Determining The Best Entity On Comparison Mining Using Sequential Rule Approaches

Nu Nu War
University of Computer Studies, Mandalay
nuwar99@gmail.com

Abstract

With the increased amount of information rapidly available on the World Wide Web, Internet users that want to know opinions about products are becoming difficult to determine which product (entity) is the best on many product sites. When the product manufacturers are interesting how the product compares with those of competitors, opinion mining on comparative sentences becomes very important. Mining on comparative sentences is called comparison mining. The purpose of this paper is to get the best entity from superlative relations in the comparison mining. This paper focuses on mining comparative (opinion) words and determines the best entity on comparative sentences from the product reviews data set. Determining the best entity depends on just one feature that has same nature or application domain. This paper mentions a rule- based approach that integrates two sequential rule mining techniques that utilizes POS tagging. Determining the best entity on comparative sentences is effective and time saving, not only for individuals but also for organizations such as business intelligence units.

1. Introduction

Most of research has been studied sentiment analysis from the user generated content (e.g. customer reviews, forum posts, and blogs) on the web in many research areas. Opinions can be expressed on anything, e.g. a product, a service, a topic, an individual, an organization or an event. Opinions are very important, whenever someone needs to make decision, not only individuals but also organization.

Consumers want to choose and decide by the opinions that other consumers have commented when purchasing a product, using a service, finding opinions on political topics, and many other decision making tasks. Whenever a new product comes into market, the product manufacturer also wants to know consumer opinions on the product. So it is necessary to determine and mine the best product on the comparison products.

A comparative opinion expresses a preference relation of two or more entity based on some of their shared features. It is conveyed using comparative or superlative form of an adjective or adverbs, e.g., "Cake tastes better than bread". Comparisons are one of the most convincing ways of evaluation. Evaluating an entity is to directly compare it with some other similar entity. It needs to be same domain and shared one feature. It isn't necessary to be considering more than one shared feature in system.

The aim of this research is thus to identify the best entity in each comparative/superlative sentence. An observation about comparative/superlative sentences is that in each such sentence there is usually a comparative

word (e.g., "more", "less", “better”, “worse” and –er word) or a superlative word (e.g., "most", "least", “best”, “worst” and –est word).

The entities being compared is more than two entities and appear on the sides of the comparative word to determine which entity is the best entity by using conjunction words such as "Though, Although, But, However". We identify the best entities by using these words with opposite orientation. Moreover a superlative sentence may also have one entity, e.g., “Camera X is the best”. For simplicity, we use comparative words (sentences) to mean both comparative words (sentences) and superlative words (sentences).

In paper, we study how to determine which entity is the best entity in comparative/superlative sentences commented on by its authors. We approach a proposal technique to identify which sentence is a comparative/superlative sentence and determine the best entities from comparative/superlative relations. Experimental evaluation conducts using web evaluative texts; including consumer reviews data set. There are two contributions.

(1) Studying the comparative sentences on sentiment analysis and,

(2) Determining the best entity in comparative/superlative sentence mining.

This paper is organized as follows. Section 2 contains the related works. Section 3 gives the background theory about sequential rule mining approaches. Section 4 describes the comparison mining method. Section 5 presents the details of the proposed system and conclusion remarks are given in section 6.

2. Related Works

Many researchers have been studied comparative sentiment analysis in various research areas recently.

In [1], researcher has been studied the problem of identifying comparative sentences in text documents, e.g. new articles, consumer reviews of products, forum discussions. Comparative sentence contain some indicators (comparative adverbs and comparative adjectives). Many sentences that contain such words are not comparative, e.g. "I cannot agree with you more". Similarly, many sentences that do not contain such indicators are comparative sentences, e.g."GSM phone X has Bluetooth, but GSM phone Y does not have". If so, we can assume that "GSM phone X is better than GSM phone Y". This is a challenging problem for mining comparative/superlative sentences on sentiment analysis.

