A Soft Label Strategy for Target-Level Sentiment Classification

Da Yin, Xiao Liu, Xiuyu Wu, Baobao Chang
MOE Key Lab of Computational Linguistics, School of EECS, Peking University
{wade_yin9712,lxlisa,xiuyu_wu,chbb}@pku.edu.cn

Abstract

In this paper, we propose a soft label approach to target-level sentiment classification task, in which a history-based soft labeling model is proposed to measure the possibility of a context word as an opinion word. We also apply a convolution layer to extract local active features, and introduce positional weights to take relative distance information into consideration. In addition, we obtain more informative target representation by training with context tokens together to make deeper interaction between target and context tokens. We conduct experiments on SemEval 2014 datasets and the experimental results show that our approach significantly outperforms previous models and gives state-of-the-art results on these datasets.

1 Introduction

Target-level sentiment classification aims to identify the sentiment polarities towards given targets by analyzing sentence context. For example, in the sentence *"The **food** is good but **service** is bad."*, there are two targets *"**food**"* and *"**service**"* mentioned. The sentiment towards *"**food**"* and *"**service**"* are positive and negative respectively.

Neural network models (Tang et al., 2016a; Wang et al., 2016; Tang et al., 2016b; Liu and Zhang, 2017; Ma et al., 2017; Tay et al., 2017; Chen et al., 2017; Huang et al., 2018; Gu et al., 2018) have achieved high accuracy on this task. Most of the neural network models introduce attention mechanism to find the correlation between target and context tokens. However, the combination of word-level features computed by attention weights may introduce noise into model. For instance, in *"The dish tastes bad but its **vegetable** is delicious though it looks ugly."*, these attention-based models tend to highlight some involve some other words such as *"bad"* and *"ugly"*.

Instead of using the attention mechanism, we propose a soft label approach for the target-level sentiment classification task. Intuitively, the task could be treated as a two-step process. Firstly the sentiment words that are related to the given target, called opinion words, are labeled and extracted. Then the final decision on the sentiment polarity would be made by taking all the extracted opinion words into account. However, this kind of hard label strategy, which directly determines whether a token is an opinion word or not, for labeling opinion words is non-differentiable and hinders training through normal back-propagation. Thus we use a soft labeling model to avoid the hard decision and make sure the model works in an end-to-end way.

Specifically, the soft label model is used to measure the likelihood of a context word as an opinion word at each time step. The larger the value of one word's soft label, the greater its effect on target sentiment. In fact, given a target, people are accustomed to going through a sentence from beginning to end, and to judge whether current word is highly related to the target sentiment at each step with comparison of history information till the current word in the reading process. Therefore, we implement an LSTM-based (Hochreiter and Schmidhuber, 1997) soft labeling model by a history-based approach, which utilizes history information (previous soft labels and cell states) together with representation of the current word, to decide how to pay attention to history information or current word representation based on their correlation with target representation.

Moreover, since the convolution layer (LeCun et al., 1989) does better in capturing local active features than other neural networks do and these extracted features are proved to be beneficial to text classification (Kim, 2014; Johnson and Zhang, 2015), we apply a convolution based encoder to extract these features. The distance of the features to target is also essential as texts may be long and contain several targets. The closer tokens

Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 6–15
Minneapolis, June 6, 2019.

are more likely to affect on the targets. Therefore, we adopt positional weights to scale the features with relative distance information between context tokens and the target.

Target representation is also critical to this task. Previous works, such as Tang et al. (2016a), simply take the average of target embeddings as target representation. In fact, this kind of representation does not incorporate contextual information. Words in a sentence have strong dependencies on each other. Thus it is necessary to train target representation together with context tokens to obtain more informative representation dependent on contextual information.

In summary, our contributions are as follows:

- Our model uses a soft label approach to evaluating the likelihood of a context word as an opinion word based on the history information.
- Our model leverages convolution layer, which is seldom used in the task, to extract features, and these features are accordingly weighed by positional information.
- Our model learns more informative representation of the target, instead of the average of target embeddings, and strengthens the interaction between target and context tokens in soft label computation process.
- We conduct experiments on benchmark datasets and the experimental results show that our approach significantly outperforms previous models and achieves state-of-the-art results on these datasets.

2 Related Work

Early methods mainly apply supervised learning approach with large quantities of handcrafted features (Blair-Goldensohn et al., 2008; Yu et al., 2011; Jiang et al., 2011; Kiritchenko et al., 2014), but ignore context information and deep relations between target and context tokens.

