Research and design of IT-based Kazakh terminology recognition system construction techniques

Muheyat Niyazbek[1,2] and Kuenssaule Talp[3,*]

[1] College of Computer Science and Technology, Xinjiang University, 830017, Urumqi, China
[2] Xinjiang Key Laboratory of Multilingual Information Technology, 830017, Urumqi, China
[3] College of Chinese Medicine of Xinjiang Medical University, Urumqi, 830011, Xinjiang, China

Abstract

The demonstration and realization of terminology identification systems in the IT domain is considered to be one among the most significant measures to utilize terminological resources in this field more efficiently. This article describes the research and design of an IT-based terminology recognition system for the Kazakh language. The system uses Conditional Random Fields (CRF) and manual modification methods and proceeds to analyze the patterns of term formation and related term recognition methods based on the characteristics of the technical terms itself in the IT domain.

Keywords

information technology field, terminology recognition, system design

1. Introduction

With the expansion of applications for processing Chinese language information, the requirements for terms retrieval in various fields of different languages is becoming increasingly imminent. Among them, using computer as a tool to build a platform for identifying the terminology in the field of IT in the Kazakh language is increasingly important for the construction of national language informatization such as Kazakh natural language information processing, Kazakh linguistics studies, information security retrieval, machine translating, corpus establishment, and terms repository in the IT field [1]. A term is a linguistic unit representing the primary and fundamental notions of a particular academic field, which is the representation of the field's central knowledge and facilitates people's rapid access to specialized knowledge, so how to retrieve terms automatically naturally becomes a research hotspot for related professionals. Automatic term acquisition is a major investigation assignment in the information processing domain,

ICCIC 2024: International Conference on Computer and Intelligent Control, June 29–30, 2024, Kuala Lumpur, Malaysia

* Corresponding author.
muheyatn@xju.edu.cn (M. Niyazbek); kuenssauletalp@163.com (K. Talp)
0009-0002-2051-6103 (M. Niyazbek); 0009-0007-1507-0844 (K. Talp)

and it has a significant usage in the fields of lexicography, ontology structuring, machine translating, and other fields. Terminology extraction is one of the pivotal technologies for constructing and extending large-scale ontology engineering in automatic or semi-automatic ways. In the past few years, the awareness of the importance of methods for identifying terms have been acknowledged and a lot of researches have been carried out, while the widely used methods for extracting terms are primarily classified into statistics-based approaches, methods based on machine learning, linguistic rules and combined hybrid methods. The system presented in this article is designed by combining linguistic rules with Conditional Random Fields (CRF) and manual modifications. In the IT domain, it is expected that through the conception and realization of the system for identifying Kazakh terminology, we will do our part in the excavation, inheritance, and innovation of national culture and national scientific and technological educational work, as well as in the security, stability, and prosperity of the community.

2. System design

The resulting framework is designed on the basis of an electronic corpus of various texts obtained from various Kazakh language websites and IT textbooks for primary and secondary schools, as well as a cooked corpus with completed word extraction, affix extraction, and lexical annotation, obtained after lexical analysis of the original corpus by various linguistic corpus tools now in use in multilingual IT laboratories. After inputting the cooked corpus into the rule-based system, it is additionally refined using a term dictionary and a rule-based term collocation system to produce the final list of candidate terms and the annotated term corpus [2-4]. Then manually modify the candidate term labeling corpus to generate the training corpus. The detailed flow of the process is shown in Figure 1.

