International Journal of Database Theory and Application
Vol.7, No.5 (2014), pp.29-38
http://dx.doi.org/10.14257/ijdta.2014.7.5.03

Recognizing Comparative Sentences from Chinese Review Texts

Wei Wang, TieJun Zhao, GuoDong Xin and YongDong Xu

Department of Computer Science and Technology,
Harbin Institute of Technology,
Harbin, China
{wangwei}@hitwh.edu.cn, {tjzhao, gdxin }@hit.edu.cn, {ydxu}@insun.hit.edu.cn

Abstract

Comparisons play an important role in making decisions by referring to the comparative opinions of opinion holders in earlier customer reviews. Recognizing comparative sentences from review texts contributes to opinion mining and information recommendation. Our objective is to automatically recognize comparative sentences from Chinese text documents. In this paper, an effective approach is proposed based on comparative patterns to recognize comparative sentences in Chinese. Our experiments on customer-generated product reviews show that the proposed approach is effective.

Keywords: *Comparative Sentence; Comparative Pattern; Review Text*

1. Introduction

Comparisons are one of the most distinctive thinking and evaluation ways. A large number of valuable information is contained in comparative sentences, e.g., contrasts on two or more entities based on their shared features, comparisons between different attributes of the same entity, the customer's preferences etc. Therefore, recognizing comparative sentences from text documents can assist people to gain information quickly and make the appropriate decision-making.

Comparisons and direct opinions are two kinds of different evaluation ways that express author's preferences in review text. For example, "诺基亚的通话质量很好(The call quality of Nokia is very good)", which directly expresses a positive sentiment on the call quality of Nokia. "诺基亚的通话质量比三星好(The call quality of Nokia is better than that of Samsung)" ,which compares Nokia and Samsung based on their call quality and expresses a preference for Nokia. Clearly, Comparative sentences have very different syntactic structures and semantic meaning with direct opinion sentences.

Two useful clues can be observed from a subset of comparative sentences. One of them is that almost each comparative sentence includes a keyword (a word or one pair of words) indicating comparison, such as"比(than)", "相同(same)" , "最(most,-est)". The other clue is that comparative sentences have particular patterns involving comparative keywords, which called comparative patterns. These patterns are related to some key elements of comparative sentences, i.e., comparative keyword, compared entities, and comparative result.

In a typical comparative sentence, there are usually two or more comparative entities, a keyword that indicates comparison. The entities being compared are often located on both sides of the comparative keyword, for example, "宝马的发动机比奔驰的好(The engine of BMW is better than that of Benz)", which "比(than)" is comparative keyword, "BMW" and "Benz" are comparative entities that appear on both sides of "比(than)".

ISSN: 2005-4270 IJDTA

Our work aims to explore a method for automatically recognizing comparative sentences in Chinese text documents.

In this paper, we propose an approach based on comparative patterns to recognize comparative sentences from Chinese text documents. The basic process of solving the problem is to first use keyword look-up technology to recognize all candidate comparative sentences, and then build comparative patterns to filter out non-comparative sentences from candidates. In building the patterns, some other factors are considered such as relative position relations of compared entities, comparative keyword, and comparative result.

The remainder of this paper is organized as follows: Section 2 discusses the related work in recognizing comparative sentences and comparative relations. Section 3 gives the problem statement including the definition and category of comparisons. Section 4 states the proposed technique. Section 5 experiment and evaluation. Section 6 concludes our study and discusses future directions.

2. Related Work

The researches about comparative sentences are mainly in two fields, linguistics and computational linguistics. Researchers in linguistics are concerned about semantics and syntax of comparative sentences, instead of the automatic recognition technology. Shang [1] summarizes the various classification systems of comparative sentences in modern Chinese. We can see that there are not uniform opinions how to classify comparative sentences in Chinese study. Chen and Zhou [2] generalize 20 kinds of sentence patterns according to the syntactic structures of comparative sentences. Due to non-formal and limited patterns, they can not be directly applied to our task. Che [3] discusses semantic types and structure types of comparative sentences whose predicates contain comparative keywords or comparative patterns. In summary, the literatures in linguistics about how to distinguish comparative sentences from non-comparative sentences have not been found.

