Scholar Publishing Group
International Journal of Educational Curriculum Management and Research
https://doi.org/10.38007/IJECMR.2024.050115
ISSN 2790-0967 Vol. 5, Issue 1: 101-106

Research on the Construction of Low-Resource Parallel Corpus Based on Translation Plug-in Technology

Zulkar Iskander[1, a], Azragul Yusup[1, 2, b*]

[1]*Xinjiang Normal University, Urumqi, Xinjiang, China*
[2]*National Language Resouree Monitoring & Research Center of Minority Languages, Beijing, China*
[a]*2212442686@qq.com*, [b]*Azragul2010@126.com*
[*]*corresponding author*

Keywords: Parallel Corpus, Low-Resource, Translation Plug-in Technology

Abstract: Parallel corpora play a crucial role in the field of natural language processing, especially in tasks such as machine translation and cross-language information retrieval. However, with the increasing demand for more languages, the challenges are becoming increasingly apparent, especially for low-resource languages. The purpose of this paper is to summarize the current status of parallel corpus research, discuss the challenges faced by low-resource corpora, focus on plug-in-based translation techniques to automatically construct Chinese-Uyghur and Chinese-Kazakh parallel lexicons. And look forward to the future research direction.

1. Introduction

A low-resource parallel corpus is a situation where the number of corpora available for training machine translation or other natural language processing tasks is relatively small between certain language pairs. This situation may be due to a small number of language speakers, limited cultural influence, or insufficient research resources. Low-resource corpora are a challenge for many language processing tasks because they usually do not have access to large-scale data for training models compared to high-resource language pairs.

This paper focuses on computerized translation software as a breakthrough, targeting research to provide technical support for building computerized translation software and improving the translation quality of computerized translation software in the case of scarce language resources. After conducting adaptive research on domain translation, the quality of computer translation software is improved. The use of dynamic planning methods and plug-in technology to automatically obtain the bilingual dictionary of new words synchronized with the social demand can not only reduce the labor cost, but also quickly build a basic machine translation system to assist Chinese learning, this paper's research has greater academic significance and practical value.

2. Overview of Parallel Corpus Construction and Research

Parallel corpus construction and research is an active field in the international arena, and various countries and organizations are committed to collecting, organizing and utilizing parallel text data between multiple languages. Many foreign research institutes and universities have also carried out parallel corpus construction and research. These projects usually involve multiple fields, such as computational linguistics, machine translation, natural language processing, etc., and aim to utilize parallel corpora to promote the progress of related research. Some open source communities and data sharing platforms such as OpenSubtitles, OPUS, etc. also provide researchers with a large number of open-access parallel corpus resources, which cover multiple language pairs and are available in reusable formats. The European Commission supports a number of cross-language research projects aimed at advancing language exchange and cross-language technologies between European countries. These include a number of large parallel corpus projects such as Europarl, JRC-Acquis, DGT-TM, etc. These projects cover a wide range of European languages and provide important resources for machine translation and other cross-linguistic tasks. UN text data is often used to build parallel corpora, which cover multiple languages, including the official languages of the UN as well as other official working languages. UN text parallel corpora play an important role in tasks such as cross-language information retrieval, machine translation, etc. The Paracrawl project, a collaborative effort between language technology research teams from various European countries, aims to collect and collate massively parallel corpora on a global scale. The project utilizes publicly available parallel text data from the Internet and uses automated tools to crawl and process the text in order to build high-quality cross-linguistic parallel corpora.

The construction and research of parallel corpora are also developing in China, and the government attaches great importance to the construction and research of language resources, providing financial support through the National Natural Science Foundation of China, the National Social Science Foundation and other channels to promote the construction and research of parallel corpora. In addition, institutions such as the National Language Resources Monitoring and Research Center (NLRMC) have carried out a series of language resources collection, collation and research work. Domestic universities, research institutions and enterprises have carried out numerous parallel corpus construction and research projects involving machine translation, information retrieval, text mining and other fields. These projects are usually multidisciplinary and cross-field cooperation, which jointly promote the construction of parallel corpora with the help of resources and technical advantages of all parties. Domestic researchers and institutions actively utilize open data and resources on the Internet, such as news reports, online texts, social media content, etc., for the construction of parallel corpora. Through web crawlers and data mining techniques, a large amount of parallel text data is collected and organized to provide support for research and application. In addition to basic research, domestic parallel corpus construction and research are increasingly focusing on application orientation. For machine translation, text alignment, information retrieval and other practical application scenarios, the construction, optimization and application of parallel corpora are carried out to promote the landing and application of related technologies in practical scenarios.