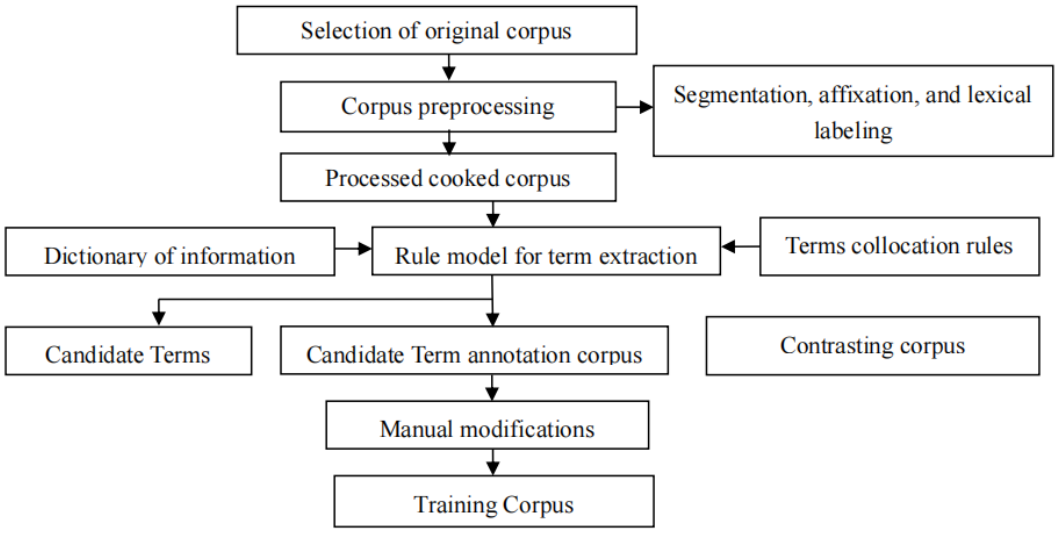

Figure 1: Workflow diagram.

2.1 The Design of Database Structure

This system is designed using CS schema for convenience and Microsoft SQL Server 2005 is applied as the backend data server for corpus storage. The tables in the database are corpus table, permission table, daily newspaper category saving table, periodical category saving table, textbook category saving table, relations table and so on. The information is recorded according to the format of these tables, such as the corpus table including number, title, save path, date of entry, number of paragraphs, the user table including user name, user password, user level, the management table including the name of the administrator, the corresponding password and the corresponding permissions and so on.

2.2 Corpus Classification

Text categorization is the process of classifying a given text into one or more categories according to its features, such as content and attributes, under a certain classification system. In such a way, the research topic of text categorization involves many issues related to natural language understanding and pattern recognition such as how to understand and classify the content of the text, and a successful text categorization system is not only a natural language processing system, but also a typical pattern recognition system. So far, text categorization of the Kazakh language is still basically done manually. Obviously, the traditional manual classification method has restricted the speed of Kazakh text classification, and it is difficult to meet the needs of social development, the research and development of a fast and accurate Kazakh text classification that can replace the manual labor is very necessary

2.3 Training corpus preprocessing

Manual annotation for a piece of Kazakh corpus is inefficient and unable to achieve a high accuracy rate, so it is more appropriate to use a dictionary-based matching method. First of all, we get a considerable amount of Kazakh corpus by surfing the website TIANSHAN.com, then we find the common specialized terms in the IT field by searching the Kazakh dictionaries, and finally, we write a program to find out the specialized terms in the dictionaries appearing in the corpus and annotate them manually.

BIO annotation is a sequential annotation method for labeling entities or lexemes in text. In BIO annotation, each word or phrase is labeled as one of three possible cases: B, I or O. Where B means that the word is the beginning position of an entity or specialized vocabulary, I mean that the word is the internal position of the entity or specialized vocabulary, and O means that the word is not an entity or specialized vocabulary. For a specialized vocabulary in a piece of Kazakh corpus, we need to mark it out using BIO annotation.

2.4 Kazakh terminology recognition

The process of Kazakh terminology recognition requires corpus acquisition and labeling. Firstly, we obtain the commonly used terms in the IT field by looking up the Kazakh dictionary, then we obtain part of the Kazakh corpus in TIANSHAN.com, find the

specialized terms in the acquired corpus by string matching, and finally conduct manual BIO annotation. Different from the Chinese model, the Kazakh model adopts BLSTM-CNN-CRF model with one more layer of CNN network structure, and the CNN layer can capture the local features in the input sequences, which improves the accuracy of the model. After the model training is completed, when the live Kazakh corpus is imported into the model, the terminology in the information domain in that corpus can be recognized.

2.5 User Management Module

The user administrative module is majorly tasked with the management of the signed-in personnel of the entire corpus language platform for managing resources, the setting and the realization of the permissions of each user. It is also responsible for the operation of adding, deleting and changing passwords for users. There are three levels of users in this system: system administrators, editors and ordinary users. The system administrator has full control over all resources, whereas other users are restricted to utilizing the language resources of the corpus according to their permissions.

3. System Functional Structures

From the aspect of system functionality, the method of random fields is used as an extraction criterion to process the term extraction issue, the term recognition in the IT field of Kazakh language is perceived as a sequential lexical annotation question, the feature quantization of term distribution is used as a feature for the training of the system, and term feature templates are trained by using the toolkit of Conditional Random Fields (CRFs).

