Automatic Term Recognition Method for Military Domain

Ani Song¹,a, Xiaoxia Jia¹,b and Wei Jiang¹,2*
¹North China Institute of Computing Technology, Beijing, 100083, China
²National key Laboratory for Complex Systems Simulation. Department of Systems General Design. Institute of Systems Engineering. AMS.,Beijing,100083,China
aemail: songani0817@163.com, bemail: 13391925003@163.com
*Corresponding author’s e-mail: jiangwei_cetc15@163.com

Abstract. With the development of military intelligence, higher requirements are put forward for automatic term recognition in military field. In view of the characteristics of flexible and diverse naming of military requirement documents without annotated corpus, the method of this paper uses the existing military domain core database, and matches the data set and core database by Aho-Corasic algorithm and word segmentation technology, so that the terms to be recognized in the data set can be divided into three types. The possible rules of word formation of military terms are summarized and phrases that conform to the rules of word formation are found in the documents as the term candidate set. The core library and TF-IDF method are used to calculate the value of the candidate terms, and the candidate terms whose value is greater than the threshold are selected iteratively as the real terms. The experimental results show that the F1 value of this method reaches 0.719, which is better than the traditional C-value method. Therefore, the method proposed in this paper can achieve better automatic term recognition effect for military requirement documents without annotation.

1. Introduction
Terms refer to words or phrases that are highly relevant to the field in a collection of documents in a specific field. As a conventional symbol for expressing professional concepts in a specific field, term recognition is in natural languages such as knowledge graph [1-2], ontology construction [3-4], machine translation [5-6], text abstract [7-8]. It plays an important role in processing [9] tasks.

In the work of military demonstration, a large number of military requirements documents were produced. The content and fields involved in various military requirements documents are quite different, and generally require multiple people to work together. Due to the different understanding and cognitive abilities of different personnel and the large number of collaborators, the collaborative writing of the entire military requirements document faces problems such as inconsistent content granularity and incorrect military terminology writing. Term recognition of military requirements documents can automatically extract key information elements, help writers quickly read the documents and locate errors, and improve the efficiency of writing military requirements documents.

Although the writing of military requirements documents follows certain terminology specifications, the military terms [10] are named flexible and diverse, and have strong long-tail characteristics. In addition, the relevant corpus of military requirements documents is small, and the cost of labeling is high. Therefore, military terminology recognition is facing serious challenges. In this paper, we use the word-building rules of terms to find candidate term sets, use the core term base
and TF-IDF to calculate the value of candidate terms, and iteratively screen candidate terms whose value greater than the threshold as true terms, thereby realizing terminology recognition for military requirements documents. This method overcomes the problem of unlabeled data sets for military requirements documents, and realizes automatic terminology recognition for military requirements documents.

The mainstream methods of term recognition are as follows.

1.1. Term recognition based on linguistics
The idea of term recognition based on linguistics believes that: terms usually appear in specific language structures and patterns. Therefore, a relatively complete set of lexical rules can be constructed by discovering strings that conform to the term pattern to automatically identify domain terms. For example, Robert [11] et al. constructed a set of universal naming rules for biological terms by analyzing the word formation of biological terms.

1.2. Term recognition based on statistics
The term recognition method based on statistics recognizes terms through the distribution frequency of words in the domain document. Compared with linguistic-based methods, this type of method is simple and efficient, does not require domain experts and annotated data sets, and has strong applicability and portability. For example, Kenneth [12] et al. proposed to identify terms through point mutual information, which is used to measure the degree of relevance of two phrases, and extract term candidates by calculating the tightness of the combination between phrases. For example, Justeson [13] et al. believe that terms often appear in documents in a specific field, so they recognize terms by word frequency.

1.3. Hybrid term recognition
The hybrid term recognition method combines several methods. The more representative method is the C-value method[14] that combines linguistics and statistics. The basic idea of C-value is to use language rules to obtain candidate term sets, and then use statistical rules to filter.

1.4. Term recognition based on machine learning
The method based on machine learning mainly combines machine learning technology for term extraction. This type of method avoids manual screening of data features but generally requires a large amount of labeled data. In order to solve the problem of large amounts of data annotation, researchers are now beginning to pay attention to unsupervised or remote supervision methods. For example, Arora [15] et al. used POS tags to tag the corpus, extracted candidate terms, and clustered them using the syntactic and semantic relationships between candidate terms to achieve unlabeled recognition. For example, Shang [16] et al. proposed the AutoPhrase method, which uses the remote knowledge base to automatically label and obtain a large number of positive and negative examples as a training set, which solves the problem of unlabeled data.

2. Materials and Methods
This section introduces the term recognition method for military requirements documents. Normally, the main steps of term recognition are shown in Figure 1:

1. Use linguistic or statistical tools to generate a set of term candidate words.
2. Use effective methods to evaluate the quality of term candidate words and sort them, so as to screen out high-quality terms.

