Linguistic Knowledge-driven Approach to Chinese Comparative Elements Extraction

Minjun Park
Dept. of Chinese Language and Literature
Peking University
Beijing, 100871, China
karmalet@163.com

Yulin Yuan
Dept. of Chinese Language and Literature
Peking University
Beijing, 100871, China
yuanyl@pku.edu.cn

Abstract

The BI (比)-structure, which highlights a contrasting characteristic between two items, is the key comparative sentence structure in Chinese. In this paper, we explore the methods of extracting the 6 constituents of the BI-structure. Previous studies are often restricted to probabilistic classification methods, where the feature used hardly embodies linguistic knowledge, therefore unintuitive. As an alternative, we propose the use of two linguistic knowledge-driven approaches, namely the POS chunking-based and TBL-based methods. The first model effectively captures grammatical restrictions over POS sequential patterns. The second model set up on new and lesser templates performs better than Brill's (1995). Experimental results show that the proposed models are simple and effective methods for Chinese comparative element extraction task.

1 Introduction

Comparison is the most representative figure of evaluation. Much of evaluative information is now available in the web, and comparative sentences prevail in Chinese web texts in increasing numbers. A significant amount of research has been conducted on automatic identification of Chinese comparative sentence and its semantic elements. However, the techniques proposed in earlier works are mostly based on statistical classification method. Due to the opaque nature of stochastic features, it is often difficult to comprehend what linguistic aspects are applied to the model.

In this paper, we present a detailed analysis on linguistic behavior of the BI (比)-structure, which is the key comparative structure in Chinese. The application of the two rule-based approaches suggested in this paper are different from previous models in that they fully use syntactic and lexical features which are intrinsic to the structure.

This paper first presents a brief literature review on the subject. The target of extraction task, i.e. Comparative Elements (CE) is then defined before demonstration of the two proposed approaches, i.e. POS chunking-based and TBL-based extraction models. Finally, we discuss the experiment's results and present our conclusion.

2 Related Work

The research on comparative sentence has been a main concern from the beginning of modern Chinese linguistics research. The different types of Chinese comparative sentences were first mentioned in *Mashi Wentong* (1898) and their classification was elaborated later by Chinese grammarians such as Lü (1942), Ding (1961) and Liu (1983). Following by their preliminary work, a series of research focused on defining syntactic and semantic structure of the Chinese comparative sentence was conducted. Li (1986) demonstrates the Chinese BI-structure simplifying rules. Shao (1990) investigates the rule of replacing and omitting elements in Chinese comparative structure.

On the other hand, Studies in Natural Language Processing mainly dealt with the identification of comparative sentence and its elements. Based on Jindal and Liu's research (2006) on comparative sentences in English, Huang(2008) and Song(2009) made a stochastic classifier based on SVMs and CRFs to tackle the Chinese comparative sentence identification and element extraction task. Besides, many models were also suggested in the fifth Chinese Opinion Analysis Evaluation (COAE2013) track. Zhou(2014) and Li(2013) made use of pattern matching technique, and Wei(2013) proposed a rule-based decision making approach based on CRF sequential tag-

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 79–85,
Beijing, China, July 30-31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

ging. Despite all these models, their performance has not shown satisfying results[1]. The identification of Chinese comparative elements especially still remains as a big challenge. A TBL-based approach, which showed good performance in Korean (Yang and Ko, 2011), would be an alternative to usual methods.

3 Task Description

3.1 Comparative Elements (CE)

We refer to Comparative Elements (CE) as entities and attributes which directly occur within comparative sentences. We defined 6 CEs as below.

e.g. 新飞度车身结构的刚度比前代提高了 164％。
The solidity of XinFeiDu's body structure has increased 164% than the previous design.

