Methods for Identifying Comparative Sentences

S.K. Saritha
Department of Computer Science
National Institute of Technology, Bhopal

R K Pateriya
Department of Computer Science
National Institute of Technology, Bhopal

ABSTRACT

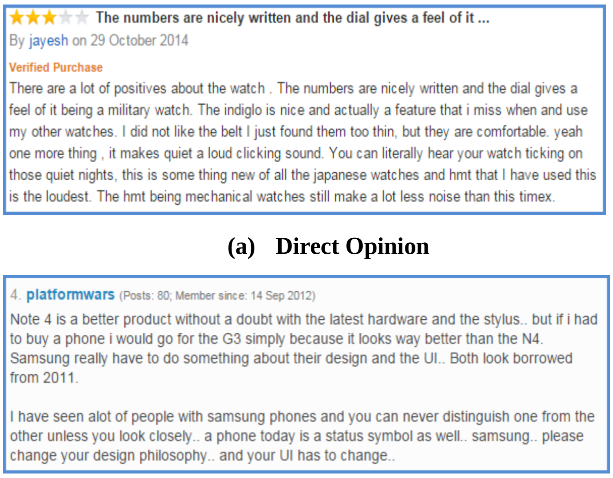
Comparative sentences are used to express the explicit classifications between two entities with respect to the degree or quantity to which they possess some gradable property. Identifying the comparative sentences is a challenging task. This gives a new research direction for the researches. In this paper various approaches which were used for the identification of the comparative sentences in the text documents are studied.

Keywords

Sentiment analysis, Sequential Rule Mining, Class Sequence Rule Mining, Machine learning.

1. INTRODUCTION

Sentiment analysis (also known as opinion mining) is a state of feeling, emotion, attitude and opinions. It refers to the use of computational linguistics, natural language processing and text analysis to identify and extract subjective information in web reviews. The reviews hold the Sentiments of the user towards the product and comparative opinion over the similar products. Comparisons are one of the most substantial ways of evaluation. The evaluation can be done in two ways: Direct and Comparisons. In direct (also call direct opinion) gives the positive or negative opinion about the product and their features. Comparisons compare the product and their features with some other similar product and their features. For example,



★★★☆☆ **The numbers are nicely written and the dial gives a feel of it ...**
By jayesh on 29 October 2014
Verified Purchase
There are a lot of positives about the watch . The numbers are nicely written and the dial gives a feel of it being a military watch. The indiglo is nice and actually a feature that i miss when and use my other watches. I did not like the belt I just found them too thin, but they are comfortable. yeah one more thing , it makes quiet a loud clicking sound. You can literally hear your watch ticking on those quiet nights, this is some thing new of all the japanese watches and hmt that I have used this is the loudest. The hmt being mechanical watches still make a lot less noise than this timex.

(a) Direct Opinion

4. **platformwars** (Posts: 80; Member since: 14 Sep 2012)
Note 4 is a better product without a doubt with the latest hardware and the stylus.. but if i had to buy a phone i would go for the G3 simply because it looks way better than the N4. Samsung really have to do something about their design and the UI.. Both look borrowed from 2011.

I have seen alot of people with samsung phones and you can never distinguish one from the other unless you look closely.. a phone today is a status symbol as well.. samsung.. please change your design philosophy.. and your UI has to change..

(b) Comparison

Figure 1: Snapshots of reviews written by the user.

Since the comparison between the products is seen in comparative sentences, it is necessary to identify these comparative sentences for business management. As and when a new product is launched in the market the companies want to know where there products stand in the market.

2. COMPARATIVE SENTENCES

Comparative is a sentence structure used when comparing two things. A significant amount of research has been done in sentiment and opinion extraction and classification over the ***subjective sentences*** whereas the comparative sentences can be ***subjective*** or ***objective sentences***.

(1) ***Subjective Sentences*** are the sentences which holds some opinion

Ex: "I Love learning Hindi"

The examples show the positive opinion of the Hindi language learner.

(2) ***Objective sentences*** are the sentences that define the phenomena that one can for example describe, generally anything that is factual and based on knowledgeable interpretation or the systematic method.

Ex. "Hindi is a language."

This example does not hold any opinion, it's a fact.

Ex: "This car is certainly **better** but it's much **more expensive**."

