Grouping of Aspects into Relevant Category Based on WordNet Definitions

Sheikh Muhammad Saqib¹, Shakeel Ahmad², Asif Hassan Syed², Tariq Naeem¹, Fahad Mazaed Alotaibi²

¹Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan, Pakistan.
²Faculty of Computing and Information Technology in Rabigh (FCITR) King Abdul Aziz University (KAU) Jeddah Saudi Arabia

Summary
Machine Translation, Information Retrieval and Knowledge Acquisition are the three main applications of sentiment analysis based on aspect word. But in some cases, there may be two aspect words having different spellings, yet belonging to the same aspect category, e.g. the words ‘costly’ and ‘expensive’ belong to the same category ‘price.’ Grouping such aspects into the same category is critical for sentiment analysis. Some have done this by using synonym, LDA (Latent Dirichlet Analysis) groups and Constrained-LDA. The WordNet dictionary is mostly used for such analysis. Although it contains a lot of synonyms, yet sometimes there are such aspects which are not synonymous, but still refer to the same broad category. For example, in the following two sentences: “this camera will not easily fit in a coat pocket” and “this camera is of large size;” aspects ‘fit’ and ‘size’ both are not synonyms yet belong to the same category. In this study, the proposed framework consists of four steps: DefString, Supporting-Words, SetWords and Intersection for Decision. The fourth stage will decide that the two aspects belong to the same category or not. After experimenting on a dataset comprising 252 pairs of same-meaning aspects and 62 pairs of different-meaning aspects, 91% accuracy and 86% F score were detected.

Key words:
Machine Translation, Information Retrieval, Sentiment Analysis, Aspect Extraction.

1. Introduction

In opinion mining field, when we want to analyze any documents e.g. summary analysis or sentiment analysis of document; besides all other efforts, it is of primary importance to find out whether the selected document belongs to the required aspects or not. Detection of aspect is also a challenging problem, but the grouping of aspect is another challenge.

A semi-supervised learning method was proposed to grouped aspect expressions into some user-specified aspect categories [1][2]. In such an analysis manual labelling is required for each category. Work based on Expectation-Maximization (EM) algorithm also required labeling and synonyms[3][4]. Labelling is an extra effort and synonym-problems are discussed at the end of this section. A method called multilevel latent semantic association was presented. Here, all the words in aspect expressions (each aspect expression can have more than one word) are grouped into a set of concepts/topics determined by LDA. Then, the detected topics can be combined with their surrounding words to obtain a particular category[5]. But most of the online product reviews have single lines, while the proposed work will determine topic/aspect using JJ or NN easily without LDA.Constrained-LDA uses a technique where two aspect expressions ai and aj share one or more words [6]. But there are also reviews which belong to the same aspect but do not share the same words.

Similarity path distance can also be useful for grouping aspect into category as used in concept similarity model [7], but in WordNet dictionary score of word “car” and “auto” is 1.0, while some word-pairs not similar in sense so they have less or none score, yet they belong to same aspect category as shown in Table-1.

Works based on synonyms also follow a very good approach for grouping aspect but in dictionary word1 is the synonym of word2 while word2 might have no synonym as word1 e.g. the word ‘picture’ has a synonym, “movie;” while the word “movie” does not have the word “picture” among its synonyms. In Table-1, G1 means Grouping Through Synonym, G2 means Grouping Through Path Distance and G3 means Grouping Through Proposed Method.

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Words</th>
<th>Path Distance</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>('picture', 'movie')</td>
<td>0.2</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>('car', 'auto')</td>
<td>1.0</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>('expensive', 'costly')</td>
<td>None</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

In Table-1:
- All those words which belong to S.No. 1 can be grouped through G1 but not through G2, while the proposed work (G3) will determine the group of those words.
- All those words which belong to S. No. 2 can be grouped through G1, G2 and also through the proposed work G3.
- All those words which belong to S. No. 3 cannot be grouped through G1 and G2, while the proposed work will
be able to group them in the same category. This is the major contribution of the work.

2. Related Work

A word in a review can be extracted as aspect and another word in another review extracted as aspect, both different words may relate to same aspect i.e. both reviews will be put in same category. This is essential because the majority of research studies to date, have investigated opinion words or examined text mining through the creation of summaries of a given document based on aspect [8], [9]. By grouping aspect, a document or sentence can easily be understood by users as well as by intelligent machines. For example, human intelligence can automatically sense and detect the meaning of a word from examining a sentence and grouped the aspects. However, in the field of artificial intelligence, efforts are continuing to be made to understand a sentence from the correct dimension or aspect given that a different word may have same aspect.

