Toward A Bilingual Legal Term Glossary from Context Profiles

Oi Yee KWONG
Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
rlolivia@cityu.edu.hk

Abstract

We propose an algorithm for the automatic acquisition of a bilingual lexicon in the legal domain. We make use of a parallel corpus of bilingual court judgments, aligned to the sentence level, and analyse the bilingual context profiles to extract corresponding legal terms in both languages. Our method is different from those in past studies as it does not require any prior knowledge source, and naturally extends to multi-word terms in either language. A pilot test was done with a sample of ten legal terms, each with ten or more occurrences in the data. Encouraging results of about 75% average accuracy were obtained. This figure does not only reflect the effectiveness of the method for bilingual lexicon acquisition, but also its potential for bilingual alignment at the word or expression level.

1 Introduction

In this study, we propose an approach for acquiring legal term translations from parallel corpora, by analysing bilingual context profiles.

Following the implementation of legal bilingualism in the 90's, Hong Kong has experienced an increasing demand for authentic and high quality legal texts in both Chinese and English. In view of this, the Electronic Legal Documentation/Corpus System (ELDoS) project was initiated in 2000.[1] ELDoS is essentially a bilingual legal document retrieval system which provides a handy reference for the preparation of legal texts, Chinese judgments in particular (Kwong and Luk, 2001; LISRC, 2001).

The data in ELDoS come from two sources: a parallel corpus of original court judgment texts, in Chinese and English, and a bilingual glossary of legal terms derived from these judgments. According to many legal professionals, different terminologies are in fact used for different genres of legal documents such as statutes, judgments, and contracts. Hence for robustness and authenticity, the glossary in ELDoS is based on the corpus rather than any existing bilingual legal dictionary.

The compilation of the bilingual glossary from the judgments is thus one of the main tasks in the project. However, identification of legal terms and relevant concepts by humans depends to a large extent on their sensitivity which is, in turn, based on personal experience and legal knowledge. So not only is the process labour intensive, the results are also seriously prone to inconsistency. More importantly, inconsistency is to be avoided in the legal domain where language use should be precise and absolute.

Naturally, one way to facilitate the process is to seek automatic means to extract the relevant bilingual terms from texts. Past studies in this area mostly dealt with English and other Indo-European languages, and only few with English and Chinese.

In this study, we start with a list of Chinese legal terms extracted by a simple but effective method tailored to the characteristics of Chinese legal texts (Kwong and Tsou, 2001). For each of these

[1] ELDoS is a joint project between the City University of Hong Kong and the Judiciary of the Hong Kong Special Administrative Region (HKSAR).

Chinese terms, we attempt to automatically identify their English equivalents by analysing the context profiles in the bilingual texts.

The rest of this paper is organised as follows. In Section 2, we review past studies on term extraction and bilingual lexicon acquisition. In Section 3, we discuss the characteristics observed for Chinese legal terms which past studies had not addressed. In Section 4, we present the proposed mechanism for acquiring bilingual legal terms, with examples for illustration. In Section 5, we report on a pilot testing of the proposed method and discuss the results, before concluding in Section 6.

2 Related Work

On monolingual term extraction, Smadja (1993) developed Xtract to learn collocation patterns within small windows, taking the relative positions of the co-occurring words into account. Lin (1998) extracted English collocations from dependency triples obtained from a parser, using mutual information to filter triples which were likely to have co-occurred by chance.

Amongst the few relevant work on Chinese, Fung and Wu (1994) attempted to augment a Chinese machine-readable dictionary by collecting Chinese character groups from an untokenised corpus statistically. They modified Smadja's Xtract to CXtract for Chinese, starting with significant 2-character bigrams within a window of ±5 characters and seeding with these bigrams to match for longer *n*-grams. The corpus is made up of transcriptions of the parliamentary proceedings of the Legislative Council (LegCo) of Hong Kong. On average over 70% of the bigrams were found to be legitimate words and so for about 30-50% of other *n*-grams. With the extracted terms, they were able to obtain a 5% augmentation for a given Chinese dictionary. On the other hand, Kwong and Tsou (2001) applied simple collocation extraction techniques on a word-segmented corpus of Chinese court judgments. They found that simple methods, with slight adjustment to accommodate for the characteristics of Chinese legal terms, are as effective and the results could supplement a manually constructed glossary from the same set of data.

