Interactive Multi-Grained Joint Model for Targeted Sentiment Analysis

Da Yin*, Xiao Liu†, Xiaojun Wan
Institute of Computer Science and Technology, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
{wade_yin9712,lxlisa,wanxiaojun}@pku.edu.cn

ABSTRACT

In this paper, we propose an interactive multi-grained joint model for targeted sentiment analysis. Firstly, different from previous works, we leverage the correlation between target and sentiment clues and deeply strengthen interaction between them because targets are highly related to the sentiment clues in a sentence. Moreover, we apply a multi-layer structure to consider multi-grained target and sentiment tagging information more comprehensively. Also, we design two specific loss functions to prevent a word from being both part of a target and a sentiment clue simultaneously, and to align the boundary information of two labeling subsystems. We conduct experiments on English and Spanish datasets and the experimental results show that our approach substantially outperforms a variety of previous models and achieves new state-of-the-art results on these datasets.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Information extraction**;

KEYWORDS

Sentiment Analysis; Joint Model; Multi-grained Model; Interaction Mechanism; Sequence Labeling; Neural Networks

ACM Reference format:
Da Yin, Xiao Liu, Xiaojun Wan. 2019. Interactive Multi-Grained Joint Model for Targeted Sentiment Analysis. In *Proceedings of The 28th ACM International Conference on Information and Knowledge Management, Beijing, China, November 3–7, 2019 (CIKM '19),* 10 pages.
https://doi.org/10.1145/3357384.3358024

1 INTRODUCTION

Targeted sentiment analysis aims to extract targets and simultaneously identify the sentiment polarities toward the extracted targets [6, 9, 40]. For example, the sentence *'The **food** is good but the **service** is bad'* has two targets *'**food**'* and *'**service**'*, of which sentiment polarities are positive and negative respectively. Most previous works pay more attention to either target extraction [17, 34, 37, 41] or sentiment classification [1, 15, 29, 30, 36] independently, and they are the subtasks of targeted sentiment analysis.

Table 1: An example of the tagging scheme applied in targeted sentiment analysis task.

	Feel	so	sad	about	Whitney	Houston
Target Extraction	O	O	O	O	B-Person	I-Person
Sentiment Classification	O	O	O	O	B-Negative	I-Negative

Previous researches treat targeted sentiment analysis as a sequence labeling task [12, 20, 23, 40, 43]. A specific series of tagging labels are proposed by combining *BIO* scheme with named entity types (e.g. person and organization) and sentiment polarity types (e.g. *positive, negative* and *neutral*). As shown in Table 1, in the sentence *'Feel so sad about **Whitney Houston**.'*, the target *'**Whitney Houston**'* is mentioned. The correct tagging results of sentiment classification should be *'O O O O B-Negative I-Negative'*. Similar to the output format of Named Entity Recognition (NER) task [31], which adopts tagging labels of *B-Person, I-Person, B-Organization, I-Organization* and *O*, the correct tagging results of target extraction ought to be *'O O O O B-Person I-Person'*.

Two kinds of methods are mainly used in targeted sentiment analysis: pipeline and joint methods. For the reason that target extraction is prior to and also the basis of target-based sentiment classification, pipeline was the usual way in the past. However, pipeline methods tend to result in error accumulation and they cannot fully take advantage of the mutual benefits of the two subtasks, as revealed by [16, 23]. Nowadays joint models become popular for their effectiveness in sharing information between two subtasks and reducing the risk of error accumulation. Previous joint models mainly use Conditional Random Fields (CRF) [13] to model the dependencies between labels [20, 23, 43]. In particular, Ma et al. [20] propose a multi-layer Bi-RNN [27] model and try to enhance the ability to predict sentiment labels from the target extraction process. It is also the current state-of-the-art model of the joint learning task.

Target extraction has a strong relationship with sentiment classification. Previous works [16, 20] merely focus on how influential the target extraction task is to the sentiment prediction. However, sentiment classification process can also make a significant contribution to target extraction. Targets are usually modified by words inferring sentiments (called **sentiment clues**, i.e., *'**sad**'* in the aforementioned sentence). The sentiment classification process involves

*† Equal Contribution.

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3358024

finding sentiment clues and matching them to target words. So by finding which words are modified by sentiment clues, the model could extract targets more accurately. Thus it is necessary to find out the connection between the targets and sentiment clues, but previous joint models lack consideration of this. To interact between target and sentiment tagging information, we introduce attention [32], which is responsible for evaluating the correlation between targets and sentiment clues, and the obtained attention scores are utilized to obtain target tagging information with predicted sentiment clues. Furthermore, we feed target tagging information into sentiment tagging information via gate mechanism to share boundary information with sentiment tagging information. The two main parts both contribute to the construction of **interaction mechanism**, which is of great importance in our model.

Moreover, the model proposed by [20] manifests the effectiveness of a multi-layer structure. The state-of-the-art model leverages two fully-connected layers on top of word representations to obtain the probability distributions of target and sentiment labels, and feeds the distributions into CRF layer in the end. However, in this task, sentiment labels contain not only boundary information but also sentiment information, and it is hard to correctly project the word representations to both kinds of information, with a simple fully-connected layer. To strengthen the classification ability of our model, a **multi-grained structure** is applied to both target and sentiment label predictions to make better use of various kinds of word representations and features. Specifically, we propose a coarse-grained tagging layer which aims to distinguish *targets* and *sentiment clues*, i.e., for each word, we represent its coarse-grained tagging information as its probability of being part of a target and a sentiment clue respectively, which is passed subsequently to the interaction mechanism and further fused with the fine-grained layer. One of the intuitions of the coarse-grained tagging layer is that the binary classification results contain boundary information, which should be shared between both target and sentiment tagging results. Thus, the essential coarse-grained tagging information can give the model an approximate range of target and beneficial for both target and sentiment tagging results. Besides, sentiment clue information that the coarse-grained tagging layer provides is important for the interaction as described before. Another intuition is that the information brought from the coarse-grained tagging layer is a great supplement for a single tagging layer and the combination of multi-grained tagging information will be more robust. Therefore, based on the principle of multi-grained structure, finally the interacted tagging information (based on the coarse-grained tagging layer), is fused with fine-grained tagging information, which is used to decide the specific tag of a word.

Lastly, boundary consistency is critical to this task. The tagging results of target extraction and sentiment classification ought to share the same boundary information. For instance, in the aforementioned sentence, *'B-Negative O'* is wrong for ***'Whitney Houston'***, as the boundaries of target extraction and sentiment classification are contradictory. Thus, the target label's probability distribution and the sentiment label's probability distribution of whether a token is part of a target should necessarily be close. Thus, a specific loss function is designed on top of the final tagging information, aiming to keep the consistency. Besides, we hope that the roles of targets and sentiment clues could be separate since it is unlikely for a token to become part of a target and a sentiment clue simultaneously. Therefore we add another loss function to alleviate the overlap of two kinds of roles.

