Research and realization of Chinese text semantic correction Based on Rule

Yefan Wu[1,a], Runbo Zhuang[1,b], Ying Jiang[1,c] and Fan Li[1,d]

[1]School of Management, Beijing Normal University Zhuhai campus, Zhuhai 519000, China.

[a]m13726279603@163.com, [b]zhuangrunbo081@gmail.com, [c]jpz6311whu@bnuz.edu.cn,

[d]519965410@qq.com

Keywords: Chinese, semantic, correction, rules, collocation

Abstract. Automatic correction is an important research field in natural language processing, it is still relatively weak in the text automatic correction technology on semantic level. This paper presents develop XML rules based on LanguageTool Chinese grammar correction, match error correction in semantic level for the content of the rule to write the corresponding rule base, and realize automatic detect the corresponding content of semantic error in the text, and put forward the corresponding modification suggestions and opinions. The experiments show a high correct rate that the process of correction the corpus, which shows that the research and implementation of semantic of Chinese text is very practical significance and valuable.

The introduction

Background and significance of research. Text is an important carrier of human social information. With the rapid development of whole society information process, the importance and urgency of text information correction is more and more obvious. The former researchers in the text of the technology has made great achievements, but their check wrong technology just based on the word level, dealing with more words, less word or wrong character correction in the text. If testing an essay about the text of population statistics in China, the correct expression is "我国人口统计（不包括台湾、香港、澳门）"("the population statistics of our country (excluding Taiwan, Hongkong, Macao") but appear the "大陆人口统计"("the population statistics of mainland") in the text. If do not found this mistake and report it out, which has violated China's political class information what appeared semantic collocation error. It is difficult to estimate the speed of network transmission, which will bring many negative effects to the society. Semantic errors, text automatic error-detection by word level is unable to achieve the error correction on semantic level, so the research and implementation of Chinese text semantic correction is a key and an essential part of Chinese text automatic correction development.

The research status at foreign. In the early 1960s, foreign carried out the study of English text automatic correction; development to today, its technology is very mature already. Because of the English text use the spaces between words and words for separators, their automatic correction is the core of the word, word-error can be of two types, one is non-word error, the other is real-word error [1]. The study found that the non-word errors in the English text accounted for 60%, the real-word error accounted for 40% [2]. The non-word errors that the string is not exist in dictionary[3]; Real-word error that string is exist in the dictionary[3], but it with the context collocation error, which cause syntactic semantic error, so the real-word error is the semantic error. Traditional real-word error detection method mainly has two types: based on the traditional method of natural language processing and the method based on statistical language model [4]. It has been studying and improving the technology of semantic correction. In recent years, Daniel Dahlmeier1 and Hwee Tou Ng present a novel approach for automatic collocation error correction in learner English which is based on paraphrases extracted from parallel corpora [6]. Their key assumption is that collocation errors are often caused by semantic similarity in the first language (L1-language) of the writer, they show the L1-induced paraphrases

outperform traditional approaches based on edit distance and so on [5]; Tiberiu Boros proposed RACAI hybrid grammatical error correction system[7], this system was validated during the participation into the CONLL'14 Shared Task on Grammatical Error Correction. Compared with the traditional method this system overcomes some shortcomings of them, such as can reduce the number of statistical errors and rules; these two methods have made great progress in text semantic correction.

The research status at domestic. Domestic research is developing rapidly, which began in early 1990s. Chinese's automatic correction is more difficulty than English; we need to use the Chinese word segmentation system because Chinese have no obvious delimiters between words in the text. Professor Zhensheng Luo, Yang-sen Zhang and so on, they do a lot of research in Chinese text automatic correction, they showed outstanding research results for us and made great contribution to Chinese automatic correction. At present, it has achieved great success in the text correction of the word level, but it is still weak in semantic level. Professor Yang-sen Zhang, et al. in the literature[8] comprehensive the Bayesian, decision tree, vector space and maximum entropy models application in Chinese word meaning eliminate gaps, and identification basis for the word sense disambiguation model's selection and application; Weihua Luo, et al. discussed the semantic level technology of the Chinese text correction[9], which not only can check the local semantic constraints, but also check the semantic collocation, the method provides a new thinking of carry out the research and implementation of Chinese text semantic correction.