Extending from monolingual collocations, bilingual translation lexicons can be acquired (e.g. Wu and Xia, 1995; Smadja et al., 1996; Fung, 1998). This is particularly useful in machine translation, and is also pertinent to our setting. Wu and Xia (1995), for instance, learned translation associations between English words and individual Chinese characters, and obtained "encouraging but unsatisfactory" results, as they claimed. They also made use of terms extracted by CXtract to learn collocation translations for English words from the bilingual LegCo proceedings, reporting a precision of about 90%.

Also using purely statistical methods, Fung (1998) discussed an algorithm, Convec, to extract bilingual lexicons from non-parallel corpora. To find the Chinese translation of an English word, she compared the context vector of the English word with the context vectors of all Chinese words for the most similar candidate. She reported a 30% accuracy if the top-one candidate was considered, and the accuracy was more than doubled if the top-20 candidates were taken.

However, we find that the above methods for bilingual lexicon acquisition are limited in at least three ways. First, they need some existing general bilingual lexicons as bridges in the extraction process. Second, since they are purely statistically based, very large corpora are required, and data sparseness is still an obstacle. For example, Fung (1998) found that the precision on term extraction from a large corpus was much higher than that from a small corpus. Notwithstanding that, the above methods are apparently restricted to single English words. As we will see in the next section, these methods would not be sufficient for the extraction of legal terms for practical uses.

3 Characteristics of Legal Terms

In this section, we compare and contrast some of the characteristics of English and Chinese legal terms. Note that we sometimes use the word "term" in a loose way, referring to expressions of various lengths instead of just single and compound words in the normal sense.

For about 150 years, the legal system in Hong Kong operated through English only. It is not until these few years that parallel Chinese versions of legal documents are produced. Hence there are few established standards on how some legal concepts in the Common Law tradition should be expressed in Chinese, and the rendition of such English terms in Chinese inevitably leads to innovative use of Chinese expressions.

Meanwhile, a legal term glossary does not only contain single-word terms, but also longer expressions for relevant legal concepts. Legal concepts are not always lexicalised. For instance, the action of filing a lawsuit against someone is lexicalised as "sue" in English or "控告" in Chinese. But apparently there is no simple term for the action resulting in the status of "assault occasioning actual bodily harm" or "毆打引致他人身體受傷" except to use the whole expression as it is.

Thus partly as a lack of cross-lingual parallel lexicalisation and partly to do with a translator's style, a concise English term can correspond to a long and complex paraphrase in Chinese, and the reverse can also be true. For example, an English term can be of a simple modifier-head structure such as "procedural irregularity", but the Chinese translation – 程序/上/不/符合/規定[2] – is more complex.

Hence we see that the compilation of a legal term glossary is much more complicated than that of a general lexicon. Very often the entries to be included are not single-word terms, and their lengths may differ considerably between Chinese and English. In this study, we therefore propose an approach for the extraction of bilingual legal terms, which makes use of the consistency observed in legal translation, and avoids the problems which are likely to be met by existing methods.

4 The Proposed Mechanism and Examples

The algorithm we propose for acquiring bilingual legal glossary models the process of corpus-based construction of bilingual dictionaries. In general, parallel corpora are used and bilingually equivalent terms are identified from analysing context profiles of parallel concordances. We also take advantage of the characteristics of bilingual legal texts and make the following assumptions:

(1) Bilingual legal texts form relatively clean parallel corpora, in the sense that the alignments are expected to be neat, with few insertions and deletions.

(2) Legal terms, be they simple or compound, tend to be translated more consistently than general terms.

Figure 1 shows a schematic representation of our proposed approach, the rationale of which is explained below.