In computational linguistics, Jindal and Liu [4] first put forward a research project about the automatic recognition of comparative sentences by computer in 2006. They adopt an integrated CSR (Class Sequential Rules) and supervised learning approach to recognize comparative sentences from English texts. Experiment results show a precision of 79% and recall of 81%. Huang *et al.*, [5] applies several supervised machine learning methods to classify a Chinese sentence into either “comparative” or not. Song *et al.*, [6] constructs a Chinese comparative pattern database and uses it to recognize comparative sentences. Alaa [7] combines POS (Part of Speech) tags and several learning methods to extract comparative statements in Arabic. Park and Blake [8] experiment a certain number of learning methods to detect comparative claims automatically from full-text scientific articles.

The work on comparisons in computational linguistics has been expanded to extract comparative relations from the identified sentences and determine the preferred objects. A comparative relation defined in [9] includes compared entities, compared features, a comparative keyword and a comparative type. Jindal and Liu [9] extract comparative relations using a new type of rules called LSR (Label Sequential Rules). Xu *et al.*, [10] builds a graphical model based on two-level CRF to recognize comparative relations and the directions of relations on mobile phone review data. Wang *et al.*, [11] constructs and uses hybrid comparative patterns to label compared entities and

compared features for Chinese comparative sentences. This paper refers to hybrid comparative patterns in [11] to recognize comparative sentences.

3. Problem Statement

This section provides the evidences related to automatic recognizing comparative sentences, including an introduction of Chinese Part-of-Speech tags, the definition of comparison, and the classification of comparison.

3.1. Part-of-Speech (POS) Tags

Since Part-of-Speech (POS) tags will be used in the subsequent discussion, here we first introduce some tags used in this paper and their POS categories. We used ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System) to execute Chinese word segmentation and POS tagging. Important POS tags and their POS categories are: n: Noun, nz: Proper Noun, x: Character String, m: Numeral, r: Pronoun, a: Adjective, v: Verb, vn: Gerund, d: Adverb, p: Preposition, u: Auxiliary, y: Modal Particle.

In English, comparisons are usually expressed using comparative words or superlative words. Chinese comparisons lack of morphology changes of comparative and superlative, but there are also certain grammatical marks (prepositions, adverbs, etc.).

3.2. Definitions

A comparative sentence describes at least one relation based on similarities or differences of two entities on an attribute. The differences are further divided into gradable (greater or lesser/superlative) and non-gradable. It should be noted that more than one comparative relation may be contained in a sentence.

A comparative relation describes a comparative opinion in a formal way. A Chinese comparative relation consists of five key elements, which can be expressed as a 5-tuple:

(KW, E1, E2, A, R)

Where KW is the keyword, E1 and E2 are two entities being compared, E1 is comparative subject, E2 is comparative object, A is the compared attribute, and R is the comparative result. The number of elements may be less than five in some sentences, which means some elements can be omitted in the specific context.

Categories of Comparisons: There are two main comparative relation categories: gradable and non-gradable [4], and each category can be further divided into several sub-categories. Gradable comparison: expresses an ordering relationship of entities on a certain attribute. It has two sub-categories, Greater or Lesser, and Superlative. non-gradable comparison: states a relation between entities without ordering. It has three sub-categories, Equality, Non-gradable difference, and implicit comparison. The categorization hierarchies of comparative relations are shown in Table 1.

Table 1. Categorization Hierarchies of Comparative Relations

Type		Example Sentence	Keyword
gradable	Greater or Lesser	“与 y 相比，X 具有更好的图像质量。” (“Compared with Y, X has better image quality.”)	与……相比 (Compared with)
	Superlative	“在所有相机中，x 性能最好。” “The performance of X is the best among all cameras.”	最(best)
non-gradable	Equality	“X 的外观与 Y 相近。” “The appearance of X is similar to that of Y.”	与……相近 (similar to)
	Difference	“X 在材质上不同于 Y。” “X is different from Y in the materials”.	不同于 (different from)
	Implicit comparison	“X 有摄像功能，而 Y 没有这项功能” “X has camera function, but Y does not have.”	Null

4. Approach

Our objective is to automatically recognize comparative sentences from Chinese text documents, which can be regarded as a 2-class classification problem. We use comparative patterns to recognize comparative sentences. Meta patterns are first extracted by mining frequent items of POS sequences including keywords. Comparative patterns are then built according to the Meta patterns. A keyword technology used to recognize candidate comparative sentences is also designed in this section.

4.1. Comparative Keyword

In order to build a Chinese keyword dictionary, the following three steps are performed.

1. A list of keywords was drafted for each sub-category except for implicit comparison.

(1) We first manually collected 10 seed words for each sub-category.

(2) Then we use Tongyici Cilin to find their synonyms.