The construction and research of parallel corpus for low-resource languages is an important topic in the field of natural language processing, especially focusing on those languages that lack large-scale data support. Parallel corpus construction for low-resource languages usually faces the problem of data scarcity. Researchers collect data through a variety of ways, including books, news reports, social media, government documents, and so on. In addition, language volunteers, research partners, or automated crawling techniques can be utilized to obtain data. For low-resource languages, researchers often rely on automated methods and technologies to reduce labor costs and

increase efficiency. Data quality is critical for parallel corpora of low-resource languages. Researchers need to address issues such as uneven data quality and misalignment. Sometimes, manual annotation is also required to improve data quality, especially when high quality data is needed for a specific domain or task. Therefore, this paper proposes a low-resource parallel corpus construction method based on translation plug-in technology on the basis of existing research on building bilingual corpora.

3. Research on the Construction of Chinese-Uyghur and Chinese-Kazakh Parallel Corpus

3.1 The realization of Chinese news webpage content crawling

In this study, we take the governmental websites such as Tianshan.com (http://www.ts.cn/) as an example for web content crawling, which contains several modules such as economics news and political news. By looking at the web page source code of the target website, we understand the structure and layout of the web page, and determine where the news content is located and the corresponding HTML tags. The news content includes title, time, and body content. Take a news in TIANSHAN.com news (the Standing Committee of the Political Bureau of the CPC Central Committee held a meeting chaired by Xi Jinping) as an example, there will be a lot of content in its source code, and we only need to crawl the content in the body tag, and filter out the content of some of its tags.

According to the webpage structure and requirements, choose the appropriate crawler tool. Common crawler tools include Python's Scrapy framework, Beautiful Soup library, Selenium and so on. All these tools provide powerful web parsing and crawling functions, and you can choose the right tool according to your needs. The news content crawled is usually raw HTML text or structured data. Data cleaning, de-duplication, storage and other processing can be carried out according to the needs, and the data can be organized into the format they need, which is convenient for subsequent analysis and application. In this study, the raw news corpus is obtained by crawling with Python crawler.

3.2 Data Processing

In this study, government websites such as Tianshan.com (http://www.ts.cn/) are used as an example for web content crawling, and these raw corpus will be mixed with a small amount of repetitive and useless garbage corpus, which can't be used directly or cause serious errors to the experimental results, which requires preprocessing of the corpus at a later stage. Regular expressions are used to match the content field data such as publication time, article source, number of comments, article title, comment address, body content, etc., which we want to crawl, and then saved in different classes.

Mainly crawl the news section of the website, which contains economics, politics, documents, society, technology, science, education, culture, sports, tourism, ecology, health care and other news categories, this paper will be in accordance with the class and id for news content crawling, and processing and analysis.

3.3 Chinese-Uyghur Parallel Corpus Construction Based on Translation Plugin

The processed Chinese text is segmented, split into sentences and words. For text segmentation, commonly used texts have obvious paragraph symbols, and Chinese text is no exception, and when crawling Chinese web content, it is usually crawling with paragraph as the basic unit, which also greatly reduces the burden of the later work; for text clauses, the commonly used end-of-sentence

terminators are usually used for clauses, and the commonly used end-of-sentence terminators in Chinese are ". " , "?" ", "?", "!" Uyghur participles are relatively simple compared to Chinese participles, and like English, there is a space between words in each sentence, which can also be utilized in the implementation of the participle algorithm.

Through the above steps, we have accomplished an orderly hierarchical division from the original text to paragraphs, then from paragraphs to statements, and finally from statements to words, and constructed the Chinese paragraph corpus, sentence corpus and word corpus respectively.

The process of constructing aligned bilingual thesaurus: read from the Chinese vocabulary list sequentially; call the search engine module with the program to search for the corresponding API translation plug-in provided free of charge on the Internet to align and translate into Uyghur; write the translation result into the vocabulary list and the position of the translated content to construct the Chinese-Uyghur language aligned corpus. The next step is to utilize the above mentioned to build a Chinese-Uyghur bilingual alignment corpus quickly, accurately and automatically.

3.4 Construction of Chinese-Kazakh parallel corpus based on translation plugin

The process of constructing a Chinese-Kazakh aligned bilingual corpus is similar to that of Chinese-Uyghur aligned bilingual corpus, but there is one more step of text conversion in Chinese-Kazakh machine translation.

Currently, Kazakh scripts are divided into two main categories, one of which is Slavic Kazakh script based on the Slavic alphabet. Machine translation of Slavic Kazakh script into other languages has been particularly well researched, especially by the Google Translate project. The other is the Arabic alphabet-based Kazakh script used in China. So it is necessary to translate Chinese into Slavic Kazakh and the translation of Chinese and Chinese Kazakh scripts is achieved by swapping letters between Chinese Kazakh and Slavic Kazakh scripts.

4. The Traditional Alignment Corpus Construction Method Comparison Analysis

Compared with the traditional popular alignment corpus construction methods, the biggest difference lies in the construction of different ideas, the traditional methods are usually used to directly obtain the content of the web page containing bilingual corpus, for example, the more common Chinese-English bilingual alignment corpus construction, there will be a large number of websites for learning English on the Internet, which can provide a large number of bilingual corpus of Chinese-English, and then with the relevant alignment technology methods to obtain the bilingual alignment corpus of paragraphs, phrases and words. Then, the bilingual aligned corpus of paragraphs, statements and words will be obtained by using the relevant alignment techniques.