Our approach starts with the bilingual corpus aligned up to the sentence level. 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 alignment at a level finer than sentence. Moreover, many important legal concepts are expressed in compound terms or phrases. Hence it would be desirable if these terms were located in one language first, before finding their equivalents in the other language, so that we do not need to restrict ourselves to single-word terms. Within the concordances, a given term often has higher frequency than other co-occurring words.[3] Since terms in legal texts are more likely to be consistently translated, which is another characteristic of legal translation as mentioned, the source concordances should share a comparable context profile, i.e. frequency distribution, with the target concordances. That means words in the target concordances forming the equivalent term should share a similar frequency with the source term. Hence, by analysing the context profiles, we can identify the words in the target language which are likely to be expressing the concept of the source term. The comparison of context

[2] The slashes mark word boundaries in Chinese.

[3] Excluding function words.

vectors in past studies is essentially achieving the same purpose, but in our way, we can in fact discard many irrelevant co-occurring terms as early as possible, without entering into any complicated calculations.

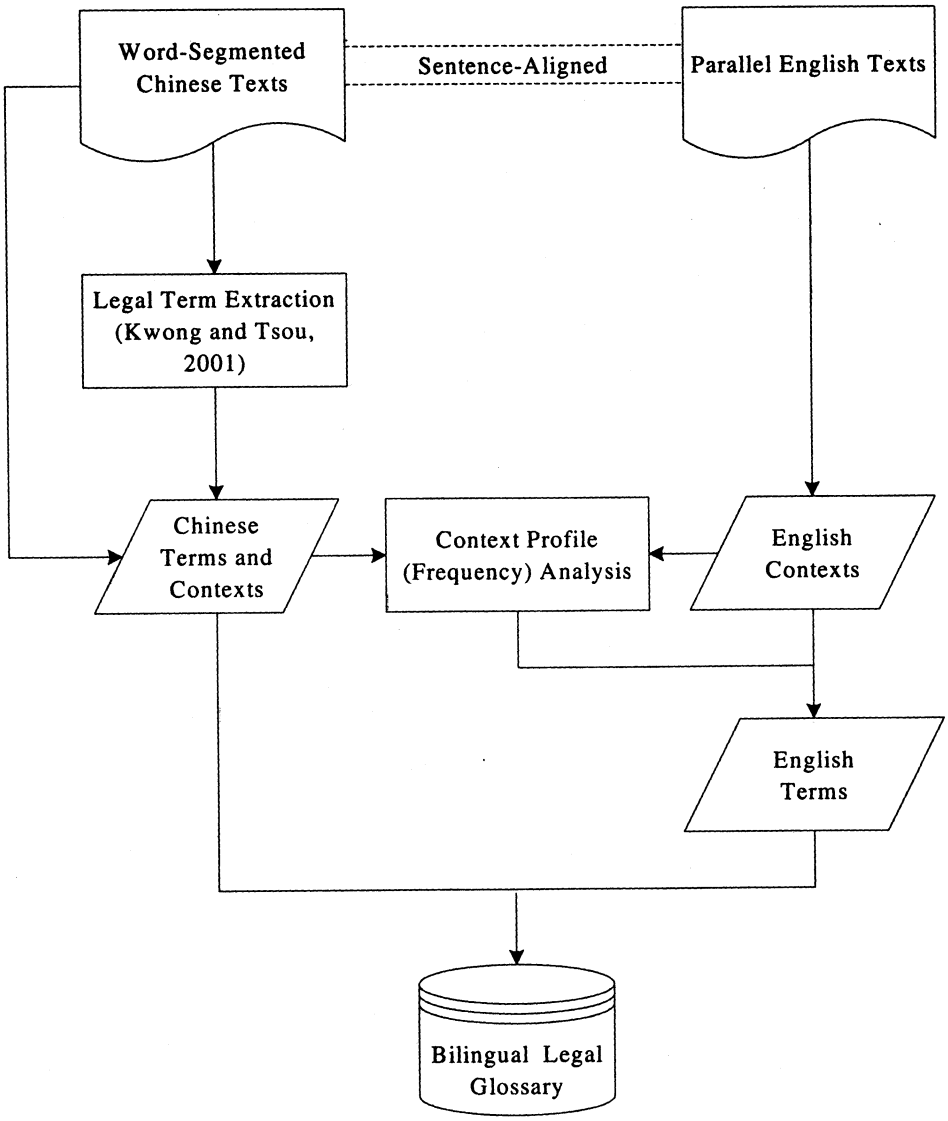

Figure 1 Schematic Representation of the Proposed Approach for Bilingual Legal Term Extraction

4.1 Observations

Examples like the following are found to support the above conjecture. The tables below show the top part of the Chinese and English context profiles (with stop words removed) of the terms 量刑基準 (starting point), 臨時協議 (provisional agreement), and 換地條件 (Conditions of Exchange) respectively. These Chinese terms were extracted automatically by the algorithm in Kwong and Tsou (2001). Some bilingual concordances are also shown. As can be seen, the top frequency words in the

English context profile often form compound terms in the English texts, or they are part of a relevant phrase spanning a small window.

