An Interrelation-based Approach to Aspect Extraction from Customer Reviews

Aye Aye Mar, Nyein Thwet Thwet Aung, Su Su Htay
Faculty of Information Science, University of Information Technology, Myanmar
{ayeayemar,nyeinthwet,suhtay}@uit.edu.mm

Abstract

Aspect extraction plays a key role in aspect-based sentiment analysis. Without knowing them, the target of sentiments expressed by customers cannot be known. If the target of the sentiments is not known, the sentiments expressed in a review are of limited use. Aspects and sentiment words are related to one another. Aspect extraction affects the performance of sentiment word extraction and also sentiment analysis. This study concentrates on aspects extraction from customer product reviews. This paper proposes an interrelation-based approach which considers the strength of interrelations between aspects and sentiment words to solve the problem of aspect extraction. The proposed approach takes into account the frequency of aspects, the weight of aspects and the weight of sentiment words associated with the aspects. Due to considering the interrelations between the aspects and sentiment words, the proposed method is expected to extract the relevant product aspects effectively and solve the major bottleneck of domain dependency.

Keywords- Aspects Extraction, Aspect-level, Feature Extraction, Sentiment Analysis, Opinion Mining, Customer Review, Text Mining

1. Introduction

Sentiment analysis of customer reviews has become a prominent research area during the last few years. Due to the rapid expansion of e-commerce, the web has become a source of grouping consumer opinions via customer reviews about the products. A collection of consumer opinions can be found in many product review websites such as amazon.com, epinions.com, cnet.com.

Sentiment analysis from customer reviews is vital for both customers and manufacturers to make the right decision. Customers inquire about the product they are interested via the reviews of other customers to decide whether he or she should buy that product or not. Alternatively, manufacturers improve the product quality and marketing campaigns based on the feedbacks of the consumers.

Sentiment analysis has been studied in three different levels of granularity: document level, sentence level and aspect level also called feature level [1]. Although sentiment analysis at document level and sentence level is useful in many applications, they cannot provide the necessary detail about the sentiment of the customers.

Among those three levels, aspect level is the most fine-grained level which extracts not only the opinions/sentiments but also the aspects/features/opinion targets. Unlike the other two levels, aspect-level sentiment analysis can decide what customers like and dislike. In the field of sentiment analysis, aspects are topics on which opinions are described. Other similar names for aspects are features, product features and opinion targets [2] [8].

Aspect-level sentiment analysis involves two tasks: aspect extraction and sentiment extraction. Aspect extraction is the most fundamental and important task of sentiment analysis because the sentiments expressed on those aspects are detected based on finding the aspects. So, this paper concentrates on the problem of aspect extraction from customer product reviews.

There are two types of aspects in aspect-level sentiment analysis: explicit aspects and implicit aspects [2]. Explicit aspects explicitly describe the name of targets in opinioned sentences e.g., "The weight of the phone is heavy". In that sentence, the weight is the aspect/opinion target and it is directly mentioned in the review. In contract, implicit aspects are not directly described in the opinion sentences and the aspects are expressed indirectly by using the implicit aspect indicator (sentiment word) e.g., "The phone is extremely light and disappears in your pocket". In that sentence, the aspect is not directly mentioned but it means weight.

Although both explicit and implicit aspects are important for sentiment analysis, explicit aspects are commonly occurred more than implicit aspects [2]. Explicit aspects extraction has been done by many previous researches but there has been limited number of research on implicit aspect extraction task.

This paper aims to solve the problem of explicit aspect extraction and proposes a method which considers the strength of relations between aspects and sentiment words. The proposed method take into account the frequency of aspects, the weight of aspects and the weight of the sentiment words associated with the aspects.

The rest of the paper is organized as follows: Section 2 summarizes the related work. The proposed system is presented in Section 3. In Section 4, the experimental evaluation matrixes are described. Section 5 concludes the paper and describes the future work of the proposed method.

2. Related Work

There have been many previous researches which have made aspect extraction for sentiment analysis using the supervised method, semi-supervised method and unsupervised methods [9]. The first and foremost attempt to aspect extraction was made by [2]. They introduced how implicit aspects differ from the explicit aspects and deal with explicit extraction problem. They used association rule mining based on the Apriori algorithm to extract frequent noun and noun phrases. They assumed that frequent noun and noun phrases as the explicit product aspects. To remove incorrect frequent aspects, they make feature pruning. Their approach considered only the frequency of noun and noun phrases so less frequent ones cannot be extracted. Our proposed method considers not only frequency of aspect candidates but also their weight and the weight of the respective sentiment words. Therefore, our approach is expected to be able to perform better than their work.

In [3], Somprasertsri proposed a supervised model to extract aspect by combining lexical syntactic features with maximum entrophy technique. In their work, they presented four different features for learning maximum entrophy. Those features include Aspects and their POS tags, Rare words, Alphanumeric feature and Dependency from syntactic parse tree. Those features were extracted from an annotated corpus. Maximum entrophy classifier was used to extract the aspects.