In [2], a technique is proposed to identify comparative sentences from reviews and forum posts, and to extract entities, comparative words, and entity features that are being compared. For example, in the sentence, “Camera X has longer battery life than Camera Y”, the technique extracts “Camera X” and “Camera Y” as entities, and “longer” as the comparative word and “battery life” as the attribute of the cameras being compared. However, the technique does not find which entity is preferred by the author. For this example, clearly “Camera Y” is the preferred camera with respect to the “battery life” of the cameras. This was a challenging problem.

The [3] aims to solve [2] problem, which is useful in many applications because the preferred entity is the key piece of information in a comparative opinion. For example, a potential customer clearly wants to buy the product that is better or preferred. However, those paper still no study of identifying the best entities for more than two entities in comparative sentences on sentiment analysis.

In [4], a system discussed a technique that automatically found the people who hold opinions about that topic and the sentiment of

each opinion. The system contained word sentiment classifier for determining word sentiments and combining sentiments within a sentence.

In [5], a bootstrapping technique which uses a small set of given seed opinion words to find their synonyms and antonyms in WordNet is approached. The [5] studied the problem of feature-based opinion summarization of customer reviews of products sold online and determined whether opinion is positive or negative by using opinion summarization system.

In [6], Korean comparison mining system proposed to classify comparative entities and predicates for Korean language. Comparison mining system can automatically provide a summary of comparisons between two (or more) entities from a large quantity of web documents.

In [7], an architecture, implementation, and evaluation of a Web blog mining application, called the BlogMiner are presented. System extracted and classified people's opinions and emotions (or sentiment) from the contents of weblogs about movie reviews. A blog mining system proposed a technique that extracts movie comments from web blogs. Web crawling and sentiment analysis are used in mining process.

In [8], Social Network based evidences is explained so that it can be exploited for the task of Opinion Detection and propose a framework for extracting opinions from blogs. Research is based on machine learning or lexical based approaches. The tasks of opinion prediction predict the sentiment of bloggers for a certain topic.

3. Theoretical Background

3.1. Class Sequential Rule

A class sequential rule (CSR) is a rule with a sequential pattern on the left and a class label on the right of the rule. Unlike classic sequential pattern mining, which is unsupervised, Sequential rules mine with fixed classes. The new method is thus supervised. Class sequential rules (CSRs) use in classification of sentences.

Let S be a set of data sequences. Each sequence is labeled with a class y. Given a labeled sequence data set D, CSR mining finds all class sequential rules in D. Let Y be the set of all classes, $I \cap Y = \{\}$. Thus, the input data D for mining is represented with $D = \{(s_1, y1), (s_2, y_2), ..., (s_n, y_n)\}$, where si is a sequence and $y_i \in Y$ is its class label. In our context, Y = {comparative, non-comparative}.

A class sequential rule (CSR) is an implication of the form, $X \rightarrow y$, where X is a sequence, and $y \in Y$. A data instance (s_i, y_i) in D is said to cover the CSR if X is a subsequence of s_i. A data instance (s_i, y_i) is said to satisfy a CSR if X is a subsequence of s_i and $y_i = y$ [9].

3.2. Label Sequential Rule

A label sequential rule (LSR) is a rule with a sequential pattern to extract the comparative relations. A label sequential rule (LSR) is of the following form, $X \rightarrow Y$, where Y is a sequence and X is a sequence produced from Y by replacing some of its items with wildcards. A wildcard, denoted by a '*', matches any item.

The input data is a set of sequences, called data sequences. These replaced items are usually very important and are called labels. The labels are a small subset of all the items in the data. LSR predict whether a word in a comparative sentence is an entity (e.g., a product name), which is a label. [9].

4. Comparison Mining

Comparisons started with subjectivity classification as a sentence classification

problem. Comparisons are not concerned with an object in isolation. Instead, it compares object with others.

Comparisons can be subjective or objective. For example, Opinion sentence is "Cycle X is very ugly". Subjective comparative sentence is "Cycle X is much better than Cycle Y". Objective comparative sentence is "Cycle X is 2 feet longer than Cycle Y".

Comparative sentences use different language structure from opinion sentences. It is also known as sentence-level sentiment classification. At this level, most of researches applied supervised learning methods such as naïve Bayesian, Support Vector Machine (SVM).