(1) 量刑基準 (Starting point)

Context profiles:

Chinese Collocations	Frequency	English Collocations	Frequency
量刑基準	6	point	6
過	4	starting	6
明顯	3	excessive	3
月	3	Ma	3
重	3	manifestly	3

Concordance examples:

Chinese Sentence	English Sentence
8年的量刑基準絕非明顯過重	the **starting point** of 8 years is no way manifestly excessive
至於偷竊罪名的18個月量刑基準	As regards the **starting point** of 18 months for the theft count

(2) 臨時協議 (Provisional agreement)

Context Profiles :

Chinese Collocations	Frequency	English Collocations	Frequency
臨時協議	14	agreement	21
協議	12	Provisional	14
訂立	5	entered	4
規定	4	parties	3
正式	4	payment	3

Concordance examples:

Chinese Sentence	English Sentence
鑑於雙方在簽署臨時協議後所作的行爲	in the light of the conduct of the parties after their entry into the **provisional agreement**
臨時協議第5條規定付款辦法如下	Clause 5 of the **Provisional Agreement** provided for payment as follows

(3) 換地條件 (Conditions of Exchange)

Context profiles:

Chinese Collocations	Frequency	English Collocations	Frequency
條件	5	Conditions	7
換地條件	4	Exchange	4
副本	3	number	2
項	3	provided	2
特別	3	General	2

Concordance examples:

Chinese Sentence	English Sentence
該附表列出一換地條件第10485號的核簽副本及數份其它文件	That schedule listed an attested copy of the **Conditions of Exchange** No.10485 and a number of other documents
換地條件包括若干特別條件	The **Conditions of Exchange** contained a number of special conditions

4.2 The Algorithm

Hence we suggest an algorithm as follows:

Step 1 Run the Chinese term extraction algorithm on the word-segmented Chinese half of the bilingual corpus.

Step 2 For an extracted Chinese (compound) term, mark it as a single unit in the original corpus and retrieve its concordances (source concordances).

Step 3 Retrieve all corresponding, aligned sentences from the English half of the corpus (target concordances). Words should be counted in their lemmatized forms.

Step 4 Delete all stop words from both the Chinese and English concordances.

Step 5 Perform a word frequency analysis from the concordances and rank the results.

Step 6 Define a frequency threshold as *T*source frequency* and a small window size *w*. For a given *T*, pick the words in the context profile of the target concordances above the threshold. Locate these words in the original concordances and mark off their co-occurring patterns within a window of size *w*. The longest string spanning over *w* forms a candidate translation of the original Chinese term.

Our method is thus different from those described in Section 2 in the following regards:

(1) It is not restricted to single-word terms. Starting from compounds in the source language, it looks for equivalents in the target language.

(2) Although more evidence would be desirable from a large corpus, the method does not inherently require a large corpus to start with. As the examples above illustrate, it works well even with only a few concordances.

(3) No prior knowledge source (e.g. online word lists, existing bilingual dictionaries, etc.) is required.

5 Pilot testing and discussion

A pilot testing of the method proposed in the last section was done. To start with, ten Chinese legal terms (all compound terms) were randomly selected from those extracted automatically by Kwong and Tsou (2001). The samples contain terms of different lengths and structures. The same set of corpus data, which consists of about 100K Chinese characters and their corresponding English portions of authentic Hong Kong court judgments, was used in the current study. Testing was done with different values for the parameters T (0.8 and 0.9) and w (n, $n+1$, $n+2$ and $n+3$, where n is the number of English words crossing the frequency threshold).

With the selected Chinese terms, the algorithm described in Section 4 was run. Accuracy was measured in terms of the amount of candidate translations extracted being the correct candidates. The results are summarized in Table 1.