Neural network models have achieved high accuracy on this task. **AE-LSTM** and **ATAE-LSTM** (Wang et al., 2016) simply concatenate target embeddings to context word embeddings to make connection between targets and contexts. However, both models described above do not obtain target representations based on context-aware information. Inspired by the TNet (Li et al., 2018), which learns deep representations for targets, we propose a model which could strengthen the interaction between target and context tokens.

Recently, most of the previous state-of-the-art models leverage attention mechanism to evaluate the correlation between the tokens in one sentence. **IAN** (Ma et al., 2017) adopts two separate LSTM layers and an interactive attention mechanism. Hazarika et al. (2018) classifies the sentiment polarities of all the targets in one sentence simultaneously with attention mechanism to model inter-target dependencies. **MemNet** (Tang et al., 2016b), **RAM** (Chen et al., 2017), **TRMN** (Wang et al., 2018) and **IARM** (Majumder et al., 2018) introduce deep memory network and multi-hop attention model over sentence-level memories to incorporate target information into sentence representations. Specifically, **TRMN** and **IARM** attach importance to the interaction between targets and contexts, and inter-target relations, which contain the information of relationship between multiple targets in one sentence, respectively. Different from them, our model adopts a novel and effective soft label approach in an intuitive way.

There are few works (Xue and Li, 2018; Huang and Carley, 2018) applying CNN, which is considered to be good at text classification, on target-level sentiment classification. **GCAE** (Xue and Li, 2018) and **PG-CNN** (Huang and Carley, 2018) are both CNN-based models and adopt gate mechanism to make interaction between target and context tokens. To improve the effectiveness of convolution layers, our model further adopts positional weights, which take relative distance information into account.

3 Model

Target-level sentiment classification task is to decide which sentiment is expressed towards a target: positive, neutral or negative.

Our model is illustrated in Figure 1. It is divided into four parts: (1) a Bi-LSTM (Schuster and Paliwal, 1997) layer to get context-aware representations, (2) a convolution based feature extractor, (3) computation of soft labels, and (4) sentiment classification using the soft labels and positional weights.

We introduce the following notations: $\boldsymbol{s} = [w_1, w_2, ..., w_n]$ denotes a sentence which consists of n words. $w_i \in \mathbb{R}^{d_0}$ is the embedding of the i-th word. $\boldsymbol{t} = [t, t+1, ..., t+m-1]$ denotes the posi-

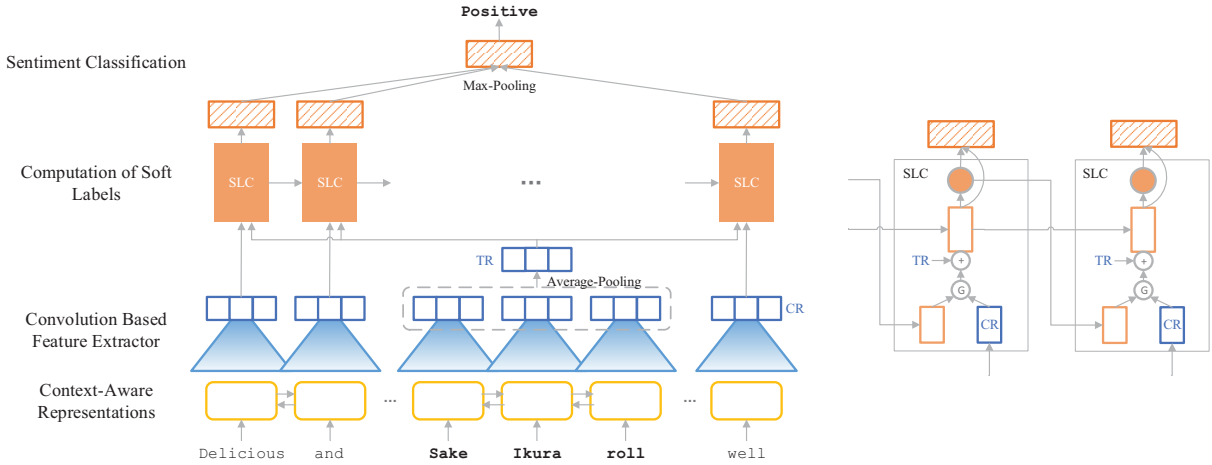

Figure 1: Overall architecture of the proposed method. We use the sentence *"Delicious and good-looking Sake Ikura roll, and sashimi tastes good as well."* as an example. The term "SLC" indicates soft label computation. "TR" indicates target representation and "CR" represents context representation.

tion of the target tokens, where $t \geq 1, t+m-1 \leq n$. The length of the target is m.

