Fine-grained Opinion Extraction with Mixed Network Model

Pan ZHANG, Hong-rong CHENG, Teng-yuan CAI and Qi-he LIU

School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, P.R. China

*Corresponding author

Keywords: Opinion mining, Entity extraction, Network model.

Abstract. With the development of opinion mining, the relevant areas are being valued more than before. As the foundation of opinion mining, the performance of opinion extraction is significant for the opinion mining results. Among the present methods of opinion extraction, many of them take advantages of the relation features. Unlike other models which use relations in an independent way, in this paper we propose a mixed network model which interactively uses two relations as well as other lexical and syntactic features to solve the extraction problem. A series of experiments show that our model outperforms other baseline models and has a further potential of improvement.

Introduction

The extraction task is fundamental for opinion mining, also called sentiment analysis. The trend of opinion extraction is generally from coarse-grained document-level to sentence-level, and then to fine-grained entity-level. Standard machine learning methods like SVM are used in [1,2] to analyze reviews on a document-level analysis. The researches such as [3,4] study about the subjectivity feature on sentence-level analysis. And in [5,6] bring up the joint model to solve extraction task on entity-level.

The critical task in entity-level is to extract entities from a whole sentence. Rule-based methods once were the main approaches for opinion extracting [8]. Recently there has been a growing interest in the statistical-based methods and statistical-rule-mixed methods. In these methods, the opinion relations between entities are proven to be a kind of valuable feature for extraction [6]. In this paper, we adopt the ways of Y&C [5] to define two kinds of relations. Here is an example from MPQA corpus [9]:

*Sally (holder) is satisfied (opinion) with her new camera (target)*

In the above example, the opinion, the target of the opinion and the holder of the opinion annotated by subscripts are the three categories of entities. The two kinds of relations are IS-ABOUT and IS-FROM, the former one indicates that an opinion is about a target, the latter one indicates that an opinion is from a holder.

In this paper, we use relations between entities to create a mixed network model as Figure 1 shown, it combines several features to extract opinion entities, with a preliminary input from a CRFs sequence labeling process. In Figure 1, the vertexes such as T1 to T3 are words in sentences. Each cluster circled by dotted line indicates an opinion expression, which can contain one or more words. The edges between clusters are either the IS-FROM or IS-ABOUT opinion relations, edges within clusters indicate that the words can form an entity. Since the model uses the two relations interactively, we call it relation based mutual extraction model (RMEM).

We perform experiments on MPQA 3.0 dataset, results of serials experiments show that our model outperforms the preliminary extraction method which considered no relations.
Related Work

If the entities can be identified accurately, it could provide much valuable information for the follow-up analysis. For instance, the detection of opinions and their corresponding targets of one product is of much value to reflect the views of customers in different aspects of product.

There are many previous researches devoted to extract entities in sentences. In Liu and Xu [11], a network model is brought up to extract entities of the product reviews. In [12], a small graph is used for locating the verb expressions. In [5], Y&C proposed the first joint inference model in which a sequence labeling method was used to extract entities roughly, then relation features are added, and an ILP formulation was used to jointly evaluate whether an entity should be extracted. The [13] proposed a MLN model to tackle the extraction problem. Their idea are similar to Y&C, but they used MLN model instead of the ILP one. The RMEM we proposed has the same input as [5,13], which is the preliminary extraction result of the CRFs, but we our model connects two kinds of relations which can provide more information.

Proposed Model

In this section we propose our model RMEM. The RMEM mainly comprises two parts:

Preliminary extraction

We adopt Conditional Markov Fields (CRFs) to preliminary extraction task, it is seen as a sequence labeling problem. We extract entities with traditional lexical and semantic features like POS tags and subjectivity. The sentences are formed by words with order, thus making them sequences. In a conditional probability model \( P(y|x) \), \( Y = (y_1, y_2, \ldots, y_n) \) is the sequence of labels, \( X = (x_1, x_2, \ldots, x_n) \) is the sequence of input, each \( x \) is a word. In the learning stage, given training data, the conditional probability model \( \hat{P}(y|x) \) is obtained by maximum likelihood estimation. In the predicting stage, given input \( X \), the model would give the output \( y \) with \( \max \hat{P}(y|x) \). The conditional probability can be computed through the forward-backward algorithm.

