Construction of Chinese-English Parallel Corpus of Chinese Laws and Regulations and Term Extraction Assisted by CAD Virtual Reality Technology

Qingguo Zhao[1*] and Jun Wang[2]

[1]School of General Education, Xi'an Eurasia University, Xi'an 710065, China, zhaoqingguo@eurasia.edu
[2]Experiment Center, Zhengzhou Electric Power Technology College, Zhengzhou 451450, China, wangjunqfry612@126.com

Corresponding author: Qingguo Zhao, zhaoqingguo@eurasia.edu

Abstract. Computer-aided term extraction based on corpus is recognized as a translation technology with high ROI, which can effectively improve the translation efficiency and ensure the consistency of term translation, and is gradually attracting international attention. Compared with other branches of linguistics, most scholars of legal language still stick to the traditional intuition and introspection method to conduct qualitative research on language, but few use massive corpus and empirical research method to conduct research on legal texts. In this paper, a parallel Corpus of Chinese laws and regulations in Both English and Chinese is constructed. On this basis, Chinese and English terms are extracted by computer software, and a comparative glossary of laws and regulations in Chinese and English is made to construct the corpus and extract computer-aided terms. This paper presents an algorithm for automatic extraction of term dictionaries from English-Chinese parallel corpora. The aligned bilingual corpus is used, and the Chinese language is segmented. The part-of-speech tagging (POS) of English corpus and Chinese corpus is carried out by English and Chinese POS tagging tools. Generate candidate sets of nouns and noun phrases from bilingual corpus. Then the translation probability between each English candidate term and its related Chinese translation is calculated. Finally, the word with the highest probability is selected as the Chinese translation of the English candidate word by greedy algorithm.

Key words: parallel corpus; CAD virtual reality technology; laws and regulations
DOI: https://doi.org/10.14733/cadaps.2023.S1.46-55

1 INTRODUCTION

Bednarek and Carr [1] consider that a term is a term for a general concept in a particular field of expertise. In many fields of natural language computer processing, such as information retrieval, information extraction, text classification and other tasks, the basic unit is usually word type term or phrase type term, which cannot be separated from the automatic processing of term 1. At the same time, Dmitrijev and Kogan [2] hold that bilingual term extraction plays an important role in the compilation of bilingual term dictionary, the construction of bilingual ontology, machine translation and cross-language information retrieval. Bilingual core terms are one of the key resources for bilingual term recognition and extraction. With bilingual core term pairs as seed pairs, we can find larger bilingual term pairs by machine learning. Domain documents, on the other hand, contain a lot of jargon. Document keywords are often valid candidates for core terms. Therefore, how to make full use of the existing large-scale professional field classification resources to extract the core terms of Chinese and English comparison is a very meaningful work. In this paper, Chinese laws and regulations are selected as candidate core terms, and key techniques such as keyword extraction and term degree calculation are used to identify core terms in Chinese and English by using the corpus of Chinese and English professional field classification. Then, based on the parallel corpus of Chinese and English professional fields, the core terms are automatically generated by using bilingual alignment technology. Corpus usually refers to language materials collected for language research and preserved in electronic form, which are collected from naturally occurring samples of written or spoken language and used to represent a particular language or language variant. Gong [3] thinks that bilingual parallel corpus generally refers to the corpus composed of the original text and its corresponding translated text, and its auxiliary function for translation is more obvious. Compared with traditional corpora, parallel corpora emerged later. It has only been about 20 years since the establishment of the Canadian Hansard Corpus (including The English and French version of the Canadian parliamentary debate Corpus), The first preliminary parallel Corpus in the world. However, due to the great potential application value of parallel corpora in linguistics and natural language processing, such as language contrastive studies, translation studies, translation teaching, translation technology development, bilingual lexicography, the construction of parallel corpora has been rapidly developing in the world. Friedemann et al. [4] think that parallel corpus refers to a bilingual corpus composed of the original text and its corresponding translation language. The degree of bilingual alignment or correspondence can be divided into text, paragraph, sentence and word. Text alignment is too simple, similar to bilingual reading texts, while Chinese-English word or character alignment is too complex and requires too much technology, which makes it difficult to achieve. At present, it is only at the theoretical level without any successful case reports. At present, the corresponding level of parallel corpus is generally at paragraph and sentence level. Ammar et al. [5] think that edit the candidate terms extracted by software, delete useless phrases (such as "relevant", "mouth goods", etc.) and general terms from the extracted results, leaving effective terms with characteristics of laws and regulations. When there is no candidate translation or candidate error for the term, select "concordance" and add the search result as the translation. After the adjustment and screening is completed, the terms are exported into TXT documents. At this time, computer-aided terminology extraction has been completed and can be used for subsequent analysis [6-7].