How can a company improve the quality of a product, place, etc. from the huge amount of reviews? Many studies have been carried out with regard to sentiment analysis, which is about determining the sentiment orientation of a review or comment [10][11][12][13]. Sentiment orientation means that a positive opinion will be an exact positive, and a negative opinion will be an exact negative [14]. The view, assessment or feeling of a person towards a product [15], aspect [16], or service is known as a sentiment [17][18][19]. Such a feeling, which is either positive or negative, can be assigned a score.

Most of the work in sentiment analysis is based on binary classification, which means that reviews or blogs are divided into “positive” and “negative” classes [20][21]. The classification of text sentiments can be done in two ways, i.e. through machine learning and score-based approaches [22][23]. Machine learning uses training data [24], while the other method uses several attributes of an entity to determine the scores. In the score-based approach, opinions can be oriented as positive or negative [25][26].

The work of [27] used a combined approach of SentiWordNet and lexical resources to determine the scores for slangs. A lexicon-based approach for extracting sentiment orientations of opinions has been used for scoring [28][29]. Studies by [30][31] used lists of positive and negative words to determine the polarity of a sentence by creating a training matrix and random forest classifier based on supervised learning. A sentiment analysis can be performed using different methods [32][33][34][35], with each method having an improved accuracy with respect to the previous one. Although a lot of work is involved in sentiment orientation [36] with the use of adjectives, frequent nouns and noun phrases, sentiment shifters, handling of ‘but’ clauses, decreased and increased quantity of an opinionated item; high, low, increased and decreased quantity of a positive or negative potential item; desirable or undesirable facts; deviations from the norm or a desired value range; and the production and consumption of resources and waste, etc., these are very important for determining the polarity of a document or sentence [37][38][29]. However, a large amount of online data is generated every day with unprecedented speed and size. Most of the available information on the Internet is in text and unstructured forms, i.e. online reviews, blogs, chats, and news. An aspect-based sentiment analysis, which can be carried out by using only particular aspects [29][39][40][41], requires less effort compared to a sentiment analysis of an object with respect to all aspects. Reviews are rated according to an object, so there should be a direct method to determine whether a review is positive or negative. LSI (Latent Semantic Indexing) is better for such a purpose [42][43]. LSI [44] has been used for the clustering of documents and for concept representations. An extended method based on LSI is able to filter unwanted emails in Chinese and English [45]. To improve accuracy of sentiment analysis, lot of work has also been done on words sense disambiguation[46][47][48]. Machine learning approaches, also called corpus-based approaches, do not make use of any knowledge resources for disambiguation [49]. Most accurate WSD systems to date exploit supervised methods which automatically learn cues useful for disambiguation from manually sense-annotated data [50], [51].

Now to detect an entity from an opinion has been done using different methods. Author [52] has used two sets i.e. the set of seed entities Q and the set of candidate entities D to determine which of the entity in D belongs to C. For entity extraction [53][54] have used a method distribution similarity by comparing the similarity of the surround words of each candidate entity with those of the seed entities; and then ranking the candidate entities based on the similarity values. Topic modeling has also been used for entity extraction. Topic modeling is an unsupervised learning method that assumes each document consists of a mixture of topics and each topic is a probability distribution over words[55]. Latent Dirichlet Analysis LDA and Probabilistic Latent Semantic Analysis (PLSA) have been used for detection of topic from a document/documents [56]. To search a required document from a huge amount of published articles is very time consuming and laborious work, topic modeling offers a computational tool to find relevant topics by capturing meaningful structure among the collections of documents [57][58]. Grouping aspect into categories has also improved aspect based sentiment analysis [59].
3. Grouping of Aspect into Relevant Category Based on WordNet Definitions

In our previous work [58], we have determined how to extract aspect words from a review. Now these extracted words will be grouped with already detected categories. If category found, then review will be grouped in the same category; otherwise the review will be placed in a new group with its aspect name as shown in Figure-1.

Following is the work to determine the category to group the Aspect-1 and Aspect-2 as shown in Figure-2.

3.1. DefString

This is a string which is the combination of all definitions of Aspect from WordNet dictionary. Here DefString1 contains all the definitions of Aspect-1, while DefString2 contains all definitions of Aspect-2.

3.2. Supporting-Words

Tag the DefString, extract only Adjectives (JJ) and Nouns (NN) from Tagged string. In Figure-2, Supporting-Words1 are the collection of ‘JJ’ & ‘NN’ from DefString-1 and Supporting-Words2 are the collection of ‘JJ’ & ‘NN’ from DefString-2.

