Boundary Identification of Preposition Structure with Preposition "Dao" for Information Processing

Yinghua Guo[1,*], Yun Tian[1], Xiaoxia Wang[1]

[1]School of Humanities, Shandong Agricultural and Engineering University, Jinan, China

*Corresponding Author

Keywords: Preposition, Boundary, Identification, Part of speech string

Abstract: For information processing, determining the boundary of preposition structure with preposition "Dao" can simplify the structure of a sentence and reduce the complexity of syntactic analysis. In this paper, the structure is identified by using boundary features and theory of part of speech string mutual information of preposition structure through analysis of lots of corpus, and the specific operation steps are summarized, so as to provide reference for the study of other prepositions.

1. Introduction

Chinese information processing can be divided into automatic word segmentation, part of speech tagging, syntactic analysis, pragmatic analysis and semantic analysis, etc. Syntactic analysis is the core of natural language understanding. Identifying the prepositional phrase with preposition "Dao" is to identify the prepositional phrase with preposition "Dao" in the Chinese text as a whole without analysis of its internal structure, which is a shallow syntactic analysis. The prepositional phrase boundary is identified and analyzed to simplify the structure of the sentence with prepositional phrase and reduce the complexity of syntactic analysis, laying a foundation for further chunk analysis and complete syntactic analysis. As for preposition boundary identification, many references in theory and method were provided by predecessor, based on which the boundary of the preposition structure with preposition "Dao" is identified in this paper with the corpus *Writer's Digest* (160 issues totally) made from 1990 to 2002.

2. Preposition Structure with Preposition "Dao"

Prepositions are words that express location, time, orientation, quantity and degree, etc., and they are positioned before the noun composition to constitute a preposition structure. The preposition structure is positioned before a verb for an adverbial modifier and after a verb for a complement. The preposition "Dao" can be used as an adverbial modifier or a complement. "Dao" is generally used to express time, location, direction and beginning-end. The preposition structure with preposition "Dao" analyzed in this paper is mainly as follows.

2.1 "Dao" + Np

"NP" here means a word, phrase and sentence expressing time.

(1) Mao Yuanzhi met his uncle Mao Zedong on the day of [arriving] in Yan'an.

(2) [Later], she performed better with age.

"Dao" expresses a time trend to modify the composition after it as an adverbial modifier, and "Dao" is used as a preposition.

2.2 (S) + "Dao" + Np + Vp

(S) + Dao + NP + VP means "Dao" + noun composition + verbal composition or subject + "Dao" + "noun composition + verbal composition. The noun composition here expresses time, location or orientation.

(3) [Go] to the rear.

  DOI: 10.25236/icallh.2020.016

(4) There must be a letter under the door of the office early [when] Ms. C comes to work on Monday.

2.3 Verb + "Dao" + Np

(5) They first flew [to] Shanghai to visit Japan Marine Corps Headquarters.

(6) They usually go to work in the company at noon and work [till] midnight.

The words after "Dao" in (5) express the location composition while in (6) express the time composition and quantity structure. These compositions, time compositions and quantity structures that express location and orientation are not the action objects. They are used for supplementary explanations of the time or duration of the action and its quantity and only as complements. If deleting "Dao", the preposition structure does not exist, and the sentence becomes not smooth or changes the original meaning. In addition, the words after "Dao" are noun compositions expressing degree, and "Dao" in these sentences cannot be removed, so "Dao" is a preposition.

2.4 Adjective + "Dao" +Np

(7) But he is too kind [to] be tolerated.

"Dao" in (7) is not closely combined with the adjective before it, and the speech is paused before "Dao"; the words after "Dao" are used to explain the adjective before it. Without "Dao", the meaning of the sentence will be difficult to be understood, so "Dao" in the sentence of adjective + "Dao" + NP is a preposition.

3. Boundary of Prepositional Phrase with Preposition "Dao"

In order to identify the boundary of the prepositional phrase "Dao + X", you need to understand what is the boundary of prepositional phrase first. The so-called boundary identification is to indicate the left and right boundaries of a preposition structure, i.e. finding the whole preposition structure in the sentence with correct segmentation and part of speech tagging.