Term extraction technology is a very important topic in information processing. According to different corpus and research purpose, different research methods can be adopted. The Champollion system is designed to extract bilingual collocation dictionaries from aligned English and French corpus by first extracting meaningful continuous and discontinuous collocations from The English corpus and then extracting corresponding translations by calculating the mutual information between English collocations and candidate French translations in aligned corpus. Based on the establishment of a large bilingual parallel corpus, a phrase database of more than

one million items is constructed by using the key technology of big data processing, the methods and ideas of artificial intelligence, and the needs of batch term generation and discovery, and an application for automatic extraction and discovery of compound terms is developed [8-10].

2 CONSTRUCTION OF PARALLEL CORPUS OF CHINESE LAWS AND REGULATIONS IN ENGLISH AND CHINESE

At present, most of the researches based on corpus-based in China choose the general direction and angle, such as Chinese-English parallel corpus. Also built one of the few some of corpus for special purposes, such as Shanghai jiao tong university, yanshan university of Shakespeare's drama English-Chinese parallel corpora is a dream of red mansions translation of parallel corpus, lu xun's novels Chinese-English parallel corpus of shaoxing liberal art &science college, etc., can say to the laws and regulations as the research object of corpus is rare, Only Chinese Laws and Regulations Chinese-English Parallel Corpus (PCCLD). At present, there is no parallel corpus specially established on the basis of laws and regulations in China. At the same time, although Chinese-English parallel corpus and computer-aided term extraction have become popular in the academic world, there is still no complete system of relevant corpus in the legal field, and there are few researches on its specific aspects. The whole algorithm process is shown in Figure 1.

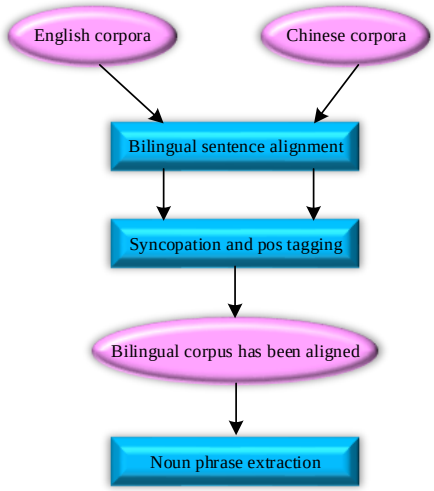

Figure 1: The whole algorithm process.

The first step of corpus construction is to collect the corpus. In order to ensure the poor progress of corpus, we can collect customs announcements, customs decrees, joint announcements, joint decrees and laws and regulations in Both Chinese and English from the official website of General Administration of Customs of China and law firm website, covering corpus published between 1995 and 2015. When collecting corpus, laws and regulations (such as import and export trade law) that are not within the scope of customs law but closely related to customs are also included in corpus, and the corpus with too few words and tables are deleted.

There are some problems in the corpus collected from the Internet, such as garbled characters, chaotic punctuation marks (the whole corner and half corner are not unified, and English punctuation and Chinese punctuation are mixed), and repeated corpus. To solve these problems, we use auxiliary software "text collator" to normalize the text processing. With the help of the replacement function of the text collator (the operation interface is shown in Figure 2), replace all full-corner characters in the corpus with half-corner characters and delete redundant Spaces and

garbled characters. At the same time, I wrote a corpus cleaning log as an individual to save the problems generated in the operation for future inquiry and reference.

In order to meet the standards of international corpus documents and systematically classify the corpus to distinguish its name, issuing department, source website, issuing date and other important information for traceability, a unified corpus marking format is adopted to mark the corpus one by one. The uniform marking format is: < corpus name > < publication date > < corpus in Chinese or English (represented by C/E) > < publication department > < source website > < collector > < collection date >.

