Conceptual Collocations of Words in Comparative Sentences

Xiangfeng Wei, Quan Zhang, Yi Yuan
Institute of Acoustics,
Chinese Academy of Sciences /
21 North 4th Ring Road, Haidian District,
Beijing, 100190, China
{wxf,zhq,yuan}@mail.ioa.ac.cn

Zhejie Chi
University of Chinese Academy of Sciences /
No.19A Yuquan Road, Beijing 100049, China
chizhejie@sina.com

Abstract

A comparative sentence is a kind of sentence for describing the comparative result after comparing two objects. There are many conceptual collocations of words in comparative sentences, so it is necessary for studying the types and characteristics in Chinese comparative sentences. This paper summarized some usual conceptual collocations in ten kinds of comparative sentences. It proposed a method of recognizing comparative sentences according to these conceptual collocations. Our system, which was built for recognizing and analyzing comparative sentences, got 80.95% F-value in the test set.

1 Introduction

A comparative sentence is a kind of special sentence in human language. Comparative sentences express the objects, range and the result of the comparison by the statement. For the characteristics of comparative sentences, many linguists have summarized and researched them. Some divided comparative sentences into comparison of the same, comparison of difference, and comparison of the supreme (Rui et al., 2009). Some classified them into similarity, competition, change and other types (Rui et al., 2009). And some summarized 28 kinds of comparative sentences (Jun and Xiaobing, 2005). These studies provide guidance for the acquisition and understanding of comparative sentences by human. Based on linguistic studies, researchers in Natural Language Processing(NLP) proposed a method for identifying Chinese comparative sentences based on support vector machines (SVM) (Xiaojiang et al., 2008), and a method for retrieving comparative relations based on conditional random field (CRF) (Gaohui et al., 2010). Some did the task based on information entropy and syntactic tree with semantic roles (Jianjun 2011). For English comparative sentences, Jindal is the first person who used SVM and CSR (Class Sequential Rules) algorithm to identify comparative sentences (Jindal and Liu, 2006a), and later used the LSR (Label Sequential Rules) to retrieve comparative elements; both tasks got good effects (Jindal and Liu, 2006b). Feldman used rules to extract the names and attributes of products from internet forum, and to compare different products (Feldman et al., 2007). Sun and etc. studied how to retrieve the relationship between two comparative objects through Web search engine (Jiantao et al., 2006).

According to Chuanjiang (2005), comparative sentences are divided into three main types based on their concepts: mutually comparing, comparing within a set and comparing with a standard. The most basic comparison is comparing between two objects, for example, "A is better than B". In Chinese comparative sentences, it can form different sentences, such as "A 比 B 优"(A BI B YOU, A is better than B) , "与 B 相比 A 优"(YU B XIANG BI A YOU, Compared with B, A is better), and "A 和 B 一样优秀"(A HE B YI YANG YOU XIU, A and B are the same excellent). If the comparison within a set is a kind of internal comparison in a collection, generally there will be the word that means the most "最 (ZUI)". Sometimes it uses the negative expression in this kind of sentence, such as "世界上没有哪个国家比美国更强大(SHI JIE SHANG MEI YOU NA GE GUO JIA BI MEI GUO GENG QIANG DA, No country in the world is more powerful than the United

States)". Standard comparison is a comparision with a standard, that is to say that one of the two comparative objects is the standard. It usually uses a word like "符合(FU HE, consistent)", "相符(XIANG FU, conform to)", such as "信封的尺寸符合国家统一标准"(XIN FENG DE CHI CUN FU HE GUO JIA TONG YI BIAO ZHUN, The size of an envelope is consistent with the national standard).

This paper firstly described the sentence patterns of the comparative sentences, and then summarized the conceptual collocations in comparative sentences. A method to identify whether a sentence is a comparative was proposed according to these conceptual collocations and the structure characteristics of comparative sentences. The system for recognizing comparative sentences, which was built on the methods in this paper, got the result of 80.95% F-value in the test set of the Fourth Chinese Opinion Analysis Evaluation (COAE2012).

