Opinion Based Word Extraction of Supervised and Unsupervised Models for Alignment

Gaidhani Shalaka Dinkar¹, Dr. M. U. Kharat²
PG Student¹

Department of Computer Engineering
MET’s Institute of Engineering BKC Adgoan, Nashik, Maharashtra, India

Abstract:
In today’s electronic era, consumers are often forced to go through many on-line reviews in order to make product choice. The Web provides ocean of opinions about products. So, the user needs to do proper analysis of the reviews, which is done by aligning the reviews efficiently. Previously, alignment was carried out using syntactic pattern matching in an unsupervised environment. Moreover, this alignment did not prove to be efficient and the alignments which were made propagated error. So, to overcome such supervised word alignment model is introduced. Opinion relations are identified as an alignment process. In order to estimate the confidence of each candidate, a graph-based co-ranking algorithm is exploited. Finally, candidates with higher confidence are extracted as opinion targets or opinion words.

Index Terms: Opinion, targets, opinion words, alignment models

I. INTRODUCTION

Due to the rapid growth of e-commerce, more and more products are sold online. As the common users are being comfortable with the Internet, an increasing number of users are writing reviews. It has become necessity to analyze these reviews. Reviews are opinions of the users, so opinion mining is carried out. In opinion mining the two fundamental subtasks are extracting opinion targets and opinion words. Opinion targets are objects about which users express their opinions, and opinion words are words which show opinions’ polarities. Proper alignment of the extracted opinions is the main task.

In order to precisely mine the opinion relations among words, a method based on a monolingual word alignment model (WAM) can be used. The corresponding modifier an opinion target can be estimated through word alignment.

```
Fig. 1
```

In above Fig. 1, the opinion words “colourful” and “big” are aligned with target word “screen”. Compared to earlier nearest-neighbour rules, the WAM does not confine identifying modified relations to a limited gap; therefore, it can capture more complex relations.

1. Standard Word Alignment model (WAM) is seldom trained in an unsupervised manner, that results in alignment quality that may be unsatisfactory. However, alignment quality can be improved by using supervision. Thus, a technique called Partially Supervised Alignment Model (PSWAM) is employed. For example, in Fig. 2, “kindly” and “courteous” are erroneously identified as modifiers for “foods” if the WAM is performed in a completely unsupervised manner. However, by using some syntactic patterns, we can claim that “courteous” should be aligned to “services”. Through the PSWAM, “kindly” and “courteous” are correctly associated to “services”.

```
Fig. 2
```

2. To ease the problem of error propagation, use co-ranking, particularly, a graph, named as Opinion Relation Graph, is constructed to model all opinion target/word candidates and the opinion relations among them. A random walk based co-ranking algorithm is proposed to estimate each candidate’s confidence on the graph. High-degree vertices are penalized to weaken the impacts of unrelated opinions. Meanwhile, prior knowledge of candidates is calculated. Finally, candidates with higher confidence than a threshold are extracted.

II. LITERATURE SURVEY

In the existing system partially supervised alignment model (PSWAM) is used to identify opinion relations and calculate the opinion associations between opinion targets and opinion words. Extraction of opinion targets and words [3] is the main concept for relevance. Reviews that are given online can be domain specific or independent, so for extraction of opinion words IEDR method is used [2] which spots the dissimilarity in opinion feature statistics across two corpora, one domain-related corpus and one domain-independent corpus. Monolingual model is used to confine opinion relations and calculate the opinion associations between opinion targets and opinion words [4]. Mining Hu and Bing Liu [2] Proposed a system which performs the summarization in two main levels:
feature extraction and opinion orientation identification. The inputs to the system are the product name and an entry page for all the reviews of the product. The output is the overall summary of the reviews. Ana-Maria Popescu and Oren Etzioni [3] introduced OPINE, an unsupervised information extraction system that mines reviews to build a model of vital product features, their evaluation by reviewers, and their relative quality across products. F. Li, C. Han ET AL [4], focused on object feature based review summarization. Disparate of the previous work with linguistic rules or statistical methods which formulated the review mining task as a joint structure tagging problem a new machine learning framework based on Conditional Random Fields (CRFs) is proposed. Kang Liu, Liheng Xu, Jun Zhao[3] proposed a approach for extracto opinion targets based on word based translation model (WTM). At first, they applied WTM in a monolingual scenario to mine the associations between opinion targets and opinion words. Then, a graph based algorithm is exploited to extract opinion targets.