In addition, the parallel corpus of laws and regulations can also be composed of the following parts:

2.1 Legal Language Parallel Corpus and Legal Lexicology

Corpus-based legal lexicology focuses on the frequency, context and proximity of a word. For example, frequency is one of the most important stylistic features. In legal English, a large number of professional terms, archaic words, loanwords, words juxtaposed with instructions and mandatory are used to reflect the formal, serious and stuffy stylistic characteristics of law. Corpus can carry out collocation word statistics and word pattern statistics of these words by indexing tool. It provides a reliable quantitative basis for studying the characteristics of legal vocabulary. Because the corpus in corpus provides a large amount of information related to the original text, such as author, era, genre, word category code, etc., it is easy to determine the usage of a word or phrase in a specific genre or variant.

2.2 Legal Language Parallel Corpus and Legal Semantics

Legal semantics is the basis of the research on the characteristics of legal language, and its research is closest to the research on the ontology of legal language. Legal semantics studies the meaning and change of legal words, especially the causes and laws of the change of words from the social and historical perspectives, such as the enlargement and contraction of meaning, the ascending and descending of meaning, the transfer of meaning, and the relationship between semantic and syntactic structures. Using corpus, using machine automatic analysis and artificial analysis to study the connection mode between legal vocabulary and grammatical structure, that is, the relationship between legal vocabulary and their grammatical environment, and the structure between grammatical structure and their legal vocabulary, and thus distinguish synonyms or synonyms of grammatical structure. Some specific words in corpus can also be assigned to analyze the fuzziness or redundancy of English and Chinese legal words.

2.3 Legal Language Parallel Corpus and Legal Sentence Jurisprudence

Material statistics shows that grammatical features in the different language domain system distribution characteristics, so the use of corpus, can be in different level study of legal language syntax, including morphological characteristics, language classes and syntactic structure, namely in parallel with the syntactic tagging corpus database research various part-of-speech tags portfolio model and the quantitative analysis of all kinds of sentence patterns. Legal language researchers can in corpus investigation of legislative language, judicial language and other legal structure of the frequency distribution, investigation of syntactic structure and the relationship between language other analysis level, and the relationship between the language and non-language factors, in the empirical way to syntactic resource usage pattern of qualitative quantitative analysis. It can also be specific to the tense and voice of verbs, modal verbs or grammatical features of clauses.

2.4 Legal Language Parallel Corpus and Analysis of Legal Texts

The aim of discourse analysis is to describe linguistic features beyond the scope of sentences. Determining and analyzing the features of legal discourse is much more difficult than determining and analyzing the lexical and syntactic features, but it is very important to describe legal language because many lexical and syntactic features can only be perfectly explained through their functions in the discourse. In the past, text analysis generally used text fragments as examples, but this kind of analysis generally did not use quantitative analysis method, nor did it study the similarities and differences of text features between different texts, different texts and different registers. The corpus approach can process a large number of texts, accurately describe the discourse features of the selected register and the extent to which the text conforms to the general discourse pattern of its register, which makes up for the shortcomings of previous discourse analysis. The parallel legal corpus can be used to analyze the three variables of legal language: language field, tenor and pattern from the perspective of sociology, starting from register theory and using interactive computer technology.

2.5 Legal Language Parallel Corpus and Legal Translation

By using parallel corpus, legal corpus can be annotated with code, and many hyper-linguistic information can be annotated, such as the translator's situation (including translator's name, gender, nationality, occupation, translation direction, etc.), translation method, translation type, source language, source book situation, publishing house, etc. All of these are important information for the investigation of legal translation and translation style/style, because the translator's choice of the type of translation, the choice of translation strategy, as well as his expression in the preface, postscript and notes, may reveal his translation motivation, style or orientation. In addition, the corpus, we can also carry out analysis on more than a translator's translation of a more translator's translation of the original analysis, analysis of men and women translators translation, more can be based on a lot of corpora in language expression of the translator's personal preference form (e.g., parts of speech/tag ratio, sentence length, word frequency and sentence patterns, tie-in way, narrative structure, etc.) analysis, Find more convincing stylistic/stylistic representations in translation.

3 TERM EXTRACTION ASSISTED BY CAD VIRTUAL REALITY TECHNOLOGY

The algorithm of automatic term dictionary extraction from English-Chinese parallel corpus mainly includes three parts: bilingual sentence alignment; Part-of-speech tagging of English corpus and segmentation and part-of-speech tagging of Chinese corpus; Glossary extraction.

