The Development and Construction of Bilingual Machine Translation Auxiliary Tool between Chinese and Kazakh Languages

Muheyat Niyazbek1,3, Kuenssaule Talp2*, Jing Sun1,3

1School of Information Science and Engineering Xinjiang University, 830046, Urumqi, China
2College of Chinese Medicine of Xinjiang Medical University, 830011, Urumqi, China
3Key Laboratory of Multilingual Information Technology, Xinjiang University, 830046, Urumqi, China

Abstract: This paper introduces the design and construction process of a bilingual machine translation auxiliary tool between Chinese and Kazakh languages. The tool uses the Jieba word segmentation tool to segment the input sentence, and then translates it according to the two language correspondence and rule-based dictionary method. This paper introduces the development and implementation process and overall framework of said auxiliary tool, and the role of some functional modules of it.

1. Introduction
Machine translation is an important field in natural language processing. At present, machine translation has been widely used and has achieved great development.

At the present stage, there is still a gap and demand market for machine translation technology between Chinese and Kazakh languages in China. The aim of this paper is to design and implement a machine translation assistant tool between Chinese and Kazakh languages for the convenient translations of both civilian users as well as researchers in related fields.

The design concept of this translation aid is to use the principle of electronic dictionary, together with the text retrieval method, to cut sentences into sections and then translate them at the scale of single words. This method can greatly improve the efficiency of individual term translation in machine translation. The user can obtain the translation between corresponding terms by finding the target word or sentence and performing the relevant operation steps.

2. System Design
When the user enters a character string, the system first determines whether the string is a long sentence or a single word, and if the character string is not a long sentence but a single word, the corresponding word will be translated directly.

If the character string is a long sentence, the string will be segmented by Jieba word segmentation tool to split the long sentence into groups of single words and then being translated respectively until the entire sentence is translated.

The corpus used in the translation process is a bilingual vocabulary database collected manually in advance. If a word that is not collected in the vocabulary database is entered, the system will
prompt that the word is not collected in the glossary.

In this system, the method of judging whether a character string is a single word is to first segment the character string using Jieba word segmentation tool, and determine is the original character string a single word or a sentence containing multiple words by the number of divided character strings. The processing workflow is shown in Figure 1:

![Processing Workflow Diagram]

3. System Architecture
The vocabulary-based machine translation auxiliary tool for Chinese-Kazakh bilingual is mainly composed of four large modules: vocabulary processing module, vocabulary management module, setting module and additional function module.

Each module contains different functional blocks. The vocabulary processing module includes vocabulary translation and vocabulary comparison functions. The vocabulary translation block mainly performs further pre-processes and translation on the input text. The vocabulary comparison block contains vocabularies that have been translated and aligned, stored as data in the database; The vocabulary management module includes sections for vocabulary adding, deleting and browsing; the setting module includes sections for file reading and text editing; the additional function module
includes sections for system help and developer information. The basic structure of the system is shown in Figure 2:

![Figure 2 System Basic Structure](image)

3.1. Vocabulary Translation and Editing

After users finish inputting the vocabulary to be translated in the interactive interface, they can complete the vocabulary translation operation by clicking the Translate Vocabulary button.

The vocabulary translation function can find relevant data in the system's vocabulary database and output corresponding results through the analysis method as shown in the previous paragraph, and users can use the functions of adding, browsing and deleting vocabulary to further edit, modify and improve the Chinese and Kazakh bilingual vocabulary.

The quick browsing function can realize the quick review of related vocabulary and data through text keyword search, as well as locate where the words are stored in the glossary by data number. The user interface consists of a user interface, a management interface and corresponding sub-interfaces, which are convenient for users to select languages, choose vocabulary and search for words as needed.

Through the management interface, the user can enter the background management function, and perform data maintenance and debugging operations in the glossary. The interface diagram is shown in Figure 3.
3.2. Browsing and Management
In the vocabulary database, there are Chinese and Kazakh bilingual words stored according to translation rules, which can be browsed, deleted and added.