In conclusion, there are some differences between the author and the former researchers' method. Their method is more tend to model reasoning, which calculate the probability of a sentence based on some parameters; the author's method is directly accurate detection for the sentence, it can make the accuracy rate of more than 90% in a certain situation. At the same time, you need to write lots of rules by the author's method to capture more about the content of the semantic level error.

Semantic collocation error and semantic correction

Definition of semantic collocation error. Semantic collocation error is some language errors are reflected in the semantic level. That is, there is no problem on the word level and the syntax of collocation but error at collocation on the semantic level. Such as "聪明的手"("smart hands") , this phrase is true in terms of words and grammar, but it is clear that there is a problem of semantic meaning. "聪明"("smart") can't match with "手"("hands"). In the dictionary of modern Chinese semantics clear lists each words in semantic belongs. There are a lot of same collocation properties which belong to the same semantic category. For example, "面包", "蛋糕"("bread" and "cakes") is belong to the food category which can match with "吃"("eat"). In contrast to the "石头", "砖块"("Stones" and "bricks") is a natural thing, which can't match with "吃" ("eat").

Types of semantic. The author concluded the categories according to the Chinese library classification as shown fig.1:

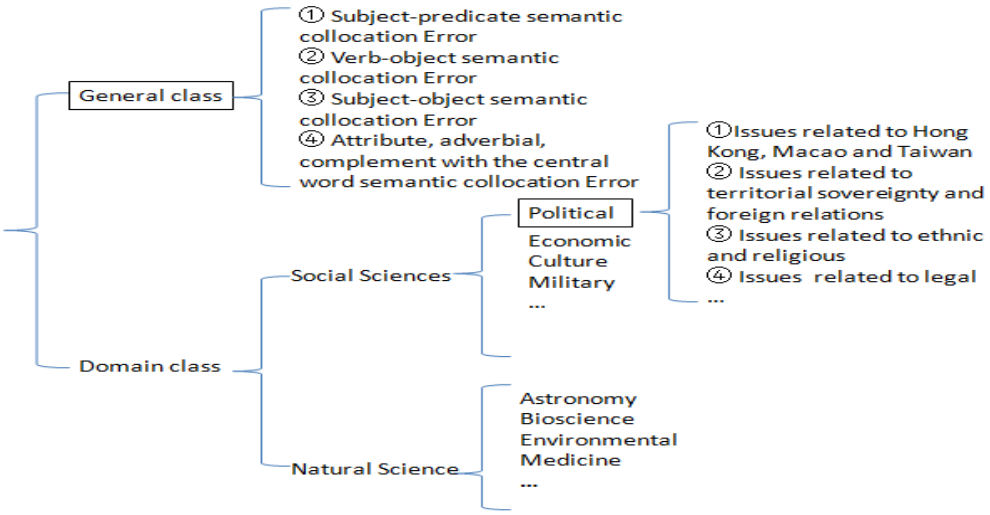

Fig. 1 Types of semantic

In fact, semantic classification is the classification of books, the author defines comprehensive literature as general class which is upper ontology[11], and defines the literature that belongs to a domain as domain class which is domain ontology[11]. The author research to the semantic error on the existing classification, through the search of a large number of literature and survey found that people who is relatively easy to make mistakes on semantic level in two areas, which is the collocation error in the general class and political class.

The error characteristics of the semantic. The author concludes that there are four major categories of semantic errors: domain, essential, concealment and severity.

(1) Domain: Semantic has the characteristic of domain, which is exist difference interpretation in the same thing, reflected in the same thing has different understanding in different areas. Then, there may be course a semantic error which the interpretation of a thing in a domain to express in different domain. For example, “李克强总经理”(“General Manager Keqiang Li”), there is no error in the commercial title, but make a serious semantic error in the field of politics, the right formulation is“李克强总理”(“Prime Minister Keqiang Li”).