3.1 Context-Aware Representations

Since words in a sentence have strong dependencies on each other, it is necessary to fetch context-aware representations to combine context information with words. In order to incorporate the context information into words, we encode them with a Bi-LSTM layer:

$$x_i = [\overrightarrow{\text{LSTM}}(w_i); \overleftarrow{\text{LSTM}}(w_i)] \quad (1)$$

We concatenate the forward and backward hidden outputs of LSTM, of which the dimension size is both d'_0, and $[;]$ denotes concatenation. We regard $x_i \in \mathbb{R}^{2d'_0}$ as the context-aware representation of word w_i, and feed it to following layers.

3.2 Convolution Based Feature Extractor

To extract the local active features, we use a convolution layer with three parallel windows, which have different sizes. Each kernel has d_1 filters. For kernel size s_j, let $W^{conv_j} \in \mathbb{R}^{d_1 \times s_j \times 2d'_0}$ be the d_1 filters for the convolution with the same size s_j, and $b^{conv_j} \in \mathbb{R}^{d_1}$ be the bias. x^{conv_j}, the output of the convolution layer is produced by convoluting W^{conv_j} with the word window $x_{i-\lfloor \frac{s_j-1}{2} \rfloor : i+\lfloor \frac{s_j}{2} \rfloor}$ at each $i \in [1, n]$ (positions out of range are padded with zero):

$$x_i^{conv_j} = \text{ReLU}(x_{i-\lfloor \frac{s_j-1}{2} \rfloor : i+\lfloor \frac{s_j}{2} \rfloor} \circ W^{conv_j} + b^{conv_j}) \quad (2)$$

where ReLU indicates a nonlinear activation function, and $\circ$ is element-wise multiplication.

Merging outputs of three kinds of kernels, the word representation is computed as:

$$h_i^E = x_i^{conv_1} \oplus x_i^{conv_2} \oplus x_i^{conv_3} \quad (3)$$

where $\oplus$ is concatenation. The dimension of h_i^E is $d'_1 = 3d_1$.

h^{target} is computed by an average-pooling layer to refine the target representation:

$$h^{target} = \frac{1}{m} \sum_{i=1}^{m} h_{t+i-1}^E \quad (4)$$

3.3 Computation of Soft Labels

Instead of using a hard label strategy and labeling explicitly context words as opinion words or not, we adopt a soft labeling model in which soft label is defined as the probability of each context word as an opinion word. An LSTM layer is applied to compute the final word representation h_i^D and the soft label l_i for the i-th word. It takes both the interacted representation produced by the convolution based feature extractor and the soft label of the previous time step as the input, in order to take history contexts into consideration:

$$h_i^D, c_i^D = \text{LSTM}(h_{i-1}^D, c_{i-1}^D, u_i) \quad (5)$$

where $h_i^D \in \mathbb{R}^{d'_1}$ is the output of the i time stamp, $c_i^D \in \mathbb{R}^{d'_1}$ is the LSTM cell state, which could be treated as long-term memory till the i-th word, and u_i is the input which will be described later.

One problem encountered here is that the history information of previous time steps may not

be closely related to the target. Consider predicting the sentiment of the target "***service***" in the sentence *"Tasty food but the **service** was dreadful!"*. When the LSTM comes to the word "*dreadful*", a simple soft label approach might indicate that the sentiment polarity is positive due to the influence of the word "*Tasty*", which in fact does not modify the target "***service***". To solve the problem, we apply a gate mechanism to determine the proportion of the history information in the input, according to the ratio of history information and current word information's correlation with the target:

$$g_i = \frac{\exp(c_{i-1}^D W^g h^{target})}{\exp(c_{i-1}^D W^g h^{target}) + \exp(h_i^E W^g h^{target})} \quad (6)$$

where $W^g \in \mathbb{R}^{d_1' \times d_1'}$ is the weight matrix. Also, we intend to strengthen the influence from target representation. Thus we further incorporate target information into the input:

$$u_i = g_i \cdot (W^D l_{i-1}) + (1 - g_i) \cdot h_i^E + h^{target} \quad (7)$$

where $W^D \in \mathbb{R}^{d_1'}$ is the weight parameter and l_{i-1} is the soft label of the $(i-1)$-th word. To reduce the dimensions of LSTM inputs, we fuse the target representation with word representations by a simple addition operation.

