language and the target language (e.g. Wu and Xia, 1995; Melamed, 1997). The acquisition of bilingual translation equivalences sometimes follows the acquisition of monolingual collocations (e.g. Wu and Xia, 1995; Smadja *et al.*, 1996). Wu and Xia (1995), for instance, made use of terms extracted by CXtract, a Chinese term extraction algorithm, to learn collocation translations for English words from the bilingual Hong Kong LegCo proceedings, reporting a precision of about 90%.

Others make use of existing bilingual dictionaries for word alignment, which is useful when the corpus is too small for statistical methods and contains many general words. For example, Ker and Chang (1997) worked with a small bilingual corpus (English-Chinese) and attempted to overcome the limitation of statistical methods by class-based rules, with reference to both an English and a Chinese machine readable dictionary.

When word alignment relies on existing lexical resources, however, the coverage tends toward the low end, probably due to the inherent limited coverage of existing resources. Huang and Choi (2000) used several linguistic resources, including bilingual and monolingual resources, for word alignment between Chinese and Korean texts. Even when they combined three algorithms together, there was still little improvement in coverage.

Using a third, pivot language as a bridge in word alignment may also improve the performance (e.g. Borin, 2000; Mann and Yarowsky, 2001). However, unlike Slavic languages or Indo-European languages, it seems difficult to imagine an effective bridge between English and Chinese.

Hybrid methods are also used (e.g. Piperidis *et al.*, 1997; Huang and Choi, 2000), and are believed to produce better alignment results.

Fung (1998), in contrast to the above which worked with parallel corpora, tried to extract bilingual lexicons from non-parallel corpora, which is more difficult. She discussed an algorithm called Convec, which compares the context vector of a given English word with the context vectors of all Chinese words for the most similar candidate. During the process, a bilingual dictionary is used to map the context words in the two languages. The method was about 30% accurate if the top-one candidate was considered, although the accuracy was more than doubled if the top-20 candidates were taken.

While most translation lexicon extraction methods do not particularly address domain-specificity, Resnik and Melamed (1997) suggested that a domain-specific translation lexicon could be obtained by filtering out general terms from the results. They, for instance, compared their extracted lexicon entries against a machine readable dictionary and discarded the terms in common.

In the next section, we will discuss our problem in this study and explain why existing methods are insufficient to solve it.

3 Characteristics of the Corpus

As mentioned earlier, we are working with a parallel corpus of bilingual Hong Kong court judgments. The following properties of the corpus render many existing word alignment and translation lexicon extraction methods limited in one way or another.

- **Small corpus size**

  The amount of bilingual Hong Kong court judgments is potentially growing, as long as there are legal proceedings. However, the part that is ready for use in this study only contains about 200K Chinese characters (about 113K word tokens upon segmentation) and about 120K English word

tokens. This size is considered small, in view of the many large corpora in general domains available for natural language processing research, and may therefore not be ideal for many statistical methods.

- **Domain-specific corpus**

  The corpus consists of legal texts, and is thus very domain-specific. It would be a good resource from which to derive a legal translation lexicon. However, a lexical approach to alignment based on existing general bilingual dictionaries might be limited, because such dictionaries do not always cover legal terminologies and their translations. Even some general terms may be translated in special ways in legal texts.

- **Different language families**

  English and Chinese are from different language families and have little resemblance of each other. As a result, lexical criteria like cognates, or alignment via a bridging language will not be applicable.

- **Unpredictable word complexity**

  In many of the studies reviewed in Section 2, the alignment was confined to single words, at least for one of the languages in question (e.g. Wu and Xia, 1995; Melamed, 1997; Resnik and Melamed, 1997; Fung, 1998; Huang and Choi, 2000). For about 150 years, the legal system in Hong Kong operated through English only. So, many legal concepts may not be as precisely lexicalised in Chinese. The lengths and complexity of legal translations between English and Chinese are therefore not necessarily correlated. Aligning legal terms should therefore take care of the varying lengths and complexity of a source term and its translation equivalence.

3.1 Two Assumptions

On the other hand, we also have some advantages from the bilingual legal texts. Legal translations are well-known for their preciseness and consistency. Thus we make the following assumptions for the current study:

1. Bilingual legal texts form relatively clean parallel corpora, in the sense that the sentence alignments are expected to be neat, with few insertions and deletions.
2. Though not necessarily one sense per discourse, legal terms tend to be translated more consistently than common terms.

Based on these two assumptions, we propose a method, which depends minimally on prior knowledge, for aligning words and expressions in the corpus and thus extracting translation equivalences, as described in the next section.

