Research on Feature Extraction Technology of Japanese Corpus Resources Based on Rule Matching

Lele Zhang*
Guangdong Ocean University, Zhanjiang Guangdong 524088 China
*Corresponding author: liu_weijy@163.com

Abstract. Existing methods of relation extraction can be divided into pattern matching-based methods, dictionary-driven methods and machine learning-based methods, among which machine learning-based methods are the mainstream methods of relation extraction at present. Based on the analysis of Snort's new features and existing rule matching methods, this paper effectively combines Snort algorithm with Japanese language features, and corrects keyword deviation caused by word frequency dependence by introducing multiple eigenvalues such as position factor and part-of-speech factor. Experimental results show that the introduction of part-of-speech factor and position factor is simple and efficient, and can effectively improve the extraction effect of Japanese keywords, which is suitable for short article keywords.

Keywords: Rule matching; Japanese corpus; Feature extraction

1. Introduction
With the rapid growth of network data, how to make people get the information they need more conveniently and quickly becomes more important. In the past, due to the limitation of research methods, the traditional Japanese research mainly focused on introspection and sentence-making by linguists. However, with the improvement of the performance of language research instruments (especially computers), large-scale statistics of language phenomena can be carried out, which makes Japanese research more and more sophisticated [1]. The accuracy of setting the deep case of case phrases in the analysis process has a great influence on the accurate generation of the final machine translation results [2]. The sample has overall statistical characteristics, and the study of the language materials used in practice can better reflect Japanese culture and understand standard Japanese. How to effectively and accurately improve the retrieval efficiency of article content has always been a research hot spot. The keywords of an article often reflect the theme of the article and are highly related to the content. It is very important for information retrieval to extract keywords accurately and quickly.

Corpus refers to a database for storing linguistic factual materials (including written and spoken languages). In recent 30 years, with the continuous development of computer technology and network technology, the construction of corpus has entered an unprecedented development period, and many large-scale corpora have been built one after another. Snort is a portable open source NIDS, which can complete the functions of real-time traffic analysis and testing IP packet login on the network, complete protocol analysis and content matching, and can be used to detect various attacks and
In essence, Snort is an intrusion detection system based on rule matching, that is, for each intrusion, its characteristic values are extracted and written into detection rules according to the specifications, thus forming a rule database. Among them, rule matching is the core processing module of Snort, and if its processing speed can be improved, the overall performance of Snort will be effectively improved.

2. Snort's Rules

In the implementation of Snort, there are four relatively time-consuming operations: capturing data packets from network transmission medium, analyzing data packet structure, matching rules and checking each data packet. Rule header corresponds to RTN (Rule Tree Node), which contains some public information such as action, protocol, source (destination) address and source (destination) end, and data flow. Snort links these rules with the same conditions into a set and describes them with RTN structure. Otherwise, if a single pattern matching algorithm is adopted, serial matching with OTN nodes is carried out downwards; If multi-pattern matching algorithm is adopted, parallel matching is performed with OTN nodes downward. The goal of Modern Japanese Written Language Balanced Corpus (BCCWJ) is to build a balanced corpus, providing users with language samples with wide coverage, strong representativeness and sufficient quantity, which can fully reflect the use of modern Japanese written language [4]. From different perspectives, the same corpus can be classified into different categories. For example, Brown Corpus is not only a written corpus, a multi-style corpus, a sampling corpus, a synchronic corpus and an uncoded corpus, but also a closed corpus.

The rule header linked list is shown in Figure 1. On the basis of the rule header linked list, the condition parameters are combined into a new top-down linked list to construct a cross linked list structure. Cancel the original list of rule header functions in RTN, and move them up to List Head. Parameter chain includes: address parameter chain and terminal 13 parameter chain. As mentioned above, the address chain can be divided into several classes. The processing function only processes the first parameter in a certain class, and other parameters refer to the processing results of this parameter.

