Some Considerations on Guidelines for Bilingual Alignment and Terminology Extraction

Lawrence Cheung, Tom Lai, Robert Luk[§], Oi Yee Kwong, King Kui Sin, Benjamin K. Tsou

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

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

Abstract

Despite progress in the development of computational means, human input is still critical in the production of consistent and useable aligned corpora and term banks. This is especially true for specialized corpora and term banks whose end-users are often professionals with very stringent requirements for accuracy, consistency and coverage. In the compilation of a high quality Chinese-English legal glossary for ELDoS project, we have identified a number of issues that make the role human input critical for term alignment and extraction. They include the identification of low frequency terms, paraphrastic expressions, discontinuous units, and maintaining consistent term granularity, etc. Although manual intervention can more satisfactorily address these issues, steps must also be taken to address intra- and inter-annotator inconsistency.

Keyword: legal terminology, bilingual terminology, bilingual alignment, corpus-based linguistics

1. Introduction

Multilingual terminology is an important language resource for a range of natural language processing tasks such as machine translation and cross-lingual information retrieval. The compilation of multilingual terminology is often time-consuming and involves much manual labour to be of practical use. Aligning texts of typologically different languages such as Chinese and English is even more challenging because of the significant differences in lexicon, syntax, semantics and styles. The discussion in the paper is based on issues arising from the extraction of bilingual legal terms from aligned Chinese-English legal corpus in the implementation of a bilingual a text retrieval system for the Judiciary of the Hong Kong Special Administrative Region (HKSAR) Government.

Much attention in computational terminology has been directed to the development of algorithms for extraction from parallel texts. For example, Chinese-English (Wu and Xia 1995), Swedish-English-Polish (Borin 2000), and Chinese-Korean (Huang and Choi 2000). Despite considerable progress, bilingual terminology so generated is often not ready for immediate and practical use. Machine extraction is often the first step of terminology extraction and must be used in conjunction with rigorous and well-managed manual efforts which are critical for the production of consistent and useable multilingual terminology. However, there has been relatively little discussion on the significance of human intervention. The process is far from being straightforward because of the different purposes of alignment, the requirements of target users and the corpus type. Indeed, there remain many problematical issues that will not be easy to be resolved satisfactorily by computational means in the near future, especially when typologically different languages are involved, and must require considerable manual intervention. Unfortunately, such critical manual input has often been treated as an obscure process. As with other human cognitive process (T'sou et al. 1998), manual terminology markup is not a straightforward task and many issues deserve closer investigation.

In this paper, we will present some significant issues for Chinese-English alignment

and term extraction for the construction of a bilingual legal glossary. Section 2 describes the background of the associated bilingual alignment project. Section 3 discusses the necessity of manual input in bilingual alignment, and some principles adopted in the project to address these issues. Section 4 provides an outline for further works to improve terminology management, followed by a conclusion in Section 5.

2. High Quality Terminology Alignment and Extraction

2.1 Bilingual Legal Terminology in Hong Kong

The implementation of a bilingual legal system in Hong Kong as a result of the return of sovereignty to China in 1997 has given rise to a need for the creation and standardization of Chinese legal terminology of the Common Law on par with the English one. The standardization of legal terminology will not only facilitate the mandated wider use of Chinese among legal professionals in various legal practices such as trials and production of legal documentation involving bilingual laws and judgments, but also promote greater consistency of semantic reference of terminology to minimize ambiguity and to avoid confusion of interpretation in legal argumentation.

In the early 90's, Hong Kong law drafters and legal translation experts undertook the unprecedented task of translating Hong Kong Laws, which are based on the Common Law system, from English into Chinese. In the process, many new Chinese legal terms for the Common Law were introduced. On this basis, an English-Chinese Glossary of legal terms and a Chinese-English Glossary were published in 1995 and 1999 respectively. The legal terminology was vetted by the high level Bilingual Laws Advisory Committee (BLAC) of Hong Kong. The glossaries which contain about 30,000 basic entries have become an important reference for Chinese legal terms in Hong Kong. The Bilingual Legal Information System (BLIS) developed by the Department of Justice, HKSAR provides simple keyword search for the glossaries and laws that are available in both Chinese and English. Nevertheless, the glossaries are far from being adequate for many different types of legal documentation, e.g. contracts, court judgments, etc. One major limitation of the BLIS glossary is its restricted coverage of legal terminology in the Laws of Hong Kong, within a basically prescriptive context as when the laws were studied at the time of its promulgation. There are other important bilingual references (Li and Poon 1998, Yiu and Au-Yeung 1992, Yiu and Cheung 1996) which focus more on the translation of Common Law concepts. These are almost exclusively nominal expressions.

