Detection on Inconsistency of Verb Phrase in TreeBank

Chaoqun Duan, Dequan Zheng, Conghui Zhu, Sheng Li
MOE-MS Key Laboratory of Natural Language Processing and Speech
Harbin Institute of Technology, Harbin, China 150001
{cqduan,dqzheng,chzhu,lisheng}@mtlab.hit.edu.cn

Hongye Tan
Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education
Shanxi University, Taiyuan, China 030006
hytan_2006@126.com

Abstract

Annotating linguistic data is often a complex, time consuming and expensive endeavor. Even with strict annotation guidelines, human subjects often deviate in their analyses, each bring different biases, interpretations of the task and levels of consistency. The aim of this paper is to explore a way to find out the inconsistencies in the corpus TreeBank which is used for syntactic analysis through the procedure we study the inconsistencies of verb phrase tagging in the corpus TreeBank. At the same time, we can analyze the inconsistencies of verb phrase tagging which are found in the corpus TreeBank in order that we can find a way to improve the consistency of verb phrase tagging automatically which is effective to improve the quality of corpus.

1 Introduction

Most empirical work in Natural Language Processing (NLP) is based on supervised machine learning techniques which rely on human annotated data of some form or another. But the construction of a corpus is a complicated work. Especially for individuals, it's a more hard assignment. Generally, a large-scale and high quality corpus comes from a team and it requires working in teams and different people is responsible for a particular part of the corpus respectively. Due to that the work is cut into several parts and distributed to different persons, inconsistencies may be generated. Because everyone has an individual understanding about the same case and different people may make the different annotations. All of these may cause inconsistencies, and even errors. When we train our models with a corpus which may contains inconsistencies even errors, the models will not represent the real distribution of the problems precisely. So the work to find the inconsistencies in the corpus and to correct them is useful to improve the precision of the models, which can help us obtain more accurate results in natural language processing.

2 Related Work

At present, the research on corpus consistency is mainly concentrated on the consistency of word segmentation and part-of-speech (POS) tagging. Liu Bo, Zheng Jiaheng and Zhang Hu proposed a method to handle the consistency of word segmentation which is based on the combination of statistics and rules. They also introduced a number of strategies to handle different kinds of inconsistency [5]. Zhang Hu and Zheng Jiaheng put forward a method to check the consistency of part-of-speech (POS) tagging on the foundation of the analysis of the part-of-speech (POS) tagging which is based on the classifications of ambiguity words [11].

Besides the research on the inconsistency of word segmentation and part-of-speech (POS) tagging, some people focused on the research about the ambiguity of structure or function in Chinese corpus gradually as well. The Chinese ambiguities of structure in high frequency are divided into three basic types based on the analysis of structural ambiguity. By analyzing the ambiguities of structure, Yang Sichun and Chen Jiajun found out the causes of structural ambigui-

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 70–75,
Wuhan, China, 20-21 October 2014

ty, and proposed some strategies to remove them, especially the solution based on examples [12].

This paper aims to find the inconsistencies of verb phrase tagging in the corpus TreeBank. As we all know, some errors always exists in corpus as a kind of inconsistency. So when we find out the inconsistencies in the corpus TreeBank we will lay a foundation of finding out errors in the corpus TreeBank.

3 Terminology

- TreeBank

In this paper, we use the corpus which is named Chinese Treebank 7.0 (CTB7.0). There are 2,448 text files in this release, containing 51,447 sentences, 1,196,329words, 1,931,381 hanzi (Chinese characters). The data is provided in four different formats: raw text, word segmented, word segmented and POS-tagged, and syntactically bracketed formats.

In Chinese Treebank 7.0 (CTB7.0), the frequency of verb phrases is in second dgree, which is only less than the frequency of noun phrase. In addition, the usage of Chinese vocabulary is very flexible and a word always can act as a variety of components of the grammar in different context, especially verb, which causes a lot of grammatical ambiguity in syntactic analysis. So, we choose the verb phrases to find out the inconsistent tagging.

- Verb Phrase

The verb (including verb compound) and aspect sequence forms the verbal head that takes zero or more complement to form a verb phrase.

- Verb Head

The verb (including verb compound) and aspect sequence forms the verbal head.