(3) After manual pruning, we generated a list of keywords for each sub-category.

2. The lists built above were merged into a keyword dictionary including 324 words.

3. Then we added 7 words (e.g. but, however etc) to the keyword dictionary for implicit comparison.

We have compiled a keyword dictionary containing 331 words. Then we use keywords to extract all comparative-sentence candidates, i.e., the sentences that do not

contain any keyword are ruled out. In this experiment, we obtained a recall of 99.54% and a precision of 27.62%. The high recall shows that these keywords can cover almost all comparative sentences. In order to improve the precision, we construct comparative patterns to exclude the non-comparative sentences that include keywords.

4.2. Comparative Pattern

For building comparative patterns, we need consider the structure of a comparative relation. As mentioned earlier, a comparison relation consists of five key elements, among which comparative keyword, two compared entities and comparative result is important to generate comparative patterns. The focus of our attention is their relative position relationship and POS. For example, "等离子电视在动态显示方面比液晶电视好（Plasma TV is better than LCD TV in the dynamic display）", which "比(than)" is comparative keyword, "Plasma TV" and "LCD TV" are compared entities that appear on both sides of "比(than)". After omitted attribute, this comparative relation can be formalized as <E1, KW, E2, R>. We then replace each element with its POS tag. For each keyword, the actual keyword and the POS tag is combined to form a single item, *i.e.*, <{n}{比(than)/v}{n}{a}>.

4.2.1. Meta Patterns

For finding such patterns, called Meta pattern, we generate a POS sequence for each comparative sub-sentence in training set. Each POS sequence is then simplified by using heuristic rules being manually compiled. The algorithm that extracts Meta schema is as follows:

1. Preprocessing stage. We perform Chinese word segmentation and POS tagging for all comparative sentences in training set.
2. A POS sequence is generated for each comparative sub-sentence.

(1) For each comparative sentence in training set, we find all keywords in it and replace their POS tags with a special symbol $.

(2) Inspecting each sub-sentence in a comparative sentence, if it contains any keyword, and then generates a POS sequence for the sub-sentence.

3. Using heuristic rules simplify each POS sequence to the simplest form. Several typical Meta patterns are shown in Table 2. For the example of section 4.2, < $ n a > is its Meta pattern.
4. The high-frequency sequences are extracted as meta patterns.

Part of Heuristic Rules:

R1 : <adjective noun> -> noun
R2 : < noun noun > -> noun
R3 : < gerund noun> -> noun
R4 : <adverb adjective> -> adjective

Table 2. Typical Meta Patterns

Type	Meta Pattern	comparative pattern
1	< $ n a > < $ x a > < $ m a >	比/p /n /a (than/p /n /a) 比/p /x /a (than/p /x /a)

		比/p /m /a (than/p /m /a)
2	< $ n > < $ x > < $ m >	相比/v /n (compared with /n) 相比/v /x (compared with /x) 相比/v /m (compared with /m)
3	< $ n $ > < $ x $ > < $ m $ >	和/cc 一样/a (the same... as) 与/p 区别/n (different...between)

In pattern sequences, the types of comparative entities have noun(n), string(x), and numeral(m). The types of comparative result are mainly adjective and verb. Type description of key elements in Meta pattern is shown in Table 3.

Table 3. Type Description of Key Elements

Type	Form	Meaning	Sample	Denote
Entity	*nz*	proper noun	索尼 (Sony)	product brand
	x	string	FX01	product model
	m	numeral	5230	
	n	noun	相机(camera)	product type
Result	*a*	adjective	好(good)	
	v	verb	缩小(reduce)	

4.2.2. Comparative Patterns

We can get several formats of comparative relations from Meta pattern.

Format 1: <{E1}, {比|有|… (than | as |…)}, {E2}, {R}>;

Format 2: <{E1}, {优于|等于|…(superior to | equal to |…)}, {E2}>;

Format 3: <{E1}, {不如|比不上|…(inferior to | can not compare with|…))}, {E2}, [{R}]>;

Format 4: <{相比|对比|…(compared with| contrast with |…)}, {E2}, {,}, {E1}, {R}>;

Format 5: <{E1}, {和|与|…(and| and|…)}, {E2}, {比较|如出一辙|各有千秋…(compare |same | each has its merits)}, [{R}]>;

Format 6: <{和|与|…(and| and|…)}, {E2}, {相比|一样|…(compare | same|…)}, {,}, {E1}, {R} >.