新飞度	刚度	比	前代	提高	164％
XinFeiDu	solidity	BI	previous design	increased	164%
SUB	DIM	BI	OBJ	RES	EXT

CE (label)	Definition
Subject Entity (SUB)	An element of comparison, i.e. topic of the sentence.
Comparative Marker (BI)	Comparative sentence marker, which is BI(比) in Chinese Bi-structure[2].
Object Entity(OBJ)	An entity that is being compared to. It is often the complement of Bi-prepositional phrase.
Dimension (DIM)	Shared property of entities being compared.
Comparative Result (RES)	The relation between entities being compared. It is often the syntactic head of comparative predicate.
Comparative Extent (EXT)	Relative difference in degree or quantity between entities in terms of DIM.

Table 1: Comparative Elements (CE) in BI-structure

Our task is to automatically extract these 6 CEs from the sentences. Note that these elements cannot simply be determined by syntactic criteria. They are involved with semantic category to some extent, but we do not use additional semantic features such as semantic role labels or lexical taxonomies in this paper.

3.2 Corpus

The corpus used in this experiment consists of 1,036 Chinese BI-structure sentences, coming from the open dataset of COAE 2013 Task 2 (Tan et al. 2013). The sentences are a collection of customer reviews and opinions from different Chinese websites pertaining to cars and electronics.

1) Preprocessing: We first conduct word segmentation and POS tagging by using ICTCLAS[3]. Second, we had to manually revise to avoid any errors because of the informal language used on the web. Three annotators were appointed to revise typos. In addition, 3,000 word-size domain-specific lexicons[4] are also utilized to guarantee the quality of word segmentation and POS tagging.

2) CE labeling: The 6 types of Comparative Elements (CE) in the 1,036 sentences were manually annotated with the corresponding CE labels of Table 1. This task was done by three trained annotators of Chinese linguistics major. Their work was double-checked by one another, and any inconsistencies between annotators were discussed before reaching an agreement. The annotated corpus was then transformed to IOB format.

4 Two methods of Comparative Elements Extraction

We now present two different proposed techniques. Model 1 uses basic part-of-speech chunking-based method and Model 2 employs Transformation-Based Error-Driven Learning (TBL) (Brill, 1995) for identifying CEs.

4.1 POS chunking-based CE extraction

4 elements of CEs, i.e. BI, OBJ, RES and EXT, form a regular sequential pattern across the sentences. First, OBJ generally occurs as complements of BI-prepositional phrase, which is mostly a noun phrase. Second, RES and EXT usually form predicates, modified by the BI-prepositional phrase, i.e. [[比 OBJ]$_{prep}$ [RES EXT]$_{pred}$]. Noticing this pattern, we can define chunk patterns with regular expressions as below.

Punctuation as delimiter, the sentence is divided into small clauses

If the clause contains"比/p", the following chunk rules are applied to create chunks.

[1] In COAE 2013 Task 2 (Chinese Comparative Element Identification), the best performance F1-score was 0.35 (Tan et al. 2013:25).

[2] There are other comparative markers such as 比不过, 不如, 优于 which are sometimes combined with RES morphologically. However, BI(比) is the only comparative marker in the scope of this paper.

[3] http://ictclas.nlpir.org/downloads

[4] Acquired from Sougou's open database. http://pinyin.sogou.com/dict/

```
Rule1. BI: {<P><.*>*}
Rule2. RES: <.*>}{<V>|<A[DN]*>|<D>
Rule3. RES: <D>|<V>|<VSHI>|<VYOU>|
        <A[DN]*>{}<V>|<A[DN]*>|<D>
Rule4. RES: <.*>{}<.*>*?<UDE1>[^<W
        J>][^$]
```

Label the items in the first chunk as BI and OBJ, and label the second chunk as RES.

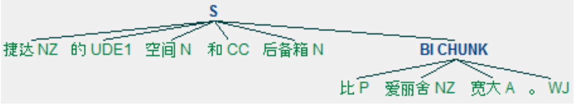
For RES chunks, Rule 5 is applied to chunk EXT.

```
Rule 5. EXT:<A>|<V[N]*>|<Y>{(<MQ>|<
          M>|<Q>|<RY>|<X>|<D>)<.*>*}
```

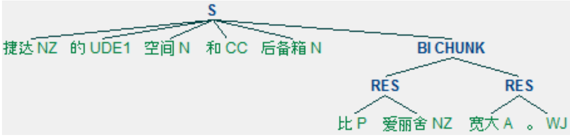
Table 2: chunking-based CE extraction process[5]

We now give a step-by-step illustration of the actual extraction process of Table 2.

