Research on several construction strategies of multilingual resource database

Lingchun Meng[*], Shuying Kong, Wenyue Li, Lei Cui
Center of Network and Information, Guangdong University of Foreign Studies, Guangzhou, China

ABSTRACT

In order to realize the effective use of information resources in different languages, it is necessary to establish a corresponding information resource database. Based on cross-language retrieval, this paper analyses the construction strategy of multilingual resource database, and takes language acquisition and processing as the perspective. The construction of the resource database provides a reference. In view of the construction of multilingual resource database, this paper makes an analysis from three aspects: resource collection and pre-processing, resource storage and retrieval, and multilingual translation technology, and summarizes its implementation methods. The collection, storage and retrieval of multilingual resources should adopt a distributed architecture. The resource library builds a language encoding and transcoding system for efficient collection and preprocessing of multiple languages. The translation between languages adopts machine translation technology mainly based on neural network machine translation.

Keywords: Multilingual, resource database, pre-processing, retrieve, machine translation

1. INTRODUCTION

At present, China is already the second largest economy, and the political, economic and cultural exchanges with various regions of the world are becoming more and more frequent. With the advancement of the “One Belt, One Road” and “Internet +” national strategies, the dissemination of international information tends to be multilingual. Since “the Belt and Road Initiative” was proposed, the four-beam and eight-column frame has basically taken shape, and the detailed description of its fine brushwork requires an accurate grasp of the national conditions and public opinion of the countries along the route. As the construction of “the Belt and Road” moves from a freehand framework to a detailed description of meticulous brushwork, the differences and dislocations in the cultures of the countries along the route, the diversification of opinions, and the differences in interest demands will be highlighted[1].

There are more than 1,000 classifications of world languages, and there are more than 50 languages in the countries along “the Belt and Road”. Since each language covers different political, economic, cultural and other information, the ethnic customs of different countries are very different, and the occurrence, presentation, Content, subject, carrier, etc. are different from ordinary information, and the importance of relevant information acquisition and analysis is self-evident.

In the current increasingly complex international situation, timely and accurate access to the required network information lays the foundation for the next step of research on public opinion hotspots, information monitoring, and intelligence content analysis. Mining information in different languages and establishing information mapping between multiple languages is conducive to the integration and development of information resources in multiple languages. The multi-source and timeliness of information resources have brought new challenges. Using the database construction of multi-language information resources to realize the knowledge of ourselves and others is a powerful guarantee to promote cooperation and exchanges.

2. RELATED RESEARCH RESULTS

In the face of various data types, different data sources, and different structures in the era of big data, efficient data collection and storage is the basic work of data information. Using various data mining and data analysis methods can effectively integrate valuable information resources. It has always been the goal of resource database construction, and it is also an important basic work in related disciplines. Therefore, it has always been a research hotspot to meet the actual

[*] mlc@gdufs.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125064A
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2661784

needs of each user from different fields and to establish a resource database covering all aspects of information needs in multiple languages.

Sun et al. constructed the standard and normative system of information organization of China's multilingual public digital cultural service platform from three aspects: object data standard, metadata standard and knowledge organization standard[2]. Academician Wushouer Slamu focused on the impact of the combination of emerging technologies and multilingual intelligent information processing on education, the characteristics and advantages of national language informatization in the construction of "the Belt and Road", and the integration of production, education and research in multilingual intelligent information processing. Many valuable suggestions have been put forward[3].

Si et al. selected more than 30 kinds of databases as the analysis objects, and conducted systematic analysis from the dimensions of resource organization system, resource type and content, database construction method, and multi-language processing problems. They proposed to adopt a multi-dimensional resource organization method and focus on the depth of resource content processing, strengthening the research on domestic and foreign cooperation mechanisms, and attaching importance to the collection of non-universal languages resources and cross-language retrieval research, etc., to provide a reference for the construction of the "One Belt, One Road" thematic database[4]. Strategies for the design and development of cross-language retrieval functions for shared databases: A question-and—document translation method based on neural network machine translation should be used to realize multiple retrieval functions, use visualization technology to present retrieval results, and provide multilingual retrieval interfaces and resources[5].