Mixed network model

We implement our mixed network model as Figure 1 showed. We regard the results of the preliminary extraction as the vertexes, and add the opinion relations as edges to create a mixed network model, then we use a random walk on this model to get a more precise result. The two relations the model mixed are IS-FROM and IS-ABOUT relations, which denotes the connections between entities. We extract IS-FROM and IS-ABOUT relations using two different ways as described in experiment part. The extraction is done by a binary classifier. Candidates in preliminary extraction results are formed.
to opinion-target pairs and opinion-holder pairs. These two kinds of pairs become candidates respectively corresponding to IS-ABOUT and IS-FROM relations. The pairs that match human annotation are regarded as positive examples.

The network is regarded as a graph $G = (V, E)$, to quantify whether a word span should be extract as an entity, the words are given scores, the higher the score, the more likely the word span is an entity. In the network model, the words within an entity have different initial scores from words outside entity. Our purpose is to let the entities find each other through their relations, the practice is to make a random walk on the graph, which means to let one category of entities to decide the score of the other category. The training data we generate from the MPQA corpus are organized in opinion-holder pairs and opinion-target pairs. The opinion entities are connected both with holders and targets, thus they are co-decided by both targets and holders. We propose the following iterative formulas to update the scores of the vertexes:

\begin{align*}
S_{\text{holder}} &= \mu_1 \times W_h \times S_{\text{opinion}} + (1 - \mu_1) \times S_{\text{holder}} \quad (1) \\
S_{\text{target}} &= \mu_2 \times W_t \times S_{\text{opinion}} + (1 - \mu_2) \times S_{\text{target}} \quad (2) \\
S_{\text{opinion}} &= \mu_3 \times W_h \times S_{\text{holder}} + \mu_4 \times W_t \times S_{\text{target}} + (1 - \mu_3 - \mu_4) \times S_{\text{target}} \quad (3)
\end{align*}

The formulas show the entities cross decide each other’s value, the $\mu$ is a parameter to adjust how much influence one entity would have on the other, $W$ is the weight matrix that record the relations between them, there are IS-FROM ($W_h$) and IS-ABOUT ($W_t$) matrix. The overall process are shown as Table 1.

### Table 1. Mixed Network Model.

<table>
<thead>
<tr>
<th>Input:</th>
<th>The preliminary extracted entities from CRFs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IS-ABOUT and IS-FROM relations (can be extracted through different methods)</td>
</tr>
<tr>
<td>Output:</td>
<td>The entities (a group of words/vertexes)</td>
</tr>
<tr>
<td>Procedure:</td>
<td>1: Create network model, each vertex is assigned with an initial value</td>
</tr>
<tr>
<td></td>
<td>2: Update the value of each vertex with our proposed iterative formula Eq.1,2,3.</td>
</tr>
<tr>
<td></td>
<td>3: After run a certain number of iterations, stop updating.</td>
</tr>
<tr>
<td></td>
<td>4: Extract the groups of vertexes whose values are higher than the threshold as entities</td>
</tr>
</tbody>
</table>

## Experiment

### Dataset and Evaluation Metrics

We implement our experiment on the version 3.0 MPQA dataset. This corpus consists of news documents in several different topics. The annotation we use in this paper are those ones that denote opinion expressions and relations between them. The three kinds of annotations that denote holder, opinion, and target are “agent”, “direct-subjective”, and “sTarget”, the IS-FROM and IS-ABOUT relations are generated form “attitude” and “targetFrame”.

The evaluation metrics we use for both entity extraction and relation extraction is the PRF (precision, recall, f-measure) evaluation system. Since an expression could contain more than one word, the border of word spans are hard to define, so we adopt the overlap criteria, which means that if a span overlaps with a gold standard then it is considered to be correct.

