Effective Use of Chinese Structural Auxiliaries for Chinese Parsing *

Yun Jin[a], Qing Li[b], Yingshun Wu[a], and Young-Gil Kim[a]

[a]Natural Language Processing Research Team, ETRI,
138 Gajeongno, Yuseong-gu, Daejeon, 305-764, Korea
{Wkim1019, suni, kimyk}@etri.re.kr

[b]School of Economic Information Engineering, Southwestern University of Finance and Economics,
55 Guang hua chun Road, Chengdu, 61130, China
liq_t@swufe.edu.cn

Abstract. Natural language parser is usually faced ungrammatical input, such as mistyping or error POS tags. If the parser uses language dependant explicit linguistic knowledge to detect and correct grammatical errors, it is useful for parser. In this paper, we propose a method that uses the Chinese structural auxiliary knowledge to detect and correct ungrammatical Chinese parsing errors. We focus on three error types: miss segmentation, miss POS tags, and miss typing. Experimental results show that appropriate use of evident Chinese structural auxiliary knowledge indeed helps to correct parsing errors and further to improve Chinese parsing performance.

Keywords: Chinese structural auxiliary, Chinese parsing, parsing error, abuse

1 Introduction

Language parsing is an essential part for various kinds of natural language understanding applications including question & answering system, machine translation and so on. Most of language parsers are usually constructed based on standard grammar rules with an assumption that texts to be parsed are error free. Unfortunately, this is not always true. Typos and irregular expression are frequently occurred in documents, especially in emails, messengers and blogs. In addition, the inaccurate preprocessing steps (e.g. parts-of-speech, text segmentation) of a parser bring more errors. Due to the agglutinative characteristics of Chinese language, these kinds of errors occur frequently and should not be ignored in Chinese language parsing. In this paper, we propose to apply Chinese structural auxiliary knowledge to detect and correct ungrammatical errors for Chinese language parsing.

A structural auxiliary is an unstressed form word, which performs the grammatical functions of structure in Chinese language. The structural auxiliaries are the most frequently words used words in Chinese. Based on our study, average occurrence of structural words in Chinese news articles is 1.25 per sentence.

Although structural auxiliaries are occurred extremely frequently in Chinese, there are only three words to perform the grammatical functions of structure. That is, "的(de), 得(de), 地(de), 之(zhi)[1]". In particular,

* The work reported in this paper was supported by the IT R&D program of MKE, "Development of Machine Translation Technology for Korean/Chinese/English Spoken Language and Business Documents".

Copyright 2009 by Yun Jin, Qing Li, Yingshun Wu, and Young-Gil Kim

[1] 之(zhi): is excluded in this paper because of its rareness in modern Chinese.

- “的” is used after an attributive to identify its attributiveness, e.g. 我的书 (my book), 中国的历史 (the history of China), or at the end of a nominal structure to form a noun phrase, namely “的-phrase”, e.g. 开车的 (man who drives), 红色的 (something red);
- “地” is used after an adjective adverbial, e.g. 愉快地学习(study happily), 轻轻地落下来(fall gently down);
- “得” is used after a verb or an adjective to introduce a completion, e.g. 跑得快(run quickly), 好得很(very good).

In this paper, we call them as CSAs (Chinese Structural Auxiliaries) for short. CSAs are one of most important language features in Chinese, however, its essential ambiguity and abuse brings several issues in natural language processing (NLP). In particular,

- Multiple grammatical functions of CSAs make the grammatical analysis quite difficult in natural language processing (NLP). For example, the auxiliary “的” not only can be a structural auxiliary “我的衣服(my clothes)”, but also can be a tense auxiliary “他的情况，我是知道的(his situation, I knew)” or pronoun “送牛奶的(milkman)”. In addition, due to the transliteration of English into Chinese, it could be a part of foreign words such as “的里雅斯特(Trieste)”.
- CSAs are usually used wrongly or improperly in Chinese because of the careless writing style of the Chinese. They are misused in place of each other. For instance, in the phrase “现金流量非常的充沛(vey abundant cash flow)”, the word “的” is used improperly to be replaced as the correct word “地”. Liu (2006) has pointed out that there are 28.2% structural auxiliary errors in primary and secondary school textbooks biased errors corpus.

