

Partially Supervised Word Alignment Method for Co-Extracting Opinion Target and Opinion Words with Polarity Classification

Ramya V. V.¹, Shahad P.²

M Tech Student, Department of Computer Science & Engineering, M.E.A Engineering College, Perinthalmanna,
Kerala, India¹

Assistant Professor, Department of Computer Science & Engineering, M.E.A Engineering College, Perinthalmanna,
Kerala, India²

ABSTRACT: In recent years web technologies lead to an large quantity of user generated opinion in online systems. This large amount of opinion on web make them viable for use as data sources. As e-commerce is more and more popular, the number of customer reviews related to product receives grows rapidly. Hundreds or even thousands number of reviews are in online. It is very difficult to read and to make an informed decision on whether to purchase the product. In this case Mining opinion targets and opinion words from online reviews is an important tasks. Mainly two steps are involved to extract the opinion targets and opinion words from online reviews. First detecting opinion relations among words, it is based on alignment model. Alignment process identifying online opinion relations. Next step calculate the confidence of each candidate. Finally, extracted opinion targets or opinion words with higher confidence.

KEYWORDS: Opinion mining, Opinion target, Opinion word, Syntactic pattern, Confidence, Tagging, Supervised method

I. INTRODUCTION

Large number of reviews are in online, it is very difficult to read and to make an informed decision on whether to purchase the product or not. Opinion mining divided into two main categories first one Supervised method and another Unsupervised method. In supervised approaches, the opinion extraction task usually regarded as a sequence labeling task. The main limitation of these two methods is that labeling training data for each domain was impracticable and time consuming. Mining opinions target and opinion word from online reviews, it is very useful for customers and manufacturers. Customers can obtain summary of product information and direct supervision of their purchase actions. Manufacturers can improve the quality of their products in a timely fashion and obtain immediate feedback of product.

For example:

“This phone has a bright and big screen, but
its resolution is very disappointing.”

In the above example, positive opinion about the phone's screen and a negative opinion about its resolution, it is not the overall sentiment. To fulfil this aim, to extract both opinion targets and opinion words [1] from online review. An opinion targets are defined as nouns or noun phrases that mean opinion targets usually are product attributes or features. Opinion targets are “screen”, “resolution” in the above example. An opinion words are defined as adjectives that means that words express users' opinions. “colourful”, “disappointing” and “big” are three opinion words in the above example. Nearest-neighbour rules [2], [3] and syntactic patterns [4] approaches are used in previous methods.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

Nearest-neighbour rules only consider nearest adjective/verb and noun/noun phrase pair. This approach cannot obtain precise results.

In this paper introduce a new approach, a word alignment model. It precisely extract the opinion relations among words and also it can capture more complex word relations, such as long span relations. “colorful” and “big” are opinion words ,that aligned with the opinion target word “screen”. The word alignment model can identify an opinion target and its corresponding modifier. Compared to previous nearest-neighbor rules identifying modified relations to a limited window. Several intuitive factors, such as word co-occurrence frequencies and word positions, into a unified model for indicating the opinion relations among words. Thus, we expect to obtain more precise results on opinion relation identification. The alignment model used in [5] has proved to be effective for opinion target extraction. However, for opinion word extraction, there is still no straightforward evidence to demonstrate the WAM’s effectiveness. Compared to traditional methods word alignment model yield better results for mining opinion targets and opinion words from online reviews.

Extracting opinion targets and opinion words is regarded as a word alignment process then proposed to calculate each candidate’s confidence . Finally, opinion targets and opinion word with higher confidence than a threshold are extracted.