(2) Essential: Can‘t only look at the language on surface of the expression, it is need to determine the content of what you want to express in real context through combination context. For example, Propaganda to Taiwan, Generally do not use “解放前”(“before the liberation”) and “解放后”(“after the liberation”), you can use “中华人民共和国成立前（后）”(“before the founding of the people's Republic of China(after)”) “新中国成立前（后）”(“before the founding of new China(after)”). Therefore, it needs to be combined with context to determine if there is error on semantic level.

(3) Concealment: Semantic error is not easy to detect because of it does‘t exist the problem on the word level and the syntax of collocation. For example, “学得很忙”(“studied very busy”),all kinds of ingredients are complete but there is collocation error on semantic level in this sentence[17], you can use “好”(“well”) as a supplement to “学”(“study”) but not “忙”(“busy”).There is no rule to following like this kind of semantic collocation error, so it is not easy to detect the error on semantic level which existence concealment.

(4)Severity: If transmitted any content which contains the error on semantic level through the network or any other communication channels, it will bring unnecessary trouble to others and ourselves, but also bring many negative effects to the society. For example, Taiwan people's daily use of the Chinese dialects, media coverage may not be called it “台语”(“Taiwanese”), such as various publications or sites can use “闽南语歌曲”(“Southern Fujian Dialect song”) but do not use the “台语歌曲”(“ Taiwanese song”). People should regulate the language when it comes to the issue of Taiwan, otherwise it will bring some unnecessary trouble. which may be faced with forced to shut down if a publication occurred this error.

In this regard, the author thinks that is urgent on the research and implementation of semantic level correction in Chinese text.

The idea of semantic correction. Step 1: information collection, collect the information abot the general class and political class. General information which is the information involves all areas, the author main research collocation errors. Such as semantic collocation error on Subject-predicate and Verb-objec, etc.; Political class information, author research which expression is wrong refers to the relevant national, Hong Kong, Macao and Taiwan, and other political information.

Step 2: integrated information. Standardization and integrated information according to the collected, screening information there is existence of errors on semantic level and also can be written by rule.

Step 3: select a proper word segmentation tool. Different from other languages such as English, Chinese is basic on unit of the word; there are no obvious segmentation marker between words, need to use the segmentation tool for segmentation word so that computer can distinguish the different part of speech of words. The author chooses to use ICTCLAS4J word segmentation system[12];most of the rules are according to how to be divided the sentence because of the different segmentation results in different parts of speech. So it's important to choose the right word segmentation tool.

Step 4:formulate proper XML rules. By principle of LanguageTool[13] and use the way of establishment of the Chinese grammar rules library as reference object, the rule base is set up by using the semantic correction information as the rule content, which to realize the function of tips and modifications contents of the correction.

Step 5: detect whether written rules is effective. Rule detects text by matching to one or several words; sentence will be marked out wrong as long as matching success. So the author testing the rule through example reflects the key words collocation. The detection results can be divided into three kinds, ① the content of the sentence error is detected; ② detect the content of the sentence is not expected to want the wrong content; ③ the error can't be detected. Modify the rule if the results of ② or ③ , if modify rule to the result of ① which you can join the rule into rule base and be determined first, or delete the invalid rules.

Step 6: Select appropriate corpus to test the rule base. Corpus is a collection of a large amount of text. The appropriate corpus is the main content appropriate, for example, the author mainly selected news, micro blog and other related content corpus, with the help of detection corpus to revise and improve the preparation of the XML rule base. For example, the author mainly selected news, micro blogging and other related content of the speech corpus, with the help of detect the corpus to modify and improve the preparation of the XML rule base.

The method of semantic correction. The author who customization the XML rule method is based on the LanguageTool [14], the main achieve semantic proofing is use of extraction words, word or part of speech and other feature information in the context [15], and use the rules to detect a sentence whether exist semantic error in the use of the key words collocation model . For example,"三峡大坝下游大量的蔬花水柏枝和中华蚊母的发现，打破了国外部分专家称三峡生物不可复制的说法"((Downstream of the dam height hydrosols bai zhi and the mosquito mother found that broke the foreign part experts claim three gorges creatures cannot copy").Which semantic collocation error between the "打破"("broke") and "说法"("claim"). So we develop appropriate XML rule which choose "打破"("broke") and "说法"("claim") these two key words into the rule. The sentence will be detected and marked if the word "说法"("claim") appears in front of the word "打破"("broke"), and also put forward the corresponding modification suggestion.