- Verb Compounds

Although compounding is highly productive in Chinese, it is still considered to be a lexical process. Therefore compounds are treated in a similar fashion as simple monolithic verbs. The challenge is to clearly identify compounds and distinguish them from situations where a phrasal projection is necessary. Due to the lack of a clear standard between compounds and phrases in Chinese, we will adopt the following working criteria for verb compounds where there is a sequence of verbs: (1) they share the argument structure, (2) they share aspect markers, (3) they share modifiers, (4) and they do not fall into the clearly defined raising or control structures.

A classification of verbal compounds are shown in Table 1.

Tags	Explanation
VCD	coordinated verb compound
VCP	verb compounds formed by VV + VC
VNV	verb compounds formed by A-not-A or A-one-A
VPT	potential form V-de-R or V-bu-R
VRD	verb resultative compound
VSB	verb compounds formed by a modifier + a head

Table 1. a classification of verbal compounds

- Aspect Maker

In Chinese, the particles (e.g.了 (le), 着 (zhe), 过 (guo)) are named as aspect maker.

- Inconsistency

In Chinese Treebank 7.0 (CTB7.0), we can find the phenomenon that a verb phrase may have different annotations in different place while they are in the same context. We define this phenomenon as inconsistency.

```
(VP(VV取得)
    (NP-OBJ(NN突破性)
            (NN进展)))
```

```
(VP(VV取得)
    (NP-OBJ(ADJP(JJ突破性))
            (NP(NN进展))))
```

Figure 1. An example of the different tagging of verb phrase in Chinese Treebank 7.0

In figure 1, we can see the annotations of “取得突破性进展” are different. In the top table, the “突破性” was tagged as JJ while in the bottom table, it wsa tagged as NN.

4 Research Method

In this paper, we find out the inconsistencies by comparing the tagging of verb phrase. In this section, we mainly describe the method and the result of experiment.

4.1 The Method Based on Comparison of Tags of Verb Phrases

We divided all of the verb phrases into different categories based on the Chinese characters which were consisted of the verb phrase. Then we compared the annotations of verb phrases which belonged to the same category each other. If we found the different tagging of verb phrases in one category, there might be inconsistency in it.

Procedure.

Our goal is to find out the verb phrases in corpus that they shared the same Chinese characters while their tagging are different.

Step1: Finding verb phrases.

Firstly, we found all of the verb phrases in corpus and divided them into different categories based on the Chinese characters which were consisted of them. At the same time we recorded their provenance which contained the index of the text and the sentence. An example is shown as follows.

eg1: 一百亿 元 人民币

TAGS:(VP(VP(NP-PRD(QP(CD)(CLP(M)))(NP(NN)))))
TAGS: (VP (NP-PRD(QP(CD)(CLP(M)))(NP(NN))))
TAGS: (VP(VP(NP-PRD(QP(CD)(CLP(M)))(NP(NN)))))
TAGS: (VP(NP-PRD(QP(CD)(CLP(M)))(NP(NN))))

Step2: Finding verb phrases that appear more than once.

Secondly, after dividing all of the verb phrases into different categories, we kept the categories that contain more than one verb phrases and removed the categories that contain only one verb phrase.

eg2: “公开、公平、公正”

TAGS: (VP(PU)(VA)(PU)(VA)(PU)(VA)(PU))
TAGS: (VP(PU)(VA)(PU)(VA)(PU)(VA)(PU))
TAGS: (VP(PU)(VA)(PU)(VA)(PU)(VA)(PU))
TAGS: (VP(PU)(VA)(PU)(VA)(PU)(VA)(PU))

Step3: Finding categories appearing different.

Thirdly, after dividing all the verb phrases into different categories based on the Chinese characters which were consisted of them, we found out the categories in which there were different tagging of verb phrases.

eg3. “抓大放小”# 4#

TAGS: (VP(PP(-NONE-*T*-1))(VP(PU)(VV)(PU)))
TAGS: (VP(PU)(VV)(PU))
TAGS: (VP(PP(-NONE-*T*-1))(VP(PU)(VV)(PU)))
TAGS: (VP(PU)(VV)(PU))

Step4: Eliminating the Influence of Omitted Structure.

Fourthly, we eliminated the influence of omitted structure. In the results of step 3, we found some differences were only caused by the omitted structure. Omitted structure was related to parsing, and we didn’t take care of this temporarily. So we eliminated the categories in which the differences were only caused by omitted structure.

eg4. “抓大放小”# 4#

TAGS: (VP(PP(-NONE-*T*-1))(VP(PU)(VV)(PU)))
TAGS: (VP(PU)(VV)(PU))

In the example above, the difference are only caused by omitted structure. So we should eliminate the category.

