The Kazakh terminology recognition system in the IT field

Muheyat Niyazbek¹, Kuenssaule Talp²

¹College of Information Science and Engineering, Xinjiang University, Urumqi, China, 830046
²College of Chinese Medicine of Xinjiang Medical University, Urumqi, Xinjiang, China 830011
Muheyatn@xju.edu.cn , Corresponding author: Kuenssaule Talp@163.com

Abstract: The design and development of the Kazakh language terminology recognition platform in IT field is one of the important means for more effective use of term resources in the field. Used a conditional random field (CRF) and manual modification method in the platform and based on the characteristics of the Kazakh language itself, this paper analyzes the rules of terminology formation and related terminology identification methods as well as describes the basic framework of the term recognition platform and the workflow and main functions of the training corpus, terminology recognition, feature template and term extraction in the system.

1. Introduction
With the continuous expansion of Chinese language information processing applications, the search requirements for technical terminologies in different languages and fields are becoming more and more urgent. The construction of the terminology recognition platform in the field of Kazakh information technology with computer as a tool plays a role in the national language informationization of Kazakh natural language information processing, Kazakh linguistics research, machine translation, corpus construction, and IT domain terminology database construction. The terminology is a language unit that represents the basic concepts of a specific subject area. It is the embodiment of the core knowledge in this field. It is convenient for people to quickly acquire professional knowledge. How to automatically obtain terminology has naturally become a hot topic of research. The term automatic extraction is an important research task in the field of information processing. It has important applications in the fields of dictionary editing, ontology construction and machine translation. The term extraction technique is one of the key technologies for the automatic or semi-automatic construction and expansion of large-scale ontology engineering. In recent years, the importance of the term extraction method has been recognized and a great deal of research has been conducted. At present, the term extraction ideas adopted are mainly divided into statistical methods based, machine-learning based, rule-based linguistics and a combination of those methods. We use a conditional random field and manual modification method in conjunction with linguistic rules. In order to contribute to the inheritance and development of national culture, as well as the prosperity and development of national education, science and technology and society

2. System design
The system is based on various Kazakh websites and electronic texts of various texts obtained in the...
primary and secondary school information technology textbooks. The lexical analysis of the original corpus is obtained through the various language corpus tools currently used in the multilingual information technology laboratory to acquire maturing corpuses that has finished word extraction, affix extraction, and part of speech tagging. Enter the maturing corpus in the rule-based Kazakh information technology domain term extraction system, and filter through the term dictionary and the term collocation rule library to obtain the final term generation candidate term and candidate terminology corpus. The candidate term annotation corpus is then generated as a training corpus by manual modification. The specific process of the system is shown in Figure 1.

3. System function module

From the perspective of system function, the extraction conditions of Kazakh information technology term extraction problem are treated by the method of random field, and the Kazakh information technology domain term recognition is regarded as a sequence part-of-speech tagging problem, and the characteristics of the Kazakh information technology domain term distribution are quantified. The characteristics of the system's training using the conditional random field (CRF) toolkit to train the terminology feature template in the field of information technology in Kazakh. The whole system can be divided into two parts: term tagging corpus and CRF pattern recognition. The term tagging corpus subsystem includes preprocessing part, generating training corpus part, term identifying part, term extracting part, delimiting rule part, etc. The other CRF pattern recognition subsystem includes model parameter part, feature selection part, feature template selection part, and the like. System function module shown in Figure 2.
3.1 Generate training corpus
The language materials stored in the IT terminology corpus have appeared in the actual use of the language, and are the basic resources for carrying the language knowledge by the electronic computer as the carrier. Real corpus needs to be processed to become a useful resource. Taking the maturing corpus in the system as input, the term is extracted from the given document according to the linguistic rules, and the training corpus is generated after further modification. The term itself can be a word or a phrase. The Kazakh IT domain has a variety of term structures. Some terms are composed of one word or two words, and some terms are composed of different additional components or nests. The form of terms can be nouns + nouns, adjectives + nouns, nouns + verbs, etc. The whole system is divided into term extraction, training corpus, term recognition and exit system. In the process of generating training corpus, the module also includes opening XML file, opening term file, marking term in XML file, saving file file, etc. Perform related operations, such as opening the termbase file for XML annotation, etc. The interface also includes the previous, next or previous paragraph, next paragraph and other options, each option has different phased operation steps, the specific operation interface of the generated training corpus module is shown in Figure 3.

3.2 Term extraction
Because of the differences between word terms and multi-word terms, the terms of different languages are different, such as nouns + nouns, adjectives + nouns, nouns + verbs, etc. As a result, the term extraction will be delimited according to the language characteristics and the composition of the terms. The term extraction module is mainly about extracting the terms in related materials. After entering the page, the interface is divided into left part and right part. On the left part, open, extract, save, and exit file, term statistics and the like can be operated. And extracted terms and number of extractions are displayed on the right part. The system's term extraction architecture diagram detailed operation interface is shown in Figure 4 below.

Figure 2 System function module
3.3 Terminology recognition

The module includes three parts: training, testing, and analysis. Different operating interfaces enter from different parts. After entering the training corpus, you can see options such as adding corpus, feature extraction, model training and the like. Each operation can continue to operate accordingly. The test module includes test corpus, term recognition, save results, and quick test. The analysis module can display the number of correctly identified terms, misidentified terms, terms marked by the system and the undetermined term, accuracy, recall rate, F value, etc. The term identification method is pre-selected, that is, the candidate term is first selected, and through analysis and observation, the part-of-speech rule table of the IT term is written, and the rule and the already-labeled word are used. The text is matched, and the corresponding word or phrase is extracted as a candidate term. The system's terminology identifies the detailed operation interface as shown in Figure 5 below.