On the basis of research and analysis on the construction status, advantageous resources, and construction framework of China's "One Belt, One Road" Network thematic databases, Yan and others summarized its main characteristics and development strategies[6].

Wang et al. established a general multilingual data collection and cleaning platform, obtained text data in nine languages including Italian and Russian, and studied language coding analysis and cleaning technology. Zhang established an Internet-oriented bilingual corpus collection system based on Hadoop distributed computing platform[7].

At present, there are relatively few resource databases focusing on multiple languages, especially non-universal languages. Looking at the development of information resource platforms, the construction of fragmented information resource databases focusing on different countries, languages, and topics needs to continue to advance.

3. SEVERAL CONSTRUCTION STRATEGIES

3.1 Resource collection and pre-processing

Multi-source heterogeneity is one of the main characteristics of massive resources. Information from different sources and structures has obvious differences in content and presentation. One is to solve the coding and language of information resources from different sources and different languages, and the other is to solve the automatic acquisition of a large number of domestic and foreign media website data and social media data[8]. This is the primary problem facing the system. In the data collection stage, in order to meet the needs of different numbers of languages and large-scale collection, the constructed collection platform includes corresponding language coding analysis, format determination, language material applicability determination, rough text data extraction, corpus deduplication, etc.

The collection of resources should adopt a distributed technology architecture, uniformly schedule all collection tasks in an efficient and scalable way, and use load balancing technology to obtain information resources. The collection platform should make full use of each download resource to monitor the status and distribute the download tasks in a timely manner. The platform adopts incremental collection technology to filter invalid information and obtain new resources in an efficient and timely manner.

Since different languages need to be collected, the codes of these languages are different, and there are situations where different languages are mixed, so language coding analysis is required. The system should support multiple languages and multiple encodings, and can automatically perform encoding conversion. It can perform multi-layer acquisition of target information sources, and collect target resources layer by layer by specifying the number of layers of the source to be captured.

For the real-time monitoring and automatic collection of the content of the target source, the update frequency of the source should be analysed, the collection interval should be adjusted automatically, and the adaptive collection strategy should be adopted. Supports real-time collection of streaming data using technologies such as FlumeNG and Kafka. The concept of list page and content page is introduced[9], and the collection task is automatically adjusted according to the update frequency of the list page and the priority of information sources, so as to control the overall load and achieve timely and efficient data collection.

The acquisition program passes the application protocol authentication, calls the API interface provided by the source, and returns the corresponding format file according to the set analysis. Each source provides a different number of API calling interfaces, and the program implements multi-threaded programming to iteratively obtain the desired source data[10], as shown in Figure 1.

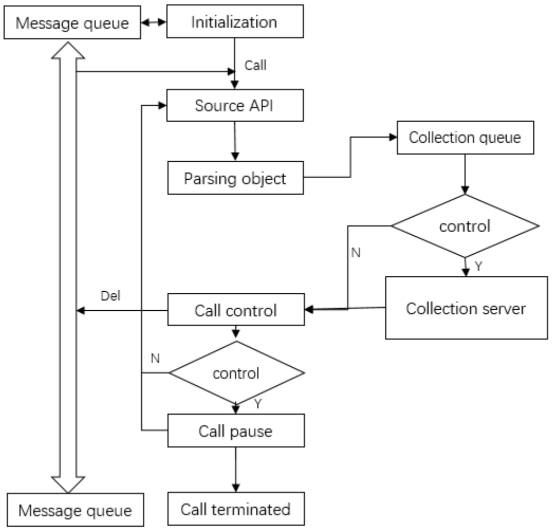

Figure 1. Source collection program flow.

Efficient information filtering and other pre-processing technologies enable the intelligent analysis system to process massive data quickly and in real time. Sensitive features indicative of spam can be extracted with high quality. Based on the optimization and improvement strategy of traditional Naive Bayes in spam text filtering, it can make up for the single limitation of the data set in the sample data, and also propose an improvement strategy of feature weighting in the feature classification strategy[11]. The system model has fast self-updating ability, thus ensuring the continuity of the good performance of the system.