Based on the co-occurrence information of bilingual core terms in bilingual parallel corpus (title alignment, abstract alignment), the correlation of bilingual core terms is calculated to obtain the alignment information of bilingual core terms. After obtaining the aligned corpus, we performed part-of-speech tagging for the English corpus and part-of-speech tagging for the Chinese corpus. Our goal is to extract a glossary of terms from the aligned and annotated corpus, so we first select the words and phrases in the English corpus that are likely to be terms. After observing the corpus, we select suitable part-of-speech patterns for the candidates of English terms. Keywords extraction methods can be divided into four categories, namely, statistics-based methods, this method does not need complex training process, simple and easy; The method based on linguistics mainly improves the quality of keyword extraction from lexical analysis, syntactic analysis, semantic analysis and discourse analysis. Based on machine learning method, statistical parameters are obtained by training the training data, and keywords are extracted from samples: mixed method, that is, the comprehensive application of the above methods or the integration of some heuristic knowledge. This paper adopts machine learning method based on maximum entropy classifier to extract keywords automatically. Keyword extraction based on maximum entropy classifier is regarded as a classification problem.

The maximum entropy model was proposed by E. Jaynes in 1957, which is the basis of the maximum entropy classifier adopted in this paper. The basic idea is to model all known factors, while ignoring all unknown factors. In short, the goal of the maximum entropy model is to obtain a probability distribution that satisfies all known facts simultaneously and is not affected by any other unknown factors through calculation.

Since the maximum entropy model does not depend on conditional independent features, the selection of features with good classification effect can be carried out arbitrarily and the influence of these selected features on each other can be ignored. In addition, compared with SVM and other classifiers based on spatial distance, the maximum entropy model is much easier to model multi-class classification problems, and each category of each event can obtain a more reliable probability value. Combined with the advantages of high training efficiency of maximum entropy, it has been successfully applied in many natural language processing fields such as network information extraction and syntax parsing. When modeling the maximum entropy model, it is not necessary to care how to use each selected feature. Moreover, it is relatively flexible in feature selection and very convenient to replace. Therefore, the selection of appropriate feature terms to describe the natural language features of Chinese and English bilingualism is the key to practical application.

4 THE SIMULATION RESULTS

In this paper, the corpus of legal professional literature is collected as the training set, and the academic papers in legal professional literature are used as the corpus to calculate terminology degree and extract and align bilingual core terms. The Chinese data part (including Chinese title, Chinese abstract and Chinese keywords) constitutes the corpus of Chinese professional field, which contains more than 460,000 Chinese bibliography information records. Similarly, the corpus of English and Chinese professional field is obtained, which contains more than 130,000 English bibliography records. The corpus of Chinese and English professional fields contains 23 categories, and the average number of documents contained in each category is 17,638. The specific distribution is shown in Figure 2. It can be seen from Figure 2 that classified corpora are disequilibrium corpora. For example, economic category (identified as F) and culture, science, education and sports category (identified as G) have a high proportion, and the sum of the two categories accounts for about 50% of the whole corpus, while individual categories contain few documents. Such as comprehensive (expressed as Z) acquisition of expertise the category of the classification of parallel corpora distribution as shown in Figure 3, parallel record each category contains an average of 4733, alignment is: the title in both English and Chinese to sentence alignment, abstract in English and Chinese to paragraph alignment, keywords list in both English and Chinese can be further processed to word alignment. In this paper, keywords are supplemented or re-selected for each document by keyword extraction technology, and these keywords are taken as candidate terms.

Using the method of term degree calculation, we calculate the term degree of Chinese and English candidate terms respectively from the corpus of Chinese and English professional field classification, and finally get the core words of each field after certain threshold control and manual inspection. According to the parallel classification corpus of professional field, the core terms of each field in Chinese and English are finally obtained through the bilingual core term alignment technology. Manually check the Core terms in English and Chinese in 26 professional fields of top-10top-50top-100top-200top-500, and calculate the correct rate of core terms alignment. Figure 4 shows the distribution of alignment accuracy of bilingual core terms. As can be seen from Figure 4, the average accuracy of the top 10 pairs of core terms in each field is 70%, and the average accuracy of the top 200 pairs of Chinese and English core terms is more than 50%, and the average accuracy of the top 500 pairs is about 45%. Among the 26 categories, the term alignment accuracy of category F is higher, and the average accuracy of the top 200 pairs of core terms is 81%, including GB, I, etc., while the accuracy of category VZ is relatively low.