![Figure 1](image_url)

**Figure 1** RTN linked list structure of rule header in improved scheme

Two-way maximum matching method is fast, open in vocabulary, simple in format and easy to expand, and because word segmentation errors are mostly concentrated on common words, it does not affect the extraction of domain feature words, and the correct rate of word segmentation results can be fully accepted for extracting feature words [5]. Taken together, the elements of the rules are "logical and"; At the same time, each rule in the rule base is a big "logical OR" relationship. Its characteristics are both flexible and powerful, and its flexibility means that it can formulate corresponding detection rules and options according to different behaviors. It not only has a wide detection range, but also defines what response should be made when an attack is detected. Each OTN (Optional Tree Node) node contains all options corresponding to a rule, and
also contains a set of function pointers, which are used to match these options. When the packet header and OTN are completely matched, a rule is triggered, an alarm is generated, and then the next packet is checked. That is to say, the semantics of idioms is not a simple superposition of the semantics of constituent elements, but an integrated semantics formed under the restriction of certain laws.

3. Bilingual Collaborative Training Based on Japanese Corpus

3.1. Bilingual collaborative training

Collaborative training is a semi-supervised learning framework, and it is also a semi-supervised learning framework based on bootstrapping. It uses views under two different independent features to train two classifiers. Each classifier classifies its own data, and adds the top N data with the highest confidence in each classification result to each other's training corpus. One is a mathematical model based on statistical theory and corpus to explain language phenomena; The other is the method of introducing language rules induced by linguists into language processing [6]. Combining the parts of speech and the position of words in the text, we calculate the comprehensive weight and sort them from high to low, and extract words as keywords. Special software must be used, otherwise it is difficult to retrieve strings larger than one word. It is difficult to make accurate predictions because of the different needs of various language studies. If there is a large enough Japanese tree database as a support, the extraction of verb lattice frame will become simpler. In reality, it is difficult to obtain a large enough Japanese tree database, but there are many large enough Japanese students' corpora. With the continuous expansion of corpus and the improvement of application technology, it seems that it has become a general trend for linguistic research to rely on corpus. However, when we use corpus, we can't ignore the role of researchers' introspection and intuition. Add a pointer to the OTN structure to point to the option parameter linked list corresponding to this rule. Besides the parameter content, the specific parameter structure also has a horizontal pointer (pointing to the next parameter in this rule) and a vertical pointer (pointing to the next parameter of this plug-in). A student directory is built under this corpus, and a sub-corpus is built by students, and irrelevant or inappropriate content is screened out by teachers to form a collaborative work corpus with shared resources. However, the entries in the cut-off list have a great influence on the extraction of domain words, especially when researchers focus on the algorithm improvement of feature extraction and feature weight calculation, but ignore the features themselves (i.e., entries) for extraction and weight calculation. Through annotation mapping, find the corresponding examples in the Japanese corpus, and add them to the training corpus of Chinese and Japanese classifiers respectively, and then continue to train and iterate until all the examples are iterated [7].

3.2. Label

From the point of view of developing natural language processing system, most of the language theories established by linguists so far only describe individual language phenomena, and can not solve the problems of the whole language system. Accuracy is the ratio of the number of relevant documents retrieved to the total number of all relevant and irrelevant documents retrieved, also called precision rate. The ratio of the number of relevant documents retrieved to the total number of all documents retrieved and not retrieved in the document library is also called recall rate. Multi-word query is especially aimed at the situation that there are many ways to write a word, so as to query the sentence containing each keyword according to the user's input of multiple keywords. The first step of annotation mapping is annotation. Firstly, a series of entities are found in Japanese sentences of parallel bilingual texts by using named entity recognition technology, and every two entities form an entity pair, which is generally called an instance. Then mark the example. When rules are matched, the order of matching subfunctions is followed first, and then only the conditional parameters in different rules are processed in each function, instead of calling the same function over and over again, which can save the system time of repeatedly calling functions.
After starting Snort for some time, the active rule set has been separated. Then we should consider how to dynamically maintain such an active rule set in real time. Up to now, the corpus of idioms has only marked parts of speech, word collocation, etc., but has not made a deep analysis of the semantic formation mechanism of idioms. At present, language theories generally introduce the concept of words to describe the connection between words. However, there are many linguistic phenomena that cannot be explained in this way. The designated position in collocation query refers to the words or parts of speech where the user defines the first three or the last three lexical positions of keywords to meet higher search requirements. The above work mainly uses existing tools to help detect and mark the relationship between examples in English sentences in parallel bilingual texts. Japanese is different from European and American languages. Usually, words in European and American languages are separated by spaces, with clear units and easy retrieval. However, Japanese is not the case. Only by dividing words by assigning codes can we handle them conveniently.