Comparisons are related but also quite not same from sentiments and opinions. They have different semantic meanings and different syntactic forms. In general, a comparative sentence expresses a relation based on similarities or differences of more than one entity. Mining of comparative sentences basically consists of identifying what features and entities are compared and which entities are preferred and best by their authors.

Comparison mining tasks are:

(1) Identifying comparative sentences, and classifying into different types or classes.

(2) Extracting comparative opinions including entities, comparative words and entity features.

(3) Determining preferred entities and opinion orientations from comparative sentences. There is no research which entity is the best entity from superlative relation in comparison mining.

4.1. Units of Opinion

An entity is a product, a service, an individual, an organization, or an event that is being compared.

Feature of an entity can have a set of components (or parts) and a set of attributes (or properties) that is commented on.

In evaluative document D, an opinion holder is a person or an organization that holds the opinion. Opinion holder expresses opinion (positive, negative, and neutral) on one feature or several features of one entity. Opinion mining task is to fetch all these information.

An opinion can be identified on any feature of the entity and also on the entity itself.

Superlative relation:

A superlative relation is the following:

<Comparative/superlative word, features, Entity S_1, Entity S_2, Entity S_3, Type, Conjunction>.

Comparative/superlative (opinion) word is the keyword used to express a comparative/superlative relation in the sentence. If a word is positive (or negative), then its comparative or superlative form is also positive (or negative), e.g., “good”, “better” and “best”.

Entity S_1, Entity S_2 and Entity S_3 are set of entities being compared.

Type is non-equal gradable, equative or superlative.

4.2. Kinds of Comparative Sentences

We classify comparisons into four main types. The first three types are gradable comparisons and the last one is non-gradable comparisons. The gradable types are defined based on the relationships of greater or less than, equal to, and greater or less than all others.

(1) Non-equal gradable: Relations of the type greater or less than that express a total ordering of some entities with regard to their shared features. For example, the sentence, “Camera X’s battery life is longer than that of Camera Y”, orders Camera X” and “Camera Y” based on their shared feature “battery life”.

(2) Equative: Relations of the type equal to that state two objects as equal with respect to some features, e.g., “Camera X and Camera Y are about the same size”.

(3) Superlative: Relations of the type greater or less than all others that rank one object over all others, “Camera X’s battery life is the longest”.

(4) Non-gradable: Sentences which compare features of two or more entities, but do not explicitly grade them, e.g., “Camera X and Camera Y have different features”. There are three main types:

(i) Entity A is similar to or different form entity with regard to some features, e.g., "Coke tastes differently from Pepsi".

(ii) Entity A has feature f_1 and entity B has feature f_2 (f_1 and f_2 are usually substitutable), e.g., "desktop PCs use external speakers but laptops use internal speakers".

(iii) Entity A has feature f, but entity B does not have, e.g., "CDMA phone X has an earphone, but CDMA phone Y does not have".

The first three types are gradable comparisons. Gradable comparisons can be classified further into two types:

(1) Adjectival comparisons and

(2) Adverbial comparisons.

Adjectival comparisons involve comparisons of degrees associated with adjectives, e.g., in "John is taller than Mary", and "John is the tallest in the class". Adverbial comparisons are similar but usually occur after verb phrases, e.g., "John runs faster than James", and "John runs the fastest in the class". This paper focuses on non-equal gradable and superlatives types. It doesn't consider other comparative types [10].

5. Proposed System Overview

This research intends to present determining which entity is the best entity from the comparative and/or superlative sentences on sentiment analysis.

In framework, online reviews data are used as training and evaluation data. Sequences are generated from input sentences in training data that includes reviews on five products such as two digital camera (Canon G3, Nikon coolpix 4300), cellular phone, MP3 player, DVD player. The reviews are downloaded from www.amazon.com. In review data set, we classify 308 comparative sentences and 2857 non-comparative sentences.

In comparative sentences, strong patterns that involve comparative words are used as attributes in learning. To discover these patterns, class sequential rule (CSR) mining is used. Each training examples has a pair (s_i, y_i), where s_i is a sequence and y_i is a class, $y_i \in$ [comparative, non comparative]. To find pattern, sequence database are built with the POS tags as the following example. The standard Penn Treebank POS tagging scheme is used for this research. The POS tags and categories that are important are: NN: Noun, NNP: Proper Noun, VBZ: Verb, present tense, 3rd person singular, JJ: Adjective, RB: Adverb, JJR: adjective, comparative, JJS: adjective, superlative, RBR: Adverb, comparative, RBS: Adverb, superlative.