Collection of information

Specification information content of political class. The author divides political information into nine categories and extract a example from all kinds of standard content as follows：

(1)Issues related to Hong Kong, Macao and Taiwan

In doing the country's population statistics, can't use the "大陆人口统计"("the population statistics of mainland"), the right argument is "我国人口统计（不包括台湾、香港、澳门）"("the population statistics of our country (excluding Taiwan, Hongkong, Macao)").

(2) Issues involving territorial sovereignty and foreign relations

“钓鱼岛不属于中国”(“Diaoyu Islands not belong to China”) this involves the issue of territorial sovereignty, the Diaoyu Islands belong to China.

(3) Involving the name of the government organizations and state leaders

To avoid emergence the “李克强总经理” (“General Manager Keqiang Li”), it should be the “李克强总理” (“Prime Minister Keqiang Li”).

(4)Involving the use of “苏联”(“Soviet Union”) and “前苏联”(“Former Soviet Union”)

“前苏联”(“Former Soviet Union”) can be used in particular, Generally use of the “苏联”(“Soviet Union”) in the context where will not be ambiguous.

(5)Involved in guidelines and policies about the party and the state

“邓小平南巡讲话”(“Xiaoping Deng speeches during the inspection Tour in the South”) was abolished, instead of “邓小平南方谈话”(“Xiaoping Deng South Talk”).

(6)Involving issues relate to the national and religious

Not called the nation which is the minority nation branch or tribe. Can be called “＊＊人”(“**people”). Such as “摩梭人”(“the people of Mosuo”)and “撒尼人”(“the people of Sani”), can’t be called “摩梭族”(“the nationality of Mosuo”) and “撒尼族”(“the nationality of Sani”).

(7)Involved in some international organizations

The members of WTO (World Trade Organization) and the APEC (Asia Pacific Economic Cooperation) can only be called “成员”(“member”) or “成员方”(“members”), and cannot be called a “成员国”(“member states”), because of in the WTO and APEC has some “Separate customs area”, they are not a sovereign state, such as the WTO's members include Chinese Taipei, Hong Kong China, Macao China; APEC members include Chinese Taipei and Hong Kong, China.

(8)Involving issues of legal

The parties to a criminal case shall not use the “罪犯”(“criminal”) before the court shall pronounce the crime, and shall use the “犯罪嫌疑人”(“suspect crime suspect”).

(9)Related to important people, time and historical facts of the event

“第二次世界大战后初期，日本一跃成为世界第二经济大国”(“Japan became the world's second economic power after the Second World War”), this sentence does not conform to historical facts, Japan as a defeated country by the United States after the end of the Second World War, its economic development in the eighty’s.

Specification information content of general class. The author divides the general information into four categories and extract a example from all kinds of standard content as follows：

(1) Subject-predicate semantic collocation error

“他一进教室，同学们的眼睛都集中到他的身上。”(“The eyes of the students focused on him when he entered the classroom.”), which “眼睛”(eyes”) can’t match with “集中”(“focused”),must be replaced “眼睛”(“eyes”) with “目光”(“gaze”).

(2) Verb-object semantic collocation error

“尽管灾区各级财政都有困难，但各个部门还是积极筹措资金，采取发放补助等方式，提高干部群众的经济状况来缓解他们的生活压力。”(“Despite the financial department difficulties of the disaster areas, various departments which actively take grant to raise funds, and so on, raise cadres and the masses economic situation to ease their lives.”) ,which “提高”(“raise”) can’t match with “状况”(“situation”),you can replace “提高”(“raise”) with “改善”(“improve”) or replace “经济状况”(“economic situation”) with “生活水平”(“living standard”).

(3) Subject-object semantic collocation error

“造纸是我国古代的四大发明。”(“Papermaking is the four great inventions of ancient China.”), should be added “之一”(“one of”) at the end of the sentence.