II. RELATED WORK

Previous methods mainly focused on sentiment analysis. Two approaches are used for this task, Supervised method and Unsupervised method. In the case of supervised methods, opinion target extraction as a sequence labeling task[2] but they require a large amount of annotated data for training and thus are less applicable compared with unsupervised methods. Mining opinion targets and opinion word is a fundamental and important task for opinion mining. To this end, there are two kinds of methods, syntax based[7] and alignment based methods. Syntax based methods usually exploited syntactic patterns ,this approach to extract opinion targets, many parsing errors are generated in this method when dealing with online informal texts. The most adopted techniques have been Nearest-Neighbour rules[3][5] and syntactic patterns[4]. Nearest Neighbour rules only consider the nearest adjective/verb to a noun/noun phrase in a limited window size as its modifier. Clearly, this strategy cannot obtain precise results. May parsing errors are generated in Syntactic method. Accordingly several syntactic patterns were designed. However, online reviews are informal writing styles, including grammatical errors, punctuation errors and typographical errors. This makes the existing parsing tools, which are usually trained on formal texts such as news paper reports, prone to generating errors. Syntax-based methods, which depend on parsing performance, suffer from parsing errors and often do not work well.

In addition, many research focused on Double Propagation[5],[7] extraction methods. it exploited syntactic based relations among words to extract opinion words and opinion targets iteratively. The main limitation is that the patterns based on the dependency parsing tree could not extract all opinion relations in reviews. Opinion target extraction is based on Skip-Tree CRF model [8], it exploited three structures including linear-chain structure, conjunction structure and syntactic structure. The main limitation of this supervised method was the need of labeled training data. If the labeled training data is insufficient, the trained data would have unsatisfied extraction performance. Labeling sufficient training data is time consuming.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

III. PROPOSED METHOD

Here propose a simple method for extracting opinion target and opinion word from online review based on word alignment model. In this paper, mine opinion targets and opinion word from online reviews is a challenging and important task in opinion mining. first , to mine opinion relations in sentences through partially-supervised word alignment model. Then, a graph-based algorithm is to estimate the confidence of each opinion target and opinion word (candidate), and the candidates with higher confidence than threshold will be extracted as the opinion targets and opinion word.

- Opinion target: Object about which users express their opinions. Nouns or noun phrases are opinion target.
- opinion words: The words that are used to express users opinions. Adjectives are opinion words.

For example:

"This phone has a bright and big screen, but its LCD resolution is very disappointing."

In the above example, three opinion words are "bright", "big" and "disappointing" and two opinion targets are "screen" and " resolution". Figure 1 represent the architecture system.

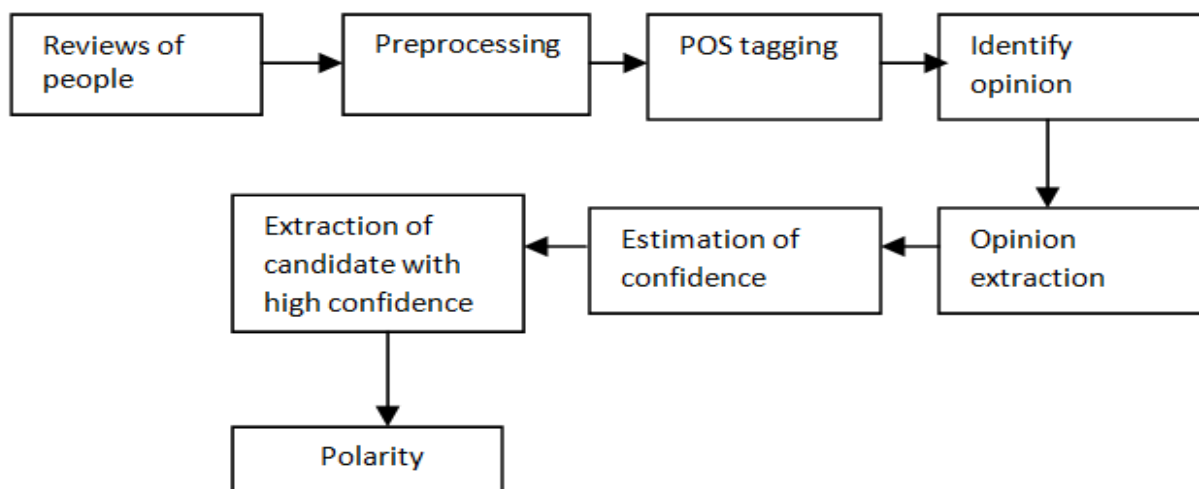


Fig. 1 System architecture

The entire system mainly includes following steps.