2 Sentence Patterns of Comparative Sentences

The basic elements of a comparative sentence include: comparative objects and the results of comparison. Comparative objects may be people, things, events and their attributes. The results of comparison can be divided into the same (or almost the same) and the different. The result of the different usually show which object is better and which is worse (the words or concepts of ‘better’ and ‘worse’ may be different according to the objects or their attributes). Based on the range and the nature of the comparative objects, the comparative sentences are divided into: mutual comparative sentences, within a set comparative sentences and standard comparative sentences. Reference comparative sentences appear more usually in mutual comparative sentences. In a reference comparative sentence, an object is as the reference point; another object may be better or worse than it. According to the literature (Chuanjiang 2005), we listed ten kinds of sentence pattern of comparative sentences as the followings:

（1）jD00J = DB1+jD0+DB2

（2）jD000J = jDBC+jD0

（3）jD001J = jDB+jD0+DC

（4）jD021J = DB1+ljlv+DB2+jDC

（5）jD022J = l02+DB2+ljlv+DB1+jDC

（6）jD01J = Cn+DB0+jDC

（7）jD011J=Cn+jlv116+DB+ljlv+DB0+jDC

（8）jD012J = Cn+jlv116+ljlv+DB0+jDBC

（9）jD02J = DBC+jD0+DB02

（10）jD020J = DBC+l02+DB02+jD0

Among them, (1), (2) and (3) are mutual comparative sentences. In the sentence patterns DB1 is the first comparative object, and DB2 is the second comparative object, and jD0 is the results of comparison. DB1 or DB2 can be the specific person, object, or event, and may be their properties. The first comparative object is often an object and its attribute, and the property of the second object is often omitted. For example:

Example 1: 骏捷和 F3 的车内空间和外观毫不逊色于A3。(jD00J = DB1 + jD0 + DB2)
(JUN JIE HE F3 DE CHE NEI KONG JIAN HE WAI GUAN HAO BU XUN SHE YU A3.)
(The interior space and look of F3 and Junjie are favorable as A3.)

In Example 1 "The interior space and look of F3 and Junjie" is the first comparative object DB1, "are favorable" is the result of the comparison jD0, "A3" is the second comparative object DB2.

Example 2: 速腾 1.4T 手动与宝来最高配的车价差不多。(jD000J = jDBC + jD0)
(SU TENG 1.4T SHOU DONG YU BAO LAI ZUI GAO PEI DE CHE JIA CHA BU DUO.)
(The prices of manual 1.4T Sagitar and high-end Bora are almost the same.)

In Example 2, the two comparative objects and their attributes together become jDBC "The prices of manual 1.4T Sagitar and high-end Bora", and "are almost the same" is the result of the comparison jD0.

Example 3: 闰土和我仿佛年纪。 (jD001J = jDB + jD0 + DC)
(RUN TU HE WO FANG FU NIAN JI.)
(Runtu and I seem to the same age.)

In Example 3, the comparative object jDB is "Runtu and I", "the same" is the result of the comparison jD0, and “age" is the two objects' attribute.

Sentence patterns (4) and (5) are reference comparative sentences. They appear more often in language than other mutual comparative sentences. In this kind of sentence, an object is as the reference, the other is compared with it. Chinese reference comparative sentences often use "than", "compared with ......" and other words, for example:

Example 4: 这车子比 q7 和卡宴还要炫。(jD021J = DB1 + 1jlv + DB2 + jDC)
(ZHE CHE ZI BI Q7 HE KA YAN HAI YAO XUAN.)
(This car is more dazzling than Q7 and Cayenne.)

In Example 4 "This car" is the first comparative object DB1, "比(BI, compare or than)" is the special concept for comparison in Chinese as logic concept ljlv, "Q7 and Cayenne" as the reference for comparison namely DB2, "is more dazzling" is the result of the comparison namely jDC.