3.3. SetWords

Now convert Supporting-Words as a set, then we can take all operations of sets on it. Here SetWords1 is a set of Supporting-Words1 and SetWords2 is a set of Supporting-Words2.

3.4. Intersection for Decision

Now take intersection of these two sets SetWords1 and SetWords2. If resultant set is not empty, it means Aspect-1 and Aspect-2 both belong to the same category otherwise both belong to a different category.

Here we are taking two pairs, first pair (“expensive”, “costly”) belongs to the same aspect, while the second pair (“beautiful”, “large”) belongs to different aspect. In Table-2, total elements from intersection is 2 (non-empty) it means first pair belongs to same aspect. In Table-3, total elements from intersection is 0 (empty) it means second pair belongs to a different aspect.
Table 2: Example based on First Pair (Same Aspect)

<table>
<thead>
<tr>
<th>Processes</th>
<th>Aspect-1 (expensive)</th>
<th>Aspect-2 (costly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DefString</td>
<td>high in price or charging high prices</td>
<td>entailing great loss or sacrifice having a high price</td>
</tr>
<tr>
<td>Supporting-Words</td>
<td>[(u'high', 'JJ'), (u'price', 'NN'), (u'high', 'JJ')]</td>
<td>[(u'great', 'JJ'), (u'loss', 'NN'), (u'sacrifice', 'NN'), (u'high', 'JJ'), (u'price', 'NN')]</td>
</tr>
<tr>
<td>SetWords</td>
<td>Set {'high', 'price', 'high'}</td>
<td>Set {'great', 'loss', 'sacrifice', 'high', 'price'}</td>
</tr>
<tr>
<td>Intersection</td>
<td>Set {'high', 'price'}</td>
<td></td>
</tr>
<tr>
<td>Decision</td>
<td>Total Element = 2, it means “expensive” and “costly” belongs to same category</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Example based on Second Pair (Different Aspect)

<table>
<thead>
<tr>
<th>Processes</th>
<th>Aspect-1 (beautiful)</th>
<th>Aspect-2 (large)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DefString</td>
<td>delighting senses exciting intellectual emotional admiration (of weather) highly enjoyable</td>
<td>garment size large person above average size number quantity magnitude extent fairly large important effect;</td>
</tr>
<tr>
<td>Supporting-Words</td>
<td>[(u'intellectual', 'JJ'), (u'emotional', 'JJ'), (u'admiration', 'NN'), (u'enjoyable', 'JJ')]</td>
<td>[(u'garment', 'NN'), (u'size', 'NN'), (u'large', 'JJ'), (u'person', 'NN'), (u'average', 'JJ'), (u'number', 'NN'), (u'quantity', 'NN'), (u'extreme', 'JJ'), (u'important', 'JJ'), (u'effect;')]</td>
</tr>
<tr>
<td>SetWords</td>
<td>Set {'intellectual', 'emotional', 'admiration', 'enjoyable'}</td>
<td>Set {'garment', 'size', 'large', 'person', 'average', 'size', 'number', 'quantity', 'extent', 'large', 'important', 'effect;'}</td>
</tr>
<tr>
<td>Intersection</td>
<td>Set {}</td>
<td></td>
</tr>
<tr>
<td>Decision</td>
<td>Total Element = 0, it means “beautiful” and “large” belongs to different category</td>
<td></td>
</tr>
</tbody>
</table>

4. Results and Discussions

A complex example from [14] has been solved easily from proposed method i.e. “This camera will not easily fit in a coat pocket” and “this camera is of large size” here fit and size both are not synonyms but belongs to same aspect category as depicted in S. No. 1 of Table-4. To check the accuracy of proposed model, we have collected 252 synonyms (same meaning) pair of words and 62 words which have not shared same meaning[60][61]. Some of the samples are shown in Table-4. In this table, G1 means Grouping Through Synonym, G2 means Grouping Through Path Distance and G3 means Grouping Through Proposed method. All pairs are detected correctly by the proposed method G3, i.e. (beautiful, 'large') and ('chair', 'car') both belong to different aspect so proposed method decided correctly.