We implement a CRFs model for preliminary extraction. For relation extraction, we implement the Adj method in [5] and the relation extraction method in [13]. These methods are used in previous
researches which have similarity to our work, thus they are suitable for comparison. We assemble the two relation extraction methods to mixed network model as our baselines.

**CRFs:** To preliminarily extract entities, we discard some badly annotated sentences, and organize the rest into a format that the words are separately put in lines, each word line contains the word itself, its POS tag, the subjectivity of this word (obtained from the WordNet Subjectivity Lexicon) and the label of this span. The labels are represented in the “IOB” format, which is a common tagging format used by many researches. To implement CRFs model, we use CRF++. We create a customized template file to make each word connect to their features and the adjacent words. Only one neighbor front and back of the current word is considered in this template. The hyper-parameter c is set to 2.25 after running a 10-fold cross validation. All the other parameters are set to default.

**Adj method in [5]:** The main practice of this method is to assume the adjacent entities are more likely to have relations (adjacency rule). We iterate the documents and match entities to their neighbors, the opinion-target pairs and opinion-holder pairs we get are results of relation extraction.

**Relation Extraction Method in [13]:** The main practice of this method is to adopt both lexical features and dependency path to train the relation classifier. We generate the dependency path and distance (the words’ number between two different spans) from the document using Stanford coreNLP [14]. Because we chose a different corpus from [13], our classifier is trained by LIBSVM [15] instead of LIBLINEAR in [13] to fit the corpus’s size.

<table>
<thead>
<tr>
<th>Adj method in [5]</th>
<th>IS-ABOUT</th>
<th>IS-FROM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>31.3</td>
<td>23.0</td>
<td>26.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>86.4</td>
<td>78.0</td>
<td>81.9</td>
</tr>
</tbody>
</table>

To further extract entities with our mixed network model, the CRFs result is used as input. A relation weight matrix is used to represent the relations between entities. Since the extraction of two different relations are separated, so in the Eq.4 W_h and W_t cannot be merged together. We load the different relation extraction results to our network model separately to get entity extraction results. Then we compare their performances:

<table>
<thead>
<tr>
<th>Opinion</th>
<th>Target</th>
<th>Holder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>CRFs</td>
<td>41.2</td>
<td>63.5</td>
</tr>
<tr>
<td>Network Model + Adj method in [5]</td>
<td>63.3</td>
<td>36.4</td>
</tr>
<tr>
<td>Network Model + Relation Extraction Method in [13]</td>
<td>62.6</td>
<td>67.1</td>
</tr>
<tr>
<td>Network Model + Gold Standard</td>
<td>77.8</td>
<td>38.4</td>
</tr>
</tbody>
</table>

**Results**

Table 2 and Table 3 separately show the results of the two relation extraction methods and the entity extraction results. The entity extraction results are consists of preliminary extraction and network model extraction results with different relations features.

When we look at Table 2, the difference in data shows that relation extraction method in [13] has a much better performance than Adj method. It is reasonable because the former one adopt dependency path features. And we can see that the IS-ABOUT relation is more difficult to extract than the IS-FROM relation, one possible reason is that a target can be expressed in many more ways than a holder.

The Table 3 gives us the information about how the model works. The two relation extraction methods are not implemented completely the same as in their original paper, and they are run in a different corpus, so their performance are not as good as in the reference paper, but they can still provide some improvement to the preliminary extraction. This meets the intuition because the preliminary extraction lack the help of relation features. One other discover is that the better extracted
relation can lead to a better entity extraction result. The best result of target and holder are extracted by network model with gold standard relations. The increase in number indicates that a practical way to raise the performance of our network model would be to apply more accurate relations.

Conclusions
In this paper, we propose a relation based mutual extraction model (RMEM) to extract opinion-related entities. The extraction is divided into two parts. The preliminary extraction by CRFs and the further extraction by network model. From experiments we know our model can effectively utilize the information carried by the lexical features as well as opinion relation features. Our model can achieve better performance than the methods that only adopt sequence labeling or relation classifying methods. For further work, we plan to explore and search for more accurate features that can indicate the entities, and to load these features to our model to reach a better result.

References