As seen in Table 1, the results are in fact very encouraging. The algorithm correctly identifies the English equivalents of many Chinese terms under test. The extracted terms are not restricted to any particular length or structure. In most cases, the results are similar with T set at 0.9 or 0.8. However, it is still marginally better with a higher T, to include only the most salient words. As for the variation of w, a wider window seems to introduce more noise, but that also seems to depend on the length and complexity of the term in question. Generally speaking, the optimal combination in our experiment is 0.9 for T and $n+1$ for w, which results in an average accuracy of over 75%.

In addition, we observe the following interesting phenomena and problems, which call for further refinement of the algorithm as well as post-processing steps to clean up the results. We will discuss below how our method might be improved.

- Pattern Generalisation

Some generalisation from the translation candidates would be needed. For example, the different renditions found for "上訴得直", including "allow the appeal", "appeal be allowed", "allowing the appeal", and "appeal is allowed", are essentially the variants of the same English V-O pattern, namely "allow appeal". To make an informative bilingual glossary, we need both the root form as well as the more frequent form found in real data, i.e. the corpus per se.

- Further Significance Testing

Although T could be varied, it is possible that words other than the relevant ones also cross the frequency threshold. On the one hand, these words, although not part of the correct translation, are very strong collocates of the term in question. On the other hand, these words might in fact be frequent throughout the corpus, and their association with the term in question is not significant. As a result, even though our method can get rid of most irrelevant words at an early stage, the significance of the remaining ones and their association strength are still worth attention. Our samples on "司法管轄權" give an ideal

illustration. The correct translation for the term is "jurisdiction", but it always co-occur with "court", which is nevertheless extremely abundant in the whole corpus.

- Other Supplementary Parameters

Sometimes more than one translation candidate would be found within the same concordance line, but the original Chinese term only appeared once on the Chinese side. For instance, in the first example below, the correct translation, "lawful order", was found twice on the English side where there was only one "合法命令" in the Chinese sentence. On the other hand, in the second example below, two different candidate English terms, "Court of Appeal" (incorrect) and "allowed the appeal" (correct), were found for the Chinese term "上訴得直". In these cases, it is apparent that some ways have to be established to decide on the exact correspondences. Relative position might be one criterion, although not always reliable with two languages so different in nature. Alternatively, since we would not just focus on one or two terms in the whole corpus, it is very likely that "Court of Appeal" had already been identified as the translation for another term: "上訴法庭". So by cross checking with other terms and their translations, we might be able to filter out some invalid candidates.

And his refusal to answer constituted disobedience of a *lawful order* . His superior's order to answer was a *lawful order* . But in substance the case against him ran thus .	但實質上指控他的案是：他的上司命令他要回答問題的命令是合法的，而他拒絕回答，構成了不服從合法命令的行爲。
Having heard arguments which were much fuller than those put before the trial judge , the *Court of Appeal* (Hon Chan CJHC , Leong and Stuart - Moore JJA) unanimously , with each member giving a reasoned judgment , *allowed the appeal* and reversed the direction to discharge the applicants on the trafficking counts .	經聆聽較在原審法官席前更詳盡的論據後，上訴法庭(高等法院首席法官陳兆愷，上訴法庭法官梁紹中和上訴法庭法官司徒冕)每名法官皆宣告附有理由的判決，一致判決上訴得直，並推翻就販運罪名釋放兩名申請人的指令。

- Anaphora Resolution

In many cases, the corresponding English rendition for "強制執行公約裁決" ("enforcement of a Convention award") is identified as "enforcement of the award". This is acceptable from the perspective of bilingual alignment, and in fact it is a perfect match in this context. The accuracy of the method, therefore, does not only reflect the effectiveness of the method for bilingual lexicon acquisition, but also hints on its potential for bilingual alignment at the word or expression level. However, "enforcement of the ward" is not the precise translation for the term out of context. The reason for such a mismatch is that the definite description "the award" must be referring to some aforementioned "Convention award". Hence, to improve the precision of the term extraction process, either the anaphors have to be resolved beforehand, or discarded from the candidates.