III. PROPOSED SYSTEM

A collective extraction strategy was previously adopted. The instinct represented by this strategy was that in sentences, opinion words usually appear with opinion targets, and there are strong variation relations and associations among them (called as opinion relations or opinion associations). In earlier methods, mining the opinion relations between opinion targets and opinion words was solution to collective extraction. To this end, the most-adopted techniques have been nearest-neighbor rules and syntactic patterns. A noun/noun phrase can find its modifier through word alignment. Partially-supervised word alignment model undergoes word alignment in a partially supervised framework. After that, a huge number of word pairs, each of which is composed of a noun or noun phrase and its modifier is obtained. Then associations between opinion target candidates and opinion word candidates is calculated as the weights on the edges. For the second problem, a random walking with restart algorithm is exploited to propagate confidence among candidates and the confidence of each candidate on Opinion Relation Graph is estimated. More explicitly, the high-degree vertices are penalized according to the vertices’ entropies and integrate the candidates’ prior knowledge. In this way, extraction precision can be improved. The proposed method can be implemented as follows:

First extract Opinion targets and opinion words from given dataset.
- Capture opinion relations between opinion targets and opinion words
- Word alignment Model(WAM)
- Partially Supervised WAM (PSWAM)
- Calculating opinion association among words

Estimating candidate confidence with graph co-ranking
- Random walking
- Penalizing on high degree vertices
- Calculate candidates prior knowledge

IV. SYSTEM ARCHITECTURE

A Opinion Relation Graph that is a bipartite undirected graph G = (V;E;W) is constructed. Based on Opinion Relation Graph, to estimate the confidence of each candidate a graph-based co-ranking algorithm is implemented. The candidate with higher value than threshold is extracted as opinion word or target with the help of graph co-ranking algorithm. Confidence estimation is done by ways viz. “Random walking”, “Penalizing vertices” and “Prior knowledge of candidate”.

INPUT: Online customer reviews
PROCESSING:
- Extract opinion targets and words
- Construction of bipartite graph (opinion relation graph).
- Apply the graph based co-ranking algorithm in order to estimate confidence.
- Assign confidence to each candidate.
- Apply WAM and PSWAM.
- Calculate opinion associations.

OUTPUT: Associations and Opinion relations among words and targets.

A bipartite undirected graph G = (V;E;W) called as Opinion Relation Graph is constructed. Based on Opinion Relation Graph, a graph-based co-ranking algorithm will be implemented to estimate the confidence of each candidate. The candidate with higher value than threshold is extracted as opinion word or target with the help of graph co-ranking algorithm. Confidence estimation is done by ways viz. “Random walking”, “Penalizing vertices” and “Prior knowledge of candidate”. In which, random walking is carried out with the help of “Constraint Hill-Climbing Algorithm”. Opinion relation graph estimates the opinion relations among candidates. This is done in two phases viz. WAM and “PSWAM”. In this WAM is carried out in unsupervised environment whereas, PSWAM in partially supervised environment.

A. Word Alignment Model

A bilingual word alignment algorithm is applied to the monolingual scenario to align a noun/noun phase potential opinion targets) with its modifiers (potential opinion words) in sentences. Given a sentence with ‘n’ words

\[ S = W_1, W_2, W_3, ..., W_n \]

Word alignment \[ A = (i,a_i) | i \in [1,n], a_i \in [1,n] \]

can be obtained as,

\[ A^* = \arg\max_A P(A|S) \] (1)

Where,

\[ (i,a_i) = \text{noun/noun phrase at position } i \text{ is aligned with its modifier at position } a_i \]

B. Partially Supervised Word Alignment Model

Standard word alignment model is usually trained in a completely unsupervised manner, which may not obtain precise alignment results. Thus, to improve alignment performance, perform a partial supervision on the statistic model and employ a partially-supervised alignment model to incorporate partial alignment links into the alignment process. Here, the partial alignment links are regarded as constraints for the trained alignment model. Partial word alignment can be obtained as,

\[ A^* = \arg\max_A P(A|S, \hat{A}) \] (2)

Where,

\[ \hat{A} = \text{Partial Alignment Links}, \]

\[ A = \{(i,a_i) | i \in [1,n], a_i \in [1,n]\}, \]

\[ A^* = \text{Optimal Alignment} \]
C. Parameter Estimation for PSWAM

Alignments generated by the PSWAM must be as consistent as possible with the labeled partial alignments. To fulfill this aim, an EM-based algorithm is used. The search space for the optimal alignment is constrained on the "neighbor alignments" of the current alignment, where "neighbor alignments" denote the alignments that could be generated from the current alignment by one of the following operators:

1. MOVE operator $m_i$ which changes $a_i = i$
2. SWAP operator $s_{j1,j2}$ which changes $a_{j1}$ and $a_{j2}$

D. Constraint Hill-Climbing Algorithm

Alignments generated by PSWAM must be as consistent as possible with the labeled partial alignments. To achieve this an EM-based algorithm is used. A complex alignment model viz. IBM-2, HMM is used. This indicates that standard EM training algorithm is time consuming and impractical. To resolve this problem GIZA++ provides a Hill-climbing algorithm, which is a local optimal solution to accelerate the training process.

