Tibetan interrogative sentence recognition and classification based on phrase features

Mabao Ban[1,3,4,], *Zhijie* Cai[1,2,3,4*], *Rangzhuoma* Cai[1,2,3,4], and *Rangjia* Cai[1,3,4]

[1]College of Computer Science and Technology, Qinghai Normal University, Qinghai Xining 810016, China
[2]School of Computer Science and Technology, Southwest Minzu University, Sichuan Chengdu 610041, China
[3]Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province, Qinghai Xining 810008, China
[4]Key Laboratory of Tibetan Information Processing, Ministry of Education, Qinghai Xining 810008, China

Abstract. The recognition of Tibetan interrogative sentences is a basic work in natural language processing, which has a wide application value in terms of Tibetan syntactic analysis, semantic analysis, intelligent question answering, search engine and other research fields. Employing interrogative pronouns as a entry point to analyze the phrase features before and after interrogative pronouns, the paper proposes a method for Tibetan interrogative sentence recognition and classification based on phrase features by designing a Tibetan interrogative sentence recognition and classification model based on phrase features. Experimental results show that the recognition accuracy, recall rate and F value of this method are 98.21%, 100.00% and 99.10% respectively, and the average classification accuracy, recall rate and F value are 96.98%, 100.00% and 98.39%, respectively.

1 Introduction

With the development of computer technology, the research of Tibetan natural language processing has gradually developed from word level to sentence level. Tibetan interrogative sentence is a common sentence pattern, and its recognition and classification is one of the key technologies in Tibetan syntactic analysis, semantic analysis, intelligent question answering, search engine and other tasks.

In the recognition methods of sentences and sentence patterns, the commonly used methods are rule method, statistical method and the combination of rules and statistics, etc. There are many documents on Chinese sentence pattern recognition. Literature [1-4] employs different methods to identify and classify Chinese subjective sentences, explanatory opinion sentences, opinion sentences, and graceful sentences, all of which have achieved good experimental results. In terms of Tibetan sentence and sentence pattern recognition, because there is no obvious boundary symbol in Tibetan sentence, the current

[*] Corresponding author: 1402554093@qq.com

research mainly focuses on sentence boundary recognition technology [5-14], which provides a theoretical basis for the study of Tibetan sentence boundary recognition. The research on Tibetan sentence pattern recognition and classification technology has not been reported. The research shows that identifying different sentence patterns and classifying them can improve the performance of question answering system. Analyzed the phrase features before and after interrogative pronouns.

2 Tibetan interrogative sentence recognition and classification based on phrase features

2.1 Tibetan interrogative sentence recognition and classification model

In Tibetan written language, each interrogative sentence contains at least one interrogative pronoun with distinct structural features. Taking interrogative pronouns as the starting point, this paper designs a Tibetan interrogative sentence recognition and classification model with phrase features as shown in Fig.1.

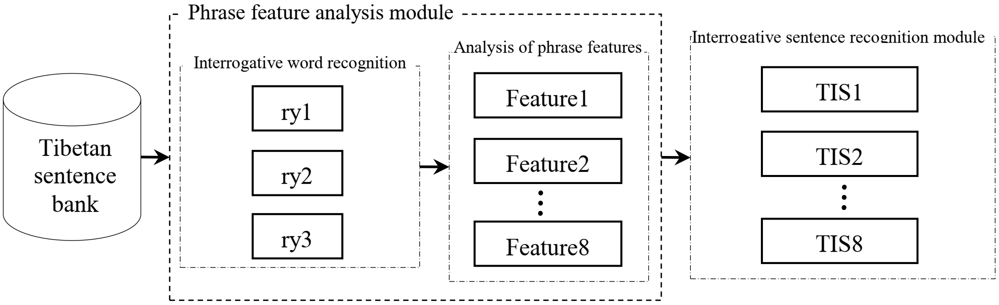
Phrase feature analysis module
Interrogative word recognition
Analysis of phrase features
Interrogative sentence recognition module
Tibetan sentence bank
ry1
ry2
ry3
Feature1
Feature2
Feature8
TIS1
TIS2
TIS8

Fig.1. Tibetan interrogative sentence recognition and classification model based on phrase features.

The Tibetan interrogative sentence recognition and classification model based on phrase features includes phrase feature analysis and question sentence recognition module. There are two parts in the phrase feature analysis module: interrogative word recognition and phrase feature analysis. In the part of interrogative word recognition, interrogative pronouns are identified by ry1, ry2, and ry3. The phrase feature analysis part obtains the phrase feature Feature1 or Feature2 or...or Feature8 of the corresponding question sentence by analyzing the phrase features before and after ry. The interrogative sentence recognition module recognizes and classifies Tibetan interrogative sentences exploits phrase characteristics.