With the output of the LSTM layer, the soft label l_i is computed as:

$$\begin{aligned} l_i &= p(e_i = 1 | h_i^D) \\ &= p(e_i = 1 | l_1, l_2, ..., l_{i-1}, h_{i-1}^D, h_i^E) \\ &= \text{sigmoid}(W^l h_i^D + b^l) \end{aligned} \quad (8)$$

where $e_i = 1$ indicates that the word should be considered as bearing sentiment towards the current target, $W^l \in \mathbb{R}^{d_1'}$ and $b^l \in \mathbb{R}$.

3.4 Sentiment Classification

Features that are close to the target often contribute more to the sentiment towards the target. Considering the impact of the distance to the target, we define the positional weights:

$$pos_i = \begin{cases} 1 - \dfrac{t - i}{\beta} & i \in [1, ..., t-1] \\ 1 - \dfrac{i - t + 1}{\beta} & i \in [t, ..., n-m] \end{cases} \quad (9)$$

where β controls the rate of decaying of the positional weights according to the distances to the target. The value of the rate is $\frac{1}{\beta}$.

Algorithm 1 Training framework of our model.

Input: Sentence $\boldsymbol{w}$, target $\boldsymbol{t}$, golden label $\boldsymbol{y}$.

$h^E, h^{target} = ComputeRepresentation(\boldsymbol{w}, \boldsymbol{t})$
for word w_i in sentence $\boldsymbol{w}$ **do**
  if $i == 1$ **then**
    $g_i = 0$
  else
    $g_i = ComputeGate(h_i^E, c_{i-1}^D, h^{target})$ (Eq.6)
  end if
  $u_i = ComputeInput(g_i, l_{i-1}, h_i^E, h^{target})$ (Eq.7)
  $h_i^D, c_i^D = \text{LSTM}(h_{i-1}^D, c_{i-1}^D, u_i)$
  $l_i = ComputeSoftLabel(h_i^D)$ (Eq.8)
end for
$\boldsymbol{p} = Predict(l, h^D, pos)$
$L = CrossEntropy(\boldsymbol{p}, \boldsymbol{y})$
Back propagate errors and update parameters θ

Then we combine the soft labels and positional weights together to take both the history contexts and the relative distances into consideration. The integrated weight of the i-th word is:

$$c_i = l_i \cdot pos_i \quad (10)$$

We put the word representations together to predict the sentiment towards the target, according to the integrated weight of each word:

$$p(\tilde{y} | \boldsymbol{w}, \boldsymbol{t}) = \text{softmax}(W^p \text{max}\{c_i \cdot h_i^D\} + b^p) \quad (11)$$

where $\tilde{y}$ is the three categories of sentiment polarity, $max\{\cdot\}$ is the max-pooling operation, $W^p \in \mathbb{R}^{3 \times d_1'}$ and $b^p \in \mathbb{R}^3$ are the prediction matrix and its bias. In summary, the whole algorithm is shown in Algorithm 1.

In training, we utilize the cross entropy loss function as the objective:

$$L = -\frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{3} y_{i,j} \log p_{i,j} + \lambda ||\theta||^2 \quad (12)$$

where T is the number of training samples, $y_i \in \mathbb{R}^3$ denotes the ground truth label of sample i, represented by one-hot vector, and $p_{i,j}$ is the predicted probability of sample i with sentiment j. θ is the set of all parameters and λ is the coefficient for L_2 regularization.

Algorithm 1 shows the overall framework of our model.

4 Experiments

4.1 Experimental Setup

We conduct experiments using the benchmark datasets of SemEval 2014 Task 4 (Pontiki et al.,

Dataset		Positive	Negative	Neutral
Restaurant	**Train**	2159	800	632
	Test	730	195	196
Laptop	**Train**	980	858	454
	Test	340	128	171

Table 1: Statistics of benchmark datasets.

2014)[1], which contain reviews about laptop and restaurant respectively and are used by previous works. The statistics of two benchmark datasets are shown in Table 1. There are three kinds of sentiment polarity: positive, negative and neutral.