6 Conclusion

Thus in this paper, we have proposed a mechanism based on bilingual context profiles for the automatic extraction of bilingual legal terms. Not many past studies discussed the problem between English and Chinese. Our algorithm, unlike other past methods, does not require any prior knowledge source and is not limited to single-word terms. The only resource needed is a sentence-aligned parallel corpus. Our pilot experiment has demonstrated the plausibility of the algorithm, with an average accuracy of about 75%, and in fact above average for many test instances. Our next step is to fine-tune the algorithm, with regard to the various points discussed in Section 5, and then apply it on a larger scale.

Chinese Term	*w\T*	0.9	0.8	Correct English Renditions
罪名成立 (12)	*n*	100%	100%	(was) convicted, convicting, conviction
	n+1	--	--	
	n+2	--	--	
	n+3	--	--	
合法命令 (11)	*n*	100%	45.5%	lawful order(s)
	n+1	81.8%	36.4%	
	n+2	54.5%	18.2%	
	n+3	36.4%	18.2%	
註冊摘要 (11)	*n*	100%	100%	memorial
	n+1	--	100%	
	n+2	--	81.8%	
	n+3	--	81.8%	
必然含意 (14)	*n*	100%	100%	necessary implication
	n+1	--	--	
	n+2	--	--	
	n+3	--	--	
開庭審理 (12)	*n*	100%	100%	hearing(s)
	n+1	--	--	
	n+2	--	--	
	n+3	--	--	
上訴得直 (16)	*n*	87.5%	87.5%	allow the appeal, appeal be allowed, allowing the appeal, appeal is allowed
	n+1	75%	75%	
	n+2	62.5%	62.5%	
	n+3	62.5%	62.5%	
專家報告 (10)	*n*	80%	80%	experts' report, report of the experts
	n+1	80%	90%	
	n+2	90%	90%	
	n+3	90%	80%	
司法管轄權 (13)	*n*	15.4%	15.4%	jurisdiction
	n+1	15.4%	15.4%	
	n+2	15.4%	15.4%	
	n+3	15.4%	15.4%	
強制執行公約裁決 (10)	*n*	0%	6.3%	enforcement of a Convention award, enforcement of Convention awards, Convention award enforcement
	n+1	6.3%	12.5%	
	n+2	12.5%	56.3%	
	n+3	56.3%	56.3%	
不導致自己入罪的特權 (10)	*n*	0%	0%	privilege against self-incrimination
	n+1	100%	100%	
	n+2	100%	100%	
	n+3	100%	100%	

Table 1 Results of Pilot Testing of Extraction Algorithm

Acknowledgements

We thank the Judiciary of the HKSAR for providing the judgment data. The author takes sole responsibilities for the findings and views expressed hereon.

References

Fung, P. (1998) A Statistical View on Bilingual Lexicon Extraction: from Parallel Corpora to Non-parallel Corpora. *Lecture Notes in Artificial Intelligence, 1529*: 1-17.

Fung, P. and Wu, D. (1994) Statistical Augmentation of a Chinese Machine-Readable Dictionary. In *Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2)*, Kyoto.

Kwong, O.Y. and Luk, R. (2001) Retrieval and Recycling of Salient Linguistic Information in the Legal Domain: Project ELDoS. *Presented in the Annual Conference and Joint Meetings of the Pacific Neighborhood Consortium (PNC 2001)*, Hong Kong.

Kwong, O.Y. and Tsou, B.K. (2001) Automatic Corpus-Based Extraction of Chinese Legal Terms. To appear in *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, Tokyo, Japan.

Language Information Sciences Research Centre (LISRC). (2001) *ELDoS Version 1.0: Installation and Operation Manual*. City University of Hong Kong.

Lin, D. (1998) Extracting Collocations from Text Corpora. In *Proceedings of the First Workshop on Computational Terminology*, Montréal, Canada.

Smadja, F.Z. (1993) Retrieving Collocations from Text: Xtract. *Computational Linguistics, 19(1)*: 143-177.

Smadja, F.Z., McKeown, K. and Hatzivassiloglou, V. (1996) Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics, 22(1)*: 1-38.

Wu, D. and Xia, X. (1995) Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon. *Machine Translation, 9(3-4)*: 285-313.