Identifying Product Features from Customer Reviews Using Hybrid Patterns

Khairullah Khan[1], Baharum Baharudin[1], and Aurangzeb Khan[2]
[1]Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Malaysia
[2]Institute of Engineering and Computing Sciences, University of Science & Technology Bannu, Pakistan

Abstract: *In this paper we have addressed the problem of automatic identification of product features from customer reviews. Costumers, retailors, and manufacturers are popularly using customer reviews on websites for product reputation and sales forecasting. Opinion mining application have been potentially employed to summarize the huge collectionof customer reviews for decision making. In this paper we have proposed hybrid dependency patterns to extract product features from unstructured reviews. The proposed dependency patterns exploit lexical relations and opinion context to identify features. Based on empirical analysis, we found that the proposed hybrid patterns provide comparatively more accurate results. The average precision and recall are significantly improved with hybrid patterns.*

Keywords: *Opinion mining, features extraction, syntactic relation, context dependency.*

Received February 3, 2012; accepted January 22, 2013; published online April 4, 2013

1. Introduction

Opinion Mining (OM) applications have been popularly employed to identify and summarize opinion presented in the reviews documents. The important problems of OM that have been attracted the researcher are mostly dependent on the following questions:

- What is the opinion of the people about certain product?
- Who presented the opinion?
- Opinion about what?

Based on the above question some authors have developed a very good OM model which has different components. The identification of each component from unstructured reviews has been addressed by different authors [2, 6, 12, 16]. Most of the works have addressed the following components of opinion.

- *Opinion Holder:* The source/reviewer who has given the opinion.
- *Target Object/Features:* The entity or attribute of the entity about which opinion is expressed.
- *Opinion:* Expression of opinion holder about the features of the products.

Reviews can be collected in two different formats i. e., structured and unstructured. The structured opinion is questionnaires about the products or items with specified options against each feature. Thus structured opinion is bounded and static. The analysis of structured opinion is simple, easy to process and analyse. But unstructured opinion is collected in the form of free text. The free text processing involves various steps to get summarize results. How to identify different components of opinion from free text is challenging area of research. In this paper our focus is on product features extraction from unstructured reviews. Every opinionated sentence in document consists of target about which opinion is expressed. Generally opinion can be expressed by a person about an entity. The entity may be a person, organization, service, product, event and etc. The entity can have a set of components, parts and attributes. Bing Liu in his book chapter defines OM as: "Given a set of evaluative text documents D that contain opinions (or sentiments) about an object, opinion mining aims to extract attributes and components of the object that have been commented on in each document $d \in D$ and to determine whether the comments are positive, negative or neutral" [13].

The main focus of this paper is on features identification problem. Every individual features of the target object is important to get complete analysis of the reviews. For example we have an entity which has different features out of which some features may be liked while other may be disliked by the reviewers. Since it is impossible to read every individual sentence and comments, therefore automatic systems are being developed automatically identify and summarize the opinions. In order to analyse and summarize the reviews it is required to automatically identify and extract those features which are discussed in the reviews. Hence features mining of products are important for opinion mining and summarization. The task of features mining provides a base for opinion summarization [7].

Several approaches have been reported for features identification that employs dependency patterns for features identification from unstructured reviews. Dependency patterns exploit grammatical structure and contextual rules to identify relevant product

features in text documents. For detail explanation about how the dependency patterns are employed to extract features, the reader may proceed to the related work given in section 2. As mentioned earlier we have proposed hybrid dependency patterns for features identification. The hybrid pattern is a combination of dependency patterns. Some of the combined patterns are derived from the existing work while other patterns have been identified through observation and empirical analysis. More explanation about hybrid patterns is given in section 3. The results and comparative analysis of the proposed hybrid patterns based algorithm are presented in section 4, while section 5 concludes the paper.