In our experiments, we use GloVe.840B.300d embeddings (Pennington et al., 2014)[2] as previous works do. Each word embedding has 300 dimensions. Out-of-vocabulary (OOV) words are randomly sampled from the uniform distribution $\mathcal{U}(-0.02, 0.02)$. Weight matrices are initialized by sampling from uniform distribution $\mathcal{U}(-0.1, 0.1)$. The kernel sizes of convolution based feature extractors s_1, s_2, s_3 are 3, 4, 5. Each kernel consists of 128 filters. The dimension of outputs of LSTM $2d'_0$ and the convolution layer d'_1 are 400 and 384 respectively. We use Adam optimizer (Kingma and Ba, 2014) with learning rate 0.003. The batch size is set to 128. In order to alleviate overfitting, we set the dropout rate to 0.5 and the coefficient of L_2 regularization to 0.00001. The hyperparameter β used to calculate positional weights is set to 40. We choose the model with the minimum loss on testing set among 100 epochs. Besides, since there exists class imbalance in SemEval dataset, we additionally show the Macro-F1 scores of each model together with accuracy metric to further investigate the effectiveness and robustness of our model.

4.2 Comparison Results

In order to evaluate the effectiveness of our model, we compare it with 10 previous state-of-the-art models. The description is below:

• **AE-LSTM** (Wang et al., 2016) encodes the context-aware words to get representation. Then it simply uses the concatenation of context-aware word representations and target embeddings to classify the sentiment. However, the target embeddings do not contain contextual information.

• **ATAE-LSTM** (Wang et al., 2016) additionally leverages attention mechanism on top of **AE-LSTM** to find out relevant words with target.

• **GCAE** (Xue and Li, 2018) is based on CNN and applies Gated Tanh-ReLU Units (GTRU) to control the information flow from the target and build interaction between targets and contexts.

• **MemNet** (Tang et al., 2016b) uses a multi-hop attention mechanism whose query of the first attention layer is target representation. The attention result and the linear transformation of target representation are summed and used as the memory and the query of the next attention layer. Output of the last attention layer is considered as the sentiment representation used for classification.

• **IAN** (Ma et al., 2017) uses two attention mechanisms to select information from contexts and targets according to the average of encoded targets and contexts separately. The concatenation of two attention results is used for sentiment classification.

• **PG-CNN** (Huang and Carley, 2018) is also based on CNN and uses gate mechanism to incorporate target information into CNN architecture.

• The model designed by Hazarika et al. (2018) classifies all the targets in one sentence simultaneously with attention mechanism and inter-target dependencies detected by a complicated two-layer LSTM structure. One LSTM layer is designed to obtain the whole sentence representation based on each target in one sentence, similar to **ATAE-LSTM**. Then the model feeds the sentence representations altogether into the other LSTM to find the inter-target dependencies.

• **RAM** (Chen et al., 2017) uses multi-hop attention mechanism on position-weighted memories and combines the attention results to synthesize important features in difficult sentence structures. The model still constructs the memories by sentence-level information as **MemNet** does.

• **TRMN** (Wang et al., 2018) is a target-sensitive memory network, where various interaction mechanisms between target and context are leveraged. The whole architecture is similar to **MemNet**.

• **IARM** (Majumder et al., 2018) also leverages recurrent memory networks with attention mechanism. The memory is built by the sentence representation based on target information as **ATAE-LSTM** does. In addition, the model con-

[1]The detailed task definition can be obtained from http://alt.qcri.org/semeval2014/task4/

[2]Pre-trained word embeddings can be obtained from https://nlp.stanford.edu/projects/glove/

Models	Restaurant		Laptop	
	ACC	Macro-F1	ACC	Macro-F1
AE-LSTM*	76.60	66.45	68.90	62.45
ATAE-LSTM*	77.20	65.41	68.70	59.41
GCAE	77.28	-	69.14	-
MemNet*	78.16	65.83	70.33	64.09
IAN	78.20	-	72.10	-
PG-CNN	78.93	-	69.12	-
Hazarika et al. (2018)	79.00	-	72.50	-
RAM*	79.38	68.86	73.59	70.51
TRMN	-	69.00	-	68.18
IARM	80.00	-	73.80	-
Ours*	**80.98**†	**71.52**†	**74.56**†	**71.63**†

Table 2: Comparisons with baselines and ablation experiments (%). The best results are in bold. The model with $*$ means its result is the average value of 5 runs. The result with † means statistical significant at the level of 0.05 with the baselines tagged by $*$.

centrates on inter-target dependencies by memory networks, instead of vanilla LSTM structure used in the model proposed by Hazarika et al. (2018).