Comparing Figure 4 with Figure 2 and Figure 3, it can be seen that the number of training samples in the field has a great impact on the alignment of bilingual core terms in the field. If the number of samples is higher, the alignment of bilingual core terms is higher, and vice versa.

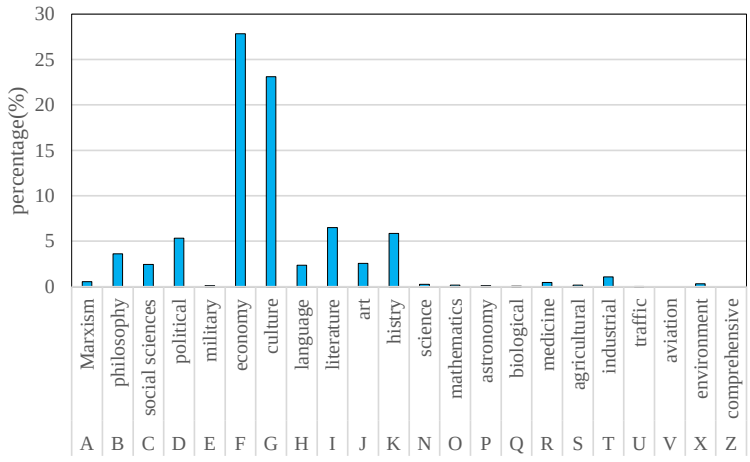

Figure 2: The specific distribution.

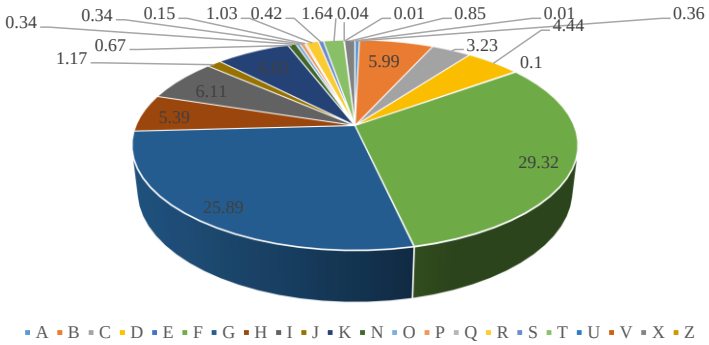

Figure 3: The classification of parallel corpora distribution.

Through this article set up the corpus of the existing parallel sentences to translate words extraction experiment, the matching rate of all the words of the sum of the average get matching rate on average, only through the existing vocabulary, the mutual translation of English-Chinese dictionary to extract 28.4% of 57.8% can be obtained to translate words, on this basis, got 62.3% of the results, different strategies to use, Finally can get 71.5% as a result, the visible, based on existing synonyms and near synonyms expansion of English-Chinese translation dictionary entry, can make full use of English has a large number of Chinese loanwords the characteristics, of the makes the translation lexicon effect obtained significantly improve, can more objectively reflect the characteristics of the bilingual aligned parallel to the words of vocabulary translation, concrete is shown in figure 5.

Using the above strategies, it can be concluded that the proportion of the number of words translated from Chinese to English to the total number of Words in English sentences is 0.7145, which is a very large proportion compared with the Chinese and English sentence pairs.