Example 5: 与前代的产品相比，它的操控感明显更好。(jD022J = l02 + DB2 + ljlv + DB1 + jDC)
(YU QIAN DAI DE CHAN PING XIANG BI, TA DE CHAO KONG GAN MING XIAN GENG HAO.)
(Compared with the previous generation, it is obviously better in sense of control.)

In Example 5, "与(YU, with)"(l02) is the concept of indicator for comparative objects, "previous generation"(DB2) is used as a reference for comparison. "相比(XIANG BI,compared)"(ljlv) is a special kind of logic concept in Chinese for comparison. "它的操控感(TA DE CAO KONG GAN, its sense of control)" is a comparative object DB1. "obviously better"(jDC) is the result of the comparison. Of course, the semantic chunks in order of appearance in a sentence pattern can be flexible. For example, Example 5 can be expressed as "它的操控感与前代的产品相比明显更好(TA DE CHAO KONG GAN YU QIAN DAI DE CHAN PING XIANG BI MING XIAN GENG HAO. Compared with the previous generation, it is obviously better in sense of control.)". The sequence of semantic chunks was adjusted from "l02+DB2+ljlv+DB1+jDC" to "DB1+l02+DB2+ljlv+jDC".

Sentence patterns (6), (7) and (8) are comparisons within sets. The comparison is within a collection of objects. One object in the collection is compared with other objects within the collection. The comparative result is generally expressed as the most. The basic statement is sentence pattern (6): jD01J=Cn+DB0+jDC. Cn is the collection. DB0 is the object for comparison. jDC is the comparison result. For example:

Example 6: 这个价位的 Andriod 手机里 v880 最好。(jD01J = Cn + DB0 + jDC)
(ZHE GE JIA WEI DE ANDRIOD SHOU JI LI V880 ZUI HAO.)
(v880 is the best in this level price of Andriod phone.)

In Example 6, the collection of comparison Cn is "this level price of Andriod phone". The object to be compared DB0 is "v880". The result of comparison jDC is "the best."

Within a set comparative sentences can be added into negative concept(jlv116). The result of the comparison in fact is also the most. For example, "世界上没有哪个国家比美国更强大(SHI JIE SHANG MEI YOU NA GE GUO JIA BI MEI GUO GENG QIANG DA. No country in the world is more powerful than the United States)". This example belongs to sentence pattern (7): jD011J=Cn+jlv116+DB+ljlv+DB0+jDC, where "世界上(SHI JIE SHANG, in the world)" is the collection Cn, "没有(MEI YOU, no)" is the negative concept jlv116, "哪个国家(NA GE GUO JIA, any country)"(DB) is other objects within the collection, "美国(MEI GUO, the United States)" is the main comparative object DB0, "更强大(GENG QIANG DA, more powerful)"(jDC) is the result of the comparison. This sentence pattern can also be expressed as "世界上没有比美国更强大的国家(SHI JIE SHANG MEI YOU BI MEI GUO GENG QIANG DA DE GUO JIA. There is not any country in the world which is more powerful than the United States)". It belongs to sentence pattern (8): jD012J= Cn+jlv116+ljlv+DB0+jDBC, where "更强大的国家(GENG QIANG DA DE GUO JIA, more powerful country)"(jDBC) is the semantic chunk merged from the comparative object and its attribute.

Sentence patterns (9) and (10) belong to standard comparative sentences. In standard comparative sentences one comparative object is a standard, such as "信封的尺寸要符合国家统一标准(XIN FENG DE CHI CUN YAO FU HE GUO JIA TONG YI BIAO ZHUN. The size of envelopes must meet to the national standard)". The sentence belongs to sentence pattern (9): jD02J=DBC+ jD0+DB02, where "The size of envelopes" is the DBC, "must meet to" is jD0, "the national standard" is DB02; Another example is "信封的尺寸要与国家统一标准相符(XIN FENG DE CHI CUN YAO YU GUO JIA TONG YI BIAO ZHUN XIANG FU. The size of envelopes must be consistent with the national standard)". This example belongs to sentence pattern (10): jD020J=DBC+l02+DB02+ jD0, where "信封的尺寸(XIN FENG DE CHI CUN, the size of envelopes)" is DBC, "与(YU, with)" is l02, "国家统一标准(GUO JIA TONG YI BIAO ZHUN, the national standard)" is DB02, "相符(XIANG FU, consistent)" is jD0, and "要(YAO, must)" is a part of jD0 which is isolated in front of jD0.