In summary, our contributions are as follows:

- Our model significantly strengthens the mutual interaction between target extraction and sentiment classification, and pays attention to how to more accurately detect targets and corresponding sentiments.
- We propose a multi-grained model to integrate tagging information and comprehensively consider the results acquired in different layers. Note that the coarse-grained tagging layer and interaction mechanism aim at explicitly sharing approximate boundary information between target and sentiment tagging results and assisting in building the connection between targets and sentiment clues.
- We design two specific loss functions to align the tagging results of two subtasks and distinguish the roles of target and sentiment clue.

2 RELATED WORK

2.1 Target Extraction

Target extraction is also called aspect term extraction [25]. Early methods concentrate on rule-based approaches, which are dependent on large numbers of handcrafted features and rules [19, 26]. The traditional machine learning method CRF [8, 9, 28] is usually applied to constrain the transition between tagging labels of *BIO* scheme. The models based on neural networks [17, 18, 34, 35, 37] are popular and competitive in this task. Recently, a CNN-based model [38] achieves state-of-the-art performance on target extraction task with domain embeddings.

2.2 Target-oriented Sentiment Classification

Target-oriented sentiment classification is also known as aspect-based sentiment classification. Traditional methods take advantage of sentiment lexicons to classify the sentiment [11, 42], while neural network models devote to constructing interaction between target and context words. [5, 21, 22, 30, 33, 36] leverage attention mechanism to evaluate the correlation between the tokens in one sentence. The current state-of-the-art model TNet [15] is based on a transformation network to strengthen the interaction between targets and contexts. There are a few works [7, 39] applying CNN, which is considered to be good at text classification. The models proposed by Xue and Li [39] and Huang and Carley [7] are both CNN-based models and adopt gate mechanism to make interaction between target and context tokens. Thus, inspired by the previous works of this task, we give deeper insight into the individual words by attention mechanism and intend to make better use of the correlation and interaction between targets and sentiment clues.

2.3 Joint Model for Targeted Sentiment Analysis

Currently, few works using joint learning methods for targeted sentiment analysis. The model proposed by Mitchell et al. [23] is based on the traditional methods, which rely on handcrafted features, such as sentiment lexicons. Meanwhile, they propose the benchmark datasets widely used in targeted sentiment analysis in

this paper. Zhang et al. [43] combine neural networks and handcrafted features in a neural and integrated way to achieve better performance. The joint model proposed by Li and Lu [14] intends to determine the sentiment scope for each word, and extract entities while predicting their sentiments within the scope. Li et al. [16] is a unified model which omits the target extraction subsystem and uses simple weight matrices to build transitions from target boundaries to target sentiment polarities. Note that the models proposed by Li and Lu [14] and Li et al. [16] are both dependent on sentiment resources. Ma et al. [20] design a hierarchical multi-layer Bi-GRU model and feed the information of extraction part into the classification part. To the best of our knowledge, there is no work taking a further step in building an information exchange process between target extraction and sentiment classification without any gold-annotated sentiment resources, and integrating multi-grained target and sentiment tagging information of two subsystems.

3 MODEL

3.1 Overview

The goal of our model is to jointly tag the word sequence with *B-{Person, Organization}, I-{Person, Organization}, O* for target extraction and *B-{Positive, Negative, Neutral}, I-{Positive, Negative, Neutral}, O* for sentiment classification. Formally, given a sentence $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n], \mathbf{x}_i \in \mathbb{R}^{d_w}$, we produce two tagging sequences: $y^{\mathcal{T}} = [y_1^{\mathcal{T}}, y_2^{\mathcal{T}}, ..., y_n^{\mathcal{T}}]$ for target extraction and $y^{\mathcal{S}} = [y_1^{\mathcal{S}}, y_2^{\mathcal{S}}, ..., y_n^{\mathcal{S}}]$ for target-based sentiment classification. The number of target extraction labels and sentiment classification labels are C_{tar} and C_{sen}.

The whole architecture of our model is illustrated in Figure 1. It has three main modules: (1) coarse-grained tagging layer (the left part); (2) interaction mechanism (the middle part); (3) fine-grained tagging layer (the right part).

The **coarse-grained tagging layer** intends to determine the basic role of words (whether a word belongs to a target and whether a word is a sentiment clue). According to Figure 1, the *coarse-grained* target tagging information and sentiment clue information are derived from binary classification on contextual word representations obtained by Bi-GRU, of which inputs are the concatenation of word-level and character-level embeddings.

Because there exists a strong connection between targets and sentiment clues as mentioned in Section 1, we design an **interaction mechanism** with attention to build the connection between targets and sentiment labeling parts. The details of our proposed interaction mechanism are shown in Section 3.3.

The **fine-grained tagging layer** attempts to decide the specific tag of a word (e.g. B-Person, I-Positive, O, etc) with deep neural networks. The *fine-grained* tagging information is based on the deep word representations, which are the hidden outputs of two stacked Bi-GRU. According to Figure 1, with the help of gate mechanism, the tagging information derived from interaction and the fine-grained layer are fused together. Then the outputs of our multi-grained model are obtained from CRF layers. The details of the multi-grained structure are shown in Section 3.2.

3.2 Multi-grained Structure

Coarse-grained Tagging Layer: Firstly, we extract semantic representations of words in the input text, which are useful for both target extraction and sentiment classification. Besides word embeddings, we also apply character-level CNN to capture the morphological features of a word. This is because there are many OOV words (e.g., proper nouns and abbreviations) in texts (especially informal texts like tweets), and they do not have pre-trained word embeddings. The character-based word representation $\mathbf{ch}_i$ for $\mathbf{x}_i$ is computed by a CNN with max-pooling layer and ReLU activation function where the window size is s_c and the number of filters is d_{ch}. The concatenation of word embedding and character-based word representation is subsequently fed into a Bi-GRU layer to build the context-based word representation $\mathbf{h}_i$ defined by the concatenation of forward and backward hidden outputs of the Bi-GRU, where $\mathbf{h}_i$ is of the length $d_h = 2(d_w + d_{ch})$.

Based on the word representations, we obtain *coarse-grained* target tagging information $\mathbf{z}^{\mathcal{T}} \in \mathbb{R}^2$, which means whether a word belongs to a target or not, and sentiment clue information $\mathbf{z}^{\mathcal{S}} \in \mathbb{R}^2$, which indicates whether the word is a sentiment clue:

$$\begin{aligned} \mathbf{z}_i^{\mathcal{T}} &= \text{Softmax}(W^z \times \mathbf{h}_i) \\ \mathbf{z}_i^{\mathcal{S}} &= \text{Softmax}(W^q \times \mathbf{h}_i) \end{aligned} \tag{1}$$

where $W^z, W^q \in \mathbb{R}^{2 \times d_h}$ are weight matrices. Specifically, we treat the first dimension of $\mathbf{z}_i^{\mathcal{T}}$ and $\mathbf{z}_i^{\mathcal{S}}$ as the probability inferring to a part of target and a sentiment clue, respectively. It is worth noting that $\mathbf{z}^{\mathcal{T}}$ is the basis of following target extraction and sentiment classification, for only words that are targets will be classified into different target types and sentiment types, and those that are not targets will simply be labeled "O" in the final results.