2.2 An analysis of the features of Tibetan interrogative sentences

Tibetan interrogative sentence is a sentence pattern classified according to the mood of the sentence. It is a sentence that asks others questions about the type and nature of the things in question [15-18]. Compared with declarative sentences, imperative sentences and exclamatory sentences, Tibetan interrogative sentences have obvious differences in mood and emotional color.However, the current technology can not identify interrogative sentences according to mood and emotional color. By analyzing the structural features of Tibetan interrogative sentences, we find that each interrogative sentence contains at least one interrogative word (called interrogative pronoun ry in part of speech marker set, also known as interrogative pronoun below). Tibetan interrogative pronouns are very clear and limited in number. In order to analyze the features of interrogative sentences, we divide interrogative pronouns into three categories. The classification of Tibetan interrogative pronouns is shown in Table 1.

Table 1. Classification of Tibetan interrogative pronouns.

Serial number	type	Interrogative pronouns
1	ry1	གམ་ངམ་དམ་ནམ་བམ་མམ་འམ་རམ་ལམ་སམ
2	ry2	ཅི་ཇི་སུ་གང་དུ་ནམ
3	ry3	ཨེ

In Table 1, except for "ནམ", all the others belong to one type, and there is no multi-category problem. The type of the interrogative pronoun "ནམ" can be judged according to its position and context. When it appears after the verb, adjective or auxiliary verb, it belongs to ry1, otherwise it belongs to ry2.

Employ interrogative pronouns as an entry point, we analyze the grammatical structure and structural characteristics of Tibetan interrogative sentences. According to the different combination characteristics of interrogative pronouns and their contexts, we can divide them into general interrogative sentence (TIS1), emphatic interrogative sentence (TIS2), specific interrogative sentence (TIS3), optional interrogative sentence (TIS4), yes no interrogative sentence (TIS5), ཨེ interrogative sentence (TIS6) , Self-questioning and self-answering questions (TIS7) and multiple interrogative pronouns (TIS8) etc eight types of interrogative sentence pattern. Among the eight types of interrogative sentences, the phrases that can be used before and after the interrogative pronouns are different, and the characteristics of the Tibetan interrogative sentence phrases shown in Table 2 are obtained through statistics.

Table 2. Tibetan interrogative phrase feature table.

Sentence types	ry types	Feature types	Phrasal features	example sentences
TIS1	ry1	Feature1	S->PFG1+(ry_f+ry)	ཁྱེད་ཀྱིས་བོད་ཡིག་ཤེས་སམ།
TIS2	ry1	Feature2	S->PFG1+(མ|མི+df_b)+ry	བདག་གིས་ཡི་གེ་བྲིས་པ་མ་མཐོང་ངམ།
TIS3	ry2	Feature3	S->ry|ry+ry_b|+PFG2 S->PFG1+(ry|ry+ry_b|+PFG3) S->NP|RP|FP|TP|VP|AP|UP+(ry+ry_b)	ཅི་ཞིག་ལ་འདྲི་བྱེད་ཀྱི་ཚིག་ཟེར།
TIS4	ry1	Feature4	S->(PFG1+(ry_f+ry))+PFG3	ཁྱེད་ཀྱིས་ཞོ་འཐུང་ངམ་འོ་མ་འཐུང་།
TIS5	ry1	Feature5	S->(PFG1+(ry_f+ry))+(མ|མི+df_b)	མི་ཏོག་པད་མ་མཛེས་སམ་མི་མཛེས།
TIS6	ry3	Feature6	S->PFG1+(ry+ry_b)	ཁྱེད་ཀྱིས་བོད་ཡིག་ཨེ་ཤེས།
TIS7	ry1|ry2|ry3	Feature 7	S->PFG1+((ry_f+ry|ry+ry_b)+PFG5+cn)+PFG3	མི་ཏོག་གང་མཛེས་ཤེ་ན་པད་མ་མཛེས།
TIS8	ry1, ry2, ry3	Feature 8	S->PFG4+(ry_f+ry|ry+ry_b)+PFG5	གང་ལ་འགྲོ་རྒྱུ་ཡིན་པ་ཨེ་ཤེས་ཟེར་རམ།