(8) In 1956/t, /w Wan/m Man/ag was about to/d finish/v his studies/n, /w after which he had to/d return/v [to/p Bulgaria /ns] /v / y, /w so they/r must/d make/v a choice/v. /w

When the sentence "/w after which he had to/d return/v [to/p Bulgaria /ns] /v / y, " is searched for information processing, "[" is tagged before "Dao" and "]" is tagged after "Bulgaria / ns" to form the whole prepositional structure "[to/p Bulgaria/ns]", and the left boundary of the preposition structure is the preposition itself. The purpose of this paper is to determine the right boundary of the preposition structure.

In order to facilitate the automatic identification, we use two concepts of "internally-related word" and "externally-related word" put forward by Wu Yunfang. The former is the last word in the preposition structure, while the latter is the first word in the right of the preposition structure. The internally-related words are divided into the part of speech of internally-related words and part of speech of internally-related words, so are externally-related words. In sentence (8), "Bulgaria" is an internally-related word, "Qu" in the Chinese text is an externally-related word, "ns" is the part of speech of the internally-related word, and "v" is the part of speech of the externally-related word. Prepositions, internally-related words and externally-related words are the keywords for identification of preposition structure.

4. Strategies for Boundary Identification

4.1 Boundary Features of Preposition Structure "Dao + X"

The internal structure of prepositional phrase is very complex, but its internally-related words and externally-related words are regular. The study above shows that X in the preposition structure "Dao + X" is a composition that expresses time, quantity, orientation, location and degree. Some internally-related words go with the externally-related words sometimes, for example, in the structure of "Dao" + NP + VP, the preposition compositions are only those expressing location and orientation.

(9) [Go to/p Germany/ns] to negotiate/v, /w

When the part of speech of NP after "Dao" is "ns" and it goes with the verb "negotiate/v", we can determine the right boundary of this preposition structure.

4.2 Analysis of Part of Speech String Mutual Information

As a concept in the information theory, mutual information can be used to measure the degree of relevance between two events. Mutual information is often used in the computational linguistics to indicate the closeness degree of relevance between two linguistic phenomena.

In this paper, the method is used to deal with relatively complex preposition structure. Although these part of speech strings are longer and complex in structures, their internal part of speech strings are regular in distribution. When the time compositions after "Dao" are analyzed, some of them are relatively complex, so the highly frequent part of speech strings will be listed.

(10) [By/p the morning/v /n of/u the 28th day/t], /w the fierce/ad attack/v that /u lasted/v /u all/m night/q began /v to be/c gentle/v. /w

As shown in the sentence above, preposition structure is not a simple time word, so we shall count its right boundary, i.e. the occurrence probability of "/u/n", i.e. the co-occurrence probability with the preposition "Dao".

5. Algorithm Design for Boundary Identification

Analyze these compositions that serve as preposition structures and find the features of these preposition structures, so as to identify the whole preposition structures.

5.1 "Dao" + Np

NP means the time and quantity compositions. Among these compositions, some words frequently exist in the corpus, and they must be right boundaries once existing.

(11) [/w By/v] /w mid-/t April/t, /w

(12) /w Stop/v chemotherapy/vn [/w by /v] May/t 1998/t, /w

In the "Dao + t + t" format, "t+t" serves as a preposition composition, and once "Dao" goes with such format, its right boundary will go after the second "t".

Considering the relative complexity of some preposition structures, this paper only extracts two items on or before the right boundary to investigate their co-occurrence frequency as the basis for boundary identification. The sequence of internal part of speech string in the preposition structure of "by/p the beginning/f of 1944/t, one/m year/q later/f /u, /w" is m/q/f/u/t/f, and the co-occurrence frequency of the last three items, i.e. u/t/f, is investigated to reduce the complexity of the rule. For example, the co-occurrence frequency of part of speech string /u/n that expresses time composition is 0.0023.

(13) [By/p the morning/v /n of/u the 28th day/t]

5.2 "Dao" + Np + Vp

In this structure, preposition structure is a composition that expresses location and orientation, which first analyzes the simple preposition structure and uses the internally-related words and externally-related words to identify the boundary.