Interaction Mechanism: As mentioned before, the correlation between targets and their corresponding sentiment clues is strong. Thus, we design an interaction mechanism to build information exchange process between target and their corresponding sentiment clues. Sentiment clue tagging information is of great help to target extraction due to the observation that if A is a sentiment-bearing word, the word A modifies tends to be part of a target. So with the help of sentiment clues, we build *interacted* target tagging information. After that, *interacted* target tagging information is fused into sentiment tagging information to share boundary information, and then form *interacted* sentiment tagging information. The brief formulation is expressed as below:

$$\mathbf{l}^{\mathcal{T}}, \mathbf{l}^{\mathcal{S}} = \text{Interaction}(\mathbf{z}^{\mathcal{T}}, \mathbf{z}^{\mathcal{S}}) \tag{2}$$

where $\mathbf{l}^{\mathcal{T}} = [\mathbf{l}_1^{\mathcal{T}}, ..., \mathbf{l}_n^{\mathcal{T}}]$ and $\mathbf{l}^{\mathcal{S}} = [\mathbf{l}_1^{\mathcal{S}}, ..., \mathbf{l}_n^{\mathcal{S}}]$ are *interacted* target tagging information and *interacted* sentiment tagging information, respectively. The dimensions of *interacted* target and sentiment tagging information are C_{tar} and C_{sen}. The interaction process will be described in Section 3.3 in detail.

Fine-grained Tagging Layer: We adopt another Bi-GRU of which the inputs are the hidden outputs of the previous Bi-GRU layer for more careful and overall consideration. We treat the hidden representation $\mathbf{h}'_i$ of this Bi-GRU layer as the sum of the forward and backward hidden outputs to integrate the bidirectional information while preserving the scale of parameters. d'_h, which is the dimension of $\mathbf{h}'_i$, is equal to d_h. We produce the fine-grained results quite straightforwardly, and more complicated and effective methods can definitely be used to improve the result.

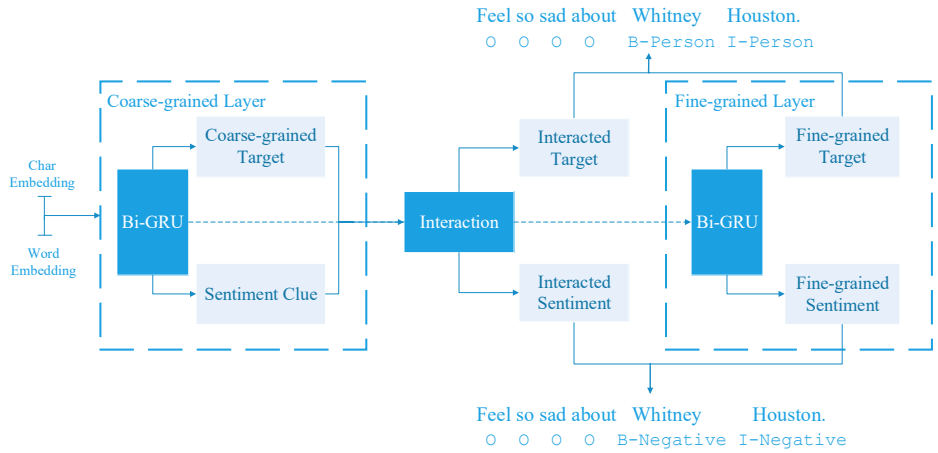


Figure 1: The overall architecture of our proposed model.

Table 2: Main variables used in the model. "TI" indicates tagging information.

Variable	Meaning
$\mathbf{z}_i^{\mathcal{T}}$	*Coarse-grained* Target TI
$\mathbf{z}_i^{\mathcal{S}}$	Sentiment Clue TI
$\mathbf{l}_i^{\mathcal{T}}$	*Interacted* Target TI
$\mathbf{l}_i^{\mathcal{S}}$	*Interacted* Sentiment TI
$\mathbf{f}_i^{\mathcal{T}}$	*Fine-grained* Target TI
$\mathbf{f}_i^{\mathcal{S}}$	*Fine-grained* Sentiment TI
$\mathbf{o}_i^{\mathcal{T}}$	*Multi-grained* Target TI
$\mathbf{o}_i^{\mathcal{S}}$	*Multi-grained* Sentiment TI

We perform transformation on top of this to get *fine-grained* target and sentiment tagging information $\mathbf{f}_i^{\mathcal{T}} \in \mathbb{R}^{C_{tar}}$ and $\mathbf{f}_i^{\mathcal{S}} \in \mathbb{R}^{C_{sen}}$:

$$\begin{aligned} \mathbf{f}_i^{\mathcal{T}} &= W^{ft} \times \mathbf{h}'_i \\ \mathbf{f}_i^{\mathcal{S}} &= W^{fs} \times \mathbf{h}'_i \end{aligned} \tag{3}$$

where $W^{ft} \in \mathbb{R}^{C_{tar} d'_h}$ and $W^{ft} \in \mathbb{R}^{C_{sen} d'_h}$ are the weight matrices.

Fusion: With the tagging information obtained from the interaction mechanism and the fine-grained layer, we fuse them to get more comprehensive predictions $\mathbf{o}_i^{\mathcal{T}}$ and $\mathbf{o}_i^{\mathcal{S}}$ for target and sentiment labeling respectively:

$$\begin{aligned} \mathbf{o}_i^{\mathcal{T}} &= \text{Gate}(\mathbf{l}_i^{\mathcal{T}}, \mathbf{f}_i^{\mathcal{T}}) \\ \mathbf{o}_i^{\mathcal{S}} &= \text{Gate}(\mathbf{l}_i^{\mathcal{S}}, \mathbf{f}_i^{\mathcal{S}}) \end{aligned} \tag{4}$$

where $\mathbf{o}_i^{\mathcal{T}}$ and $\mathbf{o}_i^{\mathcal{S}}$ are *multi-grained* target and sentiment tagging information respectively. The dimensions of $\mathbf{o}_i^{\mathcal{T}}$ and $\mathbf{o}_i^{\mathcal{S}}$ are C_{tar} and C_{sen}. Specifically, we treat the last dimension of $\mathbf{o}_i^{\mathcal{T}}$ and $\mathbf{o}_i^{\mathcal{S}}$ as the information inferring to being a non-target word and the others are related to different types of targets.

Gate$(\cdot,\cdot)$ in Equation 4 is gate mechanism. Suppose the inputs of gate mechanism are $\mathbf{a} \in \mathbb{R}^{d_a}$ and $\mathbf{b} \in \mathbb{R}^{d_b}$, the computation of all the gates used in our model is illustrated below:

$$\begin{aligned} \mathbf{g} &= \sigma[W^g \times (W^{trans} \times \mathbf{a})] \\ \text{Gate}(\mathbf{a}, \mathbf{b}) &= \mathbf{g} \odot (W^{trans} \times \mathbf{a}) + (1 - \mathbf{g}) \odot \mathbf{b} \end{aligned} \tag{5}$$

where $\odot$ is the symbol of element-wise product, $W^g \in \mathbb{R}^{d_b \times d_b}$ and $W^{trans} \in \mathbb{R}^{d_b \times d_a}$ are weight matrices. If the dimension of $\mathbf{a}$ equals to $\mathbf{b}$, we simply omit W^{trans} (in other words, $W^{trans} = I$).