These characteristics of CSAs cause a number of errors in segmentation, POS(parts-of-speech) tagging and parsing in NLP. In this paper, we propose the application of CSAs error analysis techniques to improve the natural language understanding in Chinese. The rest of this work is organized as follows. We first briefly describe what technologies can be relevant to this idea in Section 2. Three kinds of error types to be considered for Chinese language parsing are systemically studied and the approach to detect and correct these errors is presented in Section 3. We then test the performance of our approach using a known news corpus (Section 4). This paper is concluded with speculation on how the current work can be further improved in Section 5.

2 Related Work

The natural language processing issues related with ungrammatical text errors have been studied by several researchers. Atwell (1987) proposed an N-gram-based method to detect mistyping, and lack or extraneous constituents of sentences. Foster and Vogel (2004) has been collected a 20,000 word corpus of ungrammatical English sentences from a variety of written language sources (newspapers, emails, websites, etc.). Each ungrammatical sentence in the corpus is corrected, producing a parallel corpus of grammatical sentences. They use this data to evaluate a parser’s ability to produce an accurate parse for ungrammatical sentence. Gamon et al. (2007) applied speller techniques and language modeling approaches to detect and correct errors in incorrect usage of determiners and choice of the preposition. Different with previous researches on English text, we focus on the ungrammatical errors in Chinese text, which are more complicated than English due to the essential characteristic of Chinese such as agglutination.

Liu (2006) and Pang et al. (2004) studied the problem of CSAs. Instead of applying CSA to NLP, they just focused on error analysis of CSA. In this paper, we propose the application of CSAs error analysis techniques to improve the natural language understanding in Chinese.

3 Application of CSA Grammatical Functions for Chinese Parsing

CSAs are one of most important language features in Chinese, however, its essential ambiguity and abuse brings several issues in natural language processing (NLP). In this section, we present our approach to detect and correct the errors related with CSAs. We first studied the error types caused by the ambiguity or abuse of CSAs in Chinese news articles. Second, different CSAs errors are classified into such error types using support vector machines (SVM). Third, heuristic rules are adopted to correct these errors for better understanding of Chinese language in NLP.

3.1 Error Analysis of Chinese Structural Auxiliaries

In this section, we study the error types of CSAs and discuss how to utilize CSAs' grammatical functions to improve Chinese parsing.

3.1.1 Segmentation Error

Due to agglutinative characteristic of Chinese, word segmentation is indispensable to intelligent Chinese language process. Segmentation error, however, is inevadible and tends to misguide parse tree generation for NLP since no segmentation algorithms achieve the 100% segmentation accuracy. Segmentation dictionaries and machine learning techniques (Wong and Chan, 1996; Low et al., 2005; Zhao et al., 2006) are popular techniques to alleviate segmentation issues. However, none of previous researches has considered the effect of CSAs on word segmentation. In this article, we argue that the CSAs analysis should be incorporated to reduce the word segmentation errors. For example, Chinese word "丽珠得乐(Bismuth Potassinm Citrate)" is a medical name for a stomach pain killer. If segmentation dictionary does not contains this word (This is usually true for most foreign words and hi-tech words), according to traditional segmentation algorithms, it might be segmented into two different forms, "*丽珠_得_乐*" or "*丽_珠_得_乐*". If POS tagging is applied to further analyzing this word, it might obtain two wrong parsing results as "丽珠/NR 得/DE 乐/NR" and "丽/NR 珠/NN 得/DE 乐/NR". Obviously, this medical noun should be treated as an entire unit instead of segmenting it into pieces. Otherwise, it will lead to parsing errors.