As one can see that language construct in comparative sentences, example above are different from opinion sentences. Generally Comparative sentences use ***comparative adjectives*** **(**Like -er/-est**)** to describe people and things. So, identifying comparative sentences is a challenging task.

In this paper, various methods which were used by the researchers for identifying the comparative sentences from the different documents are studied.

3. RESOURCES

Due to advancement of WWW, lot of information is posted on it. Most of the information is in the textual form. This information can be used for text mining task. Figure 2 shows the state of web. A few resources where one can find the domain based data for Sentiment analysis and opinion mining task are listed below.

3.1 Blogs

A blog is an online private journal or log which is updated frequently. It's a place where one can express a lot to the world. It can have individual authors or be a collection of authors. Typically have a people who comment or respond to the blog post which is mostly domain based. It has a history or an archive of previous blog posts. One can get lot of textual information on blogs it can also have other domain content and have links to those websites.

The people post comments about the product, like they share their likes and dislikes about the product. These comments help the manufacturers, business personals and so on to know the advancement of their product in the market.

3.2 E-commerce Sites

E-Commerce or eCommerce are serving in products or services using the WWW. Humans look at what they spend on; it's very natural to expect that they want to know everything about the item for consumption they're going to buy, the purchase process, payment methods, the delivery service and so on, which helps them to make safe decision and give their view over the deal.

Figure 2: A Snapshot from Web on State of Web

3.3 Datasets

A Dataset is a collection of data. One can find dataset that are bench mark data which can be used by the researchers for validation to their research. Many datasets are available online for download for the researchers.

3.4 Review Sites

A review site is the great way of representing the reviews by the reviewers about the product, business, people or services. Most of the people are depending on these reviews to know about the product or services before buying or going to the place, as these reviews are mostly user generated reviews.

4. COMPARATIVE SENTENCE IDENTIFICATION

A few researches have been done on comparative sentences. The researchers have used different supervised and unsupervised techniques to identify the comparative sentences and the relations. A various methods which are available for identification of these sentences are discussed.

4.1 Linguistic Approach

In [1], the author proposed a linguistic approach for identifying the comparative sentences. In this author tried to categorize different types of comparative sentences on the basis of syntax and semantics. **Syntax and semantics** [2] are terms used in relation to characteristics of language. It is concerned with the structure of language. It is a matter of the logical or grammatical form of sentences, rather than what they refer to or mean. **Semantics** is concerned with the meaning of words and sentences.

Author stated that some comparative sentences use the morphemes *more/-er*, *less* and *as* explicitly for the purpose of creating collections of superiority, inferiority and equality, and so on to see that to which the object is compared.

4.2 Types of Comparative Sentences

A comparative sentence defines at least one similarity or difference relation between two aspects. A sentence may include an aspect and more than one topic on which the comparisons are made. A topic can be names of a person, a product brand, a company, a location, so on. An aspect is the part or property of the relation that is being compared. There are 4 types of comparatives [3]:

(1) ***Non – Equal Gradable:*** These are the relations of the type greater or less than which express the total ordering of some topic with regard to certain aspect. Keywords like *better, ahead, beats*, etc [4].

Ex: "Mobile X camera is better than that of Mobile Z"

(2) ***Equative:*** These are relation of the type equal to that states two aspects are equal with respect to some topic. Keywords and phrases like *equal to, same as, both, all*

Ex: "Mobile X *and Mobile Y both have good features*"

(3) ***Superlative:*** These are relations of the type greater or less than all others that rank one topic over all others. Keywords and phrases like *best, most, better than all*

Ex: "Mobile X *is the cheapest mobile available in market"*

(4) ***Non – Gradable:*** The aspects of two or more topics are compared in the sentences but they don't explicitly grade them.

Ex: "Mobile X has good camera and Mobile Z has good battery life".

The first three types are comparative which are called gradable comparatives and the last one is non – gradable comparative.

4.3 Sequential Pattern Mining Approach

Sequential Pattern mining (SPM) is a data mining technique used for finding statistically relevant patterns between the data examples where the values are delivered in a sequence. The values are presumed as discrete. Sequential pattern mining is a special case of structured data mining [5]. There are numerous computational problems which are addressed within this field, which include building efficient databases and indexes for sequence information, extracting the frequently occurring patterns, comparing sequence for similarity, and recovering missing sequence members. So, the Sequence mining problems can be classified as a string mining which is typically based on string processing algorithm and itemsets mining which is typically based on association rule mining. There are two types of Sequential Rule mining: Class Sequential Rule Mining and Label Sequential Rule Mining.