In the end, we use CRF to model the dependencies between labels. Given the *multi-grained* target tagging information $\mathbf{o}^{\mathcal{T}} = [\mathbf{o}_1^{\mathcal{T}}, ..., \mathbf{o}_n^{\mathcal{T}}]$ and sentiment tagging information $\mathbf{o}^{\mathcal{S}} = [\mathbf{o}_1^{\mathcal{S}}, ..., \mathbf{o}_n^{\mathcal{S}}]$, the output target and sentiment tagging sequences are:

$$\begin{aligned} y^{\mathcal{T}} &= \text{CRF}_{TAR}(\mathbf{o}_1^{\mathcal{T}}, ..., \mathbf{o}_n^{\mathcal{T}}) \\ y^{\mathcal{S}} &= \text{CRF}_{SEN}(\mathbf{o}_1^{\mathcal{S}}, ..., \mathbf{o}_n^{\mathcal{S}}) \end{aligned} \tag{6}$$

where $y^{\mathcal{T}}$ and $y^{\mathcal{S}}$, computed by Viterbi Algorithm [3] , are considered as the sequences with maximal probabilities among all the possible tagging sequences of target extraction and sentiment classification, respectively.

3.3 Interaction Mechanism

Figure 2 depicts the interaction mechanism in detail. The motivation of designing interaction mechanism is two-fold: (1) a word is likely to be a target when modified by a sentiment clue; (2) it is necessary for sentiment tagging information to attend to target tagging information for they ought to share the same boundary information. Therefore, the information sharing process can also be divided into two parts: (1) sentiment clue information to target tagging information; (2) target tagging information to sentiment tagging information.

A. Sentiment clue information to target tagging information: For the first motivation, we use attention mechanism to evaluate the correlation between word representations:

$$\begin{aligned} U_{i,j} &= \tanh(\mathbf{h}_i^T \times W^{att} \times \mathbf{h}_j) \\ A_{i,j} &= \frac{\exp(U_{i,j})}{\sum_{k=1}^{n} \exp(U_{i,k})} \end{aligned} \tag{7}$$

where W^{att} is a $d_h d_h$ weight matrix, and A is nn attention score matrix. With the attention scores, we further transfer sentiment clue information $\mathbf{z}^{\mathcal{S}}$ and build the *sentiment-aware* target tagging information (**SAT**) $\mathbf{sa}_i^{\mathcal{T}} \in \mathbb{R}^2$:

$$\mathbf{sa}_i^{\mathcal{T}} = \sum_{j=1}^{n} A_{i,j} \cdot \mathbf{z}_j^{\mathcal{S}} \tag{8}$$

Compared with the *coarse-grained* target tagging information, SAT gets help from the prediction of sentiment clues. For instance, consider the two words *'**Houston**'* and *'**sad**'* in the sentence mentioned in Section 1. If *'**sad**'* is predicted as a sentiment clue and *'**Houston**'* is highly linked to *'**sad**'*, *'**Houston**'* has a high possibility of being a target.

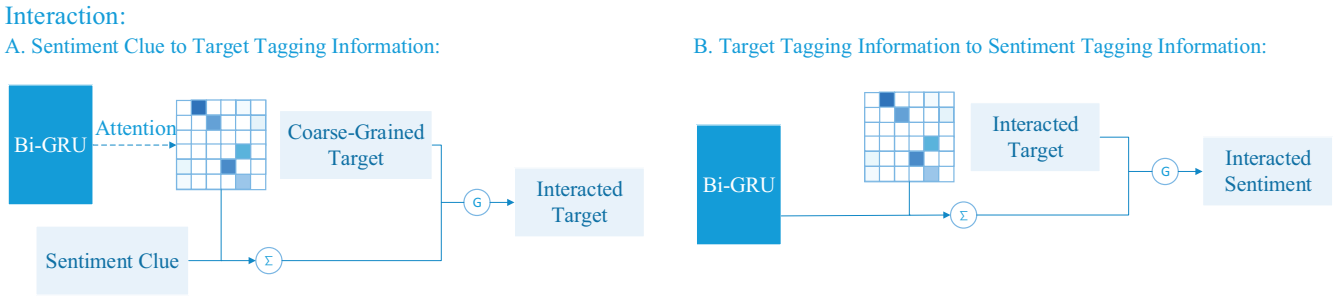


Figure 2: The interaction mechanism. The "Σ" denotes weighted sum according to attention scores, and "G" denotes the gate mechanism (Equation 5).

Then we use a gate mechanism to combine the *coarse-grained* target tagging information with SAT, and obtain *interacted* target tagging information. In order to fuse with *fine-grained* target tagging information, we transform the result with $W^l \in \mathbb{R}^{C_{tar} \times 2}$, so that the dimensions of *interacted* and *fine-grained* target tagging information are the same:

$$\mathbf{l}_i^{\mathcal{T}} = W^l \times \text{Gate}(\mathbf{z}_i^{\mathcal{T}}, \mathbf{sa}_i^{\mathcal{T}}) \tag{9}$$

B. Target tagging information to sentiment tagging information: Similarly, based on the attention score matrix, if a word w has a close correlation with a sentiment clue, the representation of the sentiment clue should take great proportion in the attention-based representation of w. The attention-based representation will help the model classify the sentiment of word w correctly.

Thus, for sentiment classification, the word representations are involved to build the attention-based sentiment tagging information by using attention score matrix A:

$$\begin{aligned} \mathbf{r}_i &= \sum_{j=1}^{n} A_{i,j} \cdot \mathbf{h}_j \\ \mathbf{att}_i^{\mathcal{S}} &= \text{ReLU}(W^r \times \mathbf{r}_i) \end{aligned} \tag{10}$$

where W^r is a $C_{sen} \times d_h$ transformation matrix and $\mathbf{att}_i^{\mathcal{S}} \in \mathbb{R}^{C_{sen}}$ is the attention-based sentiment tagging information. ReLU means activation function.

To share the boundary information between the two tagging subsystems, the *interacted* target tagging information is incorporated into the sentiment tagging information via gate:

$$\mathbf{l}_i^{\mathcal{S}} = \text{Gate}(\mathbf{l}_i^{\mathcal{T}}, \mathbf{att}_i^{\mathcal{S}}) \tag{11}$$

where $\mathbf{l}_i^{\mathcal{S}} \in \mathbb{R}^{C_{sen}}$ is called *interacted* sentiment tagging information.