4 Bilingual Term Alignment and Extraction

4.1 Task Definition

As Huang and Choi (2000) pointed out, the "alignment" problem is often not explicitly defined and apparently everyone understands what is and should be going on. They, however, adopted their own definition. In the current study, we take "word alignment" as the more viable task of "translation spotting", as in the ARCADE project.[1]

4.2 Our Approach

Our method is similar to Piperidis *et al.*'s as we share the observation that the source word and the target word should have similar frequencies if they are bilingual equivalence, except for function words. However, their method compares all source-target associations for every possible pair of words between the source sentence and the target sentence, and located the correct translation from the local maximum. We, on the other hand, adopt a simpler comparison, paying utmost attention to one term at a time.

Our approach starts with a sentence-aligned bilingual corpus. As said, bilingual corpora in the legal domain are relatively clean corpora. Sentences can often be one-to-one aligned. Given that legal terms are not always cross-lingually lexicalised in similar ways, as discussed in Section 3, term length and position in a sentence might not be reliable parameters for word alignment. Instead, we deal with one term at a time, taking all sentences containing the term into consideration; and we refer to the group of sentences containing a given term as "concordance lines" of that term. Thus within its

[1] http://www.up.univ-mrs.fr/~veronis/arcade/index-en.html

concordance lines, a given term often has higher frequency than other co-occurring words, except for function words. Since terms in legal texts are more likely to be consistently translated, translation equivalences in the target concordances should share a similar frequency with the source word. Hence, by analysing the frequency profiles, we can identify the words in the target language which are likely to be expressing the concept of the source word.

Our algorithm is as follows:

1. Extract salient compound terms from the word-segmented Chinese texts and treat them as single words in subsequent steps.
2. Mark up stop words from both the Chinese and English texts, and lemmatise the English words.
3. Scan through the Chinese texts. For each un-aligned word, retrieve all sentences containing the word to give the *source concordances*, and all corresponding, aligned sentences from the English texts to give the *target concordances*.
4. Perform a word frequency count from the concordances and rank the results.
5. Pick the words from target concordances above some frequency threshold. The longest string containing one or more of these words and spanning within a small window size in each target concordance gives a candidate of translation equivalence. (The setting of the threshold and the window size will be discussed in Section 5.2.)
6. Repeat Steps 3 to 5 until all Chinese words are processed.

Thus our method is not restricted to aligning single words. It does not inherently require a large corpus to start with. Also, no prior knowledge source like existing bilingual dictionaries is required.

5 Pilot Experiment

5.1 Test Materials

The corpus we used is a domain-specific one, consisting of parallel texts of Hong Kong court judgments, in English and Chinese.[2] Each judgment has a header containing the basic information of the case (e.g. case number, judges, etc.), the main judgment text, and a footer where judges sign. Only the main text was used in this experiment. For the current study, the Chinese texts contain about 200K characters, and about 113K word tokens and 7K word types upon segmentation. The parallel English texts contain about 120K word tokens, which correspond to about 7K word types. The corpus was aligned to the sentence level, and there were 4750 groups of aligned sentences.

5.2 Method

The algorithm proposed in Section 4.2 was applied to the sentence-aligned corpus. The sentence alignment was manually verified to ensure the idiosyncratic cases were rectified. The Chinese term extraction was done with the algorithm in Kwong and Tsou (2001). The English words were stemmed by applying the Porter (1980) stemmer. Only the Chinese terms occurring more than 5 times in the corpus were tackled. For the frequency threshold (in Step 5), we first look for words with frequency over $0.8 * source\ frequency$ (where *source frequency* is the frequency of the source term within the source concordances). If no words cross this threshold, we pick the word with the highest frequency and over $0.5 * source\ frequency$. For the window size, we took the empirically optimal $n + 1$ from Kwong (2002), where n is the number of English words crossing the frequency threshold.

5.3 Performance Measures

Alignment outcomes were classified into four types: *correct*, *partially correct*, *incorrect*, and *empty*, defined as follows:

- ***Correct***: All relevant content words are within the suggested translation equivalence, allowing incomplete verb forms or missing stop words before or after.
- ***Partially Correct***: Under- or over-aligned, with some content words outside or some irrelevant words inside the translation candidate.

[2]The authors acknowledge the Hong Kong Judiciary for providing the judgment texts.

- ***Incorrect***: Completely mis-aligned words.
- ***Empty***: No candidate translation equivalence is suggested.

5.4 Results

The Chinese term extraction yielded 683 potential compound terms (length ≥ 4 characters and without numerals), and 462 remained after human verification. Thus step 1 of the algorithm (see Section 4.2) re-segmented the Chinese texts with this list of compound terms.