The operation of eliminating the influence of omitted structure was on the result of step 3. Firstly, we arranged a device to store verb phrase in a tree data structure. Secondly, we used these devices to prune the omitted structure. Thirdly, we restored the devices pruned to verb phrases. Fourthly, repeated step 3. If after the operation, the tags in a category are all the same, it means that, the differences in this category caused only by the omitted structure and it should be eliminated.

4.2 Experimental Results and Analysis.

In this paper, we mainly research the first 612 texts in Chinese Treebank 7.0 (CTB7.0). We find out a total of 37416 groups of verb phrases and 2430 groups contain more than one verb phrase. In these 2430 groups of verb phrases there are 688 groups in which we find the inconsistency. After eliminating the omitted structure there are only 245 groups. And the 245 groups are the finally result. We find the 245 groups can be divided into five categories.

The first kind of inconsistency of verb phrase is that the verb phrase mark appears more than once in a verb phrase. There is such a phenomenon in the TreeBank corpus that in the outer layer of a complete verb phrase a "VP" symbol was marked repeatedly. We see this phenomenon as the first kind of inconsistency.

In the statistics about verb phrases, we took "VP" as the signal of a verb phrase. Thus, if there is a repetition of a verb phrase marked symbol , we will find two verb phrases at least in the result which satisfy the condition that they share the same Chinese character while their verb phrase tagging are different and the cause of difference is only due to the additional "(VP)". According to the idea, we have found some categories in the results of classification of verb phrases which satisfy the condition that in each of these categories there are two different tagging at least and one of them is the substring of the other one and their difference is only due to the additional "(VP)".

The second kind of inconsistency is that the type of verb compounds annotated inconsistently. There is such a phenomenon in the TreeBank corpus that the verb compounds share the same tags of part of speech in different sentences while the type of verb compounds annotation is different. And this is the second kind of inconsistency.

There are six kinds of verb compounds, which include VCD, VCP, VNV, VPT, VRD and VSB. In a verb phrase, the relative position of verb compounds is steady. Thus, the relative position of tagging corresponding to the verb compounds is steady as well. According to the fact, we have created a table for each category of verb phrases with the row standing for the index of the verb phrase and the column standing for the relative position of the verb compounds to storing the entire symbol of verb compounds in it and compared the values in column. We have found the type of verb compounds annotated inconsistently in some categories.

The third kind of inconsistency is that the tagging of phrases are not complete. We can find such a phenomenon in the TreeBank corpus that some words share the same tags of part of speech in different sentences and some of them are marked the tags of phrase while some of them are not. This is the third one.

In a category of verb phrases, if each verb phrase is marked completely, the quantity of symbol belonging to every phrase will be the same. So, if a verb phrase isn't marked completely, its quantity of symbols will less than others'. According to the description, we have arranged a device to store each verb phrase in a category in a tree data structure. The number of nodes of a tree is equal to the number of symbols of the corresponding verb phrase. We have found some categories contain inconsistency due to the lack of the tags of phrase by comparing the number of nodes of every tree in the same category.

The fourth kind of inconsistency is similar to the third one. It is also caused by not complete annotation. But what is different is that the fourth one is caused by the lack of functional tags. In the TreeBank corpus some words share the same tags of part of speech in different sentences and some of them are marked the functional tags while some of them are not.

In a category of verb phrases, all of the verb phrases shared the same part-of-speech and each verb phrase is marked completely, but the length of catenation of symbols in each verb phrase is different. It means that some verb phrase is lack of functional tags.

The fifth kind of inconsistency is caused by the different tagging of part of speech. In the TreeBank corpus there are many conversion words and their tagging of part of speech in different context are different. As a result, the verb phrases which contain them will be marked with different tags of phrase. So, we class the fifth one as the category that is caused by the different tagging of part of speech.

In a category of verb phrases, the inconsistency may result from the different part-of-speech tagging. According to the fact, we have arranged a device to store each verb phrase in a tree data structure. In the tree data structure, the parent node of leaf node is the part-of-speech tagging of corresponding to the leaf node. So, we can get the part-of-speech tagging of each Chinese character in a verb phrase from the tree data structure easily. After getting the part-of-speech tagging, we catenate all of them which are from the same tree data structure as a string. We have found some categories that contain inconsistency because of the different part-of-speech tagging by comparing the strings that belong to the same category.