In summary, the information of target and sentiment tagging parts is shared with each other. $\mathbf{l}^{\mathcal{T}}$ and $\mathbf{l}^{\mathcal{S}}$ are the outputs of our proposed interaction mechanism, and they will be combined with *fine-grained* tagging information to give a more accurate prediction.

3.4 Loss Functions

The loss function consists of four parts: two sequence tagging loss functions, an overlapping loss function (**OL**) and a boundary restriction loss function (**BRL**). The two loss functions of sequence labeling are derived from the CRF structure: the loss of target extraction $\mathcal{L}_{TAR}$ and the loss of sentiment classification $\mathcal{L}_{SEN}$ are computed from the probability of the ground truth label sequence over all possible label sequences.

The OL is applied to evaluate the similarity between the probability of belonging to a target and the probability of being a sentiment clue. In Section 3.2, the probability distributions $\mathbf{z}^{\mathcal{T}}$ and $\mathbf{z}^{\mathcal{S}}$, which represent whether a word belongs to a target and whether a word is a sentiment clue, are obtained in a classification form. However, there is an implicit constraint among the two distributions: a word is not likely to be identified as part of a target and a sentiment clue at the same time. This means that for each word, the product of probabilities of belonging to a target or being a sentiment clue should not be large. We choose the first dimension of two kinds of tagging information as the probabilities of being a target or a sentiment clue. The OL is defined as:

$$\mathcal{L}_{OL} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_{i,1}^{\mathcal{T}} \mathbf{z}_{i,1}^{\mathcal{S}} \tag{12}$$

The boundary restriction loss function (BRL) is set to restrict that the boundary information of two labeling subsystems are the same. Intuitively, the probability of belonging to a target in two labeling subsystems should be as similar as possible. In other words, the probability of not being a target should also be similar. Take $\tilde{\mathbf{o}}_i^{\mathcal{T}} = \text{Softmax}(\mathbf{o}_i^{\mathcal{T}})$ and $\tilde{\mathbf{o}}_i^{\mathcal{S}} = \text{Softmax}(\mathbf{o}_i^{\mathcal{S}})$ as the probability distributions of classification, we treat the sum of the first to $(C_{tar}-1)$-th dimensions of $\tilde{\mathbf{o}}_i^{\mathcal{T}}$ as the probability of being part of a target in the extraction part, and $\tilde{\mathbf{o}}_{i,C_{tar}}^{\mathcal{T}}$ as the probability of being labeled by *'O'*. Similarly, $\tilde{\mathbf{o}}_{i,C_{sen}}^{\mathcal{S}}$ indicates the probability of being labeled by *'O'* in sentiment tagging part. The BRL is calculated by:

$$\mathcal{L}_{BRL} = \frac{1}{n} \sum_{i=1}^{n} (\tilde{\mathbf{o}}_{i,C_{tar}}^{\mathcal{T}} - \tilde{\mathbf{o}}_{i,C_{sen}}^{\mathcal{S}})^2 \tag{13}$$

In addition, we set two hyperparameters β and γ to control the impact of OL and BRL. Finally, the overall loss is the sum of four separate loss functions:

$$\mathcal{L} = \mathcal{L}_{TAR} + \mathcal{L}_{SEN} + \beta \cdot \mathcal{L}_{OL} + \gamma \cdot \mathcal{L}_{BRL} \tag{14}$$

Table 3: The statistics of English and Spanish datasets. #Sen and #Tar denotes the number of sentences and targets. #Pos, #Neg and #Neu mean the quantities of three sentiment polarities. #Per and #Org represent the number of person targets and organization targets.

Dataset	#Sen	#Tar	#Pos	#Neg	#Neu	#Per	#Org
English	2350	3288	707	275	2306	1468	1820
Spanish	5145	6658	1555	1007	4096	3815	2843

4 EXPERIMENTS

4.1 Setup

Our proposed model is evaluated on the English and Spanish tweet datasets constructed by Mitchell et al. [23], and the statistics of two datasets are shown in Table 3. To show the effectiveness of our model, we use evaluation metrics of precision, recall and F1 score. Specifically, in the target extraction task, a correct target is extracted if and only if the boundary information and categories are both consistent with the gold annotated ones. Similarly, in the sentiment classification task, a correct sentiment polarity is predicted if and only if the boundary information and sentiments are the same with the gold annotated ones. We perform 10-fold cross-validation and report the average results as previous works do. For each fold, we randomly select 10% of the instances from the training set for development. The model with minimal loss on development set within 50 epochs is selected in each fold.

In our experiments, we use the pre-trained GloVe.840B.300d English word embeddings [24] and word embeddings pre-trained on Spanish tweets [2] as previous works do. The dimension sizes of English and Spanish word embeddings d_w are set to 300 and 200 respectively. According to the tagging scheme, $C_{tar} = 5$ and $C_{sen} = 7$. The batch size is set to 32. The dimension of the character embeddings d_c is 50. The embeddings of out-of-vocabulary words (OOV) and the attention weight matrix are randomly initialized from uniform distribution $\mathcal{U}(-0.1, 0.1)$. Xavier Initialization [4] is adopted to initialize other weight matrices and character embeddings. The kernel size of character-CNN is set to 3. The kernels consist of 50 filters. The hyperparameter β and γ used to tune the impact of the corresponding loss function to the overall loss are set to 1 and 0.7 respectively. Additionally, we adopt dropout to avoid overfitting and the rate is empirically set as 0.5. Adam optimizer [10], of which the learning rate is 0.001, is applied to optimize our model.

4.2 Comparison Results

First, we compare our model with pipeline methods to justify the effectiveness of the joint approach. Three pipeline solutions are described below:

• **CRF-Pipeline** [23] is based on vanilla CRF model, which first extracts targets and subsequently analyzes the sentiment polarities based on the extracted targets.

• **NN-Pipeline** [43] merely feeds word embeddings into neural networks and performs two tasks in a pipeline way.

• **DE-CNN-TNet** consists of the state-of-the-art target extraction [37] and aspect-based sentiment classification [15] models. For the reason that the datasets are for open domains, we do not introduce domain-specific embeddings the model uses.

The comparison results are shown in Table 4. We can see that our proposed model substantially outperforms all the baseline models on both datasets. Due to the good performance of individual **DE-CNN** and **TNet**, the pipeline outperforms a few joint models. However, notice that there exists a performance gap when comparing the best pipeline method with recently proposed models, suggesting that joint approaches could achieve much better performance on the task because of the shared information and avoiding error accumulation.

To further investigate the effectiveness of our model, we compare it with 11 joint models as follows. Except for **Bi-GRU**, **SS** and **E2E**, all the other joint models are equipped with CRF.

• **CRF** [23] is based on the traditional machine learning method CRF and utilizes handcrafted sentiment features.

• **NN-{Neural,Integrate}** [43] feed pre-trained word embeddings into neural networks. **NN-Integrate** additionally takes the handcrafted sentiment features into consideration.

• **Bi-GRU** applies a vanilla Bi-GRU network to represent a sentence and directly use a fully-connected layer to label the sequence.

• **Bi-GRU+CRF** is employed to learn the representation for sentences and introduces a CRF model to consider the influence of label dependencies.