2. Related Work

The related work reveals that the dependency patterns have been potentially employed by different approaches to identify product features from unstructured reviews. Following are the most polar approaches that have been formulated for identification of opinion target features. These approaches depend on the sequence of term which are based on grammar rules or dependency of words. For patterns extraction some work depend on the position of the words and some depends on the phrase patterns while some work depend on semantic relatedness.

Popescu and Etzioni [15] used an unsupervised technique to extract product features and opinions from unstructured reviews. This paper introduces the OPINE system based on the unsupervised information extraction approach to mine product features from reviews. OPINE uses syntactic patterns for semantic orientation of words for identification of opinion phrases and their polarity.

Ghobadi and Rahgozar [9] have proposed an ontology-based approach to extract the products' information. The ontology is based on product features. This approach exploits semantics of HTML documents and extracting the information automatically with dependency patterns.

Carenini *et al.* [5] developed a model based on user defined knowledge to create a taxonomy of product features. This paper introduces an improved unsupervised method for feature extraction that uses the taxonomy of the product features. The results of the combined approach are higher than the existing unsupervised technique; however, the pre-knowledge base mechanism makes the approach domain dependent.

Holzinge *et al.* [11] used domain ontologies based on tabular data from web content to bootstrap a knowledge acquisition process for extraction of product features. This method creates a wrapper for data extraction from Web tables and ontology building. The model uses logical rules and data integration to reason about product specific properties and the higher-order knowledge of product features.

Bloom *et al.* [4] described an unsupervised technique for features and appraisal extraction. The authors believe that appraisal expression is a fundamental task in sentiment analysis. The appraisal expression is a textual unit expressing an evaluative attitude towards some target. Their paper proposed evaluative expressions to extract opinion targets. The system effectively exploited the adjectival appraisal expressions for target identification.

Ben-David *et al.* [1] proposed a Structural Correspondence Learning (SCL) algorithm for domain classification. The idea depends on perception to get a prediction of new domain features based on training domain features; in other words, the author describes under what conditions a classifier trained on the source domain can be adapted for use in the target domain? This model is inspired by feature based domain classification. Blitzer *et al.* [3] extended the structural SCL algorithm for opinion target identification.

Lu and Zhai [14] proposed automatic integration of opinions expressed in a well-written expert review with opinions scattered in various sources such as blogs and forums. The paper proposes a semi-supervised topic model to solve the problem in a principled way. The author performed experiments on integrating opinions about two quite different topics, i.e., a product and political reviews. The focus of this paper is to develop a generalized model that should be effective on multiple domains for extraction of opinion targets.

Ferreira *et al.* [8] and Yi *et al.* [19] approached the problem of product features extraction through likelihood ratio test with lexical pattern. This approach produces best results. This approach provides relatively high precision.

Another popular approach for product features extraction is association mining approach that is based on dependency patterns. The association rule mining implemented by Hu and Liu [12]. This approach is further enhanced by Wei *et al.* [17] using semantic based patterns for frequent feature refinement and identification of infrequent features. This approach provides relatively high recall.

Goujon [10] presented a text mining approach based on linguistic knowledge to automatically detect opinion targets in relation to topic elements. This paper focuses on identification of opinion targets related to the specific topic. This approach exploits linguistic patterns for target identification.

3. The Proposed Methodology

OM applications are employing the noun phrases nearest to subjective adjectives in sentence to identify product features. The sequences of noun and adjectives called base noun phrase have been employed by various research work. For example,

NN, NN NN, JJ NN, NN NN NN, JJ NN NN, JJ JJ NN, where NN and JJ are nouns and adjectives; have been potentially exploited for candidate selection of product features [15, 17, 20]. Some authors have done more research on the base noun phrase patterns to improve the identification process as explained in the related work.