The comparisons with baseline methods are shown in Table 2. Our model significantly outperforms all the baselines. Except for **AE-LSTM**, **GCAE** and **PG-CNN**, the other baseline models adopt attention mechanism to evaluate the correlation between target and context words. However, the attention score for each word is distributed simultaneously according to simple computation by weight matrices. In our model, we intend to estimate the probability of being an opinion word at each time step based on the history information, such as previous soft labels and cell states, to take each word into account individually. Indeed, our model achieves significant improvements over the attention-based baseline models.

Moreover, we find that several baseline methods are based on memory networks, such as **MemNet**, **RAM**, **TRMN** and **IARM**. Note that the memories of these models are all based on the general sentence-level representations which might lose individual consideration and dilute the information of opinion words. Thus, it is better to take advantage of the history contexts and current word representation to consider each token individually instead of the overall sentence-level information.

Also, from the fact that **IAN**, which considers the interaction between target and context tokens, performs better than **AE-LSTM** and **ATAE-LSTM**, we observe the importance of interaction in this task. Though **GCAE** does not take context-aware representations into account, it still performs better than **AE-LSTM** and **ATAE-LSTM** do. It demonstrates the effectiveness of GTRU and further justifies the necessity of interaction between target and context. In our model, we emphasize the interaction when fusing the target representation with the context word representations and evaluating the correlation with targets to decide which information we should focus on more.

The convolution layer has been proved to be good at extracting local active features. However, the convolution based model **GCAE** and **PG-CNN** behave poorly in this task because vanilla convolution based models tend to find the salient features in the whole sentence rather than figure out the active features which are strongly associated with the target. Intuitively, closer words are more likely to modify the given target, and some of the previous state-of-the-art models also consider the relative position factors. Therefore, inspired by them, we apply a convolution based model combined with position information to achieve better performance.

4.3 Ablation Study

To evaluate the effect of each part in our model, we remove some important components or replace them with widely used alternatives. The comparisons with ablated tests are shown in Table 3. The results of ablation tests are the averages of 5 runs.

The biggest change from previous models is

Models	Restaurant		Laptop	
	ACC	Macro-F1	ACC	Macro-F1
Ours	**80.98**‡	**71.52**‡	**74.56**‡	**71.63**‡
with Hard Labels	78.34	68.17	73.14	69.01
with Attention	79.01	68.61	73.35	69.18
w/o Convolution Layer	79.15	68.45	73.28	69.24
w/o Soft Labels	79.34	68.37	73.62	69.52
w/o History Information	79.53	67.98	73.07	69.17
with AVG	80.29	69.65	73.84	70.31
w/o Positional Weights	80.54	69.95	73.65	70.44

Table 3: The results of ablation tests (%). The best results are in bold. **w/o History Information** indicates the soft label approach without consideration of history information. **with AVG** indicates the target representation is replaced by the averaged target embeddings. The result with ‡ means statistical significant at the level of 0.05.

that we use the soft label approach based on history information, such as previous soft labels and cell states, instead of using attention scores. To further confirm the effectiveness of the soft label strategy, we replace it with attention mechanism, which treats the target representation as a query and uses a weight matrix to compute the correlation between target and context words. The experimental results show that the accuracy drops over 1.97% and 1.21% and Macro-F1 score drops 2.91% and 2.45% respectively. It strongly proves the effectiveness of our soft label strategy and the better performance can be attributed to the careful consideration of each word at each time step. Additionally, we compare our model with **w/o History Information**, which does not feed previous time step's soft label and cell states information into the input of the current time step and simply uses a weight matrix to project the hidden outputs to the values of soft labels. The improvements show that the history information is indispensable for the task. The whole process of determining the soft label value in our model is fairly similar to the process of people reading a sentence and predicting the sentiments for targets discussed in Section 1. Besides, our model outperforms **with Hard Labels**, where the value of the label is either 0 or 1, because the soft approach can alleviate the propagation problem caused by hard decision. Moreover, our model greatly improves the performance compared with **w/o Soft Labels**. Obviously, the history-based soft label approach has great effects.