This paper proposes a term recognition method based on word formation rules and iterative screening. The method includes the following 4 links:

1. Preprocessing of military requirements documents
2. Term recognition based on Aho-Corasic algorithm [17] and word segmentation [18-19]
3. Term recognition based on rules and iterative screening
(4) Combination of two identification methods
The specific identification process is shown in Figure 2.

![Figure 1 Automatic term recognition steps](image1)

![Figure 2 The overall flow chart of automatic term recognition](image2)

2.1. Military requirements document preprocessing
Most military documents are in word format and have a certain template structure. Before using these documents, you need to perform data cleaning on them. This article selects the main text of the military requirements document, removes various sizes of headings, charts and formulas, and divides the text according to periods and question marks. The final result is a txt text format data set.

2.2. Term recognition based on AC algorithm and word segmentation
This paper chooses a string multi-modal matching algorithm, namely Aho-Corasic (AC) algorithm for term recognition. The AC algorithm was produced in Bell Labs in 1975. Specifically, given n words and then an article, the AC algorithm can be used to quickly find the position of the word in the article.

The algorithm core of AC automata is as follows. Construct all strings into a Trie dictionary tree as the search data type of AC automata. The Trie form is shown in Figure 3.

The algorithm constructs a fail pointer to jump to the character with the longest common prefix and suffix to continue matching when the current character is mismatched, as shown in Figure 4. Like the KMP algorithm [20], if the current character fails to match when the AC automaton matches, it uses the fail pointer to jump and scan the main string for matching.
The AC algorithm can accurately identify the terms that already exist in the term database, but this may cause semantic segmentation errors. Choosing the jieba[21] tokenizer for word segmentation can reduce false matches. If the word at the beginning of the longest match does not form a word with the previous word in the word segmentation result, and the word at the end does not form a word with the next word in the word segmentation result, the word is considered correct.

2.3. Term recognition based on rules and iterative screening

Terms can be divided into simple terms composed of words and compound terms composed of multiple words. Compound terms are generally composed of simple terms and other words. The terms to be identified in the data set are divided into three types. The first type is a term that can be fully identified through the core library for matching, and the second type is a part of the compound term that can be identified through the core library for matching. The three types are terms that are completely unrecognizable through the core library for matching, that is, unregistered words.

The AC and word segmentation methods can only successfully identify the first type of terms. Therefore, it is necessary to identify as many of the other two types of terms as possible through rules-based and iterative screening methods.

2.3.1. Term candidate set generation based on word-building rules

Terms often appear in specific language structures and patterns, and part of speech tagging can determine the position boundary of term candidate words to a certain extent. Through the word segmentation and part-of-speech tagging of the existing data set [22], the word formation feature analysis is carried out, and the possible word formation rules are defined. The word formation rule
table is shown in Table 1, where: N is a noun, A is an adjective, P is a preposition, and V is a verb. The word-building rules in the table believe that the longest noun phrase[23] or definite middle phrase may be a term.

| Phrase type                      | Word formation type          |
|----------------------------------|------------------------------|
| Longest noun phrase              | m*N(m>=2)                    |
| definite middle phrase           | (A||V)+N                     |
| definite middle phrase           | (A||V)+(A||N||P||V)*n+N(n>=1) |

2.3.2. selection of real terms
The terms in the term candidate set are screened, and the real terms are obtained through the term frequency-inverse document frequency (TF-IDF) [24] and the core term base.

Term frequency (TF) refers to the total number of occurrences of the candidate term a in the document. It is based on the assumption that if a phrase is specific to a certain field, the phrase will frequently appear in documents in this field. For the word a in a particular document, its TF can be expressed as:

$$TF(a) = \frac{\text{count}(a)}{\sum_{k=0}^{n} \text{count}(k)}$$

Among, the numerator is the number of occurrences of word a in the document, and the denominator is the sum of the number of occurrences of all words in the document.

Inverse document frequency (IDF) refers to the inverse proportion of the number of documents of the candidate term a in the entire corpus. It is based on the assumption that if a phrase appears in a document more frequently, it is less likely to be a domain term. For the word a in a particular document, its IDF value can be expressed as

$$IDF(a) = \log \left( \frac{|D|}{|\{ j : t(a) \in d(j) \}|} \right)$$

Among, |D| represents the total number of documents in the corpus. |{ j : t(a) \in d(j)}| indicates the number of documents containing the word a.