In 2000, the City University of Hong Kong, in cooperation with the Judiciary, HKSAR, initiated a research project to develop a bilingual text retrieval system, Electronic Legal Documentation/Corpus System (ELDoS), which is supported by a bilingually aligned corpus of judgments. The purpose of the on-going project is twofold. First, the aligned legal corpus enables the retrieval of legal terms used in authentic contexts where the essence and spirit of the laws are tested (and contested) in reality, explicated and elaborated on, as an integral part of the evolving and defining body of important precedent cases unique to the Common Law tradition. Second, the corpus covers judgment texts involving interpretation of different language styles and vocabulary from Hong Kong laws. The alignment markup also serves as the basis for the compilation of a high-quality bilingual legal term bank. To complete the task within the tight timeframe, a team of annotators highly trained in law and language are involved in alignment markup and related editing.

2.2 Need for Human Input

The legal professionals which are the target users of ELDoS have very stringent demands on terminology in terms of *accuracy*, *coverage* and *consistency*. Aligned texts and extracted terms must therefore be carefully and thoroughly verified manually to minimize errors. Furthermore, many studies on terminology alignment and extraction deal predominantly with nominal expressions. Since the project aims to provide comprehensive information on the manifestations of legal vocabulary in Chinese and English texts, the retrieval system should not restrict users to nominal expressions but should also provide reference to many other phenomena such as alternation of part-of-speech (POS) (e.g. noun-verb alternation) inherent in bilingual texts, as will be seen in Section 3.

The availability of bilingual corpora has made it possible to construct representative term

banks. Nonetheless, current alignment and term extraction technology are still considered insufficient to meet the requirements for high quality terminology extraction. In ELDoS project, many issues are difficult to be handled satisfactorily by the computer in the foreseeable future. Although human input is essential for high quality term bank construction, the practice of manual intervention is not straightforward. Indeed, the manual efforts to correct the errors can be substantial, and the associated cost should not be underestimated. The annotator must first go through the entire texts to spot the errors and terms left out by the machines. In this process, both the source and target materials have to be consulted. The annotator must also ensure the consistency of the output. As a result, guidelines should be set up to streamline the process.

3. Aspects of Terminology Alignment

The approach adopted for the manual annotation of alignment markup and the maintenance of term bank in the ELDoS project will be described. Additional caution has been taken in the coordination of a team of annotators.

3.1 Term Frequency

An important reason for manual intervention in bilingual term alignment is the relatively poor recall rate for low frequency terms. Many extraction algorithms make use of statistical techniques to identify multi-word strings that frequently co-occur (Wu and Xia 1995; Kwong and Tsou 2001). These methods are less effective for locating low frequency terms. Of the 16,000 terms extracted from ELDoS bilingual corpora, about 62% occur only once in about 80 judgments. For high quality alignment and extraction, failure to include these low frequency terms would be totally unacceptable.

3.2 Correspondence of Aligned Units

Because of the different grammatical requirement and language style, a term in the source language often differs in different ways from the corresponding manifestations in the target language. These differences could be alternation of POS and the use of paraphrastic expressions. Although many term banks avoid such variations and focus primarily on equivalent nominals or verbs, the correspondence of terms between two typologically different languages is often more complicated. For example, the English nominal ("fulfilment") is more naturally translated into Chinese as a verb ("不履行", "拒絕履行", "悔約"). More examples can be found in Table 1.

Alternation of POS

English	Chinese	POS alternation
The accused	被告	det + adj ~ noun
hold	裁決	verb ~ noun
fulfillment	履行	noun ~ verb
administration	承辦	noun ~ verb
repudiation	不履行	noun ~ neg + verb

Table 1. Alternation of POS

In some cases, there are simply no equivalent words in the target language. Paraphrasing or circumlocution may be necessary. Such correspondence is far less consistent and obvious to be identified by the computer.

