Design and application of computer conference+Thai corpus-oriented word segmentation consistency checking system

Minsu Kong¹, Fang Yu¹, *

¹Department of Foreign Languages, Kunming University, Kunming 650214 China

Corresponding author e-mail: konminsucn@kmu.edu.cn

Abstract. The construction of deep processing corpus is one of the hot topics in corpus linguistics and computational linguistics, and the quality of corpus directly affects the subsequent research based on this corpus. It is a very effective way to build a corpus of questions in restricted areas to help understand users’ questions. In view of the inconsistent results of word segmentation and part-of-speech tagging in Thai corpus, this paper puts forward a proofreading method and strategy based on computer conference, and designs a consistent proofreading system. This method extracts grammatical and semantic collocation information between words, judges candidate sequences by using the theory of support vector machine, and gives a segmentation result, thus improving the segmentation accuracy of Chinese corpus.

Keywords: Computer Conference, Thai Corpus, Word Segmentation Consistency Test

1. Introduction

With the rise of corpus linguistics research and the deepening of Thai information processing research, great changes have taken place in the field of natural language processing. Because the establishment of corpus needs manual verification, there will inevitably be some negligence and errors, which leads to inconsistent segmentation results of the same string in the same language environment [1]. The word segmentation of Thai text corpus is to automatically segment the corpus first, and then manually proofread the results of primary processing. An important aspect that restricts the processing quality of corpus is the consistency of part-of-speech tagging of multi-labeled words. Consistency of part-of-speech tagging means that no matter whether a word has several parts of speech, only the most suitable part of speech is tagged in the same or similar context.

At present, there are few researches on corpus consistency check at home and abroad. Literature [2] makes a preliminary analysis on the inconsistency of word segmentation in corpus. Literature [3] uses
the theory of support vector machine to judge candidate sequences and give a segmentation result, thus improving the accuracy of Chinese corpus segmentation. In order to improve the accuracy of syntactic analysis model, it is necessary to improve the tagging quality and build a high-quality syntactic tree library. Therefore, it is necessary to manually correct the automatic labeling results and check the consistency of the manually corrected results. Based on the statistical analysis of computer conference+Thai corpus-oriented word segmentation, this paper summarizes the main structural types that lead to the inconsistency of word segmentation results in the corpus, and sums up the rules for judging and proofreading word segmentation inconsistency, and verifies the effectiveness of the rules through experiments.

2. Analysis of word segmentation consistency

In the research of part-of-speech tagging, people are concerned about the research of part-of-speech tagging algorithm, but there is little research on inconsistent results of part-of-speech tagging. Through experiments, we found that character length has a great influence on the classification of strings. After studying the fields with \( L \geq 4 \), we found that 96.4\% of the inconsistencies belong to DEWS2, which is much higher than 67\% when \( L \geq 3 \), and the correct rate has reached our requirements. In the process of revising the specifications, we found that the formulation of the specifications is not as detailed as possible [4]. Some people may assume that the richer the examples are, the clearer they are. However, in the specific corpus, the syntactic structure will change with the same phrases and different language environments. Therefore, different types of applications must adopt a certain communication primitive to realize conversation. Protocol primitive is a formal description language of protocols, which is easy to describe and any expression without ambiguity is a suitable choice.

When a word is not in the dictionary, that is, when a new word appears, the number of occurrences of this new word can be specified as 0.5. In this paper, dynamic programming technology is adopted to find the maximum probability segmentation, without listing all possible segmentation of a sentence according to the dictionary in advance. Most complex sentences, or clauses and clauses use specific relation words, or clauses and clauses can use specific relation words. Specific complex sentence relations are marked by specific complex sentence relations words. We set up a computer conference on word segmentation proofreading. At present, there are more than 30 rules in the library, which will be expanded in the learning process. Some changes have been made on the original basis, and the specialized thesaurus in the financial field has been added. When searching the thesaurus, the domain thesaurus has higher priority than the general thesaurus.