TF-IDF combines term frequency and inverse document frequency to comprehensively determine candidate terms. The TF-IDF can be expressed as:

$$TF - IDF(a) = TF(a) \ast IDF(a)$$

Use the jieba tokenizer to segment the core terminology database, and the segmented phrases form a word set according to the word frequency. The candidate term a is segmented. If the more phrases generated after the segmentation appear in the vocabulary, the higher the frequency of the phrase in the vocabulary, the higher the degree of correlation between the term and the core term base, and the candidate term is a term. The more likely it is. The degree of relativity between the candidate term a and the core library can be expressed as:

$$\text{relativity}(a) = \sum_{i=0}^{n} TF(m_i)$$

Where \( m_i \) indicates that the phrase \( m_i \) exists in both the word set and the word segmentation result of the candidate term a, and TF indicates the frequency of the phrase \( m_i \) in the core term database.

Integrate TF-IDF and relativity to determine the value of the term, the value of the candidate term a can be expressed as:

$$\text{value}(a) = \frac{TF-IDF(a)}{TF-IDF_{\text{max}}} + \frac{\text{relativity}(a)}{\text{relativity}_{\text{max}}}$$

Where \( TF-IDF_{\text{max}} \) represents the maximum value of TF-IDF of the candidate term, and \( \text{relativity}_{\text{max}} \) represents the maximum value of relativity of the candidate term. When the value of the candidate term a is greater than the threshold, then a is considered to be a real term.
2.3.3. Incremental iterative screening of term candidate set
The obtained new term segmentation is added to the word set, and the value value of the candidate term that is not judged as a real term is recalculated. If the value value of the newly obtained term is greater than the threshold, it is judged as a new term. This is a process of incremental iteration [25] until no more new terms are generated. All new terms obtained can be added to the core library after being reviewed by military experts, thereby continuously expanding the core library.

2.4. Combination of two term recognition methods
Take AC and word segmentation as Method 1, and rule-based and iterative screening as Method 2. The two recognition methods are combined to complement each other to obtain the final recognition result. The combination rules are as follows:
(1) The terms identified in Method 2 are all regarded as final terms.
(2) If the term identified by Method 1 does not overlap with the term identified by Method 2, it is considered as the final term.

3. Experiment& Discussion

3.1. Experimental data
This article selects military requirements documents as the research object. After preprocessing the military requirements documents, a total of 1500 sentences are obtained as experimental data, and 200 of them are marked as the test set. The core term base comes from multiple channels such as web resources, military book catalogs, and domain experts, and contains a total of 8,622 core terms in the military field.

3.2. Experimental evaluation index
In order to quantitatively evaluate the effect of the method in the article on the recognition of military terms, the experimental results are evaluated through three indicators: precision (P), recall (R) and comprehensive evaluation indicators (F1).

3.3. Analysis of results
Table 2 shows the experimental results of different methods. Experiments show that the F1 value of the combined method reaches 0.719, and the recognition effect is the best. The fundamental reason is that the terms recognized by the combined method complement each other, and the effect is improved to a certain extent compared with the single recognition method. Comparing the rule-based and iterative screening method with the rule-only method, the recall rate after iterative screening is reduced, but the accuracy rate is greatly improved, so the F1 value is improved. The fundamental reason is that iterative screening will discard some candidate terms. Most of these candidate terms are not real terms, but a small part are real terms.

Table 2  The comparison of recognition results of different methods
| Method                               | P    | R    | F1   |
|--------------------------------------|------|------|------|
| AC algorithm and word segmentation   | 0.258| 0.207| 0.230|
| Based on rules                       | 0.622| 0.616| 0.618|
| Based on rules and C-value           | 0.739| 0.586| 0.654|
| Based on rules and iterative screening| 0.758| 0.595| 0.666|
| Combination                          | 0.772| 0.672| 0.719|

Table 3 shows the results of different methods for the recall rates of the three types of terms. It can be seen that the combination method has the best recognition effect on the first and second types of terms, and the recognition effect on the third type of terms needs to be improved. The fundamental reason is that the first type of terms is in the core library and can be well identified through AC and word segmentation. The value value of the second type of terms is affected by both the relativity value
and TF-IDF, because it is related to the core library, the relativity value will not be too small, so that most of this type of terms can be filtered out. For the third type of terms, the relativity value is 0, which can only be filtered by TF-IDF. The real terms with lower TF-IDF values are not recognized, so the effect is the worst. For example, the term command center belongs to the third category of terms. Its frequency in the data set is very low, and the term command center often appears in non-military scenarios, so its TF-IDF value is low, so the value is less than the threshold, Not considered as a term.