As said earlier, the corpus contains 4750 aligned sentences. We evaluated the alignment of terms in 50 randomly selected sentences. Sentence lengths range from 2 to 45 Chinese words, and the average sentence length is 18.6 words.

Of the total 932 Chinese words in the 50 sentences, 359 were filtered as stop words, and 59 occurred five times or less. Hence, there were 514 words subject to alignment. The outcomes are summarised below.

Correct	=	213 (41%)
Partially Correct	=	19 (4%)
Incorrect	=	41 (8%)
Empty	=	241 (47%)

Considering all cases where a translation equivalence was suggested, our method attained 53% *coverage* and 85% *precision* (percentage of correct and partially correct equivalences among all suggested equivalences). The *recall* (percentage of correct and partially correct equivalences among all test instances) was 45%. Our method gives relatively high coverage and comparable precision, as compared to existing lexically based methods, which is especially encouraging in view of its knowledge independence.

Figure 1 shows an excerpt of the output. Words in the aligned sentences were numbered as seen at the top of Figure 1. Stop words and words with too few occurrences were not processed. The suggested alignment was listed after the whole Chinese sentence was processed. In the figure, for example, c6 was aligned to e4-e6. The resultant alignment of the sentence is drawn in Figure 2. In fact, we can find the various types of alignment outcomes in this example. The links (except the dotted one) indicate correct alignment. The dotted link for c14 is an incorrect match. There were no suggestions for

```
Sentence No. 4660
c1=他  c2=現在  c3=向  c4=法庭  c5=尋求
c6=上訴許可  c7=以便  c8=提出  c9=針對
c10=判罪  c11=及  c12=刑罰  c13=的  c14=上訴
e1=He  e2=now  e3=seeks  e4=leave  e5=to
e6=appeal  e7=against  e8=conviction  e9=and
e10=sentence
c1,他  (stop)
c2,現在,27
      1,now,20
c3,向  (stop)
c4,法庭,213
      1,court,160
c5,尋求,62
c6,上訴許可,17
      1,appeal,22
      2,leav,17
      3,applic,13
c7,以便,38
c8,提出,325
c9,針對,71
      1,against,57
c10,判罪,5  (too few)
c11,及  (stop)
c12,刑罰,49
      1,sentenc,50
c13,的  (stop)
c14,上訴,183
      1,appeal,226
Alignment:
c2:e2/!!c4:!!c5:!!c6:e4-e6/!!c7:!!c8:!!c9:e7/!!c10:!!c12:e10/!!c14:e6/!!
```

Figure 1: Excerpt of Alignment Output

c4, c5, c7, c8, and c10. Nevertheless, there were in fact no literal correspondences in the English sentence for c4, c7, c8, and c14. So some of the empty alignments are correct. Further classifying the empty outcomes and taking the correct ones into account, the accuracy of our method is over 55% as far as the word alignment per se is concerned. The results will be further analysed in the next section.

As for the derivation of a translation lexicon, we thus obtained 213 suggested translation equivalences from the correct word alignments. These equivalences are potential entries of a bilingual legal term translation lexicon, upon further filtering and processing. Figure 3 shows some examples.

6 Discussion

In the last section, we have seen the effectiveness of our method from a pilot experiment. This finding is significant because the method overcomes the limitations of existing statisti-

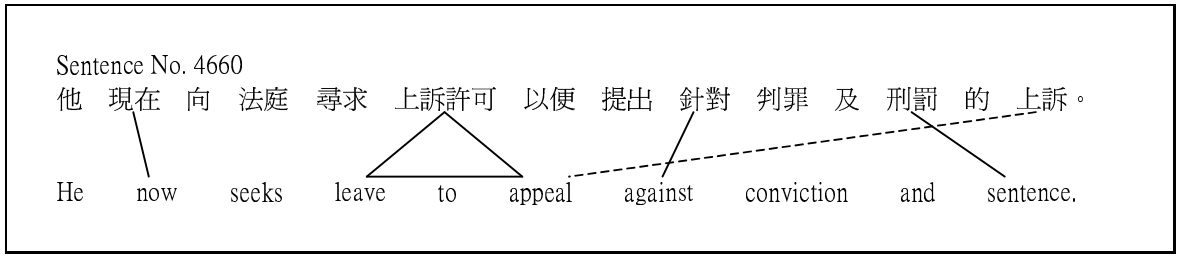

Figure 2: Graphical Representation of Resultant Alignment for Sentence 4660

cally based and lexically based methods. It does not depend on any prior knowledge source, and it works for a small, domain-specific, and parallel corpus of two very different languages. The method does not produce forced errors, and some empty alignments are actually correct. It is also able to align many personal names. In the following, we will further analyse the errors to explore ways for improving the algorithm.

