Technical Report of the CCID System for the 2th Evaluation on Chinese Parsing

Guangfan Sun
China Center for Information Industry Development, Beijing, 100044
morgan2001_sun@163.com

Abstract

This paper gives an overview of China Center for Information Industry Development(CCID) participating in the 2th Evaluation on Chinese parsing. CCID has taken part in the subtask of the analysis of complete sentences. The system participating in the above Evaluation is a rule-based Chinese parser, and its basic information is described in the paper, and its experimental situation for the Evaluation has been analyzed.

1 Introduction

Parsing is one of key issues in natural language processing, and its main task is to automatically identify the syntactic structure of sentences (syntactic units and their syntactic relations between units). The study of parsing is of critical importance for machine translation, natural language understanding, information extraction and automatic summarization of natural language processing systems. Syntactic analysis methods include methods of use of corpus annotation information in syntactic analysis and the rule-based methods such as: Shift-Reduce Parsing and Chart Parsing technology to study the Chinese syntactic structure[1]. In this paper, the Chinese parser which China Electronic Information Industry Development (CCID) uses to participate in the 2th Evaluation on Chinese Parsing is described.

2 System

The Chinese parser which CCID uses to participate in the 2th Evaluation on Chinese Parsing serves as a component of a practical Chinese-English machine translation system, and uses rule-based method, and uses statistical approach for unknown word recognition. The Chinese parser includes the following three modules: 1) Chinese word segmenting, 2) Chinese POS tagging, 3) Chinese parsing. The form of rules in the Chinese parser is production rule. The rules include general rules and specific rules. The general rules are indexed by POS or phrase types, and specific rules are indexed by Chinese word or Chinese phrase. There are multi-passes during Chinese parsing, and the result of the parsing of a Chinese sentence is a Chinese syntactic tree. The CCID's Chinese parser includes 1,930,000 entries in the basic dictionaries and 6,000 rules in knowledge base. Parts of speech and syntactic elements of the output of the CCID's Chinese parser are marked by its own set of markup symbols, and these markup symbols are mapped to parts of speech tags and syntactic component tags defined by CIPS-ParsEval-2009 by a conversion function. The CCID's tag set is mainly same as the set of CIPS-ParsEval-2009 except the used tag characters. For example, in the CCID's tag set, the tag of noun phrase is NP, and the tag of verb phrase is VP, and the tag of preposition phrase is IP; for the tags in CIPS-ParsEval-2009, the tag of noun phrase is np, and the tag of verb phrase is vp, and the tag of preposition phrase is pp.

3 Experiment

CCID participated in the 2th Evaluation on Chinese Parsing, and timely submitted parsing output of test sentences of the syntactic analysis. The Test Group returned to a very unfortunate message: "find that the results presented in the original segmentation of data are automatically

merged, so can not give the appropriate evaluation of data."

Here are two specific examples:

Original input 1:

1 显微 解剖学 是 用 光学 显微镜 和 电子 显微镜 等 显微镜 来 观察 细胞 、 组织 或 器官 等 的 构造 、 形态 的 分支 学科 ， 它 也 叫做 组织学 。

Parsing output 1:

1 [zj [dj [dj [np 显微解剖学/n] [vp [pp 用/p [np [np [np 光学显微镜/n] 和/c [np 电子显微镜/n]] 等/uO [np 显微镜/n]]] [vp [vp 观察/v] [np [np [np [np [np 细胞/n] 、/wD [np 组织/n] 或/c [np 器官/n]] 等/uO] 的/uJDE [np 构造/n]] 、/wD [np [np 形态/n] 的/uJDE [np 分支学科/n]]]]]]] ，/wP [dj 它/r [vp [vp [dp 也/d] [vp 叫做/v]] [np 组织学/n]]]] 。/wE]

Original input 183:

183 除此而外 ， 发作性 室上性 心动 过速 、 快速 心房 颤动 、 心房 扑动 等 ， 虽 属 快速性 心律 失常 ， 有时 也 必须 进行 心电图 监护 。

Parsing output 183:

183 [zj [dj [dp [dp 除此而外/d] ，/wP] [dj [np [np [np [ap 发作性/a 室上性/a] [np 心动过速/n]] 、/wD [np [ap 快速/a] [np 心房颤动/n]] 、/wD [np 心房扑动/n]] 等/uO] [pp 虽/c [dj [np [np 属/n] [np 快速性心律失常/n]] [vp [dp 有时/d] [vp [dp 也/d] [vp [vp 必须/vM [vp 进行/v]] [np 心电图监护/n]]]]]]]]] 。/wE]

Reasons for these phenomena are: "显微解剖学"、"光学显微镜"、"电子显微镜"、"心动过速"、"心房颤动"、"心房扑动"、"快速性心律失常"、"心电图监护" and some other entries have already existed as separate entries in the basic dictionaries of the CCID's Chinese parser. In parsing, these entries act as separate entries and the results also show up as separate entries. This occurs because of the larger basic dictionary(1.93 million entries), and these entries have the corresponding English translations on the expression. For a practical Chinese parser, a large number of phrases that already exist in the basic vocabularies can reduce the burden of parsing, and are useful for improving the success rate of Chinese syntactic analysis. But this adds extra burden to evaluation programs. When participating in the next Chinese parsing evaluation, some phrases that have existed in the basic dictionaries of Chinese parser will be divided to further analyze their internal syntactic structures to facilitate the evaluation process.