Since the opinionated sentences have evaluative expressions which contain features and opinion terms. It is worthwhile to mention that if either of the two clues is known then the other can be easily identified with dependency relation. Furthermore, the exiting research proved that the subjective adjectives are strong clue for opinion terms, therefore the nouns linked to subjective adjectives are more proven to product features. Based on these observations we have performed various experiments to identify potential dependency patterns for product features identification. In this paper we have proposed hybrid patterns which are based on dependency relation between opinion terms presented by subjective adjectives and product features presented by noun. The hybrid pattern is a combination of four different patterns. Hence we termed the hybrid pattern as Combined Pattern Based Noun Phrases (cBNP). The definite base noun phrase (dBNP) is derived from the existing work while linking verb base noun phrase (lBNP) and Preposition based Base Noun Pharse (iBNP) are our proposed patterns that have been identified during this course of study. Each sub pattern is explained below:

- *Definite Base Noun Phrase:* This patterns Presents Noun Phrases (BNP) with the definite article "the" before the BNP. This pattern have been employed by different authors for product features extraction [8, 19].
- *Linking Verb Based Noun Phrases:* The lBNP pattern is based on the assumptions that linking verbs between subjective adjectives and noun phrase provide best clues for opinion expressions. To get this type of pattern we employed the following regular expression:

  (1) Noun Phrase-Verb Phrase-Adjective (NPVBJJ):
  Pattern → NPVBJJ
  NP → JJ NN+*
  JJ → Adjective
  NN → Noun
  VB → Verb

  (2) Noun Phrase-Verb Phase-Adverb Adjective (NPVBRBJJ):
  Pattern → NPVBRBJJ
  NP → JJ NN+*
  RB → Adverb
  JJ → Adjective
  NN → Noun
  VB → Verb

  (3) Noun Phrase-Verb Phase-Adverb Adjective NN (NPVBRBJJNN):
  Pattern → NPVBRBJJ NN
  NP → JJ NN+*
  RB → Adverb
  JJ → Adjective
  NN → Noun
  VB → Verb

- *Preposition Based Noun Phrases:* With extensive experiments we found that if the preposition ("of/IN") comes between two BNPs then it represents entity-to-entity or entity-to-feature relation. Hence we considered this pattern for features identification. Table 1 explains the above patterns by example.

Table 1. Examples of cbnp patterns.

Pattern	Pattern Type	Example
NNVBRBJJ	vBNP	Camera/NN is/VBZ so/RB compact/JJ
NNVBRBJJ	vBNP	Camera/NN is/VBZ so/RB light/JJ
NNVBRBJJ	vBNP	Camera/NN produces/VBZ fantastically/RB good/JJ Pictures/NN
DTNNVB	dBNP	The/DT viewfinder/NN reflects/VB
DTJJNNVB	dBNP	The/DT LCD/NN sees/VB
NNINNN	iBNP	Quality/NN of/IN Photo/NN
JJIINNN	iBNP	Range/NN of/IN Lenses/NN

4. Heuristic for Features Extraction

In order to identify features through our proposed patterns we use the following steps:

- *Pre-Processing:* In this step the input documents are converted into POS tagged document. As the proposed patterns depend on sequence of terms categories therefore it is necessary to label each term with the proper part of speech. In our experiments we converted all the datasets into POS corpuses using existing part of speech tagging software.
- *Extraction of Evaluative Expressions and Product Features:* This module extracts evaluative expression using the proposed cBNP patterns as explained in the Figure 1. The heuristic first extracts the patterns using the regular expressions as given in section 3. We call this step as syntactic labelling. In the subsequent step, the algorithm determines the opinion hood of the patterns using opinion lexicon. The opinion lexicon is a list of subjective adjectives that have positive or negative polarities. The last step of this algorithm extract product features which are basically the noun phrases in the evaluative expressions obtained in the previous step.

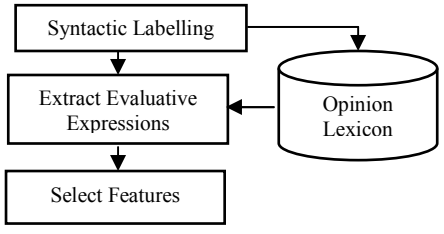

Figure 1. Product features extraction.