3.4 Feature Template

Different languages have different grammatical rules and their own characteristics. Because the term itself has strong normativeness, the identification and inclusion of terms requires a long process, so the compilation of terminology in most languages often lags behind the emergence of terms. If you can use computer-aided identification of candidates and then determine through expert participation, it will be of great help. Relevant feature templates determined according to the composition characteristics of Kazakh IT domain terms can be mainly divided into: stem, affix, part of speech and left and right words. For example: the current word (CWord), the first word on the right (RWord), the first word on the left (LPos), the suffix of the preceding word (CAffix), the right note on the right (it), etc. The term itself is a word, the term can be generalized to ordinary words and the common words can be abstracted into terms. This reflects the mutual penetration between terms and common words [8-10]. The presence of the same word may be a term in one segment and may be a multiple application of a common term in another segment.

Kazakh is a type of adhesive language. The words in the Kazakh language text are composed of a certain morphemes attached to the roots. The morphemes are divided into word formation morphemes and morphemes. Kazakh is different from Chinese and English. The Kazakh language is based on words. This is like English, but Kazakh has adhesiveness and rich context information. The change of
the form of the word is more abundant than that of English. According to the characteristics of the Kazakh IT domain term itself, the feature space of this paper is defined as:

| Number | Feature     | Signification                  | Number | Feature     | Signification                  |
|--------|-------------|--------------------------------|--------|-------------|--------------------------------|
| 1      | LWord       | The first word on the left     | 5      | LPos        | The part of speech of the first word on the left |
| 2      | CWord       | Current word                   | 6      | CPos        | The part of speech of the current word          |
| 3      | RWord       | The first word on the right    | 7      | RPos        | The part of speech of the first word on the right |
| 4      | LLPos       | The part of speech of the second word on the left | 8      | RRPos       | The part of speech of the second word on the right |

The appropriate feature templates were selected, and two main composite feature templates were constructed based on Table 1. Each information function takes values in the current word context, combines the values of the various functions into the premise of the feature, and obtains the action of the feature by marking the word, so that the feature can be extracted:

Template 1: \( [\text{RRPos}, \text{RRTE}, \text{RWord}, \text{RAffix}, \text{RPos}, \text{RTE}, \text{CPos}, \text{CTE}, \text{CWord}, \text{CAffix}, \text{LWord}, \text{LAffix}, \text{LPos}, \text{LTE}] \)

Observe the effect of one word on the left and two words on the left side of the candidate.

Template 2: \( [\text{RRPos}, \text{RRTE}, \text{RWord}, \text{RAffix}, \text{RPos}, \text{RTE}, \text{CPos}, \text{CTE}, \text{CWord}, \text{CAffix}] \)

Observe the effect of the two words on the right side of the candidate on the experimental results.

### 3.5 Experimental data

The following individual judgment indicators are used in the text: term recognition accuracy, error rate. The definition is as follows: accuracy rate \( (P) = \frac{\text{number of terms correctly recognized by the system}}{\text{total number of terms recognized by the system}} \times 100\% \); error rate \( (E) = 1 - \text{accuracy} \). The system uses the training corpus of different scales that have been marked for open testing. The test results are as follows:

| Corpus size | Test type | P(%) | R (%) | E (%) | L (%) | F value (%) |
|-------------|-----------|------|-------|-------|-------|-------------|
| 1000 sentences | Open test | 79.15 | 78.67 | 19.35 | 20.13 | 78.16       |
| 3000 sentences | Open test | 80.27 | 79.79 | 18.73 | 19.21 | 80.03       |
| 5000 sentences | Open test | 81.06 | 80.57 | 17.94 | 18.43 | 81.01       |

### 4. Conclusion

The construction of the Kazakh terminology recognition system in the IT field is a long-term, engineering-intensive project that requires massive amounts of data. At present, the construction of the Kazakh terminology recognition system in the IT field is in the initial stage, and only the collection of the original data and the sorting of the basic information are realized. If you want to meet the different needs of Kazakh linguistic information research, the technical content of processing and analysis of corpus tools needs to be continuously improved. Therefore, the construction of the system still has a long way to go and requires relevant personnel to continue to make unremitting efforts.

Jiebangren

### References

[1] Muheyat Niyazbek1, Kunsaule Talp. Research on Automatic Kazakh Information Technology Term Extraction Using CRF Model [J]. Journal Southwest Chna Normal University (Natural Science Edition) 2016, 52 (1): 53-56
[2] Muheyat Niyazbek1, Kunsaule Talp. The Discussion of several issues in the Kazakh information processing of the status [J]. Intellegent computer and application, 2011, 1(4): 45-46.

[3] Liu Jian, Tang G Huifeng Liu Wuying. An Extraction Method for Chinese Terminology Based on Statistical Technology [J]. China Terminolgy, 2014, 16(5): 10-14.

[4] Zhang Rong. Research on Term Definition Based on Identification [J]. China Terminolgy, 2014, 16(4): 5-8

[5] Muheyat Niyazbek1, Kunsaule Talp. Research of NP and VP Attachment Disambiguation based On Mixed Strategy [J]. Journal Southwest China Normal University (Natural Science Edition), 2014, 39(7): 41-46.

[6] Chen yu, zhu jian feng, Wu yu jian, zao yu yun. New Term Expansion Method Based Domain Ontology [J]. Computer Engineering, 2011, 37(7): 24-27