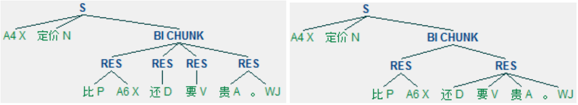
Step 1 Detecting BI (比)-Chunk
For clauses containing the comparative marker BI (比), we create a chunk that begins with it (Rule 1). We call it BI (比)-Chunk.



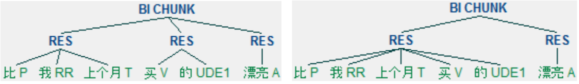
Step 2 Splitting into Two Phrases
BI (比)-Chunk can be divided into two chunks, i.e. BI-prepositional phrase and predicate because they belong to very distinctive syntactic categories. The former cannot appear independently, usually taking a noun as its object. The latter functions as the predicate, and is mostly an adjective or a verb[6]. Therefore, we designed Rule 2 to split them.



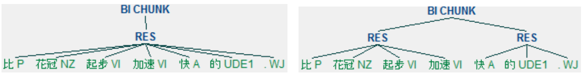
Step 3 Merging Incorrectly Separated Predicates
In most cases, however, the predicate is a complex phrasal structure. Therefore, Rule 2 incorrectly splits chunk that should have not been separated. To solve this, we employ Rule 3 to merge incorrectly divided elements of the predicate group.



Step 4 Dealing with DE (的)-Structure
In Chinese, DE (的) is often used to mark modification[7]. It can be attached to various types of syntactic categories and modify the following word. DE (的)-structure can be simplified as [[XP 的] NP]. When a verb or adjective phrase takes the position of XP, the same error as in Step 3 occurs. To tackle this problem, we use Rule 4 that enunciates the unity of modifying elements occurring at the position of XP.

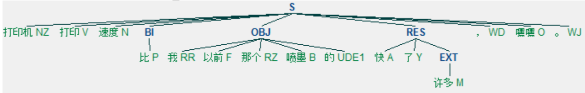


Note that DE (的) is not necessarily restricted to modification marker. When occurring at the end of the sentence, it simply marks a subjective tone. Rule 4 makes use of punctuation tag (WJ) to discern this modal particle of DE (的) from modification marker.



Step 5 CE Labeling
After successfully extracting the two chunks (BI-prepositional phrase and predicate) following the above mentioned 4 steps, we label each item in these chunks with BI, OBJ and RES tags.

Step 6 (optional) EXT Identification
Comparative Extent (EXT) usually begins with numerals, following the head of the predicate. Rule 5 detects possible EXTs in RES chunk.



4.2 TBL-based CE extraction

The advantage of using the POS chunking-based method is that it allows direct capture of linguistic information. However, (a) it requires painstaking process of manual rule construction; (b)

[5] For specifying chunk rules intuitively, we directly quote NLTK's description of chunking operator (Bird et al. 2009).

(1) <T> represents for any token tagged with T.
(2) {<pattern >} represents for Chunk Rule, which means creating a chunk with the given regex pattern within curly braces.
(3) <pattern1>}{<pattern2> represents for Split Rule, which means splitting a chunk into two chunks based on the specified pattern.
(4) <pattern1>{}<pattern2> represents for Merge Rule, which means merging two chunks together based on the specified pattern.

[6] Strictly speaking, verb and adjective are also able to occur in BI- prepositional phrase. Such a case will be handled in Step 4.

[7] It may be an inadequate way of defining DE (的) because of its flexible and diverse nature. Exceptional cases will be discussed in 5.1. See Zhu(1961) for further details.

and an error in any step could damage the performance of the whole CE extraction process.