The pre-processing process of collecting resources in multiple languages is to represent them through text, perform grammatical, semantic and syntactic analysis on them, and convert them into structured information that can be recognized by the database. However, due to differences in writing, morphological, grammar, and syntax between different languages, as well as the immaturity of non-universal language information processing technologies and tools, the common language natural language processing technologies and tools are not fully applicable to non-universal languages[8]. Improving the natural language processing technology of non-universal languages is an important way to improve the quality of pre-processing.

At present, some languages can be segmented by using the rule-based word segmentation method. For example, for Japanese, Korean, English, Thai, Russian, etc., using word segmentation models such as IK Analyzer and Me Cab can complete the basic word type division, and can achieve a certain degree of part-of-speech parsing. Based on the ClassicAnalyzer word segmentation model, it can perform word segmentation for languages such as German, Spanish, Italian, and French whose grammatical structure is a comprehensive language, and perform multilingual text analysis. On the basis of the maximum matching algorithm, the combined ambiguity detection and cross-ambiguity detection and full segmentation algorithm are used in text segmentation, which improves the accuracy of text rough segmentation and reduces the size of the rough segmentation result set[12]. This lays the foundation for further correct word segmentation.

3.2 Resource storage and retrieval

The distributed storage system is used to manage massive data, and the unstructured files collected from the source resources are stored to realize the storage and reading and writing of massive data. Based on Hadoop, HDFS distributed storage database is adopted, and through column storage technology, nodes are added from bottom to top in a linear manner for seamless capacity expansion and expansion. It has unlimited horizontal expansion capabilities and provides more visual management tools. It supports MapReduce technology, and the data table is stored on the server cluster, providing massive data storage, so as to meet the characteristics of high efficiency and scalability. HDFS is the foundation of the entire Hadoop distributed framework. The system supports massive data storage, can provide high-throughput data access, has high fault tolerance, and the node downtime does not affect the operation of the overall system[7].

Based on a distributed full-text retrieval system, it integrates technologies such as artificial intelligence, information retrieval and data mining to build a full-text retrieval cluster. The system uses the index cluster to obtain source data, converts documents of various formats to extract text information, creates an index library and stores it on different retrieval servers, and merges the results returned by the retrieval server through the retrieval proxy server and presents it to the application layer.

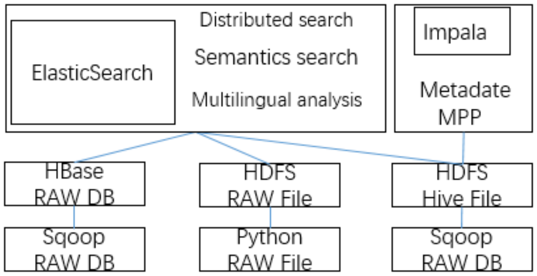

Figure 2. Multilingual intelligent storage and retrieval center.

Multilingual intelligent storage and retrieval uses Hadoop as a storage tool to build a massive and scalable data warehouse, including the original library composed of HBase, Hive, and HDFS. The ElasticSearch distributed full-text retrieval architecture is used for high-speed retrieval of a large number of indexes, including semantic analysis of tokenizers in multilingual big data storage, and conversion and query of unstructured files to generate structured indexes[13]. At the same time, it integrates the Impala standard data retrieval architecture, and uses Impala's MPP query architecture to perform high-speed queries on the data warehouse, as shown in Figure 2.

The retrieval system supports massive data, realizes the dynamic expansion of retrieval nodes, and supports multi-user high-concurrency and high-performance access. At the same time, it can flexibly allocate the data of each node, realize the load balance of each server in the cluster, and realize the reasonable allocation and scheduling of server hardware resources. The system should be loosely coupled with indexing and retrieval, and have high-reliability retrieval services to meet the retrieval requirements of retrieval engines with high availability and high concurrency.

The index organization mode of segment-by-layer merging is adopted to improve the indexing speed of massive data, so that the indexing time is linearly related to the data magnitude. The word hybrid index model is established, and the combination of word index and word index is adopted to solve the problems of low precision rate and low phrase retrieval efficiency of word index method, and avoid information loss.