- Preprocessing
- Extraction opinion
- Extract opinion with high confidence
- Identify polarity

3.1 Partially-Supervised Alignment Model

In the first step, text blocks are split into their sentences and furthermore each sentence into its words for better handling. WordNetLemmatizer is a stemming tool, it used to tag every single word and put it in its principal form. Removing Special Characters, removing stop words are included in this step.Usually nouns or noun phrases are Product features in review sentences. Raw data or Reviews are collection of sentences or text. The NLProcessor

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

linguistic parser to parse each review to split the text into sentences according to punctuation and to produce the part-of-speech tag for each word in sentences (whether the word is a noun, verb, adjective, etc). This process determine simple noun and verb groups (syntactic chunking).

Tag	Description	Tag	Description
JJ	Adjective	RBR	Comparative adverb
JJR	Comparative adjective	RBS	Superlative adverb
JJS	Superlative adjective	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
NN	Noun, singular or mass	VBG	Verb, gerund, or present participle
NNS/NNP	Noun, plural noun, singular	VBN	Verb, past participle
NNPS	Proper noun, plural	VBP	Verb, non-3rd-person singular/p
RB	Adverb	VBZ	Verb, 3rd-person singular present

Table. 1 Part-of-speech (POS) tagging

In word alignment model, which identifying opinion relations among words. All nouns/noun phrases in sentences are regarded as opinion target candidates, and all adjectives/verbs are regarded as opinion words. The syntactic patterns are not provide a full alignment, so a EM-based algorithm adopted, named as constrained hill-climbing algorithm[9], to determine the parameters. In this training process, the constrained hill-climbing algorithm can ensure that the final model is marginalized on the partial alignment links. Particularly, in the E step, their method aims to identify the alignments which are consistent to the alignment links provided by syntactic patterns, Mainly two steps are involved.

- Optimize towards the constraints.
- Towards the optimal alignment under the constraints

Some constraints are used in word alignment model as follows:

- Nouns/noun phrases (verbs/adjective) must be aligned with adjectives/verbs (nouns/noun phrases) or a null word.
- Null word means that this word either has no modifiers or modifies nothing.
- Other unrelated words, such as conjunctions ,prepositions, and adverbs, can only align with themselves.

Consider above two constraints, to obtain Fig.1 the following alignment results. where "NULL" means the null word, that means this word either has no modifier or modifies nothing From above example, the unrelated words are "This", "a" and "and". These are aligned with themselves. The word "Phone" and "has" modifies nothing and no opinion words to modify therefore, these two words align with "NULL" word.

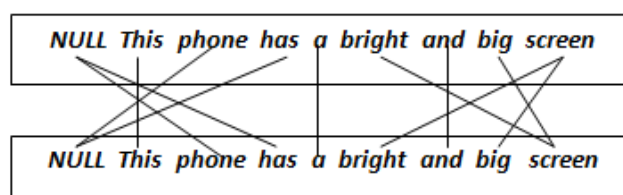


Fig .2 Opinion relations between words using the word alignment model under constrains

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

Given a sentence S with n words, $S = (w_1, w_2, \dots, w_n)$, the word alignment is,

$$A^{\wedge} = \{(i, ai) | i \in [1,n], ai \in [1,n]\} \quad (1)$$

Where (i, ai) means that a noun/noun phrase (opinion target) at position i is aligned with its modifier (opinion word) at position ai .

3.2 Opinion Associations Among Words

After alignment results [1], To get a set of opinion word and opinion target (candidates) pairs, each pair composed of a noun/noun phrase (opinion target candidates) and its corresponding modified word (opinion word candidate). Next, estimate the alignment probabilities between a opinion target W_t and a opinion word W_o ,

$$P(W_t|W_o) = \text{Count}(W_t, W_o) / \text{Count}(W_o) \quad (2)$$

Where $P(W_t|W_o)$ means the alignment probability between opinion target and opinion words. Similarly, the alignment probability $P(W_o|W_t)$ by changing the alignment direction in the alignment process. Then calculate the opinion association $OA(W_t, W_o)$ between w_t and w_o as follows:

$$OA(W_t, W_o) = (\alpha * P(W_t|W_o) + (1 - \alpha)P(W_o|W_t))^{-1} \quad (3)$$

Where α is a constant, a harmonic factor used to combine these two alignment probabilities. In this paper, $\alpha = 0.5$.