In the 245 groups of verb phrases there are 224 groups can be classed as the members of these five categories. It's about 91.43% and these five categories of inconsistency cover all kinds of inconsistency nearly. There are 63 groups belong to the first category, 9 groups belong to the second category, 26 groups belong to the third category, 51 groups belong to the fourth category, and 75 groups belong to the fifth category.

Index	Category	Example	Quantity	Percentage
1	VP repetition	(VP(VP(VV 失败))) (VP(VV 失败))	63	28.13%
2	verb compounds	(VP(VCD(VV 上市)(VV 交易))) (VP(VSB(VV 上市)(VV 交易)))	9	4.02%
3	tagging of phrase	(VP(ADVP(AD 共同))(VV 努力)) (VP(ADVP(AD 共同))(VP (VV 努力)))	26	11.61
4	functional tags	(VP(VC 为)(NP(NN 团长))) (VP(VC 为)(NP-PRD(NN 团长)))	51	22.77%
5	Different POS	(VP(VCD(VV 协调)(VV 发展))) (VP(VV 协调)(NP-OBJ(NN 发展)))	75	33.48%

Table 2. The distribution of inconsistency

5 Conclusions and Future Work

In this paper, we aim at find out the inconsistency in Chinese Treebank 7.0 (CTB7.0). Besides the method described before, we also have tried to solve this problem by to using other method which is based on the assumption that if we cluster the sentences in the corpus when we set the annotations as the conditions of similarity measurement, in the result, the small-scale clusters may represent the wrong annotations. But the result is not satisfied because of inappropriate grain size. What's more, what we have finished is inadequate. For the first method, we just consider the case that the verb phrases shared the same Chinese characters which were consisted of them while their tagging are different. We need to consider other cases in future. For the Second method, the grain sizes we have chosen is not enough.

The next jobs is to try to consider other situation to find the inconsistency in Chinese Treebank 7.0 (CTB7.0). For example, the verb phrase shared the same tags of part of speech while their tags are different. What's more, we should choose a proper grain size to remedy the method based on statistic.

Reference

Bo Liu, JiaHeng Zheng, Hu Zhang. 2008. Consistency check of segment using combination

of rule and statistics. Computer Engineering and Design, 2008(29): 1814-1816

Hui Wang. 2003. A STUDY OF CHINESE WORD SENSE DISAMBIGUATION IN MT BASED ON GRAMMATICAL AND SEMANTIC KNOWLEDGE-BASES. JOURNAL OF GUANGXI NORMAL UNIVERSITY, 2003(21):86-93

Hu Zhang, Jiaheng Zheng. 2008. Consistency Check on POS Tagging of Chinese Corpus Based on Classification. Computer Engineering,2008(34):90-92

Jiang Liu, Jiaheng Zheng, Hu Zhang. 2005. Studies on the Consistency of Word Segmented Chinese Corpus. Application Research of Computers, 2005(9):52-54

Li Wei, Hongye Tan, Jiaheng Zheng, Jian Sun. 2012. Study of Keeping Consistency of Chinese Corpus of Complete Parsing. Journal of Guangxi Normal University:Natural Science Edition, 2012(28):139-142

Maosong Sun. 1999. On the consistency of word-segmented Chinese corpus. Applied Linguistics, 1999(2):87-90

Sichun Yang, Jiajun Chen. 2005. Research on Structur al Ambiguity in Chinese Automatic Syntactic Parsing. Journal of Kunming University of Science and Technology (Science and Technology), 2005(30):45-49

Xi Miao, Jiaheng Zheng. 2006. Classified Study On Inconsistency of Segment for Chinese Corpus. Journal of Shanxi University (N at. Sci. Ed.), 2006(1):22-25

Yongping Du, Jiaheng Zheng. 2001. Design and Reali zation of the Consistence Collation System in Segment and Property of Word Notation. Computer Development & Applications, 2001(10):16-18

Yin Liu. 2002. CHINESE-ENGLISH MACHINE TRANSLATION DISAMBIGUATING WITH RULE-BASED METHOD COMBINED WITH STATISTIC-BASED METHOD. Computer Applications, 2002(22):21-23

Yili Qian, Jiaheng Zheng. 2004. Research on the Method of Automatic Correction of Chinese Part-of-Speech Tagging. JOURNAL OF CHINESE INFORMATION PROCESSING, 2004(2):30-35

Yili Qian, Jiaheng Zheng. 2004. An Approach to Im proving the Quality of Part-of-Speech Tagging of Chinese Text. Itcc, vol. 2, pp.183, International Conference on Information Technology: Coding and Computing (ITCC'04) Volume 2, 2004