In the formats above, the content in "<>" is a complete comparative format, which denotes the sequence structure of key elements in a comparative relation. The contents within "[]" are the optional items. In these formats, the format 1 matches type 1 of Meta patterns, format 2,3,4 correspond with type 2 of Meta patterns, and format 5,6 accord with type 3 of meta patterns.

4.2.3. Generalized Pattern

We have built a generalized pattern database. Generalized patterns are described as following:

Pattern 1: < keyword class1, entity class, result class >;
Pattern 2: < keyword class2, entity class >;
Pattern 3: < keyword class3, keyword class4>;
Table 4 gives some examples of syntactic categories in generalized patterns:

Table 4. Examples of Syntactic Class

Category	Example
keyword class1	比，像…(than, as…)
entity class	/n, /x, /m……
result class	/a, /v……
keyword class2	相比，对比...(*compared with ,contrast with*…)
keyword class3	和，与，跟... (and, and, with…)
keyword class4	一样, 不同… (*same,different*…)

4.2.4. Manual Rules

Some rules drafted manually also are added to pattern database about superlative comparison and implicit comparison. Because superlative comparison generally contains only a compared entity and implicit comparison hasn't obvious keywords and patterns, these rules are difficult to obtain by existing mining techniques. For instance, 是/vshi 中/f 最/d(is ... the most) is a rule of superlative comparison. The added rules include 45 superlative comparisons and 30 implicit comparisons.

5. Experimental Results

5.1. Data Sets

Our data sets consist of the product reviews on digital cameras, notebooks, automobiles, and mobile phones from the first Chinese Opinion Analysis Evaluation (COAE 2008). Reviews of each product are composed of a data set. The number of comparative sentences and non-comparative sentences in each dataset is given in Table 5.

The data sets were manually labeled by three trained annotator. They labeled in accordance with the definition of Section 3 and discussed their differences to reach an agreement.

5.2. Experimental Results

We adopt precision, recall, and F-score to verify the effectiveness of our approach. Figure 1 gives the whole results that include the precision, recall, and F-score value of different methods. In Figure 1, KWs is keywords, MR is manual rules, and G-Patterns denote generalized patterns. We use 5-fold cross validations to obtain all results, which

the datasets are randomly partitioned into 5 subsets, 4 of them are used as training set, and the final one is used as the validation set.

Table 5. Number of Sentences in Each Dataset

Data set	Comparative Sentences	Non-Comparative Sentences
Dataset 1 (camera)	527	1592
Dataset 2 (notebook)	159	791
Dataset 3 (automobile)	260	1218
Dataset 4 (mobile phone)	194	1458
Total	1140	5059

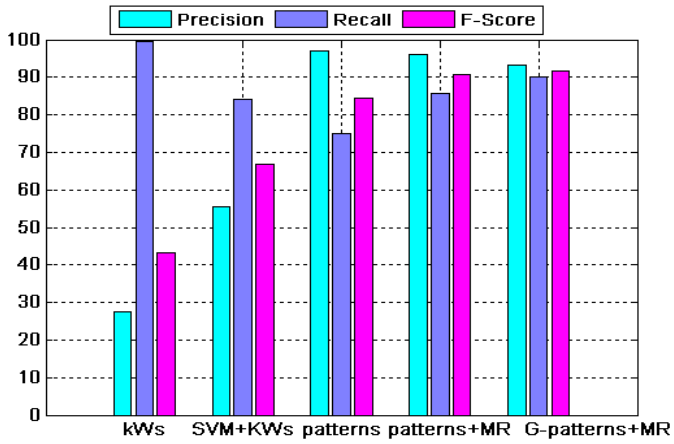

Figure 1. The Results of the Proposed Approach (%)

The Discussion of Result

(1) Keywords: We used the keyword dictionary compiled in this work to filter out the sentences that do not contain any keyword. The recall of 99.54% is achieved. This shows that these keywords can cover almost all comparative sentences. But the precision is less than 30%, which indicates that many sentences that contain keywords are not comparative sentences.

(2) SVM using keywords as features: Applying SVM to classify the sentences, we obtained the F-score of 66.82%. The SVM learning method improves the precision, but in the meantime reduced the recall.

(3) Comparative patterns: Using alone comparative patterns to classify each sentence among the comparative-sentence candidates, we achieve the precision of 96.99%, the recall of 75%, and the F-score of 84.59%.

(4) Comparative patterns and manual rules: All patterns that contain comparative patterns and manual rules are used to recognize comparative sentences. The recall and F-score values are significantly improved. The F-score of 90.6% is achieved. This shows that complicated manual rules are useful for our task.