In Table 2, the eight types of interrogative sentences cover all interrogative sentences in written Tibetan language, and they are not belong to multi-category. "ry" denotes interrogative pronouns, "ry1, ry2, and ry3" correspond to three types of interrogative pronouns in Table 1. "S" refers to a sentence, "S->X" indicates that s is composed of X, "|" represents or, "+" represents the combination of left and right sides, and "( )" indicates that the parts in brackets are combined first. "ry_f"refers to the word that may appear before the interrogative pronoun, that is, ry_f=vi|vt|up|uc|ub|ux|us|y|ad|uy. "ry_b"means the word that may appear after the interrogative pronoun. When ry=ry2, ry_b=gl|gz|gx|gj|up|ad|vt|vi|uc|ux|ub|us|uy|mj|mg|yy|y|hh|cd. When ry=ry3, ry_b=vi|vt|up|uc|ub|ux|us|y|ad. "df_b"refers to the word that may appear after the negative adverb མ or མི, namely df_b=vi|vt|up|ub|ux|us|y|ad. "NP" stands for noun phrase, "RP" for pronoun phrase, "TP" for time phrase, "VP" for verb phrase, "AP" for adjective phrase, "UP" for auxiliary phrase, "FP" for location phrase. PFG1=NP|RP|FP|TP|VP|AP|UP, PFG2=NP|FP|VP|AP|UP, PFG3=FP|VP|AP|UP, PFG4=NP|RP|FP|TP, PFG5=VP|AP|UP, cn=ཅི་ན|ཞེ་ན|ཤེ་ན.

2.3 Tibetan interrogative sentence recognition and classification based on phrase features

The interrogative pronouns and corresponding phrases in different types of Tibetan interrogative sentences have different features. Taking the interrogative pronoun ry as the entry point, the phrase features before and after the interrogative pronoun ry are analyzed, and the Tibetan interrogative phrase feature analysis algorithm (Algorithm 1) and the Tibetan interrogative sentence recognition and classification algorithm based on phrase features (Algorithm 2) are designed. The specific algorithm is as follows:

Algorithm 1. analysis algorithm of interrogative phrase features:

```
Function Feature_analysis(Sent)
ry_pos=Sent.index(ry) // Get the location index of ry
if(Sent[ry_pos]==ry3) // The interrogative pronoun ry is of type ry3
    if(Sent==PFG1+(ry+ry_b))
        Feature←Feature6;
    else if(Sent==PFG1+((ry_f+ry|ry+ry_b)+PFG5+cn)+PFG3)
        Feature←Feature7;
    else if(Sent==PFG4+(ry_f+ry|ry+ry_b)+PFG5)
        Feature←Feature8;
    else Feature←NULL;
else if(Sent[ry_pos]==ry2) // The interrogative pronoun ry is ry2
    if(Sent==ry|ry+ry_b|+PFG2 or Sent==PFG1+( ry|ry+ry_b|+
        PSG3) or Sent==NP|RP|FP|TP|VP|AP|UP+(ry+ry_b))
        Feature←Feature3;
    else if(Sent==PFG1+((ry_f+ry|ry+ry_b)+PFG5+cn+PFG3)
        Feature←Feature7;
    else if(Sent==PFG4+(ry_f+ry|ry+ry_b)+PFG5)
        Feature←Feature8;
    else Feature←NULL;
else  // The interrogative pronoun ry is ry1
    if(Sent[1:ry_pos-2]==PFG1)
        if(Sent[ry_pos-1:Sent(length)]==(ry_f+ry))
            Feature←Feature1;
        if(Sent[ry_pos-1:Sent(length)]== (ry_f+ry))+PSG3)
            Feature←Feature4;
        if(Sent[ry_pos-1:Sent(length)]==(ry_f+ry))+(མ|མི+df_b))
            Feature←Feature5;
    else if(Sent==PFG1+(མ|མི+df_b)+ry)
        Feature←Feature2;
    else if(Sent==PFG1+((ry_f+ry|ry+ry_b)+PFG5+cn)+PFG3)
        Feature←Feature7;
    else if(Sent==PFG4+(ry_f+ry|ry+ry_b)+PFG5)
        Feature←Feature8;
    else Feature←NULL;
reture Feature;
```

The function of Algorithm 1 (Feature_analysis) is to analyze the phrase features before and after the interrogative pronoun ry according to Table 2, and match the phrase features of the eight types of interrogative sentences in Table 2. If they match, return the corresponding phrase feature type, otherwise return NULL.