However, if we take a CSAs analysis of this sentence, such kind of error can be corrected easily. More specifically, knowing the syntax of "得", it can infer that the word before or after "得" should not be an noun term. Thus, using a chunking approach, we can generate a correct parse tree. By processing a simple and fast CSA analysis, we can easily solve the segmentation errors related with CSAs and consequentially improve the Chinese parsing performance.

3.1.2 POS Tagging error

POS tagging, usually as a subsequent step of segmentation in NLP, brings lots of errors in Chinese parsing. Chinese structural auxiliaries cause various kinds of errors in POS tagging. For example, "炒地 (Land speculation)", "三亩地(three acres of land)" and "注册地 (registration place)" are usually falsely tagged as "炒/VV 地/DE", "三/NU 亩/MW 地/DE", and "注册/NN 地/DE", respectively. Actually, "地" in these words are noun instead of auxiliary due to its polymorphism. Most POS tagging approaches can not differentiate its polymorph and cause POS tagging errors. Such kinds of POS tagging errors are usually occurred in compound phrases like "采得的信息(collected information)". Noticed that, because of POS tagging error, that is, mistreating "地" as an auxiliary, the parser can not find a suitable grammar rule to output reasonable parse tree.

However, if we take a CSA analysis of this sentence, such kind of error can be avoided. In CSA grammar rule, "地", as an auxiliary, should be placed before a verb or adjective, otherwise it should be a noun referring to "field". Knowing this fact, we replaced the "DE" tag

of "地" with “NN” and get a reasonable parser tree. By processing a simple and fast CSA analysis, we can easily solve the POS tagging errors related with CSAs and consequentially improve the Chinese parsing performance.

3.1.3 Language abuse error

Chinese writings are overwhelmed by Chinese structural auxiliary abuse. The reasons to explain this phenomenon can be categorized into:

- The history of Chinese language culture. The issue of division or reunion three types of Chinese structural auxiliary has been debated for a long time. Sometime, “的” and “地” are united into one auxiliary type, and sometimes are utilized separately.
- Due to the popularity of “的”, it has been inappropriately used widely in Chinese writings. According to our statistics, the usage proportion of “的” accounted for 97.42% of all CSA usage.
- Due to popularity of spelling input method in Chinese computers, the same spelling of three types of CSAs causes lots of mistyping in writing.

Figure 1 shows the distribution of CSA abuse calculated from a sample of Net ease 2007 Chinese news articles. In Figure 1, the label “SRD_BSD” denotes the misuse of “得” instead of “的”; the label “SRD_TYD” denotes misuse “得”instead of “地”; “BSD _SRD” denotes the misuse “的” instead of “得”; “BSD_TYD” denotes misuse “的”instead of “地”. As shown in Figure 1, it can be observed that the abuse ratio of CSA “地” is almost zero. It means that CSA “地” nearly causes abusing, and the proportion of abuse of CSA “得” is 7%, so the abuse of the CSA is almost caused by CSA “的”(93%). In the abuse CSA “的”, people are more confused by the pair of “的” and “地”(56%) compared with the pair of “的” and “得” (37%). Note that this statistical result is accordance with the first reason of CSA abuse.