Algorithm

Constraint Hill-Climbing algorithm mainly concentrates on optimizing constraints in order to obtain optimal alignments.

Input: Review Sentences $S_i = \{W_1, W_2, W_3, \ldots, W_n\}$

Output: Calculated alignment $\hat{a}$

Initialization: Calculate the seed alignment $a_0$ using simple model (IBM-1, IBM-2, HMM)

1. Step 1 : Optimize towards constraint
2. While $N_{ii}(\hat{a}) > 0$ do
3. If $\{a : N_{ii}(a) < N_{ii}(\hat{a})\} = \emptyset$ then
4. break
5. $\hat{a} = \text{argmax}_{a \in \mathbb{A}} \Phi(a, P_{a\text{a}})$
6. end
7. Step 2: Toward optimal alignment under the constrain
8. For $i < N$ and $j < N$ do
9. $M_{ij} = 1$, if $(i,j) \not\in A$
10. end
11. While $M_{i1j} > 1$ or $S_{j1i} > 1$ do
12. If $(a_1, a_2) \not\in A$ or $(j_1, a_2) \not\in A$ then
13. $S_{j1i} = -1$
14. end
15. $M_{i1j} = \text{argmax } M_{ij}$
16. $S_{j1i} = \text{argmax } S_{ij}$
17. If $M_{i1j} > S_{j1i}$ then
18. Update $M_{i1} \leftarrow M_{i1} + M_{j1}$
19. Update $S_{j1} \leftarrow S_{j1} + S_{j1}$
20. Set $\hat{a} := M_{i1j}(a)$
21. end
22. else
23. Update $M_{i1} \leftarrow M_{i1} - M_{j1}$
24. Update $S_{j1} \leftarrow S_{j1} - S_{j1}$
25. Set $\hat{a} := S_{j1i}(a)$
26. end
27. end
28. return $\hat{a}$

E. Calculating Opinion Association among Words

From the alignment results, a set of word pairs is obtained, each of which is composed of a noun/noun phrase (opinion target candidate) and its corresponding modified word (opinion word candidate).

V. CURRENT STATUS OF IMPLEMENTATION

A. Experimental Setup

Dataset: The real data is the CRD dataset i.e customer review dataset, that contains the reviews of numerous users regarding various different products.

Setup: JDK environment is used for implementation. The experiment is run on Windows with Intel core 2 dual processor, speed is 2.20 GHz and RAM is 1GB.

Implementation Details:-

User first has to select the required dataset, once the dataset is loaded the TF and IDF is calculated. Average Term & Document frequency is calculated from the obtained results. Next, the frequency weight of each term is calculated, this refers to the occurrence of each term in the document. Deviation of the terms is calculated, it refers to the relevance of a specific term to the different domain terms.IDR/EDR scores are calculated depending on the deviation results. Lastly the alignment of the terms is carried out.

In further to the base implementation, it is observed that the dataset of reviews are only of single domain. Only English input reviews are considered. Here for contribution other language input is also considered. User can input language in other languages too.

Selection of Dataset:- User can select required dataset as per their dataset.

Term & Document Frequency:-
Average Term & Document Frequency:

Alignment:

Expected Result

The algorithm used in PSWAM will outperform the previous method of alignment as it will be implemented in supervised scenario. The algorithm used will give the user proper aligned opinion targets and opinion words which will help user analyze the reviews more effectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Existing</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laptop</td>
<td>0.74</td>
<td>0.84</td>
</tr>
<tr>
<td>Phone</td>
<td>0.76</td>
<td>0.8</td>
</tr>
<tr>
<td>Hotel</td>
<td>0.71</td>
<td>0.76</td>
</tr>
</tbody>
</table>

VI. Conclusion

Here, a method for co-extraction of opinion words and opinion targets by using a word alignment model is proposed. Main involvement is focused on detecting opinion relations between opinion words and opinion targets. Compared to earlier methods based on nearest neighbor rules and syntactic patterns, by using a word alignment model, this method capture opinion relations more specifically and hence is more efficient for opinion target and opinion word extraction. Next, construct an Opinion Relation Graph to model all candidates and the detect opinion relations among them, along with a graph co-ranking algorithm to approximate the confidence of each candidate. The items with higher ranks are extracted. The experimental results for three datasets with different languages and different sizes prove the effectiveness of the proposed method.

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 Trans on Knowledge and Data Engineering vol.27, march 2015,