• **MBi-GRU** [20] produces word representations by a multi-layer Bi-GRU structure instead of a single Bi-GRU layer to tag a sentence.

• **HBi-GRU** [20] uses Bi-GRU to learn character-level features for each word. Then, character-level features and word embeddings are concatenated as inputs for another Bi-GRU to learn final representations for sentence.

• **HMBi-GRU+No-Target** [20] is a hierarchical MBi-GRU structure to combine character and word representations. However, the sentiment classification module does not leverage information in the target extraction module.

• **HMBi-GRU+Target** [20] also uses HMBi-GRU. Tagging information is concatenated to word representation and treated as the input of the sentiment tagging layer to exert influence on the sentiment labeling. This is the current state-of-the-art model (SoA).

• **SS** [14] introduces a new concept called sentiment scope where targeted sentiments can be decided with usage of sentiment resources. We do not take the results of target extraction because the tagging scheme is different from that of previous works and ours.

• **E2E** [16] is a unified model which leverages handcrafted sentiment features and models the transitions from target boundaries to sentiment polarities with a transition matrix. As the tagging scheme of E2E contains only the boundary and sentiment information, we incorporate target category information (e.g. person) into the labels in order to compare our model with it.

Comparing vanilla **CRF** model and neural models, we can see that neural models achieve gains on both datasets, suggesting that neural networks perform well in this task for the powerful ability to capture features. Besides, it is noticeable that **NN-Integrated** substantially outperforms **NN-Neural** method. This strongly justifies the effectiveness of sentiment features and inspires us to mine sentiment clues automatically by making better use of neural networks.

From the fact that vanilla **Bi-GRU** achieves better performance than **NN-Neural**, we could infer that the contextual information

Table 4: The comparison of pipeline, joint baseline methods and our model. * indicates the results are reproduced by ourselves. The best results are in bold.

Models	English						Spanish					
	Target Extraction			Sentiment Classification			Target Extraction			Sentiment Classification		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
CRF-Pipeline*	50.58	40.77	45.15	36.67	30.03	33.02	67.56	42.84	52.43	37.13	28.77	32.42
NN-Pipeline*	61.61	45.13	52.10	40.16	39.38	39.77	60.52	52.90	56.45	45.35	37.22	40.88
DE-CNN-TNet*	58.34	54.71	56.47	39.73	40.21	39.97	66.39	61.47	63.84	45.12	43.45	44.27
CRF*	59.55	34.06	43.33	43.09	24.67	31.38	64.17	47.03	54.28	36.92	31.45	33.97
NN-Neural	54.45	42.12	47.17	37.55	28.95	32.45	65.05	47.79	55.07	40.28	29.58	34.09
NN-Integrated	61.47	49.28	54.70	44.62	35.84	39.67	**71.32**	61.11	65.82	46.67	39.99	43.02
Bi-GRU*	58.13	43.46	49.73	45.76	32.29	37.86	65.24	53.02	58.50	46.33	37.50	41.45
Bi-GRU+CRF*	59.67	47.19	52.73	40.11	39.47	39.79	61.84	59.61	60.70	43.48	41.36	42.39
MBi-GRU	58.27	49.01	53.24	45.80	35.21	39.81	66.14	60.07	62.95	45.61	40.04	42.64
HBi-GRU	57.24	53.88	55.41	44.94	38.60	41.52	68.24	61.81	64.82	46.53	42.21	44.18
HMBi-GRU+No-Target	61.24	52.44	56.39	45.90	39.21	42.21	66.72	63.57	65.10	45.06	43.31	44.17
HMBi-GRU+Target (SoA)	60.12	53.68	56.98	46.52	39.99	42.87	68.64	63.66	66.01	48.09	43.44	45.61
SS	-	-	-	44.57	36.48	40.11	-	-	-	46.06	39.89	42.75
E2E*	64.86	60.54	62.63	44.83	35.60	39.69	64.01	52.79	57.86	**51.08**	32.57	39.78
Ours	**67.49**	**64.42**	**65.92**	**48.64**	**46.22**	**47.40**	70.55	**67.29**	**68.88**	50.16	**45.82**	**47.89**

Table 5: The comparison of our model and ablation tests. The best results are in bold.

Models	English						Spanish					
	Target Extraction			Sentiment Classification			Target Extraction			Sentiment Classification		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
Ours	**67.49**	**64.42**	**65.92**	48.64	**46.22**	**47.40**	**70.55**	**67.29**	**68.88**	**50.16**	**45.82**	**47.89**
w/o Fine-grained	45.88	47.74	46.80	37.31	32.52	34.75	57.98	56.89	57.43	42.15	37.73	39.82
w/o Coarse-grained + Interaction	57.12	53.24	55.11	44.58	39.01	41.61	67.45	62.31	64.78	45.69	41.77	43.64
w/o Interaction	62.33	57.14	59.62	46.21	41.16	43.54	64.58	62.79	63.67	46.35	42.22	44.19
w/o Interaction Part B	65.01	62.50	63.73	**49.47**	42.95	45.98	66.19	66.49	66.34	45.80	44.50	45.14
w/o Interaction Part A	66.75	62.81	64.72	48.90	44.54	46.62	68.78	65.40	67.05	49.10	43.57	46.17

makes a great contribution to the sequence labeling task. The importance of CRF can also be observed from the apparent difference between **Bi-GRU** and **Bi-GRU+CRF**. Inspired by the improvement of **MBi-GRU** over a single Bi-GRU layer, we adopt a multi-layer structure and develop a multi-grained structure based on it. Additionally, the improvement of hierarchical network **HBi-GRU** and **HMBi-GRU** over **MBi-GRU** demonstrates the usefulness of character-level features. We also find that the **HMBi-GRU+Target** model with consideration of the influence of target extraction information to sentiment classification gives superior performance compared to **HMBi-GRU+No Target**. The experimental results indicate that the interaction between target extraction and sentiment classification is of importance and indispensable. They also remind us of further considering a more complex interaction mechanism for this task.

At last, our model substantially outperforms **SS** and **E2E** which depend on gold-annotated sentiment features. This further shows the robustness and effectiveness of our model.

4.3 Ablation Study

To evaluate the effect of each part in our model, we remove some important components and compare ours with those ablated versions. Because our model consists of three main modules: coarse-grained tagging layer, interaction mechanism and fine-grained tagging layer, we attempt to remove each module and evaluate the effectiveness of them.

Our model vs. w/o Fine-grained: After removing the fine-grained tagging layer, we find the performance drops substantially on two datasets. It is partly because the *coarse-grained* and *interacted* tagging information is based on the simple binary classification, which is not related to specific label types. Moreover, the contextual representations obtained from coarse-grained tagging layer are not so informative as the deep representations in fine-grained tagging layer.