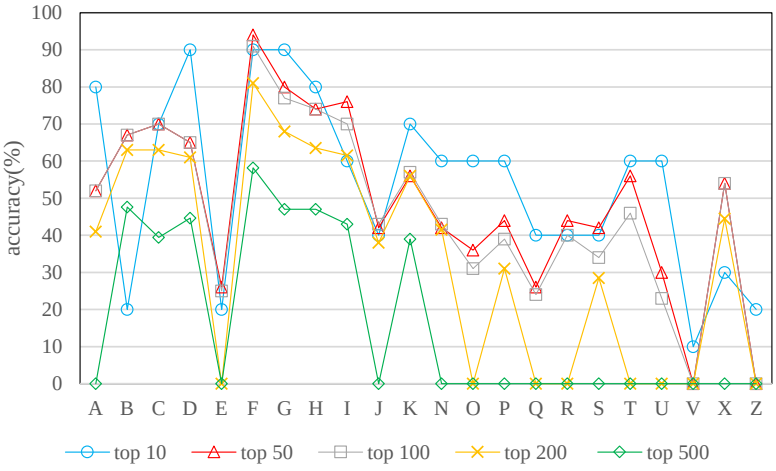

Figure 4: The distribution of alignment accuracy of bilingual core terms

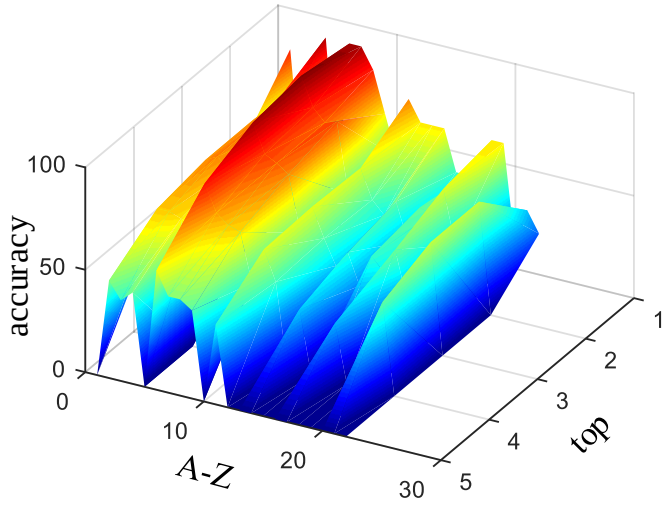

Figure 5: The accuracy.

The very high average ratio of more than 70% of the translated words in Chinese and English bilingual sentences shows that the translated words in Chinese and English bilingual sentences are a very important feature of Chinese and English bilingual alignment, which better reflects the natural language characteristics of Chinese and English bilingual. English dictionary index in organizational structure and meaning of matching algorithm based on dictionary has led to the instability of the Sino-British alignment algorithm based on vocabulary, and inefficiency of word alignment algorithm must by reducing the number of comparing each word matching to overcome as much as possible, the alignment algorithm on the adaptability of the Chinese/English bilingual improvement is more direct and reliable way. Therefore, despite the shortcomings mentioned above, the experiment proves that the translation vocabulary can well reflect the characteristics of the translation of Chinese-English bilingual sentence pairs. This scheme has great feasibility and

needs further study. The predicted value is shown in Figure 6 and the corresponding evaluation is plotted in Figure 7.

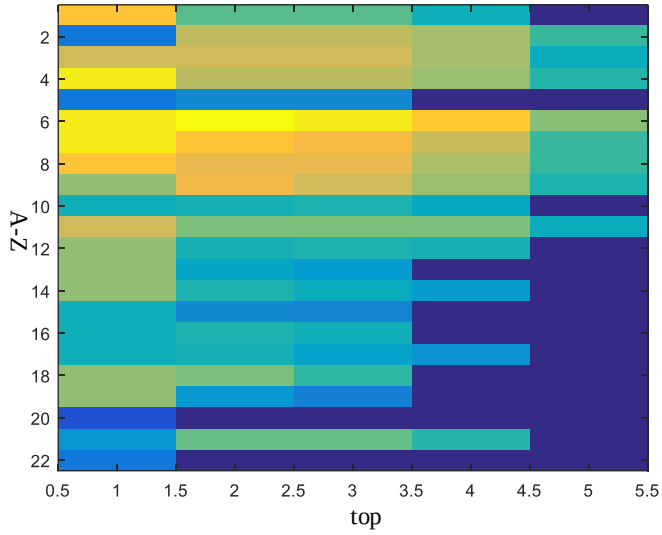

Figure 6: The predicted value.