There are two subsystems of the entire system which can be categorized into term annotation corpus and CRF pattern recognition, in which the term annotation corpus subsystem also consists of preprocessing section, creating training corpus section, term recognition section, term extraction section, delimitation rule section, etc., and another CRF pattern subsystem also comprises model parameter segment, feature selection segment, and feature template selection segment. The functional structure of the system is shown in Figure 2.

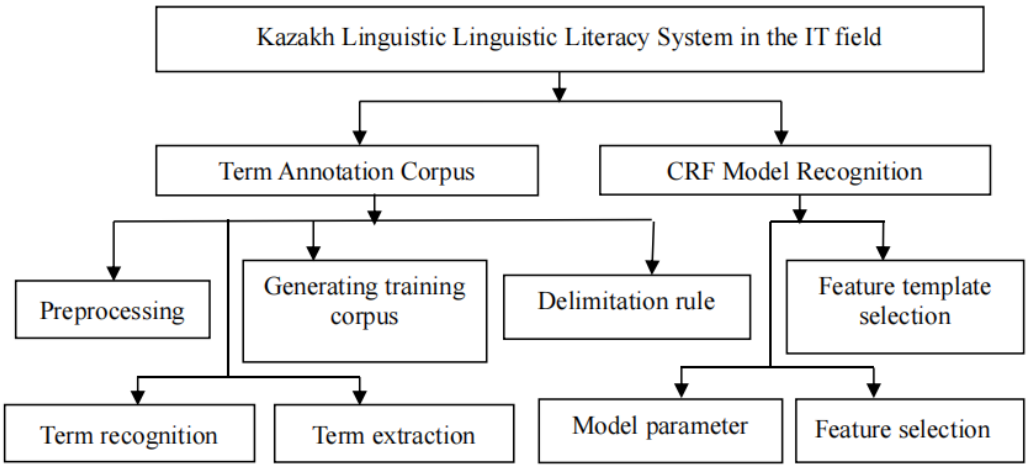

Figure 2: System Functional Structures.

3.1 Generation of the training corpus

The linguistic data stored in the Information Technology Lexical Corpus has emerged during actual language use and are the basic material for computers to carry linguistic knowledge, and the authentic corpus needs to be addressed in order to become a usable material. Using familiar corpus from the system as input, the terms are retrieved from the given text according to the grammar principles, and then further modification process is performed to generate the training corpus. The term can be either a word or a phrase on its own, and there are various structures of terms in the domain of Kazakh information technology, some of them consist of one word or two words joined together, and some of them consist of various supplementary components or nested, taking the manner in noun+noun, adjective+noun, noun+verb, and so on. The entire system is organized into sections such as term extraction, generating training corpus, term recognition and exiting the system. In the section of generating training corpus, it contains modules like opening XML file, opening terminology file, annotating terms in XML file, saving annotation file, etc., which can be used for further related operations as needed, like accessing the XML annotation file in the thesaurus [5-7]. The interface additionally contains options such as previous paragraph, next paragraph or previous paragraph, next paragraph, etc., each of which has different stages of operation procedure, and the detailed operation interface of the specific module for generating training corpus is shown in Figure 3.

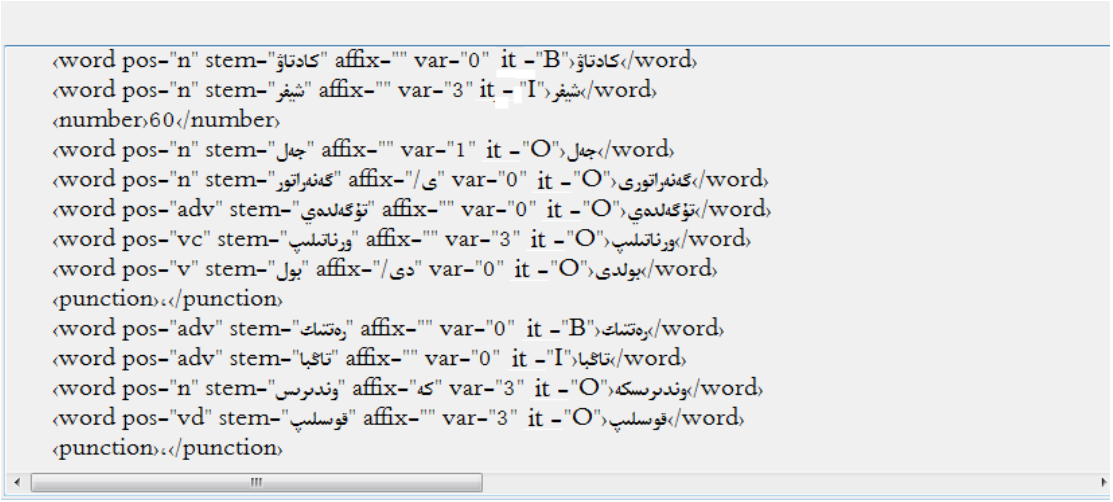

Figure 3: Interface for generating training corpus.