In the practice of corpus proofreading, the same annotated corpus sequence with the same semantic and grammatical functions often appears, and is proofread into a variety of modified sequences. Among them, the correction errors should be corrected, while some correction sequences can be justified. It is caused by the different specifications of word segmentation software. Through the statistics and analysis of corpus, it is found that this kind of inconsistency accounts for a large proportion, so it is the focus of our research. Because of its different structure and different strings, its judgment method is different from the first two, so we will not discuss it much here. In the same context, the same string has two or more different segmentation forms, which are divided and combined. According to the standard of word formation or word segmentation processing for information processing, the segmentation forms of the string should be unified. Because of the different distance from the part-of-speech mark sequence, the parts of speech of the former and the
latter words have different influence on the part-of-speech of the part-of-speech words, which is called position attribute [5].

3. Consistency check strategy of complete parsing

3.1. Strategies for finding and resolving single-layer inconsistencies

To resolve the inconsistency caused by different understandings of single-layer structure or function, we should mainly consider correcting it by regular methods. They are the objects of control protocol, which can be expressed by Token of the library in Petri net. The corresponding domain name types can be defined in the ordered domain name group [6]. When it is not in the dictionary, it will be divided into "empty/wide". After a sentence is divided into several words by the basic algorithm, if the probability of combining words exceeds a certain empirical threshold, such continuous words can constitute an old word. They usually connect clauses and do not serve as sentence components. Such as "because, so, although, but, not only, and" and so on. However, its "division" and "combination" appear in different contexts, most of which are correct, with only a few mistakes. The segmentation results of polysemous ambiguous strings are related to the adjacent words or parts of speech of adjacent words. Therefore, we use the rules based on words or parts of speech to proofread them.

The main purpose of semantic tagging of questions is to tag the semantic information expressed by questions automatically or semi-automatically, and store it in general XML format, so as to form available semantic resources of questions for practical question understanding. In order to ensure the consistency of corpus, we hope that the proofreading results of labeled corpus sequences with the same functions will converge. Because there may be many correct segmentation forms of the same string in different contexts, we can't use the regular method to generalize here, and we should correct the wrong segmentation in a targeted way. In the corresponding sentences, they are all correctly divided. Therefore, the difficulty lies in how to determine the segmentation results of these same strings before and after the corpus, which inconsistencies are correct and which inconsistencies are wrong [7].

3.2. Strategies for discovering and resolving multi-level inconsistencies

Based on the receiving semantics of adaptive travel time, the message is postmarked, and the travel time of the message from the source to the destination is counted, that is, the statistics of the actual delay value of a certain packet is used as the reference travel time for receiving control. Some Thai languages usually do not appear as words alone, but as a part of multiple words. On this basis, the author makes the following supplements: remove the pinyin of each entry, add part of speech marks to each entry in the dictionary, and allocate the total word frequency according to the proportion of each part of speech in the training corpus. Considering the priority order of structure in the structure tree, we can make objective quantitative analysis combined with the tree library, and judge the correct labeling result by using the context information.

In the statistics of academic journals, a total of 4930 inconsistent sentences were extracted, among which 4732 were inconsistent due to ambiguity and 198 were inconsistent due to prefix and suffix words. Specific proportions are shown in Table 1.

| Table 1. Inconsistent proportion |

3
Table 2. The number and proportion of inconsistency problems

| Cause of inconsistency | Combinatorial ambiguity | Intersection ambiguity | Prefix word |
|------------------------|-------------------------|------------------------|-------------|
| Number                 | 1436                    | 3017                   | 477         |
| Proportion             | 29.13%                  | 61.20%                 | 9.67%       |

If there are tagging errors in the parts of speech before and after the part of speech marker sequence, the values at the corresponding positions need to become the original negative values. These corpora are automatically segmented first and then manually proofread once. The ability of describing core words by context gradually decreases with the relative position from near to far, that is, when inferring vacant words by known context, the near context plays a more valuable role in inference than the far context. This is the scope we have to deal with. Through the simple analysis above, we can find that the rule method can not be used for the third kind of inconsistency. Therefore, we introduce the statistical method here, and solve the inconsistency more comprehensively and accurately through the complementarity of the two methods [8].