5. Results and Discussion

5.1. Datasets

We tested our proposed approach using bench mark data sets about five different products that are collected by Hu and Liu [12] from Amazon product review site. The data sets are manually annotated by the authors. These datasets have been widely reported in number of research papers for comparative analysis of product features extraction and opinion summarization. The same data set has been re-annotated by [8] due their focus study on feature extraction. The difference between these two annotations is that the Hu and Liu consider only those features about which opinion is expressed while the later one considers all the features related to the product and apply relevance scoring for targets identification. The summary of the dataset is given in Table 2. For cross validation we have compared our results with both manually annotated features.

Furthermore, we use list of positive and negative subject adjectives collected by [12] that are freely available from the author's website[1].

Table 2. Summary of datasets.

Data Sets	No. of Sentences	Manually Tag Features by [12]		Manually Tagged Features by [8]	
		Distinct	Total	Distinct	Total
APEX	739	110	347	166	519
Canon	597	100	257	161	594
Creative	1716	180	736	231	1031
Nikon	346	74	185	120	340
Nokia	546	109	310	140	470

5.2. Tools

As mentioned in section 3, our proposed algorithm has two main steps; pre-processing and features extraction. In the pre-processing we perform part of speech tagging.

In this paper we have used a state-of-the-art tool i. e. Stanford parser[2] for POS tagging. In the second step our algorithm extracts features using hybrid dependency patterns and opinion lexicon. For patterns extraction we have used a lexical tools i. e. TextSat2.0 while for the final implementation of our algorithm we developed a module in VB.Net. This module takes the input patterns extracted through TextSat2.0 and checks the subjectivity of the adjective in each input patterns to determine the opinion hood of the expressions. The module then produces a list of features from the opinionated expressions. Finally, it checks the extracted features with the list of manually annotated features in the corpus and calculates the evaluation matrices as explained in the following section.

[1]http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.
[2]http://nlp.stanford.edu/software/lex-parser.shtml.

5.3. Evaluation Criteria

To evaluate the effectiveness of our proposed features extraction algorithm we use standard evaluation measures i. e., precision, recall and f-score. To calculate these matrices, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) features were employed using the following setup.

- *TP Features:* All extracted features correctly matched with the manually annotated features called positive features.
- *TN Features:* All non-positive features not extracted by the algorithm.
- *FP Features:* All positive features included in non-positive features by the algorithm.
- *FN Features:* All positive features included in non-positive features by the algorithm.

5.4. Experimental Setup and Performance Test

As mentioned in section 5.1 we have two different annotation schemes for product features of the same datasets. For in-depth analysis and cross validation we have used both of the annotated features. In comparison we use manually annotated features and predicted features.

Table 3 shows results of features extraction using likelihood ratio test with annotation by [12]. Table 4 shows evaluation of proposed hybrid approach comparing results with manually annotated features while Table 5 shows comparison of proposed hybrid approach and likelihood ratio test approach.

Table 3. Results of pattern based approach using likelihood ratio test.

Doc	BNP			dBNP			bBNP		
	P	R	F	P	R	F	P	R	F
Apex	45.97	83.63	59.33	85.71	30.99	35.53	85.31	15.21	25.81
Cannon	44.08	87.63	58.65	85.95	39.25	53.89	86.28	29.03	43.45
Creative	38.73	88.84	53.94	88.09	42.49	57.33	90.35	36.91	52.41
Nikon	46.70	90.76	61.66	85.70	36.13	50.83	85.83	31.93	46.55
Nokia	45.19	90.07	60.19	87.53	43.26	57.90	87.44	30.50	45.22
Average	44.13	88.19	58.75	86.60	38.43	51.10	87.04	28.72	42.69

Table 4. Results of hybrid patterns based approach comparing with manually annotated features.