As mentioned before, the interaction between target and context is important in this task. Compared with the model **with AVG**, our model has

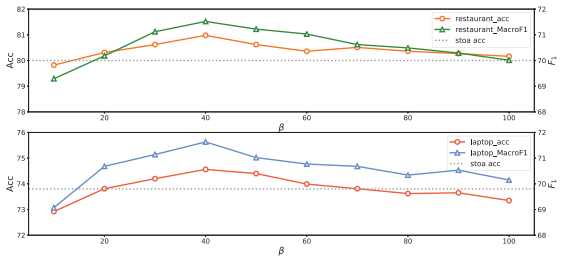

Figure 2: Effect of β on two datasets.

better performance on the two datasets for the target representation of our model contains contextual information and thus is more informative. The results indeed prove the usefulness of strengthening interaction. Lastly, without the convolution layer, the performance drops 1.83%, 1.28% on accuracy and 3.07%, 2.39% on Macro-F1 score respectively, suggesting that the convolution layer is capable of extracting active features for sentiment classification. Using relative distance information, our model greatly improves the performance of **w/o Positional Weights**. It indicates that position-aware information is beneficial to our model.

4.4 Impact of Rate of Decaying on Positional Weights

As our model involves the rate of decaying of positional weights which is controlled by β, we attempt to investigate which value is proper for β. Eq. 9 shows that the bigger β is, the slower the rate is. In our experiments, we keep the other experimental setups the same, and then vary β from 10 to 100, increased by 10. The results on two datasets are shown in Figure 2. Firstly, we notice that our model is better than most of the state-of-the-art

models on two datasets even if we do not optimize on β, suggesting that the other components of our model are effective. Besides, we observe that the performance tends to get better before β reaches 40, and there is a downward trend after it. When β equals 10, the rate of decaying is relatively fast. Since there are some long sentences in the datasets, the positional weights would lead to the loss of word information and result in worse performance. When β is large, like 100, the rate is slow and the positional weights may negligibly affect the classification process. Thus, it is necessary to choose a proper value for β.

4.5 Case Study

To further manifest the performance of our proposed model, we choose a case and show it in a heatmap form. In this case, the input sentence is *"The **dish** tastes bad but its **vegetable** is delicious though it looks ugly."* and the given target is ***"vegetable"***. There are two targets and three important sentiment words (*"bad"*, *"delicious"* and *"ugly"*) in the sentence. The challenge the model faces is to find out which sentiment word contributes more to the sentiment polarity of ***"vegetable"***. The upper part of Figure 3 is the visualization result of **with Attention** instead of using our proposed soft label strategy. We can easily find that the model attends on all the three sentiment words listed before, especially on *"bad"* and *"ugly"*, and wrongly predicts the sentiment as a negative one. It partially justifies attention mechanism's ability of extracting the sentiment words, but the wrong prediction could be attributed to the simultaneous weight distribution of attention scores and lack of individual consideration on each word.

Our proposed soft label approach is a good solution that could deal with the difficulty of matching multiple opinion words to the given target. The lower part indicates the visualization result of the value of soft labels and represents the process of soft label computation from the beginning of the sentence to the end. Besides, the proportions of history information g_i are all above 0.4, except for those of *"bad"* and *"delicious"*, which are 0.217 and 0.105 respectively. The relatively small value means that there might be a sentiment change in the place of the word. When the model browses to the word *"bad"*, as words before do not contain strong emotions, the cell states are now combined with the sentiment information of *"bad"*. When turning to *"delicious"*, the model recognizes that *"delicious"* is more relevant to the target while competing with the previous memory. Thus, its soft label's value becomes higher than that of *"bad"* and the word accounts for relatively great proportion of the cell states. Lastly, the model considers the cell states containing the information of *"delicious"* are more closely connected with the target than the word *"ugly"* is. As a result, the value of the soft label of *"ugly"* is low. Since the value of the soft label of *"delicious"* is the highest among those of all the other tokens in the sentence, the model predicts the sentiment correctly. The complex case strongly demonstrates the effectiveness of finding correct opinion words for target.