Example: “this/DT camera/NN has/VBZ significantly/RB more/JJR noise/NN at/IN iso/NN 100/CD than/IN the/DT nikon/NN 4500/CD.” It has the keywords “more” and “than”. The sequence involving “more” put in the training set as follow. (<NN}{VBZ}{RB}{more/JJR}{NN}{IN}{NN}>, comparative). Then, CSRs generates by using that sequences.

Then, CSRs are used as a classifier. A CSR simply expresses the conditional probability that a sentence is comparison if it contains the sequence pattern X. These CSR rules thus use for classification of sentences (comparative or not).

That is, for each test sentence, the algorithm finds all the rules satisfied by the sentence, and then chooses the rule with the highest confidence to classify the sentence.

After that, Label sequential rules (LSR) are used to extract the comparative entries as the following examples:

"Canon/NNP has/VBZ better/JJR optics/NNS than/IN Nikon/NNP" has $entityS1 "Canon", $feature "optics" and $entityS2 "Nikon".

The three sequences corresponding to the two entities and one feature put in the database are:

<{#start}{$entityS1,NNP}{has,VBZ}{better, JJR}{$feature, NNS}{thanIN}>

<{#start}{$entityS1,NNP}{has,VBZ}{better, JJR} {$feature, NNS}{thanIN}{entityS2, NNP}{#end}>

<{has,VBZ}{better,JJR}{$feature,NNS}{thanIN}{$entityS2, NNP}{#end}>.

After the sequence database is built, we generate the label sequential rule as follows.

Rule 1: <{*, NN}{VBZ}{JJR}{thanIN}{*, NN}>=<{$entityS1,NN}{VBZ}{JJR}{thanIN}{$entityS2, NN}>.

The generated LSRs are used to extract relation items from each input sentences or test sentences. One strategy is to use all the rules to match the sentence and to extract the relation items using the rule with the highest confidence. For example, the above rule labels and extracts "coke"as entityS_1, and "pepsi" as entityS_2 from the following sentence:

< {coke, NN} {is, VBZ} {definitely, RB} {better, JJR} {thanIN} {pepsi, NN}>.

There is no feature in this sentence. The relation word is simply the keyword that identifies the sentence as a comparative sentence. In this case, it is "better". Then we determine the best entity at next section. Proposed system design is shown in figure 1.

5.1. Identifying Best Entity: Basic Idea

Determining the best entity in a comparative/superlative sentence targets the features being compared and the comparative words. This research doesn't consider two or more features (e.g. camera resolution and sound quality) in the same nature (e.g. mobile phone).