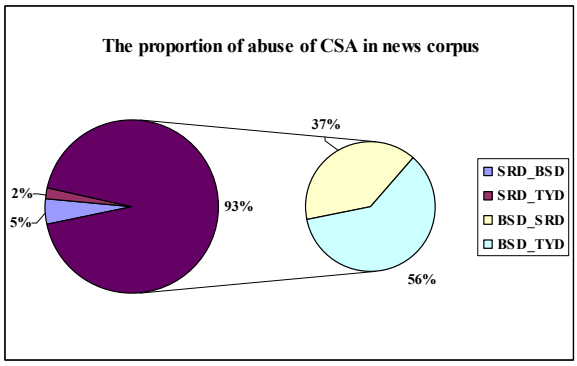

Figure1: The proportion of abuse of the CSA

Because “的” CSA can be placed most of the words, we need more sophisticated approach and need more context features to resolve abuse of CSAs. For example, we only consider “高(high)/的/多(more)” 3 words, most of the Chinese person directly recognize “的” CSA is abused with CSA of “得”, but we give more context as “收入高的多纳税(high-incomer pay more taxes)”, the people know CSA of “的” is correct. This example shows us to resolve abuse of “的” CSA, not only consider adjacent terms, but also consider long distance context features. All of those characteristics should consider into when we are building our detecting and correcting the CSA errors system.

3.2 Error Detection

Knowing the above three types of errors, we can apply support vector machine (SVM) to classify errors into these types.

The SVM is a state of the art supervised machine-learning technique proposed by Vapnik (1995) and is based on Structured Risk Minimum Principle. By the principle, when training a classification model, the aim of the learner is to optimize not just the error rate on the training data set, but also the ability of the model for predication, and the ability depend on concept VC-dimension. The SVM is being applied in many areas such as word sense disambiguation, text classifications (Song et al., 2005).

To correctly group the errors related with CSAs, we utilize adjacent terms of each CSA as contextual features for SVM. In particular, we extract *K* terms on both left and right side of each CSA in the sentence as contextual features. The context feature we used term and POS tag. In additionally, we also added some of the grammatical pattern rules used as features. This is for our data set sparseness and for reveals implicit data properties. Some of the grammatical pattern rules are show as follows in Table 1.

Table 1: Example of pattern rules

Pattern rules	Meaning	Example
{VV, AJ, AD}+的 +NN|VV	If Verb or Adjective or Adverb pre-place at "的" , then to check whether followed by a gerund	调整:NN 还:AD 在:PO 不断:VV 的:DE 变化:NN 。:PU
有+NN+的+NN|VV	If the pattern combined with "有" and Noun and "的", then to check whether followed by a gerund	有:VX 针对性:NN 的:DE 提出:NN 意见:NN
MW/AA+地+VV	If the post-place term is Verb to check whether pre-placed at "地" term have Verb	板车:NN 一:NU 排排:MW/AA 地:DE 停:VV 在:PO 路边:NN
…	…	…

3.3 Error Correction

After detecting errors related with CSA and their error types, we can correct these errors using a set of heuristic rules. Based our observation, the 3 kinds of CSA errors are have different characteristics, that is most of the language abuse errors are caused by "的", and most of the segmentation errors are caused by "得" followed by "地", and most of POS tag errors are caused by "地" followed by "得", so we can focus this characteristics to use different solution. For example, if a detected error is "的", we can first apply it abuse error correcting solution, if the solution resolve the error finish correction, but the solution can't resolved then we can apply it with segmentation error correcting solution. If a detected error is "地", we first apply it with POS tagging error correcting solution, and then apply it with segmentation error correcting solution and language abuse error correcting solution separately. In the following sub-sections, we detail introduce each CSA error correcting method.

3.3.1 Correct segmentation error

Detecting miss segmented boundary even difficult, but those errors are usually tends to following 3 different aspects.

- Grammatically conflicted with CSA, for example, "丽珠/NR 得/DE 乐/NR".
- Miss segmented single character units, for example, "*丽_珠_得_乐*".
- Important clue are appears at around, such as quotation ("丽珠得乐"系列产品), designation terms(实得集团, 梦地董事长).

Based on above aspects, if input of detected errors are have explicit clue or satisfy above patterns then we can regard it as segmentation errors and correct them.

3.3.2 Correct POS tagging error

Because there are have only two CSAs(“地” and “得”) are cause POS tagging errors, for correct POS tagging, we use pattern rule based approach.