(4) Attribute, adverbial, complement with the central word semantic collocation error

“这位奋斗在科技领域的优秀共产党员，曾经获得“全国劳动模范”、“全国五一劳动奖章”、“全国创业之星”等多项荣誉称号。”(“The struggle in the field of science and technology, outstanding communist party members, has won the model of national labor, the medal of national labor, the star of national venture and a number of honorary titles.”),which “奖章”(“medal”) and “称号”(“title”) is improper semantic collocation between the attribute and the central word.

Writing of the XML rules

Using word segmentation system. “国际贸易组织” (“International Trade Organization”) in the ICTCLAS4J[13] Chinese word segmentation system of segmentation result is: “国际/n 贸易/vn 组织/n” (“International/n Trade/vn Organization/n”). The word need to break down into three words LanguageTool system can identify, and respectively into < token > [6] of XML, the specific rules are < token > 世界 < / token >, < token > 贸易 < / token >, < token > 组织 < / token >. If the six words as a word written rules, the system is unable to identify. When the author writing a XML rules, need to detect the key words in the ICTCLAS4J Chinese word segmentation system, then write the rules. The segmentation result also shows that in the part of speech of each word, in the process of writing rules sometimes need to use the part of speech to detect errors and make the error correction rate of rules is higher.

Write appropriate rules. In Chinese collocation, “穿”(“dress”) collocate with “衣服”(“clothes”), “戴”(“put on”) collocate with “帽子”(“cap”), “穿”(“dress”) and “帽子”(“cap”) is an obvious collocation errors, as long as in the text retrieval “穿”(“dress”) collocate with “帽子”(“cap”) can detect the error. This is belongs to the verb and object collocation error of general class, the code is what we wrriten simple rules of semantic error.The detail content of rules as shown figure 2.

```
<rule id="DBDPBD_7" name="动宾语义搭配不当">
        <pattern>
                <token skip="-1">穿</token>
                        <marker>
                                <token>帽子</token>
                        </marker>
        </pattern>
        <message>“穿”与“帽子”搭配不当，您可以使用
                <suggestion>戴</suggestion>与“帽子”搭配。
        </message>
        <short>通识类动宾语义搭配不当</short>
        <example type="incorrect" corrtion="戴"></example>
</rule>
```

Fig. 2 The appropriate rule

Application testing

The purpose of text. The author test rules by corpus, the objective is to find out more potential errors in rules, because examples of test is not enough , many error rule only testing by corpus. And many semantic error sentences not are detected. Correct rate cannot improve to 100 percent; the author can constantly modify XML rules by detected results, to strive for the maximum increase the rate of error correction with rules.

The procedures of text.

The procedures as shown in fig.3.Firstly, author write a Java program for running local LanguageTool source program, java program can read the words from the corpus, and the results of test can be print. Then author select an appropriate corpus 1, detect error of corpus 1. Author can judge detected results; the results will have two situations: ① the content of the sentence errors is expected to the error of content; ② the content of the sentence errors is not expected to error. Then the two types of data collected by respectively counting, calculate correct rate in test of corpus for the first time. Then author modifying the rules by the results of test, the modified rules for testing corpus

1 again, the statistics of results for the correct rate B of second time, correct rate is improved compare with first time statistics. In n (decision according to their own conditions, the author choose 2 corpus, n = 2) corpus, constantly modify rules and improve the correct rate, effectively improve the accuracy of error detection.

The formula of calculate the correct rate: correct rate= numbers of detect the content of one sentence is not expected to want the wrong content / numbers of all detect results.