Semantic-based product aspect extraction (SPE) method was presented by Wei et.al [4]. Their work used a list of positive and negative adjective of General Inquirer to recognize sentiment words semantically and subsequently extract aspects. They applied association rule mining to detect candidate product aspects. The used the same pruning strategy of [2]. They discovered the infrequent aspects by using semantic-based refinement. Their approach relies heavily on frequency and semantic-based extraction to detect aspects. In our approach, we study the interrelation information between the aspects and sentiment words by considering their weight.

In [5], Poria et al. presented a rule-based approach to extract both explicit and implicit aspects. Rules were identified to extract aspects from product reviews. They defined implicit aspect clues (IAC) associated with implicit aspects. For identifying the aspects, they defined several dependency rules and applied WordNet and SenticNet to identify the synomyms and semantics of each IAC respectively. Their approach is limited to identify all possible rules among aspects and opinions. In addition, the association rule mining approach leads to the methodology computationally expensive.

The main goal of the proposed approach is to construct a domain-independent aspect extraction system which can extract aspects based on the interrelations between aspects and sentiment words.

3. Proposed System

The proposed system is mainly composed of three components: preprocessing, computing aspect cores for each aspect candidate and selecting the aspects from the aspect candidates. First and foremost, review dataset is preprocessed and Noun and Noun phrases are extracted as aspect candidates. Next, the system computes the aspect score for each aspect candidate and finally selects the aspects by using the threshold.

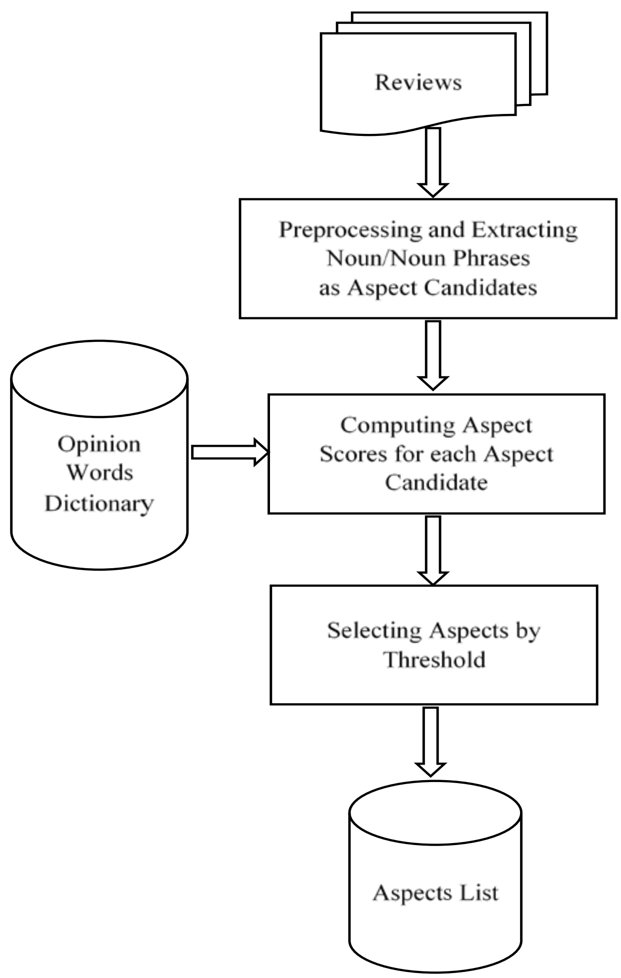

Figure 1. System flow of the proposed system

3.1. Preprocessing

Preprocessing is done before extracting the sentiment features. It includes two parts: POS tagging and stop words removal. POS tagging is done by applying the Standford POS tagger tool. POS tagging means labelling each word in a sentence with its appropriate part of speech such as noun, adjective, adverb etc.

Stop words such as verb to be, pronouns, prepositions and conjunctions do not give meaningful information for sentiment analysis. So, the stop words are removed to save the processing time.

3.2. Extracting the Aspects

Aspects are commonly found as noun or noun Phrases and they are occurred more often than the other noun and noun phrases. Accordingly to that observation, frequent noun and noun phrases are extracted as aspect candidates. To detect the relevant aspects from the aspect candidates, the proposed method computes the aspect scores by considering weight of relation between aspects and sentiment words.

The proposed method is based on the observation that there are interrelation between the aspects and the sentiment words. Interrelation information means the probability of co-occurrence of aspect and sentiment word in a sentence or in a review. The basic intuitions behind the proposed idea are as follows:

(i) An aspect is co-occurred with many sentiment words because different customers might describe their sentiments on the same aspect by using a variety of sentiment words (e.g. good camera, amazing camera, beautiful camera, lovely camera, bad camera, etc.,)

(ii) Alternatively, a sentiment word can be co-occurred with more than one aspect because customers often use some sentiment words to talk about their opinions about many aspects (e.g. good camera, good battery, good sound quality, low price, low processing time, etc.,)

Based on those assumptions, the proposed method computes the scores of aspects by considering frequency of aspects, the weight of aspects and the weight of the sentiment words associated with the aspects. The frequency of aspects is the number of occurrence of aspects in the reviews. The weight of an aspect means the number of distinct sentiment words that is co-occurred with that aspect in the reviews. Similarly, the weight of a sentiment word is the number of frequent noun or noun phrases (other aspect candidates) that is co-occurred with that sentiment words.