4.3.1 Class Sequential Rules Mining

Class sequential rule (CSR) [3] is sequential patterns which are shown in left and their class on the right. It is used for classification of sentences. CSRs are found automatically using a class sequential rule mining system. For a given labeled data set, a minimum support and a minimum confidence threshold, CSR mining finds all class sequences rules in the sequence data.

In [6] the author proposed a technique to identify the comparative sentences using the combination of CSR mining and machine learning. For classification, they experimented with two approaches:

1. By directly applying the CSRs.
2. By using a machine learning algorithm to build a classifier based on the rules.

They presented that POS tags of JJR, RBR, JJS and RBS are good indicators for identifying the keywords. They used WordNet to find the synonyms of the list of 30 words which they obtained manually through a subset of comparative sentences. After manual pruning, a final list of 69 words was recorded. Apart from this they included 9 more words and phrases such as but, whereas, on the other hand, etc. for the non – gradable comparative sentences. In this keyword strategy, the recall rate was very high and the precision was low, so they simply tried to improve the precision by including only those sentences that contained atleast one keyword and then they generated the CSR to filter out the non – comparative sentences. The sentences which did not contain any keywords were discarded. All together they obtained 83 keywords and key phrases.

The experimental results showed that, the performance of the naïve Bayesian classifier on individual dataset showed that Precision of 84 % on Reviews, 75% on Article and 73% on forums dataset. Recall was 80%, 80%, 83% and F-Score was 82%, 77% and 78% on data sets stated above. The overall result is shown in figure 3.

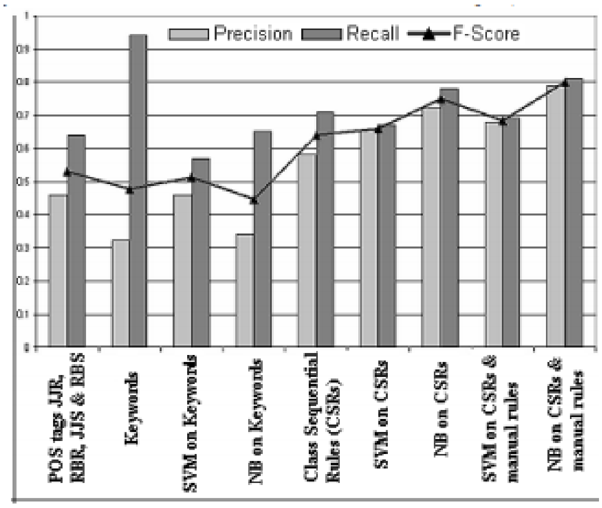
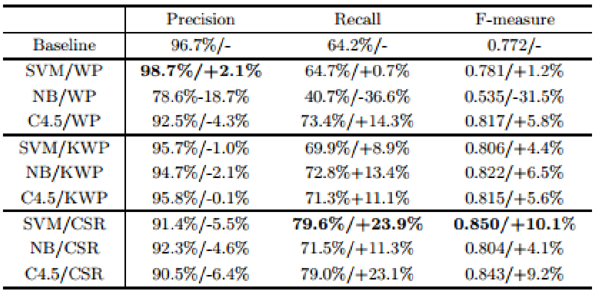


Figure 3: Shown the Precision, recall and F-score values of different approaches

In [7], the author has defined the problem of Chinese comparative sentence identification. Various Classifiers were used to classify a Chinese sentence either "comparative" or "non – comparative". Sequence and class sequence rule used to the patterns having high correlation with each class. The experiments were carried on some note book reviews form ZOL product forum [8], and then they manually labeled each sentence in the reviews. The dataset contains 1297 non – comparative and 458 comparative sentences. SVM, NB, decision tree were used for classification. The result of on various approaches is shown in figure 4. The features which include words plus their POS tags (denoted as WP), manual selected keywords plus their POS tags (denoted as KWP) and patterns obtained by Class Sequential Rule mining (denoted as CSR).