3 Identifying Comparative Sentences

When identifying whether a sentence is a comparative sentence, the characteristic of the collocation between two concepts in some sentence patterns of comparative sentences is very useful. For example, sentence pattern (4), which is jD021J=DB1+ljlv+DB2+jDC, is the most common comparative sentence in Chinese language. In sentence pattern (4), the concept ljlv is often "比(BI, than)", while the semantic chunk jDC is often the concept of j, such as quantity, quality, degree, old and new, etc., as shown in Table 1.

ljlv	j
比 (BI, than)	大/小(DA/XIAO, bigger/smaller); 好/坏/差(HAO/HUAI/CHA, better/worse/worse); 多/少 (DUO/SHAO, more/less); 快/慢(KUAI/MAN, faster/lower); 新/旧(XIN/JIU, newer/older); 长/短 (CHANG/DUAN, longer/shorter); 便宜/贵(PIAN YI/GUI, cheaper/more expensive); 逊/胜 (XUN/SHENG, not good as/better); 高/低/矮(GAO/DI/AI, taller/lower/lower); 漂亮/好看/丑 (PIAO LIANG/HAO KAN/CHOU, more beautiful/more pretty/uglier); 强/弱(QIANG/RUO, stronger/weaker); 舒服(SHU FU, more comfortable); 先进(XIAN JIN, more advanced); 轻松 (QING SONG, more relaxed); 轻/重/沉(QING/ZHONG/CHEN, lighter/heavier/heavier); 宽/窄 (KUAN/ZHAI, wider/narrower); 简单/复杂(JIAN DAN/FU ZA, simpler/more complex); 增加/减少(ZENG JIA/JIAN SHAO, more/less); 优秀(YOU XIU, excellent); 大气(DA QI, more generous); 方便(FANG BIAN, more convenient)

Table 1. The Collocation Between ljlv and j

For the second type of reference comparative sentences, sentence pattern (5), which is jD022J =l02+DB2+ljlv+DB1+jDC, the concept l02 is often "与(YU, with); 和(HE, with); 同(TONG, with); 跟(GEN, with)", while the concept ljlv is often some specific concepts of comparison, such as "比(BI, compare); 相比(XIANG BI, compare);", as shown in Table 2.

l02	ljlv
与(YU, with); 和(HE, with); 同(TONG, with); 跟(GEN, with)	比(BI, compare); 相比(XIANG BI, compare); 比较(BI JIAO, compare); 比起来(BI QI LAI, compare); 比较起来(BI JIAO QI LAI, compare)

Table 2. The Collocation Between l02 and ljlv

For mutual comparative sentences, sentence pattern (2), which is jD000J=jDBC+jD0, the two objects in the chunk jDBC for comparison are often joined by a concept like "和(HE, and)", while the semantic chunk jD0 is the concept like "一样(YI YANG, the same); 媲美(PI MEI, the same); 相同(XIANG TONG, the same); 差不多(CHA BU DUO, almost the same)", or like "区别(QU BIE, distinguish); 差距(CHA JU, distance); 竞争(JIN ZHENG, compete)" (in this case there are generally concepts like "有(YOU, has)", "没有(MEI YOU, has not)" before them in the sentence), as shown in Table 3.