Estimating Candidate Confidence [1] With Graph Co Ranking. Confidence of a candidates related to neighbor words. Confidence of a candidate (opinion target or opinion word) determined by its neighbors according to the opinion associations among them.

3.3 Graph Construction

$G = (V, E, W)$ is a Relation Graph, it is bipartite undirected graph. In relational graph G , $V = V_t \cup V_o$ denotes the set of vertices. $vt \in V_t$ denote opinion target candidates (the below nodes) and $vo \in V_o$ denote opinion word candidates (the above nodes). E denote edge set of the graph G , where $e_{ij} \in E$ means that there is an opinion relation between two vertices. It is worth noting that the edges e_{ij} is an edge between vt and vo and there is no edge between the two same types of vertices. $w_{ij} \in W$ means the weight of the edge e_{ij} . Graph based co-ranking algorithm to calculate the confidence of each candidate pair. Finally, candidates with higher confidence than a threshold are extracted as the opinion targets or opinion words from online review

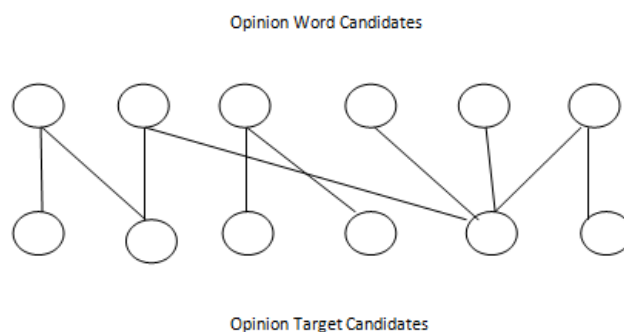


Fig. 3 Opinion relation graph

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

3.4 Candidate Confidence

Calculate the confidence of each candidate (Opinion target and Opinion word) through standard random walk with restart algorithm. Confidence of each candidate is,

$$Ct^{(k+1)} = (1-\mu) * Mt0 * Co^{(k)} + \mu * It \quad (4)$$

$$Co^{(k+1)} = (1-\mu) * Mt0 * Ct^{(k)} + \mu * Io \quad (5)$$

where $Ct^{(k+1)}$ and $Co^{(k+1)}$ are the opinion target candidate and opinion word candidate confidence, in the $(k + 1)$ iteration. $Co^{(k)}$ and $Ct^{(k)}$ are the confidence of an opinion target candidate and opinion word candidate in the k iteration. $Mt0$ is opinion associations between opinion target candidate and opinion word candidate. $mij \in Mto$ means the opinion association between the i -th opinion target candidate and the j -th opinion word candidate. where $Mt0 * Co^{(k)}$ and $Mt0 * Ct^{(k)}$ are the confidence of an opinion target or opinion word candidate is obtained through aggregating confidences of all neighboring opinion word (opinion target) candidates together according to their opinion associations. The other ones are It and Io , which denote prior knowledge of opinion target candidate and opinion word candidate.

3.5 Classification of candidates

Naive Bayes Classifier method is used to classify the candidates. Positive and negative classification method is used in this paper. Initially Positive and negative word list are exist. Seed list initially contains some of the opinion words along with their polarity. From the tagged output all the opinion words were extracted. The extracted opinion words matched with the words stored in positive list, in this case the output is positive otherwise negative. Fig 4 represent Naive Bayes Classifier.