Algorithm 2. Tibetan interrogative sentence recognition algorithm based on phrase features:

```
Input: Input_file //Input_file is Tibetan sentence bank TS
Output: Output_file //Output_file is Tibetan interrogative sentence bank TIS
read(Input_file,Sent)
while (not Eof(Input_file)) // Traverse each sentence
```

```
    {if(ry in Sent) //ry representative Interrogative pronouns
        {Feature =Feature_analysis(Sent);
         if(Feature=Feature[i];i = 1,2,…,8)
             TIS[i]←Sent; // Recognition and classification of interrogative sentences}
     read(Input_file,Sent);}
reture TIS; //TIS is Tibetan interrogative sentence bank
```

The function of Algorithm 2 is to call Algorithm 1 for each sentence containing interrogative pronoun ry in Tibetan text Input_file to obtain its corresponding phrase feature Feature, identify the interrogative sentence based on the returned phrase feature Feature, and classify it into Corresponding interrogative sentence library TIS1-TIS8.

3 Experiment and data analysis

3.1 Experimental data description and experimental design

In order to experimentally verify the performance of the feature-based Tibetan interrogative recognition and classification method, 5200 sentences including declarative sentences, interrogative sentences, exclamation sentences and imperative sentences were selected from the Tibetan corpus established by the research group as the experimental corpus. Among them, there are 1100 interrogative sentences. The distribution of various types of interrogative sentences is shown in Fig. 2.

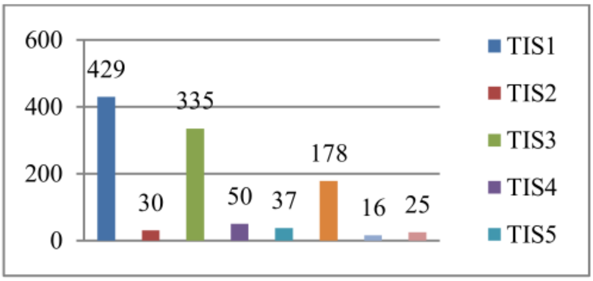

Fig. 2. Distribution of interrogative sentence types in experimental corpus.

Because the Tibetan interrogative sentence recognition and classification method based on phrase feature can recognize Tibetan interrogative sentence and classify it at the same time. In order to verify the effectiveness of the algorithm, we design two groups of experiments to test the recognition and classification performance of Tibetan interrogative sentence. Moreover, because there is no literature report on the research of Tibetan sentence pattern recognition technology, this article failed to compare and analyze the experimental data and results in other literature during the experiment. The experimental results of the two groups are shown in Tables 3 and 4, respectively.

Table 3. Experimental results of Tibetan interrogative sentence recognition.

S-N	M-N	(S-N) ∩(M-N)	P%	R%	F%
1100	1120	1100	98.21	100.00	99.10

In Table 3, S-N represents the number of interrogative sentences in the experimental corpus, M-N represents the number of interrogative sentences identified by the algorithm in this paper, and (S-N) ∩ (M-N) represents the intersection of S-N and M-N, that is, the number of interrogative sentences correctly identified by this algorithm.

Table 4. Experimental results of Tibetan interrogative sentence classification.

C	S_i	M_i	$S_i \cap M_i$	P%	R%	F %
TIS1	429	429	429	100.00	100.00	100.00
TIS2	30	30	30	100.00	100.00	100.00
TIS3	335	351	335	95.44	100.00	97.67
TIS4	50	50	50	100.00	100.00	100.00
TIS5	37	37	37	100.00	100.00	100.00
TIS6	178	178	178	100.00	100.00	100.00
TIS7	16	19	16	84.21	100.00	91.43
TIS8	25	26	25	96.15	100.00	98.04
Average				96.98	100.00	98.39

In Table 4, C represents the type of interrogative sentence, S_i represents the number of the i-th (i=1, 2,..., 8) type interrogative sentences in the experimental corpus, and M_i represents the number of the i-th type interrogative sentences identified by the algorithm in this paper. $S_i \cap M_i$ represents the intersection of S_i and M_i, that is, the number of sentences correctly classified by the algorithm in this paper.

3.2 Analysis of experimental results

Experiment 1 The experimental data in Table 3 show that the Tibetan interrogative sentence recognition method based on phrase features has achieved good recognition effect, which basically meet the practical needs. Because some declarative sentences contain not only interrogative pronouns, but also the phrasal features of these declarative sentences are the same as those of interrogative sentences, so this kind of declarative sentences are identified as interrogative sentences, and the situation of $M_i > S_i$ occurs.

Experiment 2 According to the experimental data in Table 4, the average classification accuracy, recall rate and F value of Tibetan interrogative sentences reached 96.83%, 100% and 98.38%, respectively. Except for three sentence patterns of TIS3, TIS7 and TIS8, the other classification evaluation indexes have reached 100%, which indicates that the classification accuracy, recall rate and F value of Tibetan interrogative sentence are 96.83%, 100% and 98.38%, respectively The classification of Tibetan interrogative sentences based on phrase features has also achieved good results. The reason that affects the classification of tis3, TIS7 and TIS8 is that the phrase features of these three types are the same as those of declarative sentences with interrogative pronouns.