There are a lot of resources on the Internet for bilingual corpora between international popular languages, such as the construction of Chinese-English bilingual alignment corpora mentioned above. There are a lot of English learning websites on the Internet with a large number of bilingual texts, which can be used to crawl a large number of bilingual corpora through crawler technology, and then obtain the alignment corpora through the alignment technology. However, this method is only applicable to the situation where there is a large-scale bilingual corpus in this kind of website, and there are very few bilingual corpus resources for Chinese-Uyghur, so it is difficult to obtain a large number of bilingual corpus by this traditional method, but it is possible to find a large number of Chinese corpus, which can be obtained from the website, and then generate the alignment corpus by the translation plug-in after preprocessing, which is a more suitable method for this kind of non-popular language alignment corpus. This method is more suitable for the construction of alignment corpus of this non-popular language.

5. Conclusion

In the study of low-resource parallel corpus construction, we have explored a series of methods and strategies. By analyzing the structure of web pages, using crawler tools, processing data, and utilizing machine translation models, we have constructed a rich parallel corpus, which provides important data support for the training and optimization of machine translation systems. However, we also need to realize that constructing a parallel corpus is a complex and arduous task that faces many challenges, such as data scarcity, alignment difficulties, and quality control. In future research, we need to keep exploring new methods and techniques to overcome these challenges and improve the quality and coverage of parallel corpora.

Acknowledgment

This work was supported in part by Xinjiang Uygur Autonomous Region Innovation Environment (Talents and Bases)

Construction Special Project-Natural Science Programme (Special Cultivation of Ethnic Minority Scientific and Technological Talents) Project (2022D03001). Xinjiang Normal University Young Top Talents Project; National Natural Science Foundation of China (61662081); National Social Science Foundation of China (14AZD11);

References

[1] Mao H, Yusup A, Ge Y, et al. Named entity recognition in Chinese e-commerce domain based on multi-head attention[C]//2022 9th International Conference on Dependable Systems and Their Applications (DSA). IEEE, 2022: 576-580.

[2] Wang Q, Li X. Chinese News Title Classification Model Based on ERNIE-TextRCNN[C]//Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing. 2022: 147-151.

[3] Liu Z. Development of advertising art design based on information technology[C]//International Conference on Cognitive based Information Processing and Applications (CIPA 2021) Volume 2. Springer Singapore, 2022: 3-10.

[4] Wang C, Zeng Q, Huang J. Practical Research on the Construction of Chinese-Yi-English Trilingual Parallel Corpus[J]. Lecture Notes on Language and Literature, 2023, 6(13): 71-77.

[5] Jing M. The Construction of a Multilingual Parallel Corpus for Hnewo Teyy[J]. Lecture Notes on Language and Literature, 2023, 6(16): 1-8.

[6] Lan Caiyu. Design and construction of a Chinese-English bilingual parallel corpus for Chinese medicine. Asia-Pacific Traditional Medicine,2014,10(08):1-3.

[7] Fang Lu. Research on the construction and application of English-Chinese comparable corpus. Suzhou University,2011.

[8] Niu Yitong. Research on Word Alignment Method Based on Chinese-Vietnamese Bilingual Parallel Corpus, Master's thesis of Kunming University of Science and Technology, 2017.4, Supervisor: jianmei Guo.

[9] Xue Yan. Research on Chinese-Mongolian word alignment and related technologies, Master's thesis, Inner Mongolia University, 2009.6, Supervisor: Nashun Urietu.

[10] Yusuf. Aibaidulla et al, "Contextual correlation processing in a syntactic analyzer for Viennese-centered language-driven grammars", Computer Applications and Software , 1999/6.

[11] Yusuf Aibaidulla et al, Determination of a tag set for Viennese lexical annotation for information processing, Computer Applications 2009/7

[12] Dong Meiping, Research on Automatic Acquisition and Domain Adaptation of Weihan-Chinese Machine Translation Corpus, Master's Thesis, Tsinghua University, Supervisor: Liu Yang, 2015-06.

[13] Caijenga. Research on large-scale Chinese-Tibetan (Tibetan-Chinese) bilingual corpus construction technology for natural language processing. Journal of Chinese Information,2011,25(06):157-161.

[14] Rexidan Tayi, Türgen Ibrahim. A study on sentence alignment method based on dictionary translation in Chinese-Uyghur bilingual corpus. Journal of Xinjiang University(Natural Science Edition),2009,26(03):359-363.

[15] XU Xiongfei. Research on automatic extraction of word alignment in Greater China . Jiangxi Normal University,2016.

[16] Dong Meiping. Research on the construction method of translation corpus based on non-parallel data . Tsinghua University,2015.

[17] Nurguli Aizimu put. Research on Chinese-Kazakh bilingual parallel corpus alignment method . Northeast Normal University,2013.

[18] Asimu Tohti. Research on Uyghur-Uzbek machine translation . Xinjiang University,2017.