An improved inverted table is introduced, and information such as term frequency, document frequency, inverse document frequency and document location are added to the inverted table. The index items are effectively compressed, and a secondary index is established to further improve the retrieval speed. This technology avoids that the common inverted table mechanism cannot reflect the correlation between documents and index items, and lacks sufficient information to obtain document information of phrases composed of multiple index items and document information of search terms with wildcards[14].

The application layer is designed with simple and convenient search methods and advanced Boolean search methods, which can accurately retrieve information. Combined with natural language understanding technology, the extended retrieval model provides phrase retrieval, range retrieval, wildcard retrieval, proximity query, synonym retrieval, homophone retrieval, etc.

In order to solve the problem of relevance ranking of query results, full-text retrieval should comprehensively consider the impact of item frequency and reverse document frequency on document weight, use the improved tf*idf weighting method to calculate the relevance[15], and use the improved quick sort to calculate the relevance. The two algorithms, Heap Sort and Heap Sort, provide results set return methods suitable for different application requirements.

3.3 Multilingual translation

The Internet contains a large number of resources in different languages, and various resources are timely and diverse, and the coverage of resources is wide and the number is large. Taking "the Belt and Road" as an example, there are many languages in the countries along the route, involving 65 countries and regions, and more than 50 major languages[6].

In the process of multilingual retrieval, the translation methods currently implemented include questioning translation method, document translation method, questioning-document translation method, intermediate language translation method, and non-translation method. Among them, questioning translation method, document translation method, questioning-document translation method is the current mainstream translation method[5]. The questioning translation method is to convert the searched language into the language of the document first, then perform the search of the language of the document, and display the original text of the document according to the corresponding search results. The method of document translation is to convert the original document into the searched language, but does not translate the searched language.

Based on the storage method and translation method of the current system, the retrieval method can learn from the questioning-document translation method that combines the two. First, translate the question form in the source language into the source language form consistent with the documents to be retrieved, perform single-language retrieval, and then translate the retrieval results in whole or in part into information described in the source language. This method is currently ideal for cross-language retrieval[5].

Machine translation technology uses computers to convert the source language into the target language. The development process has roughly gone through three stages, including the rule-based machine translation, the statistical machine translation model and the neural machine translation based on deep learning.

Among them, rule-based machine translation relies on the rules between languages, and has high requirements on the completeness of the summarized language rules. The statistical-based machine translation model translates according to the mathematical model of large-scale corpus to learn the conversion between languages, which requires relatively high modelling quality. In recent years, the translation performance of neural machine translation based on deep learning has improved significantly, and some progress has been made in model research, vocabulary limited research, and resource limited research[16]. However, for non-universal languages with scarce resources, the current quality of machine translation is not ideal, and it is necessary to establish large-scale, high-quality non-universal language dictionaries, bilingual parallel corpora, and semantic knowledge bases[8].

The multilingual translation neural translation system obtains the smallest unit acceptable to the model through cleaning, word segmentation, standardization, and word segmentation, and solves the problem of a large proportion of English words in multilingual vocabularies by optimizing the BPE word segmentation technology. The model building module adopts an end-to-end generative framework, which is divided into an encoding end and a decoding end as a whole. The self-attention method is used for feature extraction. At the same time, the model network is optimized by adding language vectors to solve the problem of language confusion in the model. The user interaction module adopts in the bridging method, the two multilingual models are directly connected to the English side to create a multilingual translation front-end interactive interface[17]. The system is shown in Figure 3.

By adding a network layer to distinguish languages, the system designs a suitable data processing process, builds a model training process and a front-end microservice display process. The system optimizes the multilingual data word segmentation algorithm at the data level in a targeted manner, and optimizes the discrimination of different languages in the hidden layer space at the model level. Finally, the multilingual translation system is comparable to the monolingual translation model in terms of accuracy. For most common languages, the current translation technology can learn from Google, Sogou Overseas Search, AIpatent and other machine translation technologies based on neural network machine translation. As the mainstream technology of artificial intelligence translation, it is currently applied to multilingual resource databases[18].