Our model vs. w/o Coarse-grained + Interaction: Since the interaction mechanism relies on the results of the coarse-grained tagging layer, we remove both of them simultaneously in the ablation study. Without consideration of coarse-grained tagging layer and interaction mechanism, it performs worse than our proposed model. This ablated model is similar to **HBi-GRU** structure. For this task, the interplay between target and sentiment tagging parts is of great importance. In our proposed multi-grained structure, the coarse-grained tagging layer and interaction mechanism play an

Table 6: The comparison of our model and the versions without the specific loss functions. The best results are in bold.

Models	English						Spanish					
	Target Extraction			Sentiment Classification			Target Extraction			Sentiment Classification		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
Ours	67.49	**64.42**	**65.92**	**48.64**	46.22	**47.40**	**70.55**	**67.29**	**68.88**	**50.16**	45.82	**47.89**
w/o OL	**70.09**	58.95	64.04	47.04	46.46	46.75	70.01	65.33	67.59	46.41	46.85	46.63
w/o BRL	67.49	62.11	64.69	46.09	**47.59**	46.83	69.30	66.73	67.99	46.15	**47.63**	46.88
w/o OL + BRL	66.81	61.44	64.01	45.41	46.92	46.15	68.80	66.24	67.49	46.43	46.65	46.54

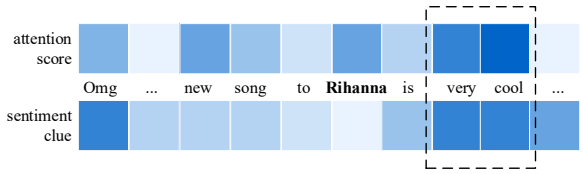


Figure 3: Each word's attention score when correlated with "Rihanna" and probability of being a sentiment clue.

important role in building a connection between target and sentiment tagging parts, with the help of word dependencies captured by attention mechanism and the detection of sentiment clues. Despite the coarse-grained tagging layer and interaction mechanism alone do not give great results, the *interacted* tagging information produced by them contains interaction information between two parts which fine-grained layer alone does not have. Thus it becomes a great supplement for better performance.

Our model vs. w/o Interaction: We also conduct experiments to investigate the availability of the interaction mechanism. First, we remove the interaction mechanism to evaluate the effectiveness of coarse-grained and fine-grained layers. We can observe that the performance declines without interaction mechanism. Moreover, compared with **w/o Coarse-grained + Interaction**, the performance is much better, suggesting that the coarse-grained layer is crucial to the task, and the integration of coarse-grained and fine-grained layers is robust enough for good performance as well.

Our model vs. w/o Interaction Part A or Part B: As shown in Figure 2, the interaction mechanism can be separated into two parts. Part A mainly focuses on transferring sentiment clue information to target tagging information; Part B aims at fusing target tagging information with sentiment tagging information. From the experimental results, the performance downgrades with the removal of Part B, with a drop of F1 score 2.19%, 1.42% on English dataset and 2.54% and 2.75% on Spanish dataset. Because the boundary information of target and sentiment tagging results need to be consistent and the previously obtained sentiment tagging information does not contain such information, it is necessary to design a gate to convey boundary information from target tagging part to sentiment tagging part. Additionally, our model substantially outperforms the model without Part A of the interaction mechanism. The performance gap results from the sentiment-aware target tagging information (SAT) described in Section 3.3. Based on the detected sentiment clues and the observation that the words modified by sentiment clues are more likely to be inferred as targets, SAT is another kind of target tagging information in the view of sentiment classification and it is proven to be effective according to the results.

We choose an example to further explain the role of SAT. The attention scores which evaluate the correlation with the target word *"Rihanna"* and each word's corresponding probability of being a sentiment clue are shown in Figure 3. We notice that the attention mechanism pays more attention to the words *"very"* and *"cool"* and the two words have a high probability of being sentiment clues. Therefore, according to the calculation of Eq. 8, SAT is capable of giving an accurate target tagging result. It also justifies the effectiveness of the sentiment clues and attention mechanism, which can correctly capture the relationship between targets and sentiment clues.

4.4 Effect of Specific Loss Functions

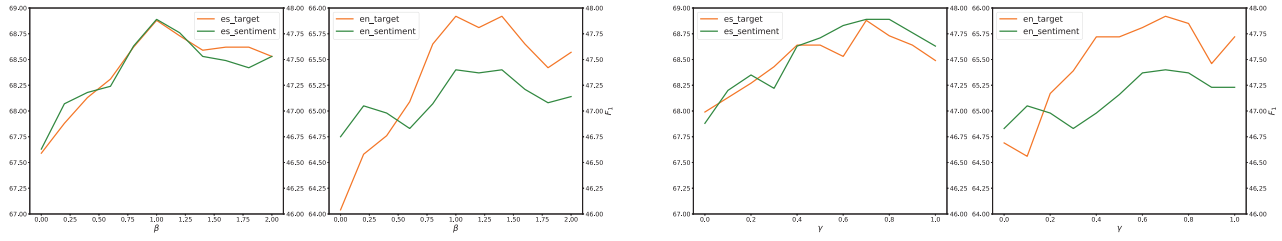
In this part, we explore the effect of the two specific loss functions proposed in Section 3.4, the overlapping loss function (**OL**) and the boundary restriction loss function (**BRL**). Also, by tuning hyperparameters β and γ which control the influence of them, we intend to investigate whether the two loss functions have positive effect.

First, we set β and γ to 0, respectively. As shown in Table 6, the performance declines when either is moved in the overall loss. Moreover, by removing both of the loss functions, the performance is even worse than just removing one of the loss functions. It further proves the necessity of the two functions.

Then, we attempt to investigate the influence of β and γ to see if they are capable of distinguishing the roles of target and sentiment clue, and keeping the boundary information consistent with each other. We conduct experiments with γ fixed to 0.7, vary β from 0 to 2 with a step of 0.2, and find that the performance is always better than all the baseline methods, suggesting that our approach is effective and stable.[1] Figure 4(a) shows a rising trend before β reaches 1 and a downtrend after it.

To be convincing, we choose one sentence and see the value of each word's probability of belonging to a target and being a sentiment clue. The probability of belonging to a target is obtained from the first dimension of *coarse-grained* target tagging information $\mathbf{z}^{\mathcal{T}}$. In the first case, as visualized in Figure 5, we notice that when β=0, both probability distributions are uniform and cannot clearly show which word is more likely to be counted as a certain category. Though the model can increase the gap between two kinds of probabilities and locate the targets and sentiment clues more accurately when β=2, the probability value is relatively low, probably because

[1]The hyper-parameters are tuned on the development set, while we show the results of cross-validation in this figure.



(a) Effect of β. (b) Effect of γ.

Figure 4: Influence of adjusting the hyper-parameters. "en" and "es" indicate English and Spanish datasets.