For correct “地” CSA POS tagging error, we checked adjacent term tag. If “地” CSA is wrong, the correct POS is Noun, so POS tag of the front located “地” CSA’s adjacent should be one of the verb(“炒:VV 地:NN”), measure word(“一:NU 亩:MW 地:NN”), noun(“注册:NN 地:NN”), and a few special pattern of “NN_AJ_NN_AJ” such as “人:NN 多:AJ 地:NN 少:AJ”, but we do not corrected front adjacent POS tag is proper noun, because this pattern is possibility of segmentation error.

For correct “得” POS tagging error, we also based on grammatical function of “得” make pattern rule to correct POS tagging errors. Because “得” have three different grammatical functions, so separately check all of those pattern rules. For example, the input is “ *竞:VV 得:DE 100 万:NU 股:MW(compete to obtain one million stocks)* ”, we check front adjacent POS tag of “得”, the term “竞” is single verb, it have complement verb property, so to correct “得” POS tag to verb. If front adjacent POS tag is AD and unigram term like “还” or “就”, we changed POS tag of “得” to auxiliary verb.

3.3.3 Correct Language abuse error

Correct language abuse error is choice one of three CSAs problem. To correct language abuse error, we use Bigram approach to calculate the probability of all co-location term with CSAs. Consider the following example of a CSA related error:

调整:NN 还:AD 在:PO 不断:VV 的:DE 变化:NN 。:PU (Adjustment still continues to change.)

Based on pre-calculated probability of front located term: *不断* with each CSA, the language abuse error corrector is consulted to provide the most likely choice of CSA:

P(不断 + 的) = 0.1722
P(不断 + 地) = 0.8278
P(不断 + 得) = 0.0000

Given this probability distribution, a correction module changed CSA with “地” to generated “*调整:NN 还:AD 在:PO 不断:VV 地:DE 变化:NN”* as output.

4 Experiment

4.1 Experimental Setting

To gauge the performance of explicit usage of CSA grammatical functions to assist Chinese parsing, we carried out a series of experiments based on a news article data set. We collected 18,617 news articles in Chinese from net ease (www.163.com) in the year of 2007, and extracted 479,749 sentences among them, and filtered 197,925 sentences that do not contain CSA sentences. The remaining 281,824 sentences are tagged with Chinese POS tags using our Chinese POS engine. Since our objective is to deal with CSA’s error, we extracted sentences collocating with “的” CSA terms also collocating with other CSAs terms from 281,824 sentences. We considered that 2053 terms are have multi-sense and could cause errors related with CSAs. In those term list, we only took 1009 terms with more than one frequency and to extracted CSA fragment in each sentence. Finally, 42,620 fragments are extracted and to used as our system data set.

To annotate the errors related with CSAs, we first use Language Model (LM) and CSA’s grammatical functions heuristic approach extract all errors related with CSAs in the sentences. Two Chinese linguists further analyze these detected errors and remaining data respectively,

and corrected both side data. By doing so, two linguists reviewed almost 60% data. The annotated data set was then divided into training (90%) and test (10%) sets.

4.2 CSA error detection

We designed a baseline system and comparison system to show the effectiveness of our system using CSA grammatical functions. The baseline system used the term (W) and POS tag (T) features, and the comparison system additionally used CSA's grammatical function.

To perform what The SVM^{light} provides four kernel methods(Linear(L), Polynomial(P), RBF(R), Tanh(T)), among them, the polynomial kernel based method performs best(P:82.8 > L:77.13 > R:39.4 > T:32.7) with window size 10 in the baseline system. Additionally, we compared the precision values of CSA error detection methods corresponding to the varying the size of K. The Figure 2 shows the results.