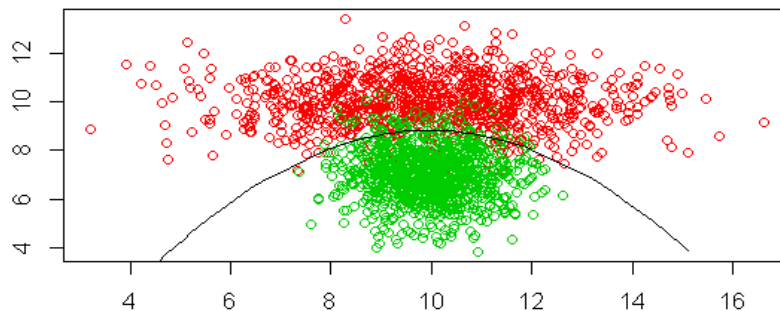


Fig. 4 Naive Bayes Classifier

IV. EXPERIMENTAL RESULTS

Many datasets are used in this approach. It includes English reviews of many products such as cameras, cars, laptops and phones etc. In the experiments, reviews or text are first segmented into sentences according to punctuation. Next, sentences are tokenized, with part-of-speech tagged using the Stanford NLP tool. Then parse English sentences. The POS tagging method is used to identify noun phrases. Precision (P), recall (R) and F-measure (F) as the evaluation metrics. The below table is a comparison table for four methods, the domain is phone.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

No	Paper	Method	Precision	Recall	F-score
1	Mining Opinion Features in Customer Reviews	Nearest-Neighbor rules	0.56	0.73	0.63
2	Walk and Learn: A Two-Stage Approach for Opinion Words and Opinion Targets Co-Extraction	Two stage method	0.73	0.71	0.72
3	Extracting Product Features and Opinions from Reviews	Syntax based methods	0.78	0.66	0.72
4	Partially supervised approach for co-extracting opinion target and opinion words from online reviews	PSWAM	0.86	0.75	0.80

Table.2 comparison of four method

V. CONCLUSION

Word alignment is more precisely and more effective method for opinion mining. This method gives an idea a novel Partially-Supervised Word Alignment Method for co-extracting opinion targets and opinion words. It is mainly focused on detecting opinion relations between opinion targets and opinion words. Next construct an Opinion Relation Graph to estimate the confidence of each candidate. The items with higher ranks or confidence are extracted out. This approach is useful both customer and manufactures. Next classify the candidates into positive and negative class using naive Bayes Classifier

REFERENCES

- [1] Kang Liu, Liheng Xu, and Jun Zhao, "Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model", *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp.636-650, March 2015.
- [2] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proc. 19th Nat. Conf. Artif. Intell.*, San Jose, CA, USA, pp. 755–760, 2004.
- [3] T. G. Moe, W. Li, Moe, and Z. Sui, "A semi-supervised method for opinion target extraction", pp. 275–276.
- [4] Q. Zhang, Y. Wu, T. Li, M. Ogihara, J. Johnson, and X. Huang, "Mining product reviews based on shallow dependency parsing", in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Boston, MA, USA, pp. 726–727, 2009.
- [5] G. Qiu, B. Liu, J. Bu, and C. Che, "Expanding domain sentiment lexicon through double propagation", in *Proc. 21st Int. Joint Conf. Artif. Intell.*, Pasadena, CA, USA, pp. 1199–1204, 2009.
- [6] B. Wang and H. Wang, "Bootstrapping both product features and opinion words from Chinese customer reviews with cross inducing", in *Proc. 3rd Int. Joint Conf. Natural Lang. Process.*, Hyderabad, India, pp. 289–295, 2008.
- [7] G. Qiu, L. Bing, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation", *Comput. Linguistics*, vol. 37, no. 1, pp. 9–27, 2011.
- [8] Fangtao Li, Chao Han, Minlie Huang, Yingju Xia, Shu Zhang, and Hao Yu, "Structure aware review mining and summarization", pp. 653–661, 200.
- [9] Qin Gao, Nguyen Bach, and Stephan Vogel, "A semi-supervised word alignment algorithm with partial manual alignments", in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 1–10, Uppsala, Sweden, Association for Computational Linguistics, pp.1-10, July 2010.
- [10] Sharma, Richa, S. Nigam, and R. Jain, "Opinion mining of movie reviews at document level", 2014.
- [11] Hemalatha, I. G. S. Varma, and A. Govardhan, "Preprocessing the informal text for efficient sentiment analysis", *International Journal*, 2012.
- [12] Xu and Liheng, "Walk and learn: a two-stage approach for opinion words an opinion targets co-extraction", *Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee*, 2013.