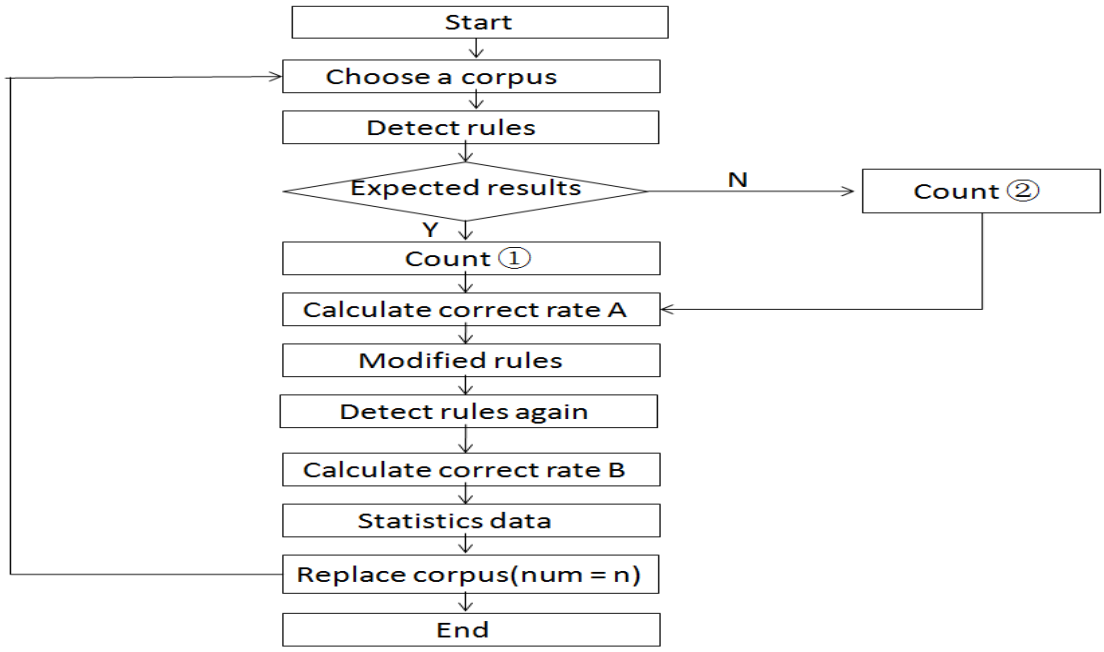

Fig. 3 The procedures of text

The test results and data statistics.the correct rate of XML rules as shown table 1 and table 2：

Table 1 The rusult of rules detection

types	Error types	Rule number	Detect the number of rules
General	subject-predicate semantic collocation error	5	3
	verb-object semantic collocation error	67	23
	subject-object semantic collocation error	4	0
	attribute, adverbial, complement and the central word semantic collocation error	4	2
Political	Issues related to Hong Kong, Macao and Taiwan	25	12
	Issues involving territorial sovereignty and foreign relations	5	4
	Involving the name of the government organizations and state leaders	10	5
	Involving the use of “苏联”(“Soviet Union”) and “前苏联”(“Former Soviet Union”)	1	1
	Involved in guidelines and policies about the party and the state	2	1
	Involving national and religious issues	14	9

	Involved in some international organizations	13	8
	Involving legal issues	6	2
	Related to important people, time and historical facts of the event	4	1
total		161	71

Table 2 correct rate of XML rules

corpus	general		political	
	First test	Second test	First test	Second test
corpus 1	59.7%	78.3%	62.3%	78.4%
corpus 2	82.3%	87.9%	86.4%	89.1%

Analysis results of the test. The author respectively detect the political class and the general class of rules , and change corpus, which due to the corpus 2 data content is too large, so the author randomly selected part of the corpus (175 m) to test the rules. The first test is use all XML rules detect the errors in the corpus; Second test is modify the rules by the first results of test, and then detect rules again in same corpus.

In testing for the first time, general class and political class of correct rate is lower. The reason is the first time writing XML rules, author write rules according to the collected content. author didn't realize that the content of detection rules but correct sentence, cause test results in more mistakenly identified as a result, such as mentioned in 4.3 some rules of changes, for example, “增加”(“increase”) and “作用”(“effect”) collocation is incorrect, but author didn't realize that specific case will appear “能增加营养或有滋补作用”(“It can increase the nutrition or have a nourishing effect”), this is the right collocation appeared in the error detection results, so author add a rule filter between two words "or" and "and" rules. The author modify the rules by examples, to exclude some error results, in the modified rules base corpus 1 second detection, the correct rate of general class and political class got obvious lifting, correct rate is around 78%.