l02	Ljlv	other concepts
与(YU, with); 和(HE, with); 同(TONG, with); 跟(GEN, with)	一样(YI YANG, the same);媲美(PI MEI, the same); 相同(XIANG TONG, the same); 无异(WU YI, the same);一模一样(YI MO YI YANG, the same); 相似(XIANG SHI, like); 相近(XIANG JIN, like); 接近(JIE JIN, like); 差不多(CHA BU DUO, almost the same); 不同(BU TONG, different)	区别(QU BIE, distinguish); 差距(CHA JU, distance); 竞争(JIN ZHENG, compete)

Table 3. The Collocation Between l02,ljlv and Other Concepts

For comparative sentences within a set, sentence patterns (6), (7), (8), they generally include the Chinese character "最(ZUI, most)" that indicates the comparative result. But conversely, not all sentences that include "最(ZUI, most)" are comparative sentences. To be a comparative sentence within a set, the sentence must include a set. The set usually acts as the supplementary semantic chunk in a sentence. And there is an indicating concept (l1) for a supplementary semantic chunk. These indicating concepts and the concepts of the superlative degree of an adjective (j60d01) are the identifying characteristics of comparative sentences within a set.

l1 or area	j60d01
中(ZHONG, in); 里(LI, in); 上(SHANG, on); 内(NEI, within); 以内(YI NEI, within); 地区(DI QU, area); 全国(QUAN GUO, around the country); 全球(QUAN QIU, around the world); 世界(SHI JIE, in the world); 市场(SHI CHANG, in the market); 方面(FANG MIAN, with respect)	最(ZUI, most); 最好(ZUI HAO, best); 最佳(ZUI JIA, best)

Table 4. The Collocation Between l1/area and j60d01

For mutual comparative sentences, sentence pattern (1), which is jD00J=DB1+jD0+DB2, the semantic chunk jD0 is the concept of a result of the comparison followed by "于(YU, than)", "过(GUO, than)", or preceded by "不(BU, not)". These concepts can form a concept for comparison, as shown in Table 5.

ljlv
好于(HAO YU, better than); 大于(DA YU, greater than); 高于(GAO YU, higher than); 好过(HAO GUO, better than); 赛过(SAI GUO, better than); 逊于(XUN YU, not better than); 不亚于(BU YA YU, as good as); 不如(BU RU, not better than); 不及(BU JI, not better than); 不次于(BU CHI YU, not as bad as); 不低于(BU DI YU, not as low as)

Table 5. The Concepts of ljlv

According to the mentioned above sentence patterns and conceptual characteristics of comparative sentences, we can set the appropriate rules and templates to identify whether a sentence is a comparative sentence. These templates and rules can improve the performance of identifying comparative sentences. To validate these sentence patterns and conceptual characteristics, we used the test set of task 2.1 (a task of identifying comparative sentences) in the Fourth Chinese opinion analysis evaluation (COAE2012). The test set is divided into cars and electronics areas. Each area has its own 3600 sentences, some are comparative sentences and some are not comparative sentences. According to the given answers, our result for identifying comparative sentences got 80.95% F-value.

Our result is higher than the average of participating systems, but is lower than the best system. The reasons, after we checking our system, are mainly in the following aspects: (1) We did not use word segmentation. Some conceptual concepts in specific words cannot act as comparative concepts. For example, the Chinese character "比(BI, than)" in Chinese words "性价比(XING JIA BI, price-performance ratio)", "百分比(BAI FEN BI, percent)", and "高宽比(GAO KUAN BI, height-width ratio)" does not act as the concept of comparison. Therefore, the sentences that contain these words may be not comparative sentences. (2) In the comparative sentences that contain the negative concept, such as "没有(MEI YOU, not)+DB2+jDC"(not good/bad as DB2), the concepts for jDC are too many to be collected, and sometimes it may be even a sentence. Therefore, the accuracy of identifying comparative sentences according to "没有(MEI YOU, not)" and other characteristic concepts will be lower than our anticipation. (3) Some comparative sentences omitted "比(BI, than)" but use the Chinese character "是(SHI, is/are)". For example, the sentence "跟STMP3770 还是有点小差别的(GEN STMP3770 BI HAI SHI YOU DIAN XIAO CHA BIE DE. There is still a little difference compared with STMP3770)". Therefore, this kind of comparative sentences cannot be identified by our system. (4) For the conceptual characteristics are acquired from words, if some words are not included in our system, then the sentences that contain the words or conceptual characteristics will not be identified. We hope our system can learn these words by machine learning or statistic training.