Paraphrasing/Circumlocution

English	Chinese
The judge **entered judgment in favour of** the respondents in respect of their claim for arrears of wages, and severance payment.	法官判答辯人的欠薪及遣散費索償勝訴。
In our view,…	我等認爲…
…**evidenced** by the Defendant's letter …	…由原告人所簽署確認的函件中可獲得證據…

Table 2. Examples of paraphrasing

Because of language differences, legal terms can be contextually realized as anaphors in the target language. Examples of such correspondence would be useful for legal drafting and translation. Again, such anaphoric relations are more accurately handled by humans.

Anaphoric Relation

English	Chinese
He was subsequently charged…	申請人後來檢控…
Liu JA dealt with that application on 14 March 1996 and dismissed **it**.	上訴法庭法官廖子明於1996年3月14日聆訊並撤銷該項申請。
Enforcement of a Convention award may also be refused if the **award** is in respect of a matter which is not capable of settlement by arbitration.	如公約裁決所關乎的事項，是不能藉仲裁解決，…則亦可拒絕強制執行該裁決。

Table 3. Examples of anaphors

3.3 Discontinuous Units

Most term extraction algorithms deal with contiguous units, e.g. *n*-gram. These algorithms would be problematical in handling discontinuous units. They include phrasal verbs (e.g. "strike out"), collocation patterns (e.g. "lodge three complaints", "被判…罪名成立"). These have to be manually added or edited. Interestingly, our preliminary study shows that over 90% of the instances of discontinuous units are found in the Chinese manifestation of English terms. Some examples are listed in Table 4.

English	Chinese
convict …	被判…罪名成立
The Court of Appeal **allowed** the tenant's appeal	上訴法庭曾判決租客上訴得直
The agreement **kept** the company **alive**	協議使公司能夠繼續運作
If the Defendant **misrepresented** to the Plaintiff that what he was signing was only…	倘若被告人錯誤向原告人表示他簽署的只是…

Table 4. Examples of discontinuous units

3.4 Selective Markup

To avoid producing "uninteresting" term alignment, restricting markup to only terms of the interested domain would be an attractive alternative to full-text alignment. In the ELDoS project, it is possible to mark up only legal terminology. Other non-legal elements can be omitted in alignment annotation. This approach has been accepted by the ELDoS client. Some examples of legal and non-legal terms are shown in Table 5.

Legal Terms			Non-legal Terms		
1	alibi evidence	不在犯罪現場證據	1	collar	領
2	order nisi	暫准命令	2	finger	手指
3	prima facie	表面	3	question	問題

Table 5. Legal vs. Non-legal Terms

However, many other terms are more ambiguous. There is often no hard and fast rule to set criteria for domain membership. Annotators would have to rely on their own individual judgement to decide whether an expression should be counted as a legal term. For example, the English words listed in Table 6 are not used exclusively in the legal domain. However, taking into account their frequency, legal context and the multiple renditions in Chinese, they are worthy of being considered as "semi-legal." What is interesting about "I" is that though the pronoun is a common pronoun, the corresponding Chinese manifestation "本席" is used exclusively in the judgments and should be regarded as legal. These examples suggest that the decision to classify a phrase as a legal term involves a great deal of complications.

Semi-Legal Terms	
impose	制定, 頒下, 判, 訂下, 訂定
terms	措詞, 措辭, 條文, 條款, 觀點
waive	放棄, 寬免, 豁免
I	本席

Table 6. Semi-legal Terms

Selective markup, however, could give rise to intra- and inter-annotator inconsistency. The vagueness of legal terms could lead to variation in the selection of the same term at different times and among different annotators. In ELDoS project, computer-aided markup tools that can instantly check candidate expressions against the term bank is an effective reference for annotators to maintain consistency. Those terms that are found in term bank should be included in the alignment. In this way, the term bank can serve as a working standard for annotators. As for new terms, our annotators have adopted the principle that whenever they have doubts as to domain membership of a new term, they should include the term in the alignment. In this way, all the candidate terms are guaranteed to be available for the term bank manager for final decision. Inter-annotator differences can also be reduced by fostering more communication among annotators such as regular review of peer work.