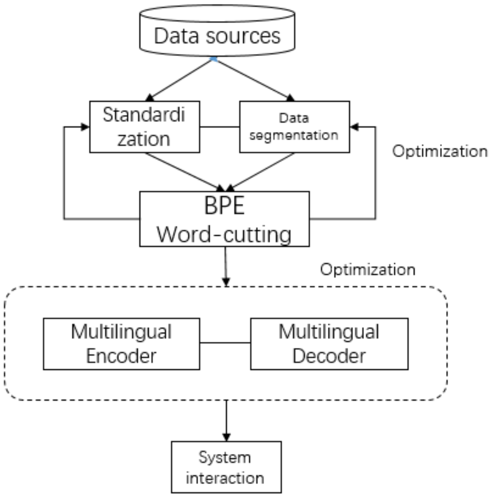

Figure 3. Multilingual interpretation neural machine model diagram.

4. SUMMARY

This paper discusses the construction ideas of multilingual resource database under the new background, and analyzes the challenges faced by resource database construction. The construction strategy of multilingual information resource database is proposed, the key issues are analyzed in detail, and the core work and key issues in the research are further clarified.

People communicate with each other in language first, and multilingual information shows different characteristics in terms of content, mode, route and area of dissemination. The use of big data, artificial intelligence and other technologies to develop and build a multilingual resource database can lay a solid foundation for in-depth analysis of information and mining of high-value public opinion information in the next step. It can not only effectively solve language communication barriers, but at the same time, with the support of the language service system, it can also help the promotion of national strategies such as the “Belt and Road”.

ACKNOWLEDGEMENTS

This work is supported by the Ministry of Education of Humanities and Social Science projects in China. (No. 20YJCZH124).

REFERENCES

[1] Qin, X. J. and Liu, H., Economic Research Guide, 451, 142(2020).
[2] Sun, G. Y. and Wu, D., Library Development, 313, 56(2022).
[3] Wei, X. F., Wang, Q., Zeng, H. J. and Shen, Y., China Educational Technology, 342, 25-30(2021).
[4] Si, L., Chen, C. and Zhou, J. Library and Information Service, 65, 4-10(2021).
[5] Si, L. and Zhou, J., Library and Information Service, 65, 20-27(2021).
[6] Yan, D. and Ma, Y. X., Research on Library Science, 407, 40-47(2017).
[7] Zhang, Z. C., [Large-Scale Bilingual Parallel Corpus Collection System Based on Hadoop], Harbin Institute of Technology, Harbin, Master’s Thesis pp 6-17(2013).
[8] Wang, L. X., Gan, S. F., Lin, N. K. and Jiang, S. Y., Journal of Intelligence, 39, 135(2020).
[9] Zheng, C., Wireless Internet Technology, 124, 110-2(2017).
[10] Kong, X. N. and Sun, H., Software Guide, 16, 187-9(2017).
[11] Peng, G., [Research on Spam Text Classification Based on Improved Naive Bayes Algorithm], Yangtze University, Jingzhou, Master’s Thesis, 2-15(2021).
[12] Li, G. H., Liu, G. S., Qin, B. B., Wu, W. J. and Li, H. Q., Computer Engineering and Applications, 753, 139(2012).

[13] Jin, G. D., Bian, H. Q., Chen, Y. G. and Du, X. Y., Journal of Software, 31, 137(2020).

[14] Duan, Y. J., [Research on Key Technologies of MicroBlog Search Research on Key Technologies of MicroBlog Search], University of Science and Technology of China, Hefei Doctor's Thesis, 2-12(2014).

[15] Xu, P. J., Li, X. F., Hui, Y. and Zhang, G. L., Journal of Jilin University (Science Education), 47, 791-4(2009).

[16] Zhang, H. Y., Liu, W. Y., Yan, H. and Wan, L., Agriculture of Henan, 569, 51-4(2021).

[17] Li, T. H., [The Design and Implementation of Neural Machine Translation System for Multilingual Mutual Translation], Beijing University of Posts and Telecommunications, Beijing, Master's Thesis, 2-9(2021).

[18] Lin, Q., Liu, Q., Su, J. S., Lin, H., Yang, J. and Luo, B., Journal of Chinese Information Processing, 33, 2-9(2019).