Detection with corpus 2 for the first time, two classes has achieved higher accuracy, general class is 82.3%, political class is 86.4%. Because of the rules was modified based on the first corpus, has effectively exclude most mistakenly content; another reason is that the content of the selected corpus 2 is different. This is from sina weibo artificial screening of positive and negative emotions, and positive, negative, contradictions in weibo microblog data content. People have relative freedom of speech on weibo, there are people from every country's comments, so there are more errors of semantic, and detect the content of the “钓鱼岛不属于中国” (“Diaoyu Islands not belong to China”), has been involved in the territorial sovereignty. But there's still exist mistakenly identified as a result, for example, not confuse the name of the ancient ethnic and the later nation “高句丽”(“Koguryo”) cannot be called the “高丽”(“Koryo”), the detected results appear people named “高丽”(“Li Gao”) and “高丽大学”(“Korea University”), which results are not expected content, no one detected result is correct and the content of error detection results is more, so author will delete this rule. After modification rules again, the general class correct rate from 82.3% to 87.9%, the political class correct rate from 86.4% to 89.1%.

But correct rate can't improve to 100%, the reason is that some statements of testing need to judge the specific context, and some is use the word part of speech to detect semantic error. For example, in the case reports “小偷”(“thief”) and “强奸犯”(“rapist”), do not use their social identity as a prefix, don't write “工人小偷”(“workers thief”) if one was once a thief in the workers; Considering there are a lot of social identities, such as the teacher, driver, engineer, but cannot listed it all, so the author decided to use the part of speech to detect before the “小偷”(“thief”) their social identity. The word segmentation results show part of speech is n (noun), then “thief” before nouns part-of-speech will be

detected, found that there are still many parts of speech is n mistakenly identified as a result, such as “一个小偷”(“a thief”) . All of these errors can't correct detection, only by artificial judgment again. So, the correct rate can't reach 100%.

Deficiency of the test. (1) Cannot be detected about the time Views error Some of the ideas with the times change will also be changed, for example, “朱元璋刚开始起义反元时，代表的是农民阶级的利益，但是后来他登基做了皇帝，就发生了转变，开始代表地主阶级的利益，对他的评价也应分时期进行”(“Yuanzhang Zhu just began to revolt against the yuan which the representative is the interests of the peasant class, but changed to represent the landlord class interests when he became emperor later, we should be period to his evaluation”), in the present, the author cannot let LanguageTool automatic change detection for some of view with the time.

(2) Comma problems lead to collocation errors cannot be detected

Some collocation between words will join the comma, for example, “春风一阵阵吹来，树枝摇曳着，月光、树影一齐晃动起来，发出沙沙的声响”(“A spring breeze blowing, branches swaying and moonlight shadows together, and make some noisy from them”), generally speaking, could not make some “响声”(“noise”) from “月光” (“moonlight”), but the LanguageTool detection a sentence by commas as a symbol of segmentation ,cannot detect the content behind the comma, so can't detect the two collocation of word problems between in two sentences.

(3) Be detected fewer rules

Few rules be detected, about several reasons, firstly, the detection of corpus may involve the content of the political and general class is not a lot, for example, with corpus 1 is the main data content of the sohu news within the classification of news corpus, news published generally more stringent, so detect error is less; secondly, it is to write the rules of error content may not be a people used to use a combination of content, such as: people usually not used the collocation of the “打破”(“broken”) and “说法”(“claim”), lead to detect errors result is no something like that ; At last, it is rules contains some predictive content, predictive content is errors in the future are likely to appear collocation error, but hasn't found such errors. For example, “李克强总经理” (“General Manager Keqiang Li”), this is relatively easy to appear collocation error, once appear, will bring serious consequences, so in order to avoid such problems, the author will write it into rules.

Summary

Summary of semantic correction. In this paper about research and realization of the two areas(general and political) of semantic correction, and the author write XML rules can detect a lot of errors in corpus, explains it is necessary for the semantic correction. Detected results, and then according to the artificial detect, constantly modified the rule for effectively raise the accuracy of semantic correction, and also got the results of correct rate more than 85%. So the research and implementation of Chinese text semantic correction is very meaningful for the actual needs.