3.5 Granularity

Term granularity is another major issue not only for machines but also for humans. The terminology list should be as simple and compact as possible to avoid redundancy of entries. For example, instead of having "allegations", "corruption", "allegations of corruption", "allegations of manslaughter" as separate entries, it is preferable to treat only "allegations", "corruption" and "manslaughter" as glossary entries. The annotators have adopted the principle that a term should be a minimal semantic unit.

Here "semantic unit" refers to single- or multi-word terms that have acquired specialized meaning or usage. For example, the phrase "great and general importance" 重大而廣泛重要性 has been used frequently as a frozen chunk, and should not be further divided into "great", "and", "general" and "importance". Similarly, "oral decision" 口頭判詞 refers to the verbal delivery of judgments in trial as opposed to written judgments. Such decisions involve the support of real-world knowledge and sophisticated semantic/pragmatic interpretation and are not easily modelled by the computer.

4. Further Works

Bilingual terminology extracted directly from the bilingual corpora bear the form as it is in the text corpora. English words with different morphological markers will give rise to multiple entries in the resulting glossary. However, from the user's point of view, verbs with the same root but different inflectional markers (e.g. "hold", "held", "holding") should be combined to form one single entry. Similarly, variants of Chinese expressions that differ simply by an optional markers 的 *de* (see Table 7) may better be treated as the same item to minimize redundancy.

English	Chinese
lawful attorney	1. 合法受權人 2. 合法的受權人
order nisi	1. 暫時命令 2. 暫時的命令
presumption in law	1. 法律上推定 2. 法律上的推定

Table 7. Chinese renditions differed by 的 *de*

Term bank management tools will be developed to process the morphological markers and combine related pairs.

5. Conclusion

In this paper, we have studied the importance of human intervention in the preparation of aligned corpus and bilingual terminology, particularly when the specialized language resource and users are involved. Their demand for high quality alignment and terminology gives rise to a number of issues that are difficult, if not impossible, to be dealt with satisfactorily by the computer. These issues include the extraction of low frequency terms, the identification of equivalent units with different POS or paraphrases, discontinuous units, consistency of term granularity, etc. Though human intervention is far more effective in attaining accurate linguistic processing, if not properly planned, it could also suffer from intra- and inter-annotator inconsistency. Regular peer review and appropriate markup checking tools will be essential to minimize inconsistency.

Acknowledgements

This study is supported through the Hong Kong Judiciary contract (#RCL/0794) and the Research Grants Council of Hong Kong under Competitive Earmarked Research Grant (CERG) No. 9040570. We specially thank the Judiciary of HKSAR for providing authentic judgments for this study. Our thanks also go to the team of alignment annotators who provided interesting examples for this paper.

References

Borin, L. (2000) You 'll Take the High Road and I'll Take the Low Road: Using a Third Language to Improve Bilingual Word Alignment. *Proceedings of COLING 2000*, pp. 97—103.

Huang, J.-X. and K.-S. Choi. (2000) Chinese-Korean Word Alignment Based on Linguistic Comparison. *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics*, pp. 392—399.

Kwong, O. Y. and B. K. Tsou. (2001) Automatic Corpus-Based Extraction of Chinese Legal Terms. *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pp 669—676, Tokyo, Japan.

Li, Z. E. and E. Poon. (1998) *English-Chinese Dictionary of Law*. Hong Kong: Commercial Press.

T'sou, B. K., H. L. Lin, S. W. K. Chan and T. B. Y. (1998) "Cognitive Variation in Cross Straits Perception of Salient Information", in B. K. T'sou, T. B. Y. Lai, S. W. K. Chan and W. S-Y. Wang (Eds.), *Quantitative and Computational Studies on the Chinese Language*, Language Information Sciences Research Centre, City University of Hong Kong, pp.31-56.

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

Yiu, T.-W. and Z.-Q. Zhang. (1996) *English-Chinese Dictionary of Commercial Law*. Hong Kong: Commercial Press.

Yiu, T.-W. and P.-K. Au-Yeung. (1992) *A Glossary of Law for Hong Kong*. Hong Kong: Commercial Press.