As observed from the pilot test, alignment errors usually occurred for one or more of the following reasons:

- Errors are inherited from the term extraction step. There were some terms which should better be separated and some potential compound terms were not extracted by the algorithm. As a result, the frequency counts would be affected.
- General words may not be frequent enough to be successfully aligned. As court judgments also contain factual description as well as legal reasoning, the factual parts may contain more general words which are not as abundant as legal terms in the whole corpus.
- The method (and its current implementation) cannot distinguish multiple occurrences of the same term in a sentence, and the first occurrence will always be suggested. Word position may need to be taken as another parameter in the future. Moreover, the method now only suggests either one translation candidate or none. It should be fine-tuned to consider multiple candidates which might exist.
- Since the method takes the whole corpus to work out the frequency profiles, it means a given Chinese term would be aligned to its most frequently found translation in the corpus. If it happens that the term is expressed in a different (and less frequent) way in the target language in a particular pair of concordance, the method will result in an empty alignment, and thus missing translation equivalences.
- Anaphors also cause many alignment errors. Since pronouns are treated as stop words, if a term in the source language is pronominalised in the target language, there is no way to properly align that term.

Terms in Chinese	**English Equivalence**
上訴人	Appellant
普通法	Common law
強制執行	Enforcement
信納	Satisfied
書面查詢	Requisition
原訴人的律師	Plaintiff's solicitors
裁斷	Findings
舊式婚姻	Customary marriage

Figure 3: Examples of Acquired Translations

As the current study makes minimal use of syntactic information (e.g. no part-of-speech tagging was done and the compound terms extracted were not governed by any particular syntactic patterns), it might be useful to explore the contribution of some syntactic processing to the alignment performance and for solving some of the above problems.

Future work also includes the refinement of the acquired translation lexicon into one for le-

gal terminology. This would require further processing of the translation equivalences suggested during the alignment process. For instance, entries in a lexicon should make sense even when out of context, so they should be free of anaphors like definite descriptions, and they should be in root form. Also, apart from filtering using a general dictionary as in Resnik and Melamed (1997), we may also filter the results against the terms found in a general-domain corpus.

7 Conclusion

In this study, we have introduced and tested a simple but effective method for word alignment and translation extraction between parallel English and Chinese legal texts, based on frequency profiles of parallel concordances. The method does not require any prior knowledge, and thus overcomes the limitations of existing statistically based and lexically based methods. It is especially designed for working with small, domain-specific, and sentence-aligned parallel corpora. About 85% precision and 45% recall were obtained from the pilot experiment. Future work will be on the fine-tuning of the algorithm and the acquired translation lexicon.

References

L. Borin. 2000. You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 97–103, Saarbrücken, Germany.

P. Fung. 1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. *Lecture Notes in Artificial Intelligence*, 1529:1–17.

W.A. Gale and K.W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics (ACL'91)*, pages 177–184, Berkeley, USA.

J-X. Huang and K-S. Choi. 2000. Chinese-Korean word alignment based on linguistic comparison. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 392–399, Hong Kong.

M. Kay and M. Roescheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

S.J. Ker and J.S. Chang. 1997. Aligning more words with high precision for small bilingual corpora. *Computational Linguistics and Chinese Language Processing*, 2(2):63–96.

O.Y. Kwong and B.K. Tsou. 2001. Automatic corpus-based extraction of Chinese legal terms. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pages 669–676, Tokyo, Japan.

O.Y. Kwong. 2002. Toward a bilingual legal term glossary from context profiles. In *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation (PACLIC 16)*, pages 249–258, Jeju, Korea.

G.S. Mann and D. Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pages 151–158, Carnegie Mellon University, Pittsburgh, USA.

I.D. Melamed. 1997. A word-to-word model of translational equivalence. In *Proceedings of the 35th Conference of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL'97)*, Madrid, Spain.

S. Piperidis, S. Boutsis, and I. Demiros. 1997. Automatic translation lexicon generation from multilingual texts. In *Proceedings of the 2nd Workshop on Multilinguality in Software Industry: The AI Contribution (MULSAIC'97)*, Nagoya, Japan.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137, July.

P. Resnik and I.D. Melamed. 1997. Semi-automatic acquisition of domain-specific translation lexicons. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, Washington, DC.

M. Simard, G.F. Foster, and P. Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, pages 67–81.

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

D. Wu and X. Xia. 1995. Large-scale automatic extraction of an English-Chinese translation lexicon. *Machine Translation*, 9(3-4):285–313.

D. Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, Las Cruces, NM.