3.2 Term extraction

Due to the differentiation of word terms, multi-word terms, etc., and the different forms of terminology in different languages, such as noun + noun, adjective + noun, noun + verb, etc., the terminology extraction will be based on the characteristics of the language and the composition of the terminology structure to define the rules of abstraction. The module is mainly for the relevant information in the term extraction, into the page after the left and right interfaces, the left side can conduct the document open, extraction, save, exit, term statistics and other operations, the right side shows the extracted terms and the number of extracted information. The detailed operation interface of the system's terminology extraction architecture is shown in Figure 4 below.

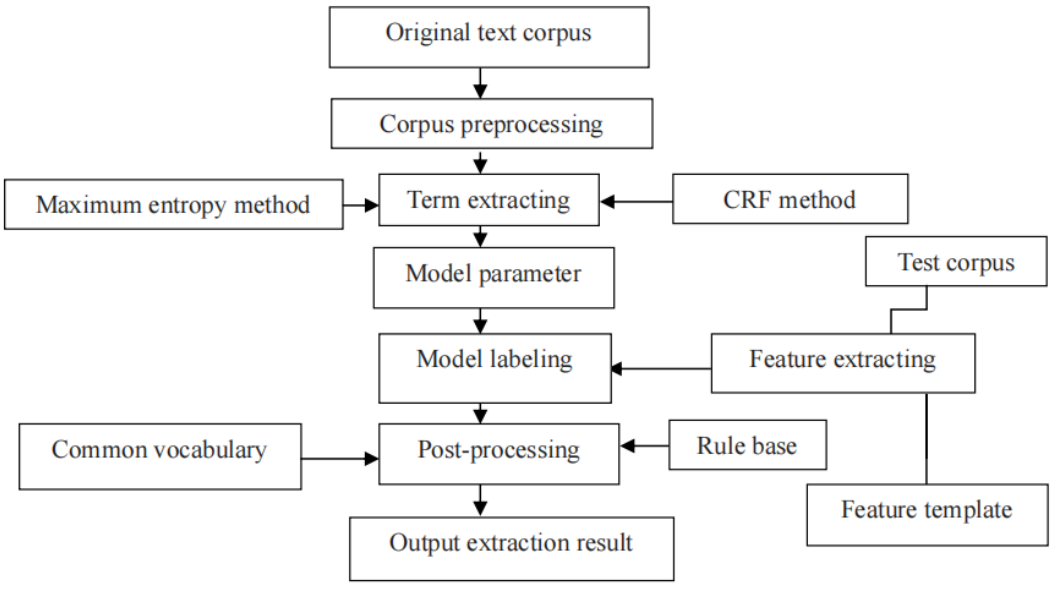

Figure 4: Architecture diagram for automatic term extraction.

3.3 Term Recognition

The module contains 3 sections: training, testing and analyzing, and from various sections, we enter various operation interfaces. When entering the training corpus section, users can view the options of adding corpus, feature extraction, model training, etc., and in each of these options, they can continue to carry out the corresponding operations.The testing module includes test corpus, term recognition, result saving, and quick testing. In the analysis module, it counts the number of correctly identified terms., the number of incorrectly recognized terms, the number of terms labeled as terms by the system, the number of undecided terms, the accuracy, the recall, the F-value and so on can be displayed. The term recognition method is based on pre-selection, i.e. candidate terms are selected first. Although Kazakh language belongs to adhesive language, the lexical properties of information technology terms have certain regularity, by analyzing and observing, the lexical rule list of information technology terms is prepared, and then the rules are used to match with the text that has been labeled with lexical properties, The candidate term is extracted based on the corresponding word or phrase. The detailed operation interface of the term recognition system is shown in Figure 5 below.