For example: "Dao" + place name (ns) + verb

(14) Go to/p France/ns to conduct/v philosophy/n research/vn

(15) Go to/p Hong Kong/ns to serve/v as manager/n, /w

(16) Go to/p Hefei/ns to do/v carpentry /n work /a. /w

(17) Go to/p Shijiazhuang/ns to /v take care of /v it

The co-occurrence frequency of "conduct" is 0.0002; that of "serve" is 0.0002; that of "do" is 0.0002; that of "to" is 0.002.

When the internally-related words only refer to a place name and the externally-related words are the verbs above, the preposition structure must go with the verb in its right boundary.

In addition, we also summarize the vocabulary with co-occurrence of such fixed internally-

related words and externally-related words as “Dao” + location word(s) + verb, “Dao” + location pronoun (r) + verb; “Dao” + noun (n) _+ verb; “Dao” + noun (n) + noun of locality (f) + verb.

For the remaining relatively complex preposition structures, we count the internal part of speech strings of the preposition structure to find out the identification rules. For example, the co-occurrence frequency of the part of speech string u/n/n is 0.0027; that of the part of speech string u/s is 0.0113.

(18) Go to/p the special /n tent/n in/f /u the factory/n to build/v /u a collective/n dormitory/n

(19) Go to/p the field/s outside/f /u the village/n

5.3 V+ “Dao” + Np

In such structure, the sequence of internal part of speech string of the preposition structure “Dao” + NP is investigated and counted by using relevant knowledge of mutual information, and the co-occurrence frequency of the internal part of speech string u/n/n in the preposition structure of adjective + “Dao” + NP is 0.0002.

(20) As small/a as/p the cultural/n features/n of/u a/m city/n, /w

For the co-occurrence frequency of internal part of speech in the preposition structure of verb + “Dao” + NP, which of the part of speech string n/f is 0.0189, and that of the part of speech string u/n/f is 0.0078.

(21) Go to/p a/m slumdog/n /f in the southern/j part/q of Philadelphia/ns

(22) Go to/p the wall/n side/f of/u the living room/n

5.4 Specific Identification Steps

First, identify the preposition structure in “Dao” + NP, and we use it to summarize the markers of right boundary as the right boundary of the preposition structure.

Second, after the successful identification in the first step, the unidentified sentences are identified using the part of speech string /u/n to the complex preposition structure in “Dao” + VP.

Third, use the summarized internally-related words and externally-related words in the “Dao” + place name (ns) + verb for identification.

Fourth, use the internal part of speech strings u/n/n and u/s of the preposition structure for identification.

Fifth, use the co-occurrence probability of the internal part of speech strings of the adjective and preposition structures before “Dao” for identification.

Sixth, identify the preposition structure in the format of verb + “Dao” + NP.

6. Conclusion

The purpose of this paper is mainly to identify the whole preposition structure to formulate the relation vocabulary of internally-related words and externally-related words according to the boundary features of preposition structure “Dao + X”; to investigate the internal part of speech string of the complex preposition structure “Dao” + X by using mutual information, the right boundary of the preposition structure is judged based on the co-occurrence probability of part of speech and a vocabulary is counted, so as to formulate identification rules and thus provide basis for programming.

References

[1] Wang Xia. Research on automatic recognition of Chinese verb-object collocation, application of language and writing, No.1, pp.138-144, 2005.

[2] Wu Yunfang et al. Research on Modern Chinese Parallel Structure for Chinese Information Processing, Language and Letter Application, No.2, pp.143, 2004.

[3] Wu Ziying, Zheng Jiaheng. Research on the Method of Automatic Recognition of Modern Chinese Abbreviations, Computer Engineering and Design, No.16, pp.254-256, 2007.

[4] Zhou Qiang, Sun Maosong, Huang Changning. Automatic recognition of the longest noun phrase in Chinese, Journal of Software, No.11, pp.53-59, 2000.

[5] Yu Junwei et al. Automatic recognition of Chinese prepositional phrases, Journal of Chinese Information Technology, Vol.19, No.4, pp.18-24, 2004.