3.3. Mapping

After completing the tagging task of Japanese sentences in Chinese and Japanese corpora, it is necessary to map the tagging content to another corresponding language. This paper maps from Japanese to Chinese. A person’s thinking is limited, so the analysis made by linguists through introspection and their own sentences will inevitably have omissions. Although lucene retrieval engine has many advantages, because it needs to connect with database, the database is stored in traditional hard disk, and the speed difference between hard disk retrieval and memory retrieval is still thousands of times. It is old in language and cannot meet the needs of modern Japanese studies. The key lies in its ability to distinguish fields. Only by putting words in the whole classification system and comparing the differences between words in this field and other fields can this ability be better obtained. Therefore, through statistical analysis of idiom corpus, this paper obtains the frequency and part of speech of the head words and idioms, and uses cognitive model to explain the semantic generation of idioms. On this basis, the semantic function of idioms is taken as the index to classify.

Start Snort for a period of time, for example, one week or one month. It is assumed that the sequence numbers of the rules whose matching frequency is not zero are \(1,2,\ldots,i\) in turn, and \(M\) represents the matching frequency of the corresponding rules. At this time, the rules under the active rule set meet the following requirements:

\[
M_r \neq 0, \text{ among them, } s = 1, \ldots, i \quad (1)
\]

\[
M_r \neq 0, \text{ among them, } s = i + 1, \ldots, n \quad (2)
\]

Our goal is to separate and maintain such an active rule set, that is, the maximum sequence number \(s\) in the active rule set should satisfy \(s \leq i\).

The articles and sentences are loaded into memory through serialization, and searched directly in memory. Instead of searching each one next to each other, an appropriate index is created. In our text corpus, the basic type of each Japanese vocabulary is selected as the key, and the value indicates the id of the sentence where the vocabulary appears. Here, the similarity of verbs is calculated based on the statistics of lexical co-occurrence [8]. Based on the idea of the same distribution of verbs of the same type and the grammatical features of the postposition of Japanese verbs, the analytic syntax tree of raw corpus is scanned. However, there are great differences in language expression between different subjects and genres, which are not only manifested in vocabulary, but also in fixed collocation and the use of sentence patterns. In Japanese computer processing, corpus is mainly used to extract mathematical models and train and learn the system. Then, words with similar maximum length are matched in the corresponding Chinese sentences in the parallel bilingual corpus. If the same words are found only in one place, the examples in Japanese sentences are mapped to the same or similar examples found in Chinese.
4. Experiment and Analysis
In the research of Japanese computer processing, corpus is mainly used to extract language information and mathematical models needed in developing information processing software and train these software. The experimental data are 1000 pieces of Japanese news information extracted in advance, and 500 pieces of Japanese news are selected as test cases. There are about 210 short Japanese news articles in 500 Japanese news articles, which are grouped accordingly. The 500 Japanese news articles are divided into 200 articles, 100 articles, 90 articles, 50 articles, and 60 articles respectively. The purpose is to verify whether the improved algorithm described in this paper will always be better than the traditional Snort algorithm through the increase of test data.