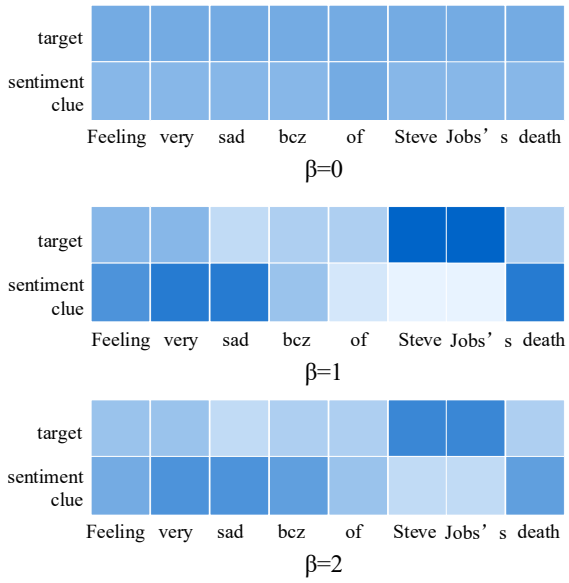


Figure 5: Case study of the influence of β. The deeper the blue is, the bigger the probabilities are.

the larger β tends to shrink the probability value to achieve a lower loss. The lower value may negligibly affect the labeling process and even introduce noise into the model. The sentence *"**Cmon Erra**! You know you are strong. So please stay strong!"* shows that this problem may lead to a wrong result. When β=2, although compared with the other words in the second case, *"**Erra**"*'s probability of belonging to a target is relatively big, the absolute value is small so that *"**Erra**"* is wrongly recognized as a non-target word. When β=1, the tagging results are more evident, especially for sentiment tagging, which is relatively hard since we do not use any gold-annotated sentiment clues, unlike **SS** and **E2E**. Though without supervision, our model precisely locates the sentiment clues such as *"**feeling very sad**"*, and also shows great performance of determining the approximate range of targets.

Similarly, to explore the influence of γ, we fix β to 1, and vary γ from 0 to 1 with a step of 0.1. According to Figure 4(b), we observe that our model still outperforms all the baseline methods. It proves

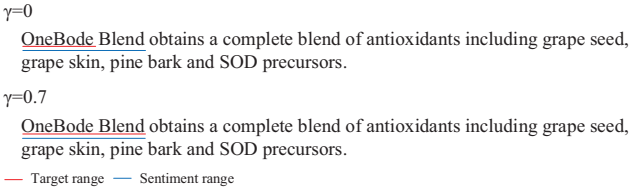


Figure 6: Case study of the influence from γ.

the robustness of our model across different values of γ. The best performance among the experiments is achieved when γ=0.7, and when γ=0, the model performs the worst, indicating the BRL is beneficial for our model. We also pick up an example and demonstrate the results in Figure 6. When setting γ=0, we find that the boundary information of two types of tagging results is not the same while it becomes consistent when γ equals 0.7. This further justifies the boundary restriction ability of BRL.

5 CONCLUSION

We propose a novel interactive multi-grained model to jointly extract targets and predict sentiment. Compared to previous works, our model builds deep interaction between targets and sentiment clues by attention and gate mechanisms. Besides, our model takes advantage of multi-layer structure and combines the information from multi-grained tagging information to take an overall consideration. Additionally, we design two specific loss functions to finely tune the labeling results. Experimental results show that our model substantially outperforms previous methods.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (61772036) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

REFERENCES

[1] Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent Attention Network on Memory for Aspect Sentiment Analysis. In *Proceedings of the*

2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 452–461.

[2] Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *5th International Workshop on Natural Language Processing for Social Media, Boston, MA, USA, December 11, 2017*. Association for Computational Linguistics, 45–51.

[3] G David Forney. 1973. The viterbi algorithm. *Proc. IEEE* 61, 3 (1973), 268–278.

[4] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 249–256.

[5] Devamanyu Hazarika, Soujanya Poria, Prateek Vij, Gangeshwar Krishnamurthy, Erik Cambria, and Roger Zimmermann. 2018. Modeling Inter-Aspect Dependencies for Aspect-Based Sentiment Analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 266–270.

[6] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 168–177.

[7] Binxuan Huang and Kathleen Carley. 2018. Parameterized Convolutional Neural Networks for Aspect Level Sentiment Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1091–1096.

[8] Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1035–1045.

[9] Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. A novel lexicalized HMM-based learning framework for web opinion mining. In *Proceedings of the 26th annual international conference on machine learning*. Citeseer, 465–472.

[10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[11] Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. 437–442.

[12] Roman Klinger and Philipp Cimiano. 2013. Bi-directional inter-dependencies of subjective expressions and targets and their value for a joint model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 848–854.

[13] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[14] Hao Li and Wei Lu. 2017. Learning latent sentiment scopes for entity-level sentiment analysis. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[15] Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation Networks for Target-Oriented Sentiment Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 946–956.

[16] Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2018. A Unified Model for Opinion Target Extraction and Target Sentiment Prediction. *arXiv preprint arXiv:1811.05082* (2018).

[17] Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect Term Extraction with History Attention and Selective Transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. 4194–4200.

[18] Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1433–1443.

[19] Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated Rule Selection for Aspect Extraction in Opinion Mining.

[20] Dehong Ma, Sujian Li, and Houfeng Wang. 2018. Joint Learning for Targeted Sentiment Analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4737–4742.

[21] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive Attention Networks for Aspect-Level Sentiment Classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 4068–4074.

[22] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, Md Shad Akhtar, Erik Cambria, and Asif Ekbal. 2018. IARM: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3402–3411.

[23] Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open Domain Targeted Sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. 1643–1654.

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[25] Maria Pontiki, John Galanis, Dimitris Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval-2014)*. 19–30.

[26] Ana-Maria Popescu and Orena Etzioni. 2007. Extracting product features and opinions from reviews. In *Natural language processing and text mining*. Springer, 9–28.

[27] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[28] Lei Shu, Hu Xu, and Bing Liu. 2017. Lifelong learning crf for supervised aspect extraction. *arXiv preprint arXiv:1705.00251* (2017).

[29] Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect Level Sentiment Classification with Deep Memory Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 214–224.

[30] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2017. Dyadic Memory Networks for Aspect-based Sentiment Analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*. 107–116.

[31] Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 142–147.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5998–6008.

[33] Shuai Wang, Sahisnu Mazumder, Bing Liu, Mianwei Zhou, and Yi Chang. 2018. Target-sensitive memory networks for aspect sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 957–967.

[34] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. *arXiv preprint arXiv:1603.06679* (2016).

[35] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled Multi-Layer Attentions for Co-Extraction of Aspect and Opinion Terms.. In *Proceedings of The Thirty-First AAAI Conference on Artificial Intelligence*. 3316–3322.

[36] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 606–615.

[37] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*. 592–598.

[38] Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2018. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. *arXiv preprint arXiv:1805.04601* (2018).

[39] Wei Xue and Tao Li. 2018. Aspect Based Sentiment Analysis with Gated Convolutional Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2514–2523.

[40] Bishan Yang and Claire Cardie. 2013. Joint Inference for Fine-grained Opinion Extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*. 1640–1649.

[41] Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction. *arXiv preprint arXiv:1605.07843* (2016).

[42] Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. 2011. Aspect ranking: identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 1496–1505.

[43] Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural Networks for Open Domain Targeted Sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 612–621.