The Algorithm 1 describes about aspect extraction process from POS tagged reviews. A number of sentiment word dictionaries are available in recent researches. SentiWordNet has a wide application in the field of sentiment analysis and it has the largest number of features, and its structure is suitable for mathematical modeling [7]. Therefore, SentiWordNet is used to detect the sentiment words from the reviews [6].

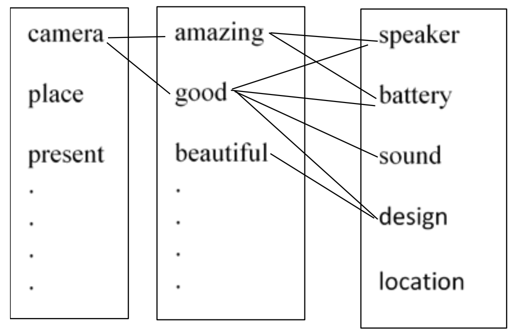

Figure 2. Relations between aspects and sentiment words.

In Figure 2, camera, place and present are aspect candidates. Amazing, good and beautiful are sentiment words which is contained in SentiWordNet dictionary and co-occurred with the aspect candidate camera. The weight of aspect camera is 2 because it is found together with two sentiment words: amazing and good. The system considers the weight of those two sentiment words so it detects other aspect candidates which are not aspect candidate camera. In that picture, the weight of sentiment word amazing is 2, the weight of sentiment word good is 4 and the weight of sentiment word beautiful is 1. To compute the aspect score of aspect camera, the system will consider the frequency of camera, the weight of camera and the weights of its associated sentiment words (amazing, good, beautiful). The more interrelation between aspects and sentiment words, the more possible for that aspect candidate to be aspect.

4. Evaluation Matrix

To evaluate the performance of the proposed method, four evaluation metrics: precision, recall, F-measure and accuracy are considered to evaluate the effectiveness of the system. These are calculated by using Eq. (1) - (4) respectively.

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$F - measure = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

where:
TP refers to the number of true positive reviews.
TN refers to the number of true negative reviews.
FP refers to the number of false positive reviews.
FN refers to the number of false negative reviews.

Algorithm 1. Aspects Extraction Algorithm

Input : POS tagged Review Document ***D, opinion-word-dictionary, threshold***
Output : List of aspects

for each aspect candidate ***a*** (N/N phrase) in ***D*** **do**
  for each ***ow*** in ***opinion-word-dictionary*** **do**
    for each sentence in ***D*** **do**
      if ***ow*** is found together with ***a*** **then**
        $Weight_a = Weight_a + 1$
        if ***ow*** in not in ***ow-list-for-a*** **then**
          add ***ow*** into ***ow-list-for-a***
        end if
      end if
    end for
  end for
  for each opinion word ***ow*** in ***ow-list-for-a*** **do**
    for each sentence in ***D*** **do**
      if N/N phrase (not ***a***) found near ***ow*** **then**
        $Weight_{ow} = Weight_{ow} + 1$
      end if
    end for
  end for
  Compute f_a (the frequency of ***a***)in ***D***
  $a\text{-}score = \log(f_a * Weight_a * \sum_{ow \in ow\text{-}list\text{-}for\text{-}a} Weight_{ow})$
  if (***a-score*** > ***threshold***) **then**
  add ***a*** into $List_{aspect}$
  end if
end for
Return $List_{aspect}$ as the list of aspects

---

5. Conclusion

In this paper, we study the nature of the aspects and sentiment words and propose an aspect extraction method based on the interrelation between the aspects and sentiment words. Since the proposed method is based on relation of the words, the method will extract the relevant aspect and solves the major bottleneck of domain dependency.

In our future work, we will make evaluation of the proposed method firstly with the product review datasets and then with the datasets from another domains. We have plan to make comparative evaluation of the proposed method by using the domain-specific sentiment dictionary instead of SentiWordNet.

6. References

[1] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[2] Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).

[3] Somprasertsri, G., & Lalitrojwong, P. (2008, July). Automatic product feature extraction from online product reviews using maximum entropy with lexical and syntactic features. In Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on (pp. 250-255). IEEE.

[4] Wei, C. P., Chen, Y. M., Yang, C. S., & Yang, C. C. (2010). Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. Information Systems and E-Business Management, 8(2), 149-167.

[5] Poria, S., Cambria, E., Ku, L. W., Gui, C., & Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. In Proceedings of the second workshop on natural language processing for social media (SocialNLP) (pp. 28-37).

[6] Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec (Vol. 10, No. 2010, pp. 2200-2204).

[7] Khan, F. H., Qamar, U., & Bashir, S. (2017). A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet. Knowledge and information Systems, 51(3), 851-872.

[8] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

[9] Rana, T. A., & Cheah, Y. N. (2016). Aspect extraction in sentiment analysis: comparative analysis and survey. Artificial Intelligence Review, 46(4), 459-483.