Table 3 Different types of term recognition results

| Method                          | The first type P | The second type P | The third type P |
|---------------------------------|------------------|-------------------|-----------------|
| AC algorithm and word segmentation | 1.000            | 0.000             | 0.000           |
| Based on rules                  | 0.604            | 0.674             | 0.519           |
| Based on rules and C-value      | 0.573            | 0.651             | 0.433           |
| Based on rules and iterative screening | 0.583            | 0.667             | 0.423           |
| Combination                     | 0.958            | 0.667             | 0.423           |

4. Conclusions

This paper proposes a method for automatic terminology recognition specifically for military requirements documents. This method uses the existing core library to match the AC and word segmentation, and divides the terms to be identified in the data set into three types. By summarizing the possible word-building rules of military terms, finding words in the document that meet the word-building rules as the term candidate set, using the core library and TF-IDF to calculate the value of the candidate term, and iteratively filtering the candidate terms with the value greater than the threshold as the real term. The recognition results of the two methods are combined as the final result. The experimental results show that the F1 value of this method reaches 0.719, which has a good recognition effect. However, the recognition effect of this method on the third type of terms is not ideal. In the next work, you can consider the use of machine learning methods to further improve the effect of term recognition.

Acknowledgments

The project is funded by National Key Laboratory for Complex Systems Simulation Fund (6142006190301) and China Postdoctoral Science Foundation (2020M683745)

References

[1] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. Semantic Web, 2016, 8(3) : 489-508.
[2] LiuQiao L Y, DuanHong L Y. Knowledge graph construction techniques[J]. Journal of computer research and development, 2016, 53(3): 582.
[3] Rani M, Dhar A K, Vyas O P. Semi-automatic terminology ontology learning based on topic modeling[J]. Engineering Applications of Artificial Intelligence, 2017, 63: 108-125.
[4] Wong W, Liu W, Bennamoun M. Tree-traversing ant algorithm for term clustering based on featureless similarities[J]. Data Mining and Knowledge Discovery, 2007, 15(3): 349-381.
[5] Bouamor D, Semmar N, Zweigenbaum P. Identifying bilingual Multi-Word Expressions for Statistical Machine Translation[C]//LREC. 2012: 674-679.
[6] Yuan Y, Gao Y, Zhang Y, et al. Cross-lingual Terminology Extraction for Translation Quality Estimation[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.
[7] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.
[8] Baralis E, Cagliero L, Mahoto N, et al. GRAPHSUM: Discovering correlations among multiple terms for graph-based summarization[J]. Information Sciences, 2013, 249: 96-109.

[9] Gong Y, Luo H, Zhang J. Natural language inference over interaction space[J]. arXiv preprint arXiv:1709.04348, 2017.

[10] Xiao-Yan Zhu. Regulation of Military Terminal [J]. Chinese technical language, 2020, 22 (04): 52.

[11] Nobata C, Collier N, Tsujii J. Automatic term identification and classification in biology texts[C]/Proc. of the 5th NLPRS. 1999: 369-374.

[12] Church K, Hanks P. Word association norms, mutual information, and lexicography[J]. Computational linguistics, 1990, 16(1): 22-29.

[13] Justeson J S, Katz S M. Technical terminology: some linguistic properties and an algorithm for identification in text[J]. Natural language engineering, 1995, 1(1): 9-27.

[14] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the c-value/nc-value method[J]. International journal on digital libraries, 2000, 3(2): 115-130.

[15] Arora C, Sabetzadeh M, Briand L, et al. Automated extraction and clustering of requirements glossary terms[J]. IEEE Transactions on Software Engineering, 2016, 43(10): 918-945.

[16] Shang J, Liu J, Jiang M, et al. Automated phrase mining from massive text corpora[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(10): 1825-1837.

[17] Aho A V, Corasick M J. Efficient string matching: an aid to bibliographic search[J]. Communications of the ACM, 1975, 18(6): 333-340.

[18] Xue N. Chinese word segmentation as character tagging[C]/International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing. 2003: 29-48.

[19] Chen X, Shi Z, Qiu X, et al. Adverahsarial multi-criteria learning for chinese word segmentation[J]. arXiv preprint arXiv:1704.07556, 2017.

[20] Knuth D E, Morris, Jr J H, Pratt V R. Fast pattern matching in strings[J]. SIAM journal on computing, 1977, 6(2): 323-350.

[21] Yi-biao Huang. Comparative study of open source Chinese chorch [D]. Zhengzhou University, 2013.

[22] Agirrezabal M, Arrieta B, Astigarraga A, et al. POS-tag based poetry generation with WordNet[C]/Proceedings of the 14th European Workshop on Natural Language Generation. 2013: 162-166.

[23] Gonzalez-Mullier I, Gros C, de France E. LEXTER, a Natural Language Processing Tool for Terminology Extraction[J].

[24] Yan Luo, Shu-Liang Zhao, Xiao-Chao Li, Yu-Hui Han, Ya-Fei Ding. Text keyword extraction method based on word frequency statistics [J]. Computer application, 2016, 36(03): 718-725.

[25] Hao Wang, Mi-Ping Wang, Xin-Ning Su. Study on Chinese Patent Terminology Extracted for Body Learning [J]. Intelligence journal, 2016, 35(06): 573-585.