4.2.1 Transformation-Based Learning

We tested an automated learning method, known as Transformation-Based Error-Driven Learning (TBL) (Brill, 1995). The basic idea of TBL is "learning from mistakes". First, the researcher may apply an initial-state annotator to the training corpus. Second, the set of user-defined templates are then used to form candidate rules. Third, each of the rules is in turn applied to the training corpus. At the same time, the net improvement of the rule is calculated and recorded for evaluation of candidate rules. Throughout the training, the process of deriving rules, scoring and selecting rules and applying them is iterated, creating an ordered sequence of transformation rules.

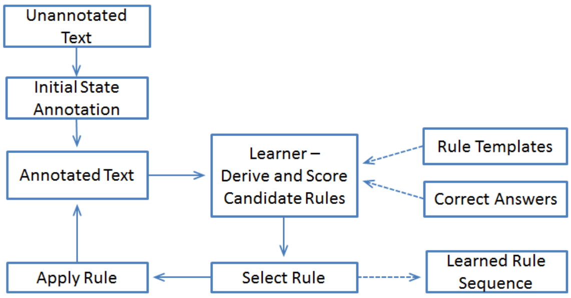


Figure 1: Learning Procedure of TBL
Modified from Brill(1995), Ramshaw and Marcus(1999)

TBL has a well-known advantage, i.e. perspicuity of linguistically meaningful rules. Different from the POS chunking-based model, the TBL-based model is more robust and possibly captures useful information that may not be noticed by the human engineer (Brill, 1995:552). Therefore, the use of TBL allows us to capture some otherwise ignored CE patterns.

4.2.2 Preliminary TBL-based CE Tagger

We treated CE extraction task as a tagging problem. Each token in training data is given an initial tag by ICTCLAS tagger. The TBL learner with user-defined templates is then trained on the training data. Consequently, we obtain the TBL-based CE tagger as a CE extraction model.

The templates play a key role in this model because the TBL-learner uses these templates to generate possible rules, which directly affect the overall performance. In order to see which type of feature contributes more to the TBL-based model accuracy rate, we divide the original template proposed in Brill (1995:553, 556) into three template subsets: (a) tag features template; (b) lexical features template; (c) both features template.

Type of features	# of templates	# of candidate rules	Accuracy (%)
(a) Tag	11	21,815	83.15
(b) Lexicon	15	59,442	85.39
(c) Tag & Lexicon	26	81,257	88.38

Table 3: CE Extraction Accuracy based on different feature templates

Since the result of (c) is the best, we take it as a standard template. Table 4 shows evaluation results using TBL-based model with 26 templates (c). We regard this score as our baseline performance.

	SUB	DIM	BI	OBJ	RES	EXT
Pre.	53.48	60.90	97.08	77.24	88.02	82.04
Rec.	19.31	38.07	97.22	80.82	69.86	69.57
F.	28.24	46.37	97.14	78.97	77.86	75.16

Table 4: The result of baseline system (%)

4.2.3 Search for Optimal Template

We now present how we obtained our proposed TBL-based CE extraction model. The baseline system is based on a relatively large amount of rules and templates. Therefore, the reduction of rules is preferable for efficient application of TBL. According to Brill (1995: 560), although the accuracy of TBL-based tagger increases with the number of transformation rules, its marginal effect dramatically decreases, and leads to computational cost. We found that 200~300 rules are desirable for our CE detecting task. As for templates, we achieved the best performance when using tag and lexicon sequences within a radius of 3 tokens as features.

For every token in BI-structure, change tag a to tag b when:	
1. The current word is ***w***.	W_0
2. One of the three preceding words is tagged *z*.	$T_{-3,-2,-1}$
3. One of the three following words is tagged *z*.	$T_{1, 2, 3}$
4. The word two after is tagged *z*.	T_2
5. The word two before is tagged *z*.	T_{-2}
6. The following word is tagged *z*.	T_1
7. The preceding word is tagged *z*.	T_{-1}
8. The preceding word is ***w***.	W_{-1}
9. The current word is ***w***, and the preceding tag is ***t***.	W_0 & T_{-1}
10. The preceding tag is ***t***, the current tag is t_2 and the following word is ***w***.	T_{-1} &T_0&W_1

Table 5: 10 proposed templates

Based on the above-mentioned templates, the TBL-based model generates a set of possible candidate rules. Table 6 lists 10 transformation rules of highest score.