Of course, compared with the quality of corpus, it is more important to use corpus reasonably and efficiently. In order to ensure the authenticity of the research results, it is necessary to scrutinize the corpus selection methods repeatedly and use various research methods reasonably. However, the continuous expansion of corpus makes researchers' workload surge. Idiom corpus can be used to systematically study the characteristics of idioms such as word frequency, language structure, meaning expression and pragmatics, and provide real, reliable and abundant data resources for various language studies, especially for quantitative or probabilistic study of idioms. Using Japanese corpus resources to search keywords, verify whether the corresponding Japanese news articles can be retrieved. Finally, calculate the accuracy P, recall R and F factors of Snort algorithm and improved algorithm. The calculation results are shown in Table 1.

| Indicators | Algorithm and number of Japanese news articles | 30 | 50 | 70 | 30 | 50 | 70 |
|------------|-----------------------------------------------|----|----|----|----|----|----|
|            | Snort algorithm                               |    |    |    |    |    |    |
|            | This paper improves the algorithm              |    |    |    |    |    |    |
| Accuracy   | 0.361                                         | 0.476 | 0.481 | 0.573 | 0.842 | 0.706 |
| Recall     | 0.4                                           | 0.38 | 0.31 | 0.561 | 0.71 | 0.692 |
| F factor   | 0.334                                         | 0.401 | 0.447 | 0.503 | 0.776 | 0.74 |

According to Table 1, when the number of Japanese news articles is the same, the accuracy p, recall r, and f factors all show that the improved Snort algorithm is superior to the original single algorithm, which proves that the improved Snort algorithm described in this paper can effectively improve the accuracy of keyword extraction to a certain extent. In addition, by calculating the time of keyword extraction, it is found that the difference between the improved Snort algorithm and the traditional Snort algorithm is less than 1 ms.

5. Conclusion
The continuous improvement of computer technology and network technology provides realistic and objective conditions for us to build a corpus of idioms. We realize that researchers' introspection and feelings are lack of positivity and objectivity, and we must combine researchers' introspection with corpus to better explain and grasp idioms. In this paper, the words and phrases with obvious domain features are used as clustering objects, and in the large-scale corpus of classification system, the feature extraction method of text classification is used to cluster the words in the domain, which achieves ideal results. The classification method in this paper organizes rules, and its performance can be improved whether it adopts single pattern matching algorithm or multi-pattern matching algorithm. In the future work, we will further study the usage rules of Japanese verbs. Further research is made on how to make better use of reliable collocation information and more features to calculate verb similarity.

References
[1] Li Zhiyi, Wang Mian, Zhao Pengwu. (2017). Research on "evaluation feature-evaluation word"
pair extraction based on conditional random field model. Journal of Information, vol. 36, no. 004, pp. 411-421.

[2] He Zhaoxia, Pan Ping, Luo Hui. (2017). Research on tampering detection technology of audio signal of timbre change. China Test, vol. 043, no. 002, pp. 98-103.

[3] Song Zhiyang. (2018). Implementation of a rule-based spam filtering algorithm. Southern Agricultural Machinery, vol. 049, no. 002, pp. 137-137.

[4] Guo Bo, Feng Xupeng, Liu Lijun, et al. (2017). Bilingual collaborative Chinese relation extraction based on parallel corpus. Computer Applications, vol. 037, no. 004, pp. 1051-1055.

[5] Sun Hao. (2018). Automatic extraction method of Japanese gender role language based on conversation corpus. Japanese Knowledge, vol. 006, no. 001, pp. 48-54.

[6] Liang Yulong, Qu Dan, Qiu Zeyu. (2018). Research on speaker perception training method based on improved i-vector. Computer Engineering, vol. 044, no. 005, pp. 262-267.

[7] Dai Baoyu. (2011). Conception of Japanese Corpus Example Dictionary Compilation. Journal of PLA University of Foreign Languages, no. 02, pp. 25-28.

[8] Hao Jinmei. (2010). Construction and application of Japanese corpus based on computer collaborative work technology. Market Modernization, vol. 000, no. 008, pp. 30-31.