	Precision	Recall	F-measure
Baseline	96.7%/-	64.2%/-	0.772/-
SVM/WP	**98.7%/+2.1%**	64.7%/+0.7%	0.781/+1.2%
NB/WP	78.6%-18.7%	40.7%/-36.6%	0.535/-31.5%
C4.5/WP	92.5%/-4.3%	73.4%/+14.3%	0.817/+5.8%
SVM/KWP	95.7%/-1.0%	69.9%/+8.9%	0.806/+4.4%
NB/KWP	94.7%/-2.1%	72.8%+13.4%	0.822/+6.5%
C4.5/KWP	95.8%/-0.1%	71.3%+11.1%	0.815/+5.6%
SVM/CSR	91.4%/-5.5%	**79.6%/+23.9%**	**0.850/+10.1%**
NB/CSR	92.3%/-4.6%	71.5%/+11.3%	0.804/+4.1%
C4.5/CSR	90.5%/-6.4%	79.0%/+23.1%	0.843/+9.2%

Figure 4: Results of different approaches on Chinese sentences

4.3.2 Label Sequential Rules

A label sequential rule (LSR) [3] is of the following form, X →Y, where Y is a sequence and X is a sequence produced from Y by replacing some of its items with wildcards, "*". A wildcard matches to any item. The support and confidence are similar to those of CSR and SPM.

In [3], the author tried to identify the comparative sentences and extracted the relations (aspects) from these sentences. For identifying the comparative sentences they used the CSR mining and set of keywords as stated in [6]. Author also tried to extract the comparative relation. To extract the relation entries/ items Label sequential rules (LSR) were used. To perform this they followed two assumptions:

(1) Checked atleast one relation is present in a sentence.

(2) Assumed that entities and features are nouns (includes nouns, plural nouns and proper nouns) and pronouns. Sometimes a noun may be used in its verb form or some action described the verb ('e.g., "Nokia costs more", where "costs" is an feature and it is verb)

They created sequence database for mining. After the sequence database was built, they generated all the frequence sequences whose minimum support was 1%. Stored those sequences which had atleast one label, and the label had no POS tag associated with it. The remaining sequences were used to generate the LSRs.

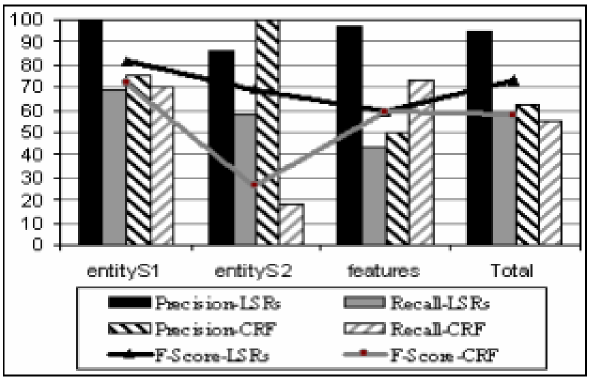


Figure 5: Shown the Precision-Recall and F-Score results of LSRs and CRF for extracting relation entries

The experimental results showed that for identifying gradable comparatives using the Naïve Bayes(NB) Classifier and CSRs, the performance gave a precision of 82% and recall of 81% (F-Score = 81%). For gradable comparative sentences the NB gave an accuracy of 87% and Support Vector Machine (SVM) gave an accuracy of 96%.

The performance measures for the extraction of relation, LSRs gave an overall F – score of 72%, while CRF gave an overall F – score of 58%. The experimental results showed that these methods are quite promising which is shown in figure 5.

4.4 Machine Learning Approaches

Supervised learning is an algorithm provided with a label for every example set, which are used to learn a mapping from the example set to labels.

In [9], goal of the author was to identify comparative sentences automatically from full text scientific articles. They introduced a set of sematic and syntactic features that characterize a sentence and then they demonstrated how these features can be used in three different classifiers: NB, SVM and Bayesian network (BN). The experiment were conducted on 122 full text toxicology article containing 14,157 sentences, of which 1,735 (12.25%) were comparisons. The experiments shown an F1 score of 71%, 69% and 74% on development set and 76%, 65% and 74% on a validation set for NB, SVM and BN, respectively.

In [10], the author proposed an automatic identification method of comparative sentences in Korean text documents. They used machine learning techniques to eliminate the non – comparative sentences from the candidates. The author classified the comparative sentences into six types, they are equality (same), similarity (similar), difference (different), greater or lesser (than), superlative (most) and predicative. Later he extracted comparative keyword form each type. Finally he setup 177 comparative keywords. 277 online documents were collected from various domain and three annotators compiled the corpus. The experimental results shown that precision, recall and F1 score for comparative keywords came up to 68.39%, 95.96% and 79.87. Comparative keywords and NB gave 85.42, 88.59 and 86.67 which is shown in the figure 6.