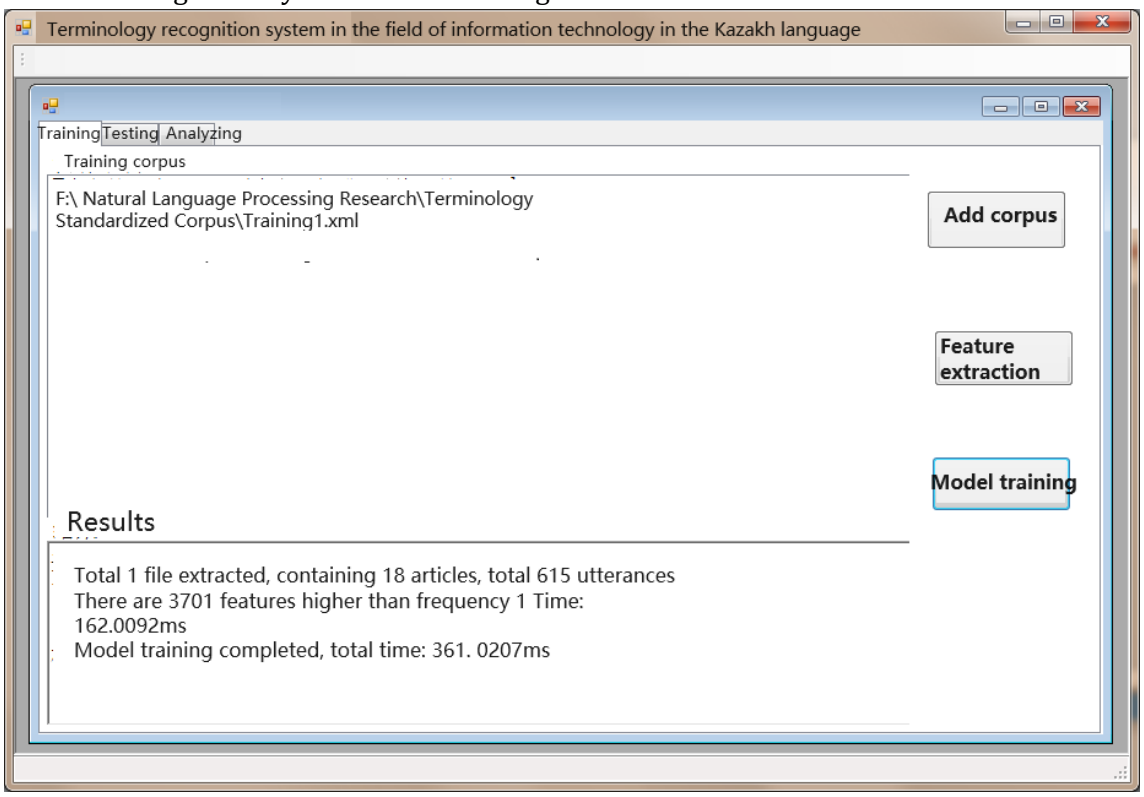

Figure 5: System term recognition interface.

3.4 Feature Template Recognition

Various languages carry distinct grammatical principles and their unique characteristics. Since the terms themselves are highly normative, the identification and incorporation of terms need to go through a lengthy process, thus, the development of terminology

dictionaries in many languages often lags behind the emergence of new terms. The use of computer-assisted identification of candidate terms, followed by their determination through expert participation would be greatly beneficial. The template of relevant features identified on the basis of the characteristics of the composition of terms in IT field in Kazakh language can be divided into the following categories: stem, affix, lexical properties and terminological annotation information for the left and right words. For instance, the current word (CWord), the first word on the right (RWord), the lexical properties of the first word on the left (LPos), the prefix of the previous word (CAffix) and the it annotation of the first word on the right (RIT). The interpenetration between terms and ordinary words is reflected in the fact that a term is itself a word, a term can be generalized into a common word, and an ordinary word can be abstracted to a term. The same word is used in different ways, it may be a term in one passage and a common word in another.

Kazakh language belongs to the adhesive type, words in Kazakh texts are formed by attaching certain morphemes to their roots, which are segmented into morphemes and lexemes. Kazakh language differs from Chinese and English in that it is word-based, and in this respect it is the same as English, but the Kazakh language is adhesive and rich in contextual information, and the morphology of Kazakh words is richer than that of English. Based on the characteristics of the Kazakh terminology in the field of IT itself, this paper defines the feature space as:

Table 1

Term recognition feature space

No	Feature	Significance	No	Feature	Significance
1	LWord	First word onthe left	5	LPos	Morpheme of the first wordon the left
2	CWord	current word	6	CPos	Morpheme of the current word
3	RWord	First word on the right	7	RPos	Morpheme of the first wordon the right
4	LLPos	Morpheme of the second word on the left	8	RRPos	Morpheme of the second woron the right

Selecting suitable feature templates, 2 major representative composite feature templates are constructed on the basis of Table 1. Each informational function derives its values from the current word context, combining each function value to form the feature's premise, and getting the role of the feature through the marking of the word, thus extracting the feature:

Template 1: [RRPos, RRTE, RWord, RAffix, RPos, RTE, CPos, CTE, CWord, CAffix, LWord, LAffix, LPos, LTE] Examine the impact of one word to the left and two words to the right of the candidate word on the experimental outcomes.