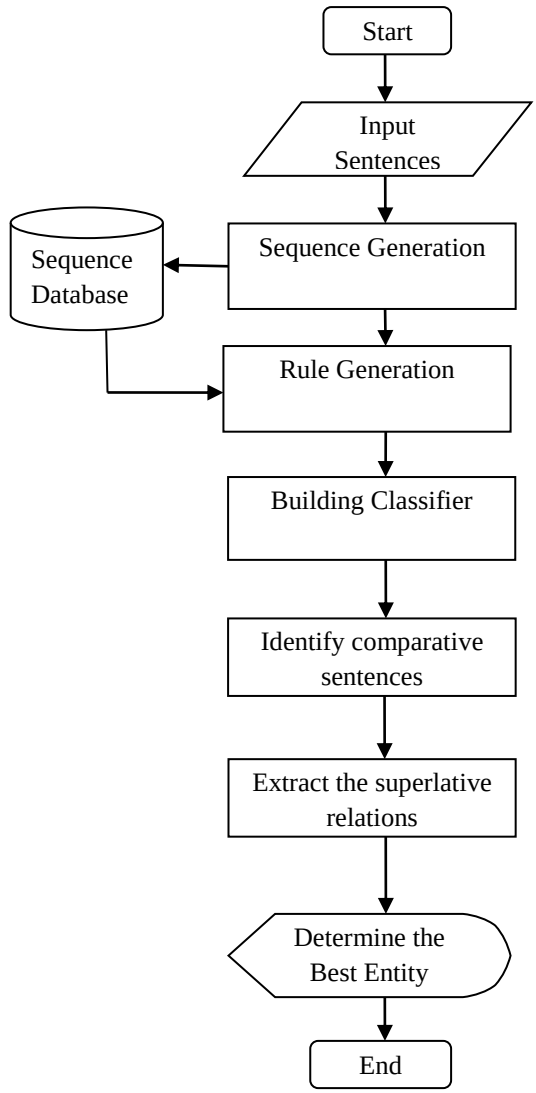

Figure 1. Proposed system design

Therefore determining the best entity depends on just one feature that has the same nature (application domain).

For example: "Picture quality of Camera X is better than that of Camera Y but it is not as good as Camera Z".

In this sentence, there are two main sentences. In the first sentence, camera X prefers rather than Camera Y with shared one feature (e.g. picture quality) due to the comparative word "better". Thus, we focus on comparative (opinion) words and same features. Sometimes, some words such as "better", "worse", "and best" indicate user preference.

The second sentence is joined conjunction word "But": So the first sentence's opinion absolutely changes. The third entity becomes the best entity with negation (not) and opinionated comparative word "good". So camera Z is the best entity. We focus on studying such opinionated comparative words.

Preferred entity is decided by a comparative word before conjunction word "But" in the comparative sentence. Conjunction rules are also used to find opinion words from large domain corpora. Two opinion words are linked by "and", their opinions are same. Two opinion words are linked by "But", their opinions are opposite. We focus on opposite words at here. How do we define and extract the best entity in comparative sentence? Some comparative word (better, worse, best) indicate user preference. It is also known as opinionated comparative. However, some comparative words based on domain are not opinionated (e.g. "longer").

Finding domain opinion words is problematic because the same word in the same domain may indicate different opinions depending on what features it is applied to. (E.g. in the camera domain), "long" is positive in the sentence "the battery life is very long". However "long" is negative in the sentence "it takes a long time to focus". So this research doesn't consider in our system such non-opinionated words.

6. Conclusion

This paper presents the idea, proposed system design and two sequential rule based approaches determining the best entity on comparative sentences. Comparison mining is useful in many applications, e.g. marketing intelligence, product benchmarking, and e-commerce. We approach two methods to identify the best entity and to extract superlative relations (entities, comparative (opinion) word, entity features) from comparative and superlative sentences.

References

[1] Jindal, N. and Liu, B. Identifying comparative sentences in text documents. SIGIR-06, 2006. Seattle, Washington, USA.

[2] Jindal, N. and Liu, B. Mining Comparative Sentences and Relations. Proceedings of NationalConference on Artificial Intelligence (AAAI'06), 2006.

[3] Ganapathibhotla, M. and Liu, B. Mining Opinions in Comparative Sentences. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 241–248, Manchester, August 2008.

[4] Kim, S. and Hovy, E. Determining the Sentiment of Opinions. Proceedings of the 20th International Conference on Computational Linguistics (COLING'04), 2004.

[5] Hu, M and Liu, B. Mining and Summarizing Customer Reviews. Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), 2004.

[6] Seon,Y. and Youngjoong, K. Extracting comparative Entities and Predicates from Texts using comparative Type classification. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1636-1644, 2011.

[7] Baloglu, A and Mehmet S, Aktas, Web Blog Mining Application for Classification of Movie Reviews, Fifth International Conference on Internet and Web Applications and Service, 2010.

[8] Missen, M, Boughanem, M and Cabanac, G, Opinion Detection in Blogs: What is still Missing? International Conference on Advances in Social Networks Analysis and Mining, 2010.

[9] Liu, B. Web Data Mining – *Exploring hyperlinks, contents and usage data*. A forthcoming book. 2006/2007.

[10] Shelia , M and Brian, M. Forming Comparative and Superlative Adjectives in English: Prescriptive Versus Psychological Rules.