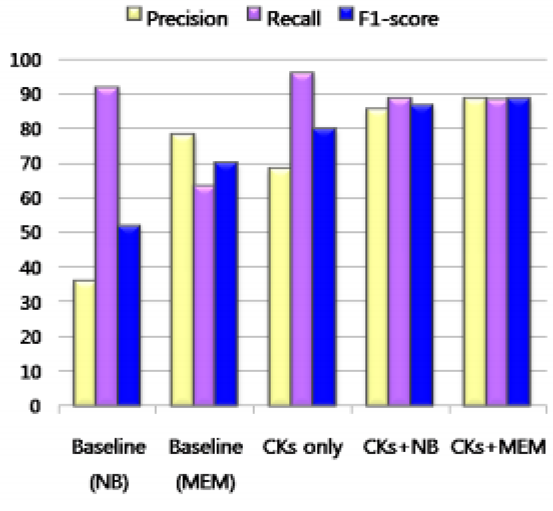


Figure 6: Shown the Precision-Recall and F-Score results

In [10], author studied the problem of comparative relation mining. (1) To understand the comparative direction in each sentence and (2) Determine the relative facts of each entity. The model was tested on Amazon reviews dataset. They employed a dictionary matching approach for entity recognition. They used the collapsed version of Gibbs sampling.

5. CONCLUSION

In this paper studied various methods for identification of comparative sentences from the text documents. Also tried to look at the methods where the relation where extracted from the comparative sentences. But most of methods require lot of manual work. This review gives a better understanding for the researchers in automation of the work where manual work can be avoided.

6. REFERENCES

[1] Christopher Kennedy, “Comparatives, Semantics of”, Department of Linguistics, Northwestern University, Evanston, IL 60208 USA, July 20, 2000. ∗To appear in the Lexical and Logical Semantics section of the Encyclopedia of Language and Linguistics, Second Edition, Keith Allen (section editor), Elsevier, Oxford.

[2] Syntax&Semantices, “http://rationale.austhink.com/ rationale2.0/ib/ exercises/tok/syntax_semantics.htm”

[3] Nitin Jindal and Bing Liu, “Mining Compartive Sentences and Relations” Proceedings of AAAI-06, the 21st National Conference on Artificial Intelligence, 2006

[4] Bing Liu, “Mining and Summarizing Opinions on Web”, http://www.cs.uic.edu/~liub/ACL06-workshop-SST.pdf

[5] Sequenctial Pattern Mining , http://en.wikipedia.org/wiki/Sequential_Pattern_Mining

[6] Nitin Jindal and Bing Liu, “Identifying Compartive Sentences in text documents” SIGIR, August 6 - 11, ACM 2006.

[7] Xiaojiang Huang, Ciaojun Wan, Jianwu Yang and Jianguo Xiao, “Learning to Identify Comparative Sentences in Chinese Text”, PRICAI 2008, LNAI 5351, pp. 187 – 198, 2008. © Springer – Verlag Berlin Heidelberg 2008.

[8] Note Book Dataset, “http://group.zol.com.cn”

[9] Dae Hoon Park, Catherine Blake, “Identifying Comparative claim sentences in full – text scientific article”, Proceeding of the 50th Annual Meeting of the Association for Computational Linguistics, Pages 1 -9, Jeju, Republic of Korea, 12 July 2012 © 2012 Association for Computational Linguistics.

[10] Seon Yang, Youngjoong Ko, “Extracting Compartive Sentences form Korean Text Documents Using Comparative Lexical Patterns and Machine Learning Techniques”, Proceedings of the ACL – IJCNLP 2009 Conference Short Papers, Pages 153 – 156, Suntec, Singapore, 4 August 2009. © 2009 ACL and AFNLP

[11] Maksim Tkachenko, Hady W. Lauw, “Generative Modeling of Entity Comparison in Text”. CIKM ‘ 14, November 3 – 7, 2014, Shanghai, China. ACM 978 – 1 – 4503 – 2598 – 1/14/11.

[12] Bing Liu, “Sentiment Analysis and Opinion Mining”, Morgan & Claypool Publishers, May 2012.