Generalized patterns and manual rules: Using generalized patterns and manual rules, the recall is increased to 90%, and the F-score reaches 91.57 %, which is the best result among these methods.

6. Conclusion

This paper solves the problem of recognizing comparative sentences from Chinese review texts, which is a 2-class classification problem. To recognize comparative sentences, an approach is proposed based on comparative patterns involving information of compared entities, comparative keyword and comparative result. A keyword technology is also designed in this paper, which used to recognize candidate comparative sentences. The experimental results show the effectiveness of the method. In our future work, we will classify comparative sentences as different sub-categories and extract comparative relations from the recognized sentences.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61173073, No.61172099, No.61073130), the Major Project of National 863 Program of China (No. 2011AA01A207), the Key Project of the National Natural Science Foundation of China (No. 60736044).

References

[1] P. Shang, "A Review on the System of Comparative Sentence", Applied Linguistics, **(2006)**, pp. 77-80.
[2] J. Chen and X. Zhou, "The Selection and Arrangement of Grammatical Items concerning Comparative Sentences", Language Teaching and Research, no. 2, **(2005)** , pp. 22-33.
[3] J. Che, "A Brief Analysis of Comparative Sentences in Modern Chinese", Journal of Hubei Normal University (Philosophy and Social Science), vol. 25, no. 3, **(2005)**, pp. 60-63.
[4] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents", In Proceedings of SIGIR'06, **(2006)** , pp.244-251.
[5] X. J. Huang, X. J. Wan, J. W. Yang and J. Xiao, "Learning to Identify Comparative Sentences in Chinese Text", In Proceedings of PRICAI'08, **(2008)**, pp. 187-198.
[6] R. Song, H. F. Lin and F. Chang, "Chinese Comparative Sentences Identification and Comparative Relations Extraction", Journal of Chinese Information Processing, vol. 23, no. 2, **(2009)**, pp. 102-107.
[7] E. H. Alaa, "Opinion Mining from Arabic comparative sentences", In Proceedings of ACIT'12, **(2012)**, pp. 10-13.
[8] D. Park and C. Blake, "Identifying Comparative Claim Sentences in Full-Text Scientific Articles", In Proceedings of ACL'12, **(2012),** pp. 1-9.
[9] N. Jindal and B. Liu, "Mining Comparative Sentences and Relations", In Proceedings of AAAI'06, **(2006)**, pp. 1331-1336.
[10] K. Q. Xu, S. S. Y. Liao, J. X. Li and Y. X. Song, "Mining Comparative Opinions from Customer Reviews for Competitive Intelligence", Decision Support Systems, vol. 50, no. 4, **(2011)**, pp. 743-754.
[11] S. G. Wang, H. X. Li, and X. L. Song, "Automatic Semantic Role Labeling for Chinese Comparative Sentences Based on Hybrid Patterns", In Proceedings of ICAI'10, **(2010)**, pp. 378-382.
[12] M. Ganapathibhotla and B. Liu, "Mining Opinions in Comparative Sentences", In Proceedings of Coling'08, **(2008),** pp. 18-22.
[13] S. Li, C. Y. Lin, Y. I. Song and Z. Li, "Comparable Entity Mining from Comparative Questions", In Proceedings of ACL'10, **(2010)**, pp. 650-658.
[14] B. Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, May **(2012)**.
[15] J. Pei, H. Pinto, Q. Chen, J. Han, B. Mortazavi-Asl, U. Dayal and M. Hsu, "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth", In Proceedings of IEEE International Conference on Data Engingeering (ICDE-2001), **(2001)**.
[16] E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions", In Proceedings of EMNLP'03, **(2003)**.

[17] Q. Xia, "A Review on Studies of Comparative Sentences of Chinese", Chinese Language Learning, no. 2, **(2009)**, pp. 58-64.
[18] S. Yang and Y. Ko, "Finding Relevant Features for Korean Comparative Sentence Extraction", Pattern Recognition Letters, vol. 32, no. 2, **(2011)**, pp. 293-296.
[19] B. Liu, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Second Edition)", Springer, **(2011)**.
[20] K. Liu, S. Wang, X. Liao and H. Xu, "Overview of Chinese Opinion Analysis Evaluation 2012", In: Proceedings of the 4st Chinese Opinion Analysis Evaluation, NanChang, China: The Professional Committee of Information Retrieval, **(2012)**, pp. 1-32.