Doc	cBNP Compared with [12] Annotation			cBNP with [8]		
	P	R	F	P	R	F
Apex	81.03	72.97	76.79	77.20	51.83	62.02
Cannon	76.28	70.54	73.29	76.68	68.07	72.12
Creative	76.98	66.48	71.34	78.14	69.83	73.75
Nikon	79.33	74.32	76.75	77.70	63.03	69.60
Nokia	81.30	74.55	77.78	80.08	64.29	71.32
Average	78.98	71.77	75.19	77.96	63.41	69.76

From the results it is clear that hybrid patterns provide consistent results on both annotation schemes. The average precision of proposed hybrid patterns based on comparison with manual features of [8, 12] are 78.98 and 77.96, respectively; which are identical.

Similarly the recall and f-measure are also identical on both schemes. The consistent results prove the validly of our proposed approach.

We have compared our results with LRT technique. This approach was initially employed by [19] for product features extraction and then extended by [15]. This approach employed dependency patterns with subsequent similarity. The LRT approach employed BNP, dBNP, and bBNP. However, the F-Measure of dBNP outperform over the other two patterns. Therefore for comparison we have selected dBNP as shown in Table 5. We implemented the LRT technique using our prototype. Figure 2 presents average f-measure of our proposed hybrid compared [8] which shows a significant improvement over LRT approach.

Table 5. Comparison of hybrid pattern based approach and likelihood approach.

Doc	Likelihood Ratio Test			Hybrid Patterns		
	P	R	F	P	R	F
Apex	85.71	30.99	35.53	81.03	72.97	76.79
Cannon	85.95	39.25	53.89	76.28	70.54	73.29
Creative	88.09	42.49	57.33	76.98	66.48	71.34
Nikon	85.70	36.13	50.83	79.33	74.32	76.75
Nokia	87.53	43.26	57.90	81.30	74.55	77.78
Average	86.60	38.43	51.10	78.98	71.77	75.19

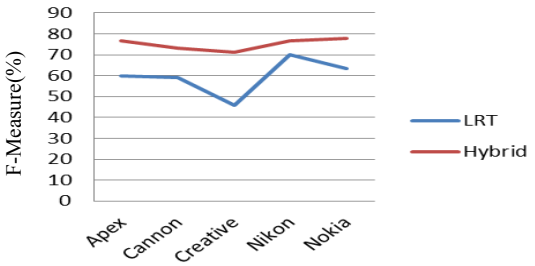

Figure 2. Comparison of hybrid and LRT method.

6. Conclusions and Future Work

In this paper we have presented pattern based product features extraction from customer reviews. Different patterns are used by several authors to identify product features for opinion mining. Some authors have proposed syntax based while some have proposed semantic based extraction. In this paper we have proposed hybrid patterns based approach which partially depends on semantic relation and partially on syntactic sequence. For semantic relation we use adjectives having polarity while for syntactic patterns we have used two existing patterns and one new pattern which are based on linking verb. Based on comparative results with existing approaches it was found that our proposed hybrid patterns outperform the existing patterns based approach.

We believe that anaphora resolution will further improve our results therefore in our future work we will extend our experiments by using anaphora resolution.

References

[1] Ben-David S., Blitzer J., Crammer K., and Pereira F., "Analysis of Representations for Domain Adaptation," *in Proceedings of Advances in Neural Information Processing Systems 19*, USA, vol. 137, pp. 1-8, 2007.

[2] Bethard S., Yu H., Thornton A., Hatzivassiloglou V., and Jurafsky D., "Extracting Opinion Propositions and Opinion Holders Using Syntactic and Lexical Cues," *in Proceedings of Computing Attitude and Affect in Text: Theory and Applications*, Netherlands, vol. 20, pp. 125-141, 2006.

[3] Blitzer J., Dredze M., and Pereira F., "Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification," *Annual Meeting-Association for Computational Linguistics*, vol. 45, no. 1, pp. 440, 2007.

[4] Bloom K., Garg N., and Argamon S., "Extracting Appraisal Expressions," *in Proceedings of Human Language Technologies/North American Association of Computational Linguists*, New York, USA, pp. 308-315, 2007.