Our corpus automatic proofreading work is carried out under the guidance of the above criteria, which improves the consistency of corpus on the basis of ensuring the correctness of corpus. Semantic tagging of questions mainly completes semantic information tagging of key words, subject words, questions, words, predicates, modal particles, predicate cases and so on. The segmentation results of polysemous ambiguous strings are related to the adjacent words or parts of speech of adjacent words. Therefore, we use the rules based on words or parts of speech to proofread them. We found that the difficulty of the problem lies in how to determine the segmentation form of inconsistent strings in the corpus. Those inconsistencies are correct and those inconsistencies are wrong, which is the first problem we need to solve. Therefore, it is more meaningful to choose the appropriate window. In addition, selecting feature items with practical significance can greatly simplify the calculation of features.

4. Design of consistency proofreading method for part-of-speech tagging

4.1. Analysis of inconsistency in word tagging

Inconsistent proofreading of part-of-speech tagging results is aimed at multi-tagged words, that is, words marked with two or more parts of speech in the corpus. In response, the questions input by users are first segmented and tagged with parts of speech, and then they are searched and matched in the semantically tagged question corpus to judge whether there is such a question. If so, the answer is directly extracted according to the question and answer tags. Therefore, we do not consider the co-occurrence of the revised model with its predecessor and successor, but only consider the statistical learning of the revised model itself. A large number of real examples of inconsistent string segmentation are obtained, and at the same time, the real context information determined as a possible segmentation form is collected, and the knowledge base of inconsistent proofreading is preliminarily established.

Data smoothing technology is used to adjust the probability estimation using the maximum likelihood rule. Firstly, it can guarantee that any probability in the model is not. Secondly, data smoothing makes the probability distribution of model parameters tend to be more uniform. The low
probability (including zero probability) is raised and the high probability is lowered. Randomly select some sentences containing words of the same category and proofread them manually. The context vector of each kind of words is calculated separately, and the context vectors of the same kind of words with the same part of speech are regarded as a class. If there is any inconsistency, the times of combination and division are weighted respectively, and the fields with small weight are changed into segmentation forms with large weight. In the analysis, if these three words are closely related, there are two structural priority relationships, namely, left priority combination and right priority combination. The synchronization and concurrent behavior of the whole conference system are described hierarchically, so that the main characteristics of the conference control channel protocol can be verified hierarchically by using the reachability and invariant analysis method of Petri net theory for the lower-level conference component Petri net model.

4.2. Proofreading strategy

In the implementation process, how to find the approximate question in the corpus for the question input by the user (called the target question) is the key to the implementation. This paper proposes two methods for calculating the similarity of question, that is, its syntactic and semantic dependence on the surrounding parts of speech. We adopt the strategy of establishing a computer conference for part-of-speech disambiguation, which consists of two parts: computer conference for feature words and computer conference for part-of-speech collocation. Inconsistent segmentation of strings in corpora constructed from real-world languages is also complicated and diverse in structure, which also shows strong regularity. A set of correction patterns used for automatic proofreading of corpus, in which the original sequences in corpus are all modified into correction sequences. The main members of the table are obtained by statistical study of the revised model.

We extracted all part-of-speech markers from 30,000 Thai corpora, and investigated them one by one. For the inconsistency of part-of-speech markers, we should judge the correct part of speech according to the context. Part-of-speech correction may lead to the change of syntactic analysis structure. Because syntactic structure and syntactic function markers are not easy to proofread automatically, it still needs to be corrected manually (Figure 1).
Because there are a lot of phenomena that the parameters are/, if they are calculated twice in this smoothing method, the distribution of probability density will tend to the words that are not counted, which is obviously unreasonable. It depends not only on some common factors, but also on the specific structure of data sets. Choosing the best hedge of k should be an adaptive strategy for data. Syntactic structures should be the same, and the order of combination should be the same. The key to judging the priority of structures in corpus is to analyze the semantic and grammatical dependence between them and the surrounding information. The number of proofreading strings is large, and the rules can not cover all possible situations. A small number of inconsistent strings can not be classified into small categories. Although the proportion is small, it has a great impact on the correct rate.