After receiving the notice that a re-evaluation can be done by the Evaluation Group to help CCID to evaluate the effectiveness of the modification of the parsing model, the following steps are carried out for the convenience of the evaluation programs:

1) Compare all words in the test task with CCID's Chinese parser, and find out the information for the words from CCID's Chinese parser, and delete all other words from the Chinese parser to avoid the situation that some Chinese words are combined when parsing.
2) Modify parsing rules that contain operations of deleting words to avoid the deletion of Chinese words in the parsing results.
3) Re-parse Chinese sentences in the test task.
4) Submit the result of the parsing to the Evaluation Group to evaluate.

The re-evaluation result is as the following:

Performance Report for Task 2-2
pos accuracy: 72.98% (19253/26381)
average of F1 of dj_sum and fj: 26.87 (%)

Label	#Auto	#Gold	#Correct
dj	3826	2290	1156
vp	5954	7397	3090
ap	532	432	267
np	5778	5199	3478
sp	0	433	0
tp	0	381	0
mp	443	614	341
mbar	47	45	29
dp	782	65	42
pp	1263	1191	546
bp	0	1	0
total	18625	18048	8949

Label	Precision	Recall	F1
dj	30.21	50.48	37.80
vp	51.90	41.77	46.29
ap	50.19	61.81	55.39
np	60.19	66.90	63.37
sp	0.00	0.00	0.00
tp	0.00	0.00	0.00
mp	76.98	55.54	64.52
mbar	61.70	64.44	63.04
dp	5.37	64.62	9.92
pp	43.23	45.84	44.50
bp	0.00	0.00	0.00
total	48.05	49.58	48.80

Label	#Auto	#Gold	#Correct
fj	450	1251	42

Label	Precision	Recall	F1
fj	9.33(%)	3.36(%)	4.94(%)

4 Discussion

Chinese parsing is an important basic research for Chinese information processing research, and gets the attention of many researchers. Current research focuses on the research on syntactic knowledge acquisition based on the corpus, and its goal is to use statistical methods from a good tree bank annotation to learn the parsing needed knowledge, and the trained parser also promotes the work of automatic/semi-automatic annotation to corpus. Statistical methods have an advantage for fine-grained knowledge of the language than the rule method, and can automatically learn knowledge from the annotated corpus, and is attractive and worthy of research.

Meanwhile, many Chinese parsers that have the background for the practical application use the rule-based approach, and, in addition to the accumulation of knowledge in the process of manual knowledge acquisition, also use statistical methods to help get the phrases from the corpus, and also include the translation equivalents acquired automatically for machine translation. An important direction of development for these systems is to find ways to learn a lot of phrase knowledge from the corpus, which can greatly reduce the difficulties encountered in the ambiguity resolution to improve the accuracy of syntactic analysis. For Chinese-English machine translation system, the difficulty will be significantly lower after adding a large number of phrases and their translation to the system, and as a result, some syntactic structure ambiguities are eliminated, and many phrases are translated as a whole and the readability of the translation also are improved.

An important development trend of natural language processing is that corpus is considered as processing objects and sources of knowledge acquisition. Rule approach has proven to be difficult to the task of processing large-scale real corpus, so the researchers turn to the help of statistical methods, and many experiments prove that statistical methods indeed have made great progress. But the statistical method has its inherent shortcomings, and statistical methods alone can hardly reach expectations of the perfect goal of natural language processing. Thus, Many researchers begin to explore ways of combination of statistical methods and rules, and have made some progress, but there is still a long way to go from the ultimate goal of natural language processing (computer can fully understand the nature of human language). The current trend of integration of empiricism and rationalism in natural language processing is a significant phenomenon, and its development will produce a lot of valuable results, and natural language processing research and applications will benefit from it.

The CCID's future research will focus on methods of automatically extracting knowledge of Chinese phrases and their translations. These methods will be mainly statistical methods, combining with some of the rules means to facilitate access to single-language knowledge and improve the correct translation rate. Progress of the research in this regard will be helpful for our practical machine translation system to improve the quality of translation. At the same time, it has a direct role in improving the quality of Chinese parser.

The paper is funded by National Natural Science Foundation of China, and the project number is: 60872118.

References

Feng Zhiwei. 2004. *The Research on Machine Translation. China Translation and Publishing Corporation.* China Translation and Publishing Corporation. Beijing, China

Zhong Chengqing. 2008. *Statistical Natural Language Processing*. Tsinghua University Press. Beijing, China

Zhao Tiejun, etc. 2000. *Principles of Machine Translation.* Harbin Institute of Technology Press. Harbin, China

Sun Guangfan, Song Jinping, Yuan Qi. 2006. *Design of bi-directional English-Chinese machine translation systems based on hybrid strategy*, Journal of Chinese Information Processing, Beijing, China.

Li Xing. 2005. *The Research on Chinese Parsing*, Master thesis, Chinese Academy of Sciences, Beijing, China.

Lu Junzhi, Chen Xiaohe, Wang Dongbo, Chen Feng. 2008. *Chinese Parsing Algorithm Based on Grammatical Function Matching*, Computer Engineering and Applications, Beijing, China.