Pass	Old tag	Context	New tag
1	P	W_0＝比	BI
2	A	$T_{-3,-2,-1}$ = BI	RES
3	NZ	$T_{-3,-2,-1}$ = BI	OBJ
4	N	$T_{-3,-2,-1}$ = BI	OBJ
5	M	$T_{-3,-2,-1}$ = RES	EXT
6	A	$T_{-3,-2,-1}$ = OBJ	RES
7	N	$T_{1,2,3}$ = BI	DIM
8	NZ	$T_{1,2,3}$ = BI	SUB
9	UDE1	$T_{-3,-2,-1}$ = BI	OBJ
10	X	$T_{-3,-2,-1}$ = BI	OBJ

Table 6: 10 Rules of highest score

With the rules given above, the model takes example (1a) as an input, and applies the rules in order of 1 → 2 → 3 → 5 → 7 → 8 → 9, producing the CE-tagged result of example (1b).

(1a) E1-471/nz ,/wd 声音/n 比/p AS4752/nz 的/ude1 好/a 多/m 了/y 哦/e

(1b) E1-471/sub ,/wd 声音/dim 比/bi AS4752/obj 的/obj 好/res 多/ext 了/y 哦/e

In addition, Examples (2-5) below illustrate some of the linguistically meaningful transformation rules that the TBL model based on 10 templates (Table 5) has captured.

(2) `VYOU->RES if Word:更@[-1]`
比/bi 同价位/obj 机型/obj 更/d 有/res 分量/res 。/wj
BI same-price model more have amount
(This product) has more amounts for the same price.

(3) `EXT->RES if Word:要@[-1]`
比/bi 捷达/obj 的/obj 要/v 多/res 得/ude3 多/ext 。/wj
BI Jetta DE should more DE much
(A car model's something) should be much more than a Jetta's.

In (2-3), "更" is equivalent to "more" in English, and "要" conveys subjective meaning of difference in degree. The proposed model makes use of them as an RES marker because they frequently occur before RES.

(4) `RZ->OBJ if Word:比@[-1]`
诺基亚/sub 5230/sub 同等/b 价位/n 下/f 比/bi 其它/obj 手机/obj 都/d 好/res
Nokia 5230 is even better than the equivalent class of other cellphones.

(5) `RES->EXT if Word:好@[0]& Pos:RES@[-1]`
比/bi 老/obj 天籁/obj 的/ude1 油漆/dim 硬/res 好/ext 多/ext 。/wj
(A car model's coating) is much stronger than the coatings of Teana.

In (4), the pronoun following BI "其它" is likely to be a constituent of OBJ. "好" is very likely to be a degree complement, i.e. EXT, if RES precedes it. Instead of functioning as RES, it stresses the degree of RES as shown in example (5).

5 Results

5.1 Result of POS Chunking-based Model

The overall performance of the chunking-based CE extraction model (Section 4.1) is as follow.

	BI	OBJ	RES	EXT
Precision	96.94	75.96	42.03	63.33
Recall	96.94	82.63	86.65	60.03
F-score	96.94	79.15	56.60	61.63

Table 7: The results of chunking-based CE extraction (%)

The CE mining process of chunking-based model is based on simplistic grammatical assumptions: (a) Only nominal elements serve as the complement of BI-prepositional phrase; (b) Predicates can be a word or a group of words (phrase) that are adjectives or verbs; (c) Within the BI (比)-Chunk, the elements occurring before the modifier marker DE (的) are all regarded as modifier. These assumptions, of course, are somewhat over-generalized, and do not fit in many real cases[8]. However, the 5 Rules applied based on these assumptions show a fair performance in Table 7 when applied to a limited scope of BI-Chunk.