The user can enter Chinese or Kazakh words to query the corresponding words, or enter the initials of both Chinese and Kazakh to look up or browse some of the words, and most of the commonly used vocabulary can be found.

Other types or specialized Chinese-Kazakh vocabulary and relevant terms can be added as required to further improve the vocabulary in the database and improve the scope and accuracy of the auxiliary translation tool. The operating interface is shown in Figure 4.

3.3. Database architecture design
The system uses Microsoft Access as the backend database, and each dictionary in the library consists of glossaries (Chinese-Kazakh glossary).

Each glossary consists of a number of different numeric fields, namely the Chinese vocabulary field and the Kazakh vocabulary field. And Microsoft SQL Server was used as the backend data service to store the data.

The vocabulary database mainly contains Chinese glossary, Kazakh glossary, vocabulary relation table, user table and management table. The user table includes user name, user password, user level and other information, and the management table includes administrator name, corresponding
password and corresponding authority, etc. The bilingual vocabulary data table is shown in Table 2-1.

| List name | Description               | Data Type | Data Length | Primary Key | Foreign Key |
|-----------|---------------------------|-----------|-------------|-------------|-------------|
| ID        | ID number                 | int       | 50          | TRUE        | FALSE       |
| Hanyu     | Chinese vocabulary        | Varchar   | 150         | FALSE       | FALSE       |
| Hayu      | Kazakh vocabulary         | Varchar   | 150         | FALSE       | FALSE       |

4. Conclusion
This paper introduces the design process of machine translation auxiliary tool between Chinese and Kazakh languages, and summarizes the layout and usage of each interface of the system, the pre-processing of text files and vocabulary terms.
The author believes that translation software will become more and more mature and perfect in the future with the continuous development of science and technology, and will gradually replace traditional human translation for the most part or even completely.

Therefore, the construction of cross-language translation assistant software systems is a necessary part of the development of a society. It is a large project that will take a lot of time and effort, and requires a huge amount of data support.

Thus, the construction of this system requires the developers to continuously improve their technical support, further improve the capability and accuracy of software translation, and make continuous efforts to improve the convenience of the software and expand the coverage of the corpus.

References
[1] Alifu. Kuerban, Abulimiti Abudureyimu, Tuergen. Yibulayin. Design of a Dictionary for Uighur-Chinese Machine Translation System [J]. Computer Engineering And applications, 2006,42(20):76-78.
[2] Muheyat Niyazbek, Kunsaule Tälp, Gulila Altenbek. On NP and VP Attachment Disambiguation Based On Mixed Strategy [J]. Journal of Southwest China Normal University (Natural Science Edition), 2014,39 (07) : 41-45.
[3] Rukeyanmu Wusimanjiang, Maierhaba Aili, Tuergen Yibulayin. The Rule-based Uygur Kazakh Machine Translation [J]. Journal of Xinjiang University(Natural Science Edition),2016,3:339-342.
[4] Zulkar, Azhar, Azragul. Natural language processing Chinese(language)-Kazakh translation Machine translation Kazakh translation [J], Computer Knowledge and Technology, 2016.(5) 166-168.
[5] Askar Yakup, Azragul, Yusup Abaydulla. Research on File Format Conversion Technology For Constructing Uyghur Language Corpus [J]. Computer Applications and Software, 2012, 29 (6) : 14-16
[6] Gulziya·Abudujieili. Design and Implementation of Chinese-KazakhDiglossia Electronic Dictionary [D]. Xia men University, 2013
[7] Liu Ri-sheng,Yang Zhen-li. An idea of the resource sharing platform of corpus [J]. Journal of Academic Library and Information Science, 2012, (2) : 46-48
[8] Muheyat Niyazbek, Kunsaule Tälp, Gulila Altenbek,. WU Shou-yong. On Design and Implementation of Kazakh Corpus Management System [J]. Journal of Southwest China Normal University(Natural Science Edition), 2012,37 (11) : 37-40.