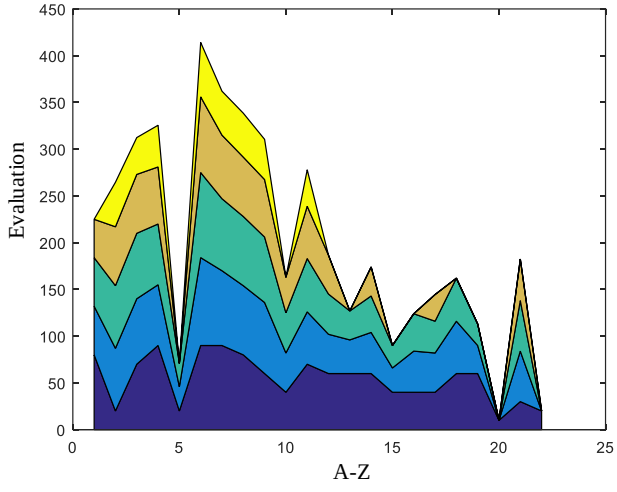

Figure 7: The evaluation.

5 CONCLUSION

In this paper, by constructing bilingual parallel corpus of Chinese laws and regulations, using computer software to extract terms, explore recorded the establish of corpus and calculation of auxiliary term extraction and the problems existing in the different phases of exporting the glossary in Chinese laws and regulations, the Chinese laws and regulations inconsistent terminology translation of terminology translation to discuss about the problem of classification, Some suggestions are put forward for the improvement of some term translation, and the general experience of corpus construction and term extraction is summarized, which can provide reference

for legal English, law and regulation translation, computer-aided term extraction based on parallel corpus of law and regulation, and term management in foreign public sector.

Qingguo Zhao, https://orcid.org/0000-0002-1242-1553
Jun Wang, https://orcid.org/0000-0001-9354-7598

REFERENCES

[1] Bednarek, M.; Carr, G.: Computer-assisted digital text analysis for journalism and communications research: introducing corpus linguistic techniques that do not require programming, Media International Australia, 181(1), 2021, 131-151. https://doi.org/10.1177/1329878X20947124
[2] Dmitrijev, A.; Kogan, M.: The Potential of Corpus Linguistics in Training Foreign Language Teachers Majoring in Computer Assisted Language Teaching, Society.Communication.Education, 10(38), 2020, 69-85. https://doi.org/10.18721/JHSS.10407
[3] Gong, W.-C.: An Innovative English Teaching System Based on Computer Aided Technology and Corpus Management, International Journal of Emerging Technologies in Learning, 14(14), 2019, 69-80. https://doi.org/10.3991/ijet.v14i14.10817
[4] Friedemann, V.; Hanjo, H.; Isabelle, G.: Computer - Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies, Law & Social Inquiry, 43(4), 2018, 1340-1363. https://doi.org/10.1111/lsi.12305
[5] Ammar, M.; Shamdeen, M.; Kasedeh, M.; Mansour, K.; Ammar, W.: Using Distance Measure based Classification in Automatic Extraction of Lungs Cancer Nodules for Computer Aided Diagnosis, Signal & Image Processing: An International Journal, 12(3), 2021, 25-43. https://doi.org/10.5121/SIPIJ.2021.12303
[6] Shelmerdine, S.; Singh, M.; Norman, W.; Jones, R.; Sebire, N.; Arthurs, O.: Automated data extraction and report analysis in computer-aided radiology audit: practice implications from post-mortem paediatric imaging, Clinical Radiology, 74(9), 2019, 11-18. https://doi.org/10.1016/j.crad.2019.04.021
[7] Bednarek, M.: Invisible or high-risk: Computer-assisted discourse analysis of references to Aboriginal and Torres Strait Islander people(s) and issues in a newspaper corpus about diabetes, Plos one, 15(6), 2020, 1-22. https://doi.org/10.1371/journal.pone.0234486
[8] Zheng, J.; Fan, W.: Different processes for translating expressive versus informative texts? A computer-assisted study of professionals' English–Chinese translation, Digital Scholarship in the Humanities, 36(3), 2021, 782-793. http:// doi.org/10.1093/LLC/FQAA052
[9] Yan, K.: Research on Computer Aided Medical Translation Technology, Basic & Clinical Pharmacology & Toxicology, 127(1), 2020, 140-144. https://doi.org/10.1007/s12108-018-9384-2
[10] Imad, A.: Use and Evaluation of Computer-Aided Translation Tools (CAT) on the Word Level from the Perspective of Palestinian Translators and Translation Trainees, AWEJ for Translation & Literary Studies, 4(1), 2020, 111-130. https://dx.doi.org/10.24093/awejtls/vol4no1.9