5.2 Result of TBL-based Model

All evaluations of TBL-based model in this paper are based on a 5-fold cross validation. The proposed TBL-based model with 10 templates shows the results below.

	SUB	DIM	BI	OBJ	RES	EXT
Pre.	53.35	62.13	97.41	77.16	87.94	79.78
Rec.	23.04	40.70	97.04	83.15	69.50	72.31
F.	31.88	48.46	97.22	80.01	77.58	75.79

Table 8: The results of TBL-based CE extraction (%)

Guided by our new templates (Table 5), the model first locates comparative marker BI (比), then searches the surroundings for the tag/lexical features while gradually narrowing its scope. As a result, the 10 templates enable an effective detection of elusive CE instances such as those in example (2-7).

[8] Under many circumstances in Chinese, a noun (or noun phrase) can also serve as predicate; Transferred-designation(转指) "XP 的" construction can also act as subject other than as a modifier.

Model	SUB	DIM	BI	OBJ	RES	EXT
POS chunking-based	-	-	96.94	79.15	56.60	61.63
TBL-based (baseline, 26 Templates)	28.24	46.37	97.14	78.97	77.86	75.16
TBL-based (proposed, 10 Templates)	31.88	48.46	97.22	80.01	77.58	75.79

Table 9: Comparison between models (f-score, %)

Table 9 compares the scores of two CE mining methods, the POS chunking-based and the TBL-based approach. Compared to the TBL-based model, the POS chunking-based model is unable to extract SUB and DIM. Because these two elements frequently occur outside a BI-prepositional phrase, it is hard to capture their irregular occurrence positions in the sentence. In contrast, the TBL-based model is able to detect SUB and DIM. However, their identification rate is relatively low.

Nevertheless, our proposed TBL-based model outperforms the baseline system by using much smaller templates. It shows we found a simple and more expressive set of rule templates.

Moreover, the proposed TBL-based model achieved an increase of 21% for RES and 14% for EXT f-score in comparison with the POS chunking-based model. This improvement mainly benefits from the proper use of both tag and lexical information.

(6) 市场/nz 中/f 比/p 它/rr 靓/a 的/ude1 产品/n 很/d 少/a 。/wj

There are very few products prettier than that one in the market.

(a) 市场/n 中/f 比/bi 它/obj 靓/obj 的/ude1 产品/obj 很/res 少/res 。/WJ

(b) `RES->A if Word:很@[-1]`
市场/n 中/f 比/bi 它/obj 靓/res 的/ude1 产品/n 很/d 少/a 。/WJ

(7) 花冠/nz 比/p 伊兰特/nz 贵/a 近/a 3 万/m

Corollas are more expensive than Elantras by nearly 30 thousand RMB.

(a) 花冠/nz 比/bi 伊兰特/obj 贵/res 近/res 3 万/ext

(b) `RES->EXT if Word:近@[0]`
花冠/nz 比/bi 伊兰特/obj 贵/res 近/ext 3 万/ext

As for examples (6-7), the POS chunking-based model (Section 4.1) incorrectly identifies “少, 近” as RES. As we can see in example (6a), the POS chunking-based model wrongly identifies “很少” as RES because the BI-chunk “比它靓” occurs in front of the modification marker “的”. In (7a), the model mistook “近” for RES because it cannot discern “近” from “贵” only with the tag information. In contrast, TBL-based model makes a correct decision of (6b) and (7b) based on lexical information.

6 Conclusion

In order to make the best use of meaningful features in linguistic context, we have proposed the use of two rule-based methods for Chinese comparative element (CE) extraction. The POS chunking-based model performs well with basic Chinese grammatical rules. We then use the TBL-based method to extract other linguistic patterns that the first model can hardly detect. Results showed that our TBL-based model achieved higher score than Brill’s (1995), demonstrating that our new 10 templates can effectively extract the distinct features of Chinese BI-structure as shown in examples (1-7).