Template 2: [RRPos, RRTE, RWord, RAffix, RPos, RTE, CPos, CTE, CWord, CAffix] Observe the effect of one word to the right of the candidate word and two words to the right on the experimental outcomes.

3.5 Experimental data

The article uses the following singular determination metrics: accuracy rate for term recognition, error rate. They are defined as shown below: accuracy rate (P) = number of all terms correctly recognized by the system/total number of terms recognized by the system*100%; error rate (E) = 1 - accuracy rate.
The system was tested in an open manner using annotated training corpora of different sizes. Here are the test results.

Table 2
Results of the term recognition test

Corpus size	Test Type	P(%)	R(%)	E(%)	L(%)	F-Value(%)
1200 sentences	open-ended	80.15	79.76	20.53	21.30	79.54
2900 sentences	open-ended	79.27	78.97	17.79	18.11	78.03
4900 sentences	open-ended	80.01	79.33	16.05	17.73	79.06

4. Conclusion

The establishment of a terminology recognition system is a large-scale project with a long project period and a large requirement of data. The development of the system of terminology recognition is still in its initial stage and has a long way to go. At present, only the collection of raw data and the organization of basic information on terminology in the IT field of the Kazakh language have been completed. Relevant professionals need to make unremitting efforts to refine the technological methods of corpus tool processing and analysis and to continuously improve the construction of the system in order to further meet the various needs of Kazakh-language information research.

References

[1] Q. Wang, C. Zhang, H. Ding, “Formation of Academic Discourse System Based on Terminology,” China Terminology, vol. 26, no. 1, pp. 63-67, 2024.
[2] J. Jiang, X. Qi, “Design and Realization of Online Query System of Chinese-English Terminology of Chinese Medicine,” China Terminology, vol. 24, no. 4, pp. 92-96, 2022.
[3] Z. Li, Y. Zhong, H. Wang, J. Liu, Y. Sun, “Research on Domain Term Extraction Method Based on Deep Learning and Statistical Information,” Frontiers of Data and Computing, vol. 4, no. 2, pp. 87-98, 2022.
[4] Y. Jia, D. Zhu, “Medical Named Entity Recognition Based on Deep Learning,” Computer Systems & Applications, vol. 31, no. 9, pp.70-81, 2022.
[5] X. Wang, H. Tao, “Research on Chinese Named Entity Recognition based on Deep Learning,” Journal of chengdu university of information technology, vol. 35, no. 3, pp.264-270, 2020.

[6] J. Jiang, X. Qi, “Design and Realization of Online Query System of Chinese-English Terminology of Chinese Medicine,” China Terminology, vol. 24, no. 2, pp. 92-96, 2022.
[7] X. Wei, “The Name and Reality of China Terminology: Some Thoughts on the Identification of Its Problem DoMain,” China Terminology, vol. 23, no. 2, pp. 03-10, 2021.
[8] M. Conrado, T. Pardo, S. Rezende, “A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set,” Naacl Student Research Workshop, 2013.
[9] G. Lample, M. Ballesteros, S. Subramanian, “Neural Architectures for Named Entity Recognition,” arXiv, 2016.
[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics, 2019.
[11] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep Contextualized Word Representations,” In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics, 2018.
[12] K. Takeuchi, N. Collier, “Analysis of Machine Learning Model for Technical Term Extraction in Biological Science Papers,” Ipsj Sig Notes, 2002.
[13] N. Chatterjee, N. Kaushik, “Automatic Extraction of Agriculture Terms from Domain Text: A Survey of Tools and Techniques,” 2020.
[14] T. Yang, K. Hu, “Study on clinical terminology extraction of traditional Chinese medicine based on internal aggregation and boundary degree of freedom of character strings,” IEEE International Conference on Bioinformatics & Biomedicine, 2017.
[15] C. Xu, et al., “Chinese patent terminology extraction,” Computer Engineering and Design, 2013.