Future prospect of semantic correction. In this paper, the author study semantic error correction with a small category, but the XML rules is not enough for the great data, and some problem is insolvable. For example, time views change, between the sentence and the sentence errors. In order to solve the above problems, the author should be change way of detection, use the knowledge base to detect error, so in the detection it will detect by the attribute of word, such as stone is assigned to the attribute of inedible category, but apple is assigned to the attribute of edible category, on the basis of a knowledge base can be write a Java plug-in to connect the knowledge base, when open the web page or a text document it will automatic scanning knowledge base and detection error.

Acknowledgement

This work is supported by a project granted by the National Social Science Foundation of China (Project No. 14CTQ041), a grant from Science and Technology Plan Project of Guangdong Province (Project No. 2014A080804001), and a grant of 2015 Annual Provincial School Moral Education

Innovation Project (Project No. 2015DYZD015) from the Department of Education of Guangdong Province. The corresponding author of this paper is Ying Jiang (jpz6311whu@bnuz.edu.cn).

[1] The corpus including text of classification from the laboratory of Sogou, the source of data: http://www.sogou.com/labs/dl/c.html

[1] The positive, negative and contradictory microblogging data from Sina Weibo, the sources of data: http://www.datatang.com/data/47209

Reference:

[1] Samanta, Pratip, and Bidyut B. Chaudhuri. "A simple real-word error detection and correction using local word bigram and trigram." ROCLING. 2013.

[2] Kukich, Karen. "Techniques for automatically correcting words in text." ACM Computing Surveys (CSUR) 24.4 (1992): 377-439.

[3] Zhang Yangsen, and Shiwen Yu. "Summary of the Technology on Text Automatic Correction [J]." Application Research of Computers 23.6 (2006): 8-12.

[4]Feng,Jinfeng. "Research of Chinese Text Rutomatic Error Detection" Southeast University.2011

[5] Dahlmeier, Daniel, and Hwee Tou Ng. "Correcting semantic collocation errors with L1-induced paraphrases." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[6] Bannard, Colin, and Chris Callison-Burch. "Paraphrasing with bilingual parallel corpora." Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005.

[7] Boroș, Tiberiu, et al. "RACAI GEC–a hybrid approach to grammatical error correction." CoNLL-2014 (2014): 43.

[8] Zhang, Yang-Sen, and Jiang Guo. "Analysis and comparison of 4 kinds of statistical word sense disambiguation models [J]." Journal of Beijing Information Science & Technology University 2 (2011): 002.

[9] Luo, Weihua, Zhen-Sheng LUO, and Xiao-Jin GONG. "Study of Techniques of Automatic Proofreading for Chinese Texts." Journal of Computer Research and Development 1 (2004): 036.

[10] Wu, Lin, and Yang-Sen Zhang. "Reasoning Model of Multi-level Chinese Text Error-detecting Based on Knowledge Bases." Computer Engineering 20 (2012): 006.

[11] CAO, Jin-dan, Yang MI, and Xi-long ZHOU. "Research on Domain Ontology Construction Based on Upper-level Semantic Relationships." Information Science 9 (2014): 028.

[12] Yi, Tian-Peng, and Qi-An Chen. "Comparison Research of Segmentation Performance for Chinese Analyzers Based on Lucene [J]." Computer Engineering 22 (2012): 072.

[13] Naber, Daniel. "A rule-based style and grammar checker." (2003).

[14] Jiang Ying;Zeng Jie;Lin Qihong;Guo Yingshan;Liao Wensheng. "XML Rule Customization Method of LanguageTool Chinese Grammar Proof-Reading" Library and Information Service,2014,05:86-92.

[15] Zhang, Yangsen, and Gaijuan Huang. "A Study on the Disambiguation of Chinese Words Based on Multi-Knowledge-Source." Chinese Linguistics 2 (2008): 010.

[16]Caihong,Gu analysis and modification of incorrect sentences on http://www.ccppg.com.cn/baokan/zhongguozhongxueshengbao/zuixinbaodao/2006-03-03/39133.html