Chinese comparative element mining involves techniques of various domains including coreference resolution, named entity recognition and parsing. However, the linguistic features used in this paper are limited to instances of regular (type-3) grammars. In our future work, we plan to investigate some feasible Chinese linguistic features on the level of context-free grammars.

References

Bird, Steven, Edward Loper and Ewan Klein. 2009. Natural Language Processing with Python. O’Reilly Media Inc.

Brill, Eric. 1992. A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the workshop on Speech and Natural Language*, pp.112-116.

Brill, Eric. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4): 543-565.

Ding, Shengshu. 1961. Xiandai Hanyu Yufa Jianghua. *Beijing: Commercial Press*.

Huang, Xiaojiang et al. 2008. Learning to Identify Chinese Comparative Sentences. *Journal of Chinese Information Processing*, 22(5): 30-37.

Jindal, Nitin and Bing Liu. 2006. Identifying Comparative Sentences in Text Documents. In *Proceedings of SIGIR’06*, pp.244-251.

Jindal, Nitin and Bing Liu. 2006. Mining Comparative Sentences and Relations. In *AAAI* (22): 1331-1336.

Li, Linding. 1986. Hanyu Jufa Juxing. *Beijing: Commercial Press*, pp.285-301.

Li, Yan et al. 2013. PRIS_COAE at COAE 2013 Track. In *Proceedings of the fifth Chinese Opinion Analysis Evaluation*, pp. 53-69.

Liu, Yuehua. 1983. Shiyong Xiandai Hanyu Yufa. *Beijing: Foreign Language Teaching and Research Press*, pp.833-854.

Lü, Shuxiang. 1942. Zhongguo Wenfa Yaolüe. *Beijing: Commercial Press*, pp.352-370.

Ma, JianZhong. 1989. Mashi Wentong. *Beijing: Commercial Press*, pp. 134-142.

Ramshaw, Lance. A. and Mitchell P. Marcus. 1999. Text Chunking Using Transformation-based Learning. In *Natural language processing using very large corpora*, pp. 157-176. Springer Netherlands.

Shao, Jingmin. 1990. Biziju Tihuan Guilü Chuyi. *Zhongguo yuwen*, *vol.6*.

Song, Rui et al. 2009. Chinese Comparative Sentences Identification and Comparative Relations Extraction. *Journal of Chinese Information Processing*, 23(2): 102-122.

Tan, Songbo et al. 2013. Overview of Chinese Opinion Analysis Evaluation 2013. In *Proceedings of the fifth Chinese Opinion Analysis Evaluation*, pp. 5-33.

Wei, Xianhui et al. 2013. DUTIR: Method Research of Sentiment Analysis and Elements Extraction of Chinese Short Text. In *Proceedings of the fifth Chinese Opinion Analysis Evaluation*, pp. 116-128.

Yang, Seon and Youngjoong Ko. 2011. Extracting Comparative Entities and Predicates from Text Using Comparative Type Classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1636-1644.

Zhou, Hongzhao et al. 2014. Chinese Comparative Sentences Identification and Comparative Elements Extraction Based on Semantic Classification. *Journal of Chinese Information Processing*, 28(3): 136-141.

Zhu, Dexi. 1961. Shuo De. *Zhongguo yuwen*, 12:1-15.

Appendix

ICTCLAS Part-of-Speech Tags

Tag	Part-of-speech	
A	Adjective	形容词
AD	Adverbial adjective	副形词
AN	Nominal adjective	名形词
D	Adverb	副词
E	Exclamative particle	叹词
M	Numeral	数词
N	Noun	名词
NZ	Proper noun	专有名词
P	Preposition	介词
Q	Classifier	量词
RY	Wh-pronoun	疑问代词
UDE1	"De"	的
VSHI	"Shi"	是
VN	Gerund	名动词
VYOU	"You"	有
WJ	Period	句号
X	Character	字符
Y	Modal particle	语气词