[5] Carenini G., Ng R., and Zwart E., "Extracting Knowledge from Evaluative Text," *in Proceedings of the 3rd International Conference on Knowledge Capture*, USA, pp. 11-18, 2005.

[6] Choi Y., Cardie C., Riloff E., and Patwardhan S., "Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns," *in Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, Canada, pp. 355-362, 2005.

[7] Feldman R., Fresko M., Goldenberg J., Netzer O., and Ungar L., "Extracting Product Comparisons from Discussion Boards," *in Proceedings of the 7th IEEE International Conference on in Data Mining*, Omaha, USA, pp. 469-474, 2007.

[8] Ferreira L., Jakob N., and Gurevych I., "A Comparative Study of Feature Extraction Algorithms in Customer Reviews," *in Proceedings of IEEE International Conference on Semantic Computing*, Santa Clara, California, USA, pp. 144-151, 2008.

[9] Ghobadi A. and Rahgozar M., "An Ontologybased Semantic Extraction Approach for B2C eCommerce," *the International Arab Journal of Information Technology*, vol. 8, no. 2, pp. 163-170, 2011.

[10] Goujon B., "Text Mining for Opinion Target Detection," *in Proceedings of Intelligence and Security Informatics Conference*, Athens, Greece, pp. 322-326, 2011.

[11] Holzinger W., Krüpl B., and Herzog M., "Using Ontologies for Extracting Product Features from Web Pages," *in Proceedings of the 5th International Semantic Web Conference*, USA, vol. 4273, pp. 286-299, 2006.

[12] Hu M. and Liu B., "Mining and Summarizing Customer Reviews," *in Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining*, USA, pp. 168-177, 2004.

[13] Liu B., *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data*, Springer, New York, USA, 2011.

[14] Lu Y. and Zhai C., "Opinion Integration through Semi-Supervised Topic Modeling," *in Proceedings of the 17th International Conference on World Wide Web*, USA, pp. 121-130, 2008.

[15] Popescu A. and Etzioni O., "Extracting Product Features and Opinions from Reviews," *in Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, USA, pp. 339-346, 2005.

[16] Shanahan J., Qu Y., and Wiebe J., *Computing Attitude and Affect in Text: Theory and Applications*, Springer, Netherlands, 2006.

[17] Wei C-P., Chen Y-M., Yang C-S., and Yang C-Y., "Understanding What Concerns Consumers: A Semantic Approach to Product Feature Extraction from Consumer Reviews," *Information Systems and E-Business Management*, vol. 8, no. 2, pp. 149-167, 2010.

[18] Wiegand M. and Klakow D., "Convolution Kernels for Opinion Holder Extraction," *in Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, California, USA, pp. 795-803, 2010.

[19] Yi J., Nasukawa T., Bunescu R., and Niblack W., "Sentiment Analyzer: Extracting Sentiments about a Given Topic Using Natural Language Processing Techniques," *in Proceedings of the 3rd IEEE International Conference on Data Mining*, Washington, USA, pp. 427-434, 2003.

[20] Zhang L. and Liu B., "Identifying Noun Product Features that Imply Opinions," *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers*, USA, vol. 2, pp. 575-580, 2011.

Khairullah Khan received his PhD degree in information technology from Universiti Technologi PETRONAS, Malaysia. He is an assistant professor at the University of Science and Technology Bannu Pakistan. His current research interests include data mining, opinion mining and information retrieval.

Baharum Baharudin received his Master degree from Central Michigan University, USA, and his PhD degree from the University of Bradford, UK. He is currently an associate professor at the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS Malaysia. His research interests lies in image processing, data mining and knowledge management.

Aurangzeb Khan received his BS-degree in computer Science from Gomal University D. I. Khan, Pakistan, and his Master degree in information technology from the University of Peshawar, Pakistan, and his PhD degree from the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia. He is an assistant professor at the University of Science & Technology Bannu Pakistan. His current research interests include data mining, opinion mining and information retrieval.