4 Conclusion

We have summarized the patterns of comparative sentences and analyzed some usual conceptual collocations in Chinese comparative sentences. Accordingly we put forward a method to identify comparative sentences. This method is able to identify most of the comparative sentences, but there are also some shortcomings and deficiencies. The main reason is our conceptual words are not comprehensive enough. This affected the precision and recall rate in identifying comparative sentences. There is also a need to introduce the method of statistical learning model to learn the characteristics of the conceptual collocation in comparative sentences as much as possible, and to adapt to specific comparative sentence types and varied words and concepts.

Comparative sentence is a kind of sentence that has its own features, so the study of comparative sentences analysis should focus on the sentence types and characteristics. Comparative sentences generally contain two objects or their properties to be compared and give the comparative result. In comparative sentences there would be some special comparative concepts and the basic concepts of comparative results. These concepts are very helpful to our analysis and provide us with favorable conditions. However, not all comparative sentences abide these sentence types and conceptual collocations. Some comparative sentences may have more than three objects or omit some comparative objects. Some have their own special expressions. The diversity of concepts in comparative objects and results is also a difficult problem to be solved. We believe that with the gradual mutual combination between the statistical model analysis methods and the rules analysis methods, it will improve the precision and performance of analyzing comparative sentences.

Acknowledgement

This paper is supported by Twelfth Five-Year Project Subject of National High Technology Research and Development Program (863 Program) (Grant No. 2012AA011102); State Language Commission Twelfth Five-year Research Project, The Semantic Association in Linguistic Concept Space (Grant No. YB125-53); Information Technology Project of CAS, Science and Technology Information Resource Aggregation; Information Technology Project of CAS, Data and Application Integration for Minority Language Information Processing(Grant No. Y329251431); The Knowledge Innovation Engineering Project of Institute of Acoustics, CAS (Grant No. Y154141431).

References

Chuanjiang Miao. 2005. Introduction of HNC (Hierarchical Network of Concepts) theory. Tsinghua Press, Beijing, China, 243-246.

Feldman R, Fresko M, Goldenberg J. 2007. Extracting product comparisons from discussion boards[C] //Proc of the 7th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 469-474.

Gaohui Huang,Tianfang Yao,Quansheng Liu. 2010. Mining Chinese Comparative Sentences and Relations based on CRF algorithm. Application Research of Computers,27(6):2061-2064.

Jianjun Li. 2011. Research on the Identification of Comparative Sentences and Relations and Its Application, Thesis of master degree, Chongqing University, China.

Jiantao Sun, Xuanhui Wang, Dou Shen, et al. 2006. CWS: a comparative Web search system[C] //Proc of the 15th International Conference on World Wide Web. New York, ACM Press, 467-476.

Jindal N, Liu Bing. 2006a. Identifying comparative sentences in text documents[C] //Proc of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM Press, 244-251.

Jindal N, Liu Bing. 2006b. Mining comparative sentences and relations[C] //Proc of the 21st National Conference on Artificial Intelligence. Boston:AAAI Press,1331-1336.

Jun Chen, Xiaobing Zhou. 2005. Selecting and Sorting the grammar items of comparative sentences. Language Teaching and LinguisticStudies,2005(2):22-33.

Rui Song, Hongfei Lin, Fuyang Chang. 2009. Chinese Comparative Sentences Identification and Comparative Relations Extraction. Journal of Chinese Information Processing,23(2):102-107,122.

Xiaojiang Huang, Xiaojun Wan, Jianwu Yang. 2008. Learning to Identify Chinese Comparative Sentences. Journal of Chinese Information Processing,22(5):30-37.