Table 4: Sample of Results from 314 collected pairs

<table>
<thead>
<tr>
<th>S.No</th>
<th>Words</th>
<th>Path Distance</th>
<th>Intersection</th>
<th>Total Words</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>('fit', 'size')</td>
<td>0.1</td>
<td>set({'something', 'u'size'})</td>
<td>2</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>('expensive', 'costly')</td>
<td>None</td>
<td>set({'u'high', 'u'price'})</td>
<td>2</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>('car', 'auto')</td>
<td>1.0</td>
<td>set({'u'engine', 'u'combustion', 'u'wheels', 'u'internal', 'u'motor', 'u'vehicle'})</td>
<td>6</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>('chair', 'car')</td>
<td>0.1</td>
<td>set({})</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>('smart', 'clear')</td>
<td>None</td>
<td>set({'u'quick'})</td>
<td>1</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>('interesting', 'fascinating')</td>
<td>0.1428571</td>
<td>set({'attention', 'u'interest'})</td>
<td>2</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>('administer', 'manage')</td>
<td>0.1666666</td>
<td>set({'u'charge', 'u'direct'})</td>
<td>2</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>('native', 'local')</td>
<td>0.1</td>
<td>set({'u'characteristic', 'u'particular'})</td>
<td>2</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>('object', 'thing')</td>
<td>0.125</td>
<td>set({'u'entity'})</td>
<td>1</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>10</td>
<td>('picture', 'movie')</td>
<td>0.2</td>
<td>set({'u'sound', 'u'story', 'u'form', 'u'entertainment', 'u'continuous', 'u'sequencing', 'u'illusion', 'u'movement'})</td>
<td>8</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>('beautiful', 'large')</td>
<td>None</td>
<td>set({})</td>
<td>0</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
4.1 Statistical Measures

A confusion matrix [62] is formed from the four outcomes produced as a result of binary classification. A binary classifier predicts all data instances of a test dataset as either positive or negative. Here we consider classification as Yes and No i.e. ‘Yes’ means two words belong to same aspect and No means two words belong to different aspects. This classification (or prediction) produces four outcomes: True-Yes (TY): correct same aspect prediction, False-Yes (FY): incorrect different aspect prediction, True-No (TN): correct different aspect prediction, False-No (FN): incorrect same aspect prediction. Various measures can be derived from a confusion matrix as shown in Table-6.

After analysis of proposed work on said words, obtained results are shown in Table-5. In Table-5, G1 means Grouping Through Synonym, G2 means Grouping Through Path Distance and G3 means Grouping Through Proposed Method.

Applying above values in online confusion matrix, and calculated results shown in Table-6.

Table 6: Statistical Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Derivations</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity or True Yes Rate (TYR)</td>
<td>$\text{TYR} = \frac{\text{TY}}{\text{TY} + \text{FN}}$</td>
<td>0.8095</td>
</tr>
<tr>
<td>Precision or Yes Predictive Value (YPV)</td>
<td>$\text{YPV} = \frac{\text{TY} + \text{FN}}{\text{TY} + \text{FY}}$</td>
<td>0.9189</td>
</tr>
<tr>
<td>Accuracy</td>
<td>$\text{ACC} = \frac{\text{TY} + \text{TN}}{\text{Y} + \text{N}}$</td>
<td>0.7891</td>
</tr>
<tr>
<td>F1 Score</td>
<td>$\text{F1} = \frac{2 \times \text{TP}}{\text{TP} + \text{FP} + \text{FN}}$</td>
<td>0.8608</td>
</tr>
<tr>
<td>False Yes Rate</td>
<td>$\text{FYR} = \frac{\text{FP} + \text{FN}}{\text{FP}}$</td>
<td>0.2951</td>
</tr>
<tr>
<td>False No Rate</td>
<td>$\text{FNR} = \frac{\text{FP}}{\text{FP} + \text{TP}}$</td>
<td>0.1905</td>
</tr>
</tbody>
</table>

Table 6 is depicting the results of grouping of detected aspects of product reviews and obtained improved precision (91%), accuracy (78%) and F1-score (86%) when compared to the previous methods from Table-5.

5. Conclusion

Searching of documents is not only a part of search engine, but also very important for sentiment analysis, text summarization, etc. Aspect based sentiment analysis also requires those documents related to same aspects. How these documents are identified, there is need of some words which can determine those documents which are related to these aspects. Instead of comparing these aspects with all the documents, just compare these aspects with the aspects of an individual document. If there is a list of aspect words, then there is need to group the aspects, if same aspects are defined in different words. In this study, grouping aspects into same category is investigated. Based on the experimental results through confusion matrix, it was found that results generated from the proposed method show a significant improvement from existing texts related not only to precision and accuracy, but also to FYR (False Yes Rate) and FNR (False No Rate) which are 29% and 19% respectively. Grouping aspect is not only fruitful for searching, information retrieval, clustering; but also helpful for sentiment analysis, natural language processing and word sense disambiguation.

Conflict of Interest: Sheikh Muhammad Saqib, Shakeel Ahmad, Asif Hassan Syed, Tariq Naeeem and Fahad Mazaed Alotaibi declare that they have no conflict of interest.

References