5. Test results and analysis

Firstly, the automatically marked tree database is manually corrected, and then the results of manual correction are tested for inconsistency of word segmentation, part of speech and syntactic analysis in turn. According to the different purposes of synchronization and asynchrony, the conference control part is realized by using the communication service of X and the communication interface provided by the network transport layer and the network layer. In a word, the basic idea is that if a phrase in a new sentence appears in the training library, it must be consistent with the segmentation of the training set. Set up a text file containing all relation words and their labels separately. In this paper, the positive and negative mechanical matching word segmentation method is combined with the statistical word segmentation method, and the results are refined by the knowledge and rules of place names, institutions and relational words.

In order to evaluate the effectiveness of this method in consistency checking, we developed an experimental system for consistency checking of Thai corpus based on classification. The experimental results are shown in Table 2.


| Category | Number of occurrences in corpus | Correctly processed number | Accuracy (%) |
|----------|---------------------------------|----------------------------|--------------|
| DEWS1    | 207                             | 207                        | 100          |
| DEWS2    | 8207                            | 7933                       | 96.66        |
| DEWS3    | 328                             | 267                        | 81.40        |

Through the analysis of experimental results, it is found that the model can achieve better results.

We also developed several small programs to identify new words in the test set. The first program is used to identify numbers, dates, percentages, times, loanwords, etc. The conference engine manages objects according to functional modules, and is responsible for interpreting events transmitted by the user interface through X, decoding coded data, and linking action objects to form local customer requests. The recall rate and accuracy rate are used to evaluate the consistency test results of various types. Among them, only one kind of "division" and "combination" of the same string is standard, so the number of strings that should be proofread is large. Moreover, the rules used fail to cover all possible situations, and there are a few inconsistent fields which are not classified into small classes. Although the proportion is not large, it has a great impact on the accuracy rate. As the recall rate is greatly improved, the occurrence of small probability events in part-of-speech tagging is solved to some extent.

6. Conclusion

In this paper, the phenomenon of segmentation consistency in Thai corpus is statistically analyzed and studied. According to the different characteristics of consistent strings, such as internal structural features, probability distribution characteristics, or information provided by local context mainly based on syntax, and combined with computer conferences that can be self-learning, a classification strategy combining rules and statistics is proposed. Automatic proofreading of corpus can be realized by reasonably selecting the revised schema table, and the corpus quality can be improved by four steps: preprocessing, statistics-based, rule-based and consensus-based revision. The experiment in this paper is still in the stage of closed test, and the rules are induced manually. In the future work, it is proposed to adopt clustering method, select appropriate discriminant function groups, and conduct open experiments on large-scale data. The goal is to automatically judge the consistency type of word segmentation and automatically obtain rules through machine learning.

Acknowledgments

This work was supported by Scientific Research Fund Project of Yunnan Provincial Education Department under grant no.(2020J0537).

References
[1] Huang Shuiqing, Wang Dongbo. The construction, performance and application of the subwords corpus of the People's Daily in the New Era (1)-corpus construction and evaluation[J]. Library and Information Service, 2019(22):5-12.

[2] Wang Zhigang, Fu Yiping, Cai Baiguan. Research on the consistency test method of empirical uncertainty distribution [J]. Fuzzy Systems and Mathematics, 2017, 031(003): 175-182.

[3] Tang Hongxiang, Hua Mengying. Design and application of early warning information inspection system[J]. Agriculture and Technology, 2018, 38(24):253.

[4] Xi Xuefeng, Chu Xiaomin, Sun Qingying, et al. Micro-topic structure modeling and corpus construction of Thai texts[J]. Computer Research and Development, 2017, 054(008):1833-1852.

[5] Yu Ganhua, Zhang Yajie, Huang Haijing, et al. A new method for checking the consistency of the judgment matrix in AHP[J]. Mathematics in Practice and Knowledge, 2017(22):191-200.

[6] Xu Linhong, Ding Kun, Chen Na, et al. Construction of the emotional corpus of Thai literature citations[J]. Journal of Information, 2020, 039(001): 25-37.

[7] Shan Lu, Li Guohui. Research on Time-Frequency Consistency Test Method Based on Improved Evidence Theory[J]. Automation Technology and Application, 2019, 38(01):74-78.

[8] Li Hao, Hu Min, Du Xu, et al. Design and application of mobile interactive system for large-scale academic lectures[J]. China Education Informatization, 2018, 000(010):92-96.