4 Conclusion

It is the key technology and premise of syntactic analysis, text classification and sentiment analysis to identify different sentence patterns and conduct targeted research. This paper designs a Tibetan interrogative sentence recognition model based on phrase features, analyzes the phrase features before and after the interrogative pronouns, and proposes a Tibetan interrogative sentence recognition and classification method based on phrase features. In order to verify the effectiveness of this method, we design two groups of experiments to examine the performance of interrogative sentence recognition and classification. The experimental results show that the recognition accuracy, recall rate and F value of interrogative sentences are 98.21%, 100.00% and 99.10% respectively, and the average classification accuracy, recall rate and F value are 96.83%, 100% and 98.38%, respectively.

The National Natural Science Foundation of China (618666032,61966031,61662061,61063033), National Key Research and Development Project, Projects funded by the Department of science and technology of Qinghai Province (2019-SF-129), innovation team funding project of "Yangtze River scholars and innovation team development plan" (IRT1068),Qinghai Provincial Key Laboratory Project (2013-Z-Y17, 2014-Z-Y32, 2015-Z-Y03), Key Laboratory of Tibetan information processing and machine translation (2013-Y-17).

References

1. L. Peiyu, X. Jing, F. Shaodong, et al. Subjective sentence recognition based on Hidden Markov model. Chinese Journal of information technology, 2016, **38**(4): 206-212 .
2. H. Yu, P. Da, F. Guohong. Chinese explanatory opinionated sentence recognition based on auto-encoding feature. Journal of Peking University (NATURAL SCIENCE EDITION), 2015, **51** (2): 234-240.
3. Z. Jie, W. Run. Recognition of opinion bearing sentences in microblogs based on new words extension and feature selection. Journal of information technology, 2013,32 (9): 945-951.
4. F. Ruiji, W. Dong, W. Shijin, et al. Elegart sentence recognition for automated essay scoring. Chinese Journal of information technology, 2018,**32** (6): 88-97.
5. W.M. Lengzhi. Tibetan sentence boundary recognition method based on end part of speech. Qinghai Normal University, 2016.
6. Zh. Weina, L. Huidan, Y. Xin. The Tibetan sentence boundary identification based on legal texts. National Youth computational linguistics symposium, 2010.
7. M. Weizhen, W.M. Zhaxi, Nima zhaxi. Method of identification of Tibetan sentence boundary. Journal of Tibet University (NATURAL SCIENCE EDITION), 2012, **27** (2): 70-76.
8. Z. Weina, Y. Xin, L. Huidan. Method of identification of Tibetan sentence boundary. Chinese Journal of information technology, 2013,**27** (1): 115-120.
9. L. Lin, L. Congjun, J. Di. Tibetan functional chunks boundary detection. Acta Sinica Sinica, 2013,**27** (6): 165-169.
10. C. Zangtai. Research on the automatic identification of Tibetan sentence boundaries with maximum entropy classifier. Computer engineering and science, 2012, **34**(6): 187-190.
11. L. Xiang, C. Zangtai, J. Wenbin. A maximum entropy and rules approach to identifying Tibetan sentence boundaries. Acta Sinica Sinica, 2011, **25** (4): 39-45.
12. R. Qingji, A.J. Cairang.Research on automatic recognition method of Tibetan sentence boundary. Information and computer (theoretical Edition), 2014 (8): 62-63.
13. Z. Xiji, L. Ba. Based on function words and sentence patterns Tibetan sentence extraction method. Journal of Northwest University for Nationalities: Natural Science Edition, 2018, **39** (04): 39-43.
14. Q.C. Zhuoma H.Q. Cairang C.R. Dangzhi, et al. Tibetan sentence boundary recognition based on mixed strategy. Journal of Inner Mongolia Normal University: Chinese version of natural science, 2019, **48** (5): 400-405.
15. J. Taijia. General theory of modern Tibetan grammar. Gansu Nationalities Press, 2000.
16. J. Taijia. Tibetan syntax research. China Tibetology press, 2013.
17. Anonymous. Introduction to Tibetan grammar. Sichuan people's publishing house, 2014.
18. B. Mabao, C. Zhijie, L.M. Zhaxi. Tibetan interrogative sentences parsing based on PCFG. Chinese Journal of information technology, 2019, **33** (2): 67-75.