5 Error Analysis

Though our model achieves good performance by adopting the soft label strategy, we find that our model fails to predict the sentiment correctly in some cases. For example, when predicting the sentiment of the target ***"staff"*** in the sentence *"The **staff** should be a bit more friendly."*, our model tends to classify the sentiment as a positive one because of the opinion word *"friendly"*. Actually, the modal verb *"should"* represents the implicit meaning that the staff is not friendly and the customer hopes the staff could change the attitude towards customers. Therefore, there is still a room for our model to mine the kind of implicit semantics, not only based on the explicit opinion words. Additionally, we choose to detect the sentiment of ***"startup times"*** in the sentence *"**Startup times** are incredibly long: over two minutes."* and find that our model wrongly predicts the sentiment as a positive one. Though *"long"* is usually used to praise the quality of battery, it represents negative meaning when modifying the ***"startup times"***. The fact that the same opinion word represents totally different sentiments in different contexts may lead to the error.

6 Conclusion and Future Work

We propose a soft label approach to target-level sentiment classification task. Our model benefits from the soft label strategy based on history information, positional weights to take relative distance into account, and deeper interaction between target and context tokens. Experimental results on two benchmark datasets show that our model indeed substantially outperforms previous works. In

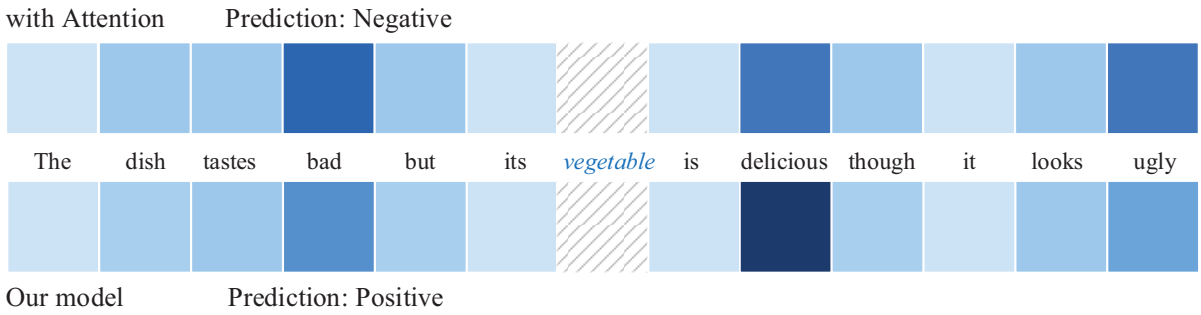

Figure 3: Case study of our proposed model and **with Attention** described in Section 4.3. The given target is *"**vegetable**"* and the sentiment towards it is positive. The deeper the blue is, the bigger the values of attention scores and soft labels are. Notice that the values of soft labels are normalized and they do not contain any position information.

the future, taking the encountered errors into account, we will do further researches on mining implicit semantics and distinguishing different sentiments expressed by the same opinion word in various kinds of contexts.

Acknowledgement

We would like to thank the anonymous reviewers for the helpful discussions and suggestions. This work is supported by National Natural Science Foundation of China under Grant No.61876004, No.61751201 and M1752013. The corresponding author of this paper is Baobao Chang.

References

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW workshop on NLP in the information explosion era*, volume 14, pages 339–348.

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*, pages 452–461. Association for Computational Linguistics.

Shuqin Gu, Lipeng Zhang, Yuexian Hou, and Yin Song. 2018. A position-aware bidirectional attention network for aspect-level sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 774–784. Association for Computational Linguistics.

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)*, pages 266–270. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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*, pages 1091–1096. Association for Computational Linguistics.

Binxuan Huang, Yanglan Ou, and Kathleen M. Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *Social, Cultural, and Behavioral Modeling - 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, July 10-13, 2018, Proceedings*, pages 197–206.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics.

Rie Johnson and Tong Zhang. 2015. Semi-supervised convolutional neural networks for text categorization via region embedding. In *Advances in neural information processing systems*, pages 919–927.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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)*, pages 437–442.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

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)*, pages 946–956. Association for Computational Linguistics.

Jiangming Liu and Yue Zhang. 2017. Attention modeling for targeted sentiment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 572–577. Association for Computational Linguistics.

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*, pages 4068–4074.

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*, pages 3402–3411.

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)*, pages 1532–1543. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris 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)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

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

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307. The COLING 2016 Organizing Committee.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224. Association for Computational Linguistics.

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*, pages 107–116.

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)*, volume 1, pages 957–967.

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*, pages 606–615. Association for Computational Linguistics.

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)*, pages 2514–2523. Association for Computational Linguistics.

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*, pages 1496–1505. Association for Computational Linguistics.