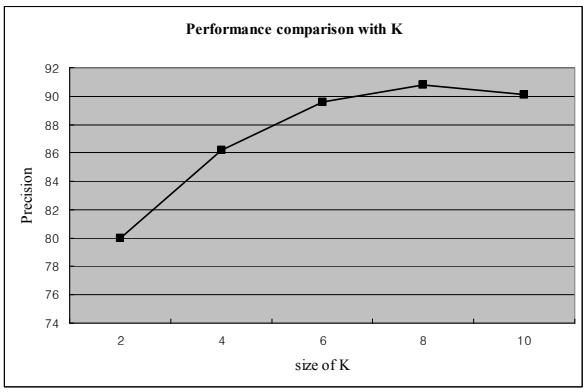

Figure 2: Precision comparison with size of K

Based on the above setting we compared our baseline system with the comparison system additionally used CSA's grammatical function. Table 2 shows the performance comparison between two methods. It appears that our CSA grammatical function knowledge helps to detect CSA errors.

Table 2: Precision comparison of the two approaches

Approach	Method	Precision
Baseline	W	81.13
	T	89.84
	WT	90.8%
Comparison	WTG	92.32%

4.3 CSA Error correction

According to the pattern rules mentioned in 3.3, we corrected each detected CSA error. Table 3 shows that correction accuracy and combination of the detection and correction accuracy. The experiment shows although the ratio of final performance decreases, this result is of enough worth.

Table 3: The accuracy of the each error case

Correction methods	Accuracy	Combined
Abuse of CSA (Ca)	74.41%	70.48%
POS tagging (Cp)	93.4%	90.26%
Segmentation (Cs)	85.78%	81.7%
Cs + Cp + Ca	83.1%	79.7%

To evaluate the final performance of Chinese parsing, we designed a baseline system and comparison system. For comparison system we used CSA detection (D) and correction (C) module, but baseline system didn't used any one of them. For evaluation, we randomly selected 200 sentences from evaluation data set, but contained CSA error sentences which are guarantee at 10 ~ 20% in those selected sentences. The result is shown in Table 4.

Table 4: Chinese parsing performance

Chinese parsing performance	Baseline	CSA D+C	Improved
Accuracy	72.4%	76.2%	5.249%

5 Conclusion

In this paper, we addressed issues related to using CSA's grammatical functions to automatically detect and correct CSA error to reach correct parsing errors in Chinese parser. The experiment shows using explicit CSA knowledge helps to reduce parsing errors. Nowadays the statistics and probability approach very popular but combine this approach with language dependent knowledge and grammatical function would better help to improve performance.

References

Atwell, E.S. 1987. How to detect grammatical errors in a text without parsing it. *Proceedings of the third conference on European chapter of the Association for Computational Linguistics*, pp.38-45.

Foster, J. and C. Vogel. 2004. Good reasons for noting bad grammar: Constructing a corpus of ungrammatical language. *In Pre-Proceedings of the International Conference on Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, pp.151-152.

Gamon, M., J.F. Gao, C. Brockett, A. Klementiev, V.B. Dolan, D. Belenko and L. Vanderwende. 2008. Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. *Proceedings of the third International Joint Conference on Natural Language Processing*, pp.449-456.

Liu, X.M. 2006. Usage Analysis of the Chinese Structural Auxiliary. *Modern Chinese*, Issue 12, 98-99.

Low, J.K., H.T. Ng, and W.Y. Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN workshop on Chinese Language Processing*, pp.161-164.

Pang, K.H. 2004. Problems in the way of the Structural Auxiliary Words and Their Application. *Journal of Shangqiu Teachers college*, 20(1), 162-163.

Song, S.K., Y. Jin and S.H. Myaeng, 2005. Abbreviation Disambiguation Using Semantic Abstraction of Symbols and Numeric Terms. *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering,* pp.14-19.

Vapnik, V.N. 1995. The Nature of statistical Learning Theory. *Springer*.

Wong, P.K and C. Chan. 1996. Chinese Word Segmentation based on Maximum Matching and Word Binding Force. *Proceedings of the 16th Conference on Computational Linguistics*, pp.200-203.

Zhao, H., C.-N. Huang and M. Li. 2006. An Improved Chinese Word Segmentation System with Conditional Random Field. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pp.162-165.