Aspect Term Extraction with History Attention and Selective Transformation*

Xin Li[1], Lidong Bing[2], Piji Li[1], Wai Lam[1], Zhimou Yang[3]
[1]Key Laboratory of High Confidence Software Technologies, Ministry of Education (CUHK Sub-Lab),
Dept of Systems Engineering & Engineering Management, Chinese University of Hong Kong
[2]Tencent AI Lab, Shenzhen, China
[3]College of Information Science and Engineering, Northeastern University, China
{lixin, wlam, pjli}@se.cuhk.edu.hk, lyndonbing@tencent.com, yangzhimou@stumail.neu.edu.cn

Abstract

Aspect Term Extraction (ATE), a key sub-task in Aspect-Based Sentiment Analysis, aims to extract explicit aspect expressions from online user reviews. We present a new framework for tackling ATE. It can exploit two useful clues, namely opinion summary and aspect detection history. Opinion summary is distilled from the whole input sentence, conditioned on each current token for aspect prediction, and thus the tailor-made summary can help aspect prediction on this token. Another clue is the information of aspect detection history, and it is distilled from the previous aspect predictions so as to leverage the coordinate structure and tagging schema constraints to upgrade the aspect prediction. Experimental results over four benchmark datasets clearly demonstrate that our framework can outperform all state-of-the-art methods.[1]

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) involves detecting opinion targets and locating opinion indicators in sentences in product review texts [Liu, 2012]. The first sub-task, called Aspect Term Extraction (ATE), is to identify the phrases targeted by opinion indicators in review sentences. For example, in the sentence "*I love the operating system and preloaded software*", the words "operating system" and "preloaded software" should be extracted as aspect terms, and the sentiment on them is conveyed by the opinion word "love". According to the task definition, for a term/phrase being regarded as an aspect, it should co-occur with some "opinion words" that indicate a sentiment polarity on it [Pontiki *et al.*, 2014].

Many researchers formulated ATE as a sequence labeling problem or a token-level classification problem. Traditional sequence models such as Conditional Random Fields (CRFs) [Chernyshevich, 2014; Toh and Wang, 2014; Toh and Su, 2016; Yin *et al.*, 2016], Long Short-Term Memory Networks (LSTMs) [Liu *et al.*, 2015] and classification models such as Support Vector Machine (SVM) [Manek *et al.*, 2016] have been applied to tackle the ATE task, and achieved reasonable performance. One drawback of these existing works is that they do not exploit the fact that, according to the task definition, aspect terms should co-occur with opinion-indicating words. Thus, the above methods tend to output false positives on those frequently used aspect terms in non-opinionated sentences, e.g., the word "restaurant" in "*the restaurant was packed at first, so we waited for 20 minutes*", which should not be extracted because the sentence does not convey any opinion on it.

There are a few works that consider opinion terms when tackling the ATE task. [Wang *et al.*, 2016] proposed Recursive Neural Conditional Random Fields (RNCRF) to explicitly extract aspects and opinions in a single framework. Aspect-opinion relation is modeled via joint extraction and dependency-based representation learning. One assumption of RNCRF is that dependency parsing will capture the relation between aspect terms and opinion words in the same sentence so that the joint extraction can benefit. Such assumption is usually valid for simple sentences, but rather fragile for some complicated structures, such as clauses and parenthesis. Moreover, RNCRF suffers from errors of dependency parsing because its network construction hinges on the dependency tree of inputs. CMLA [Wang *et al.*, 2017] models aspect-opinion relation without using syntactic information. Instead, it enables the two tasks to share information via attention mechanism. For example, it exploits the global opinion information by directly computing the association score between the aspect prototype and individual opinion hidden representations and then performing weighted aggregation. However, such aggregation may introduce noise. To some extent, this drawback is inherited from the attention mechanism, as also observed in machine translation [Luong *et al.*, 2015] and image captioning [Xu *et al.*, 2015].

To make better use of opinion information to assist aspect term extraction, we distill the opinion information of the whole input sentence into opinion summary[2], and such

*The work was done when Xin Li was an intern at Tencent AI Lab. The project is substantially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14203414).

[1]Codes are available at https://github.com/lixin4ever/HAST.

[2]Technically, opinion summary is the linear combination of the opinion representations generated from LSTM.

distillation is conditioned on a particular current token for aspect prediction. Then, the opinion summary is employed as part of features for the current aspect prediction. Taking the sentence "*the restaurant is cute but not upscale*" as an example, when our model performs the prediction for the word "restaurant", it first generates an opinion summary of the entire sentence conditioned on "restaurant". Due to the strong correlation between "restaurant' and "upscale" (an opinion word), the opinion summary will convey more information of "upscale" so that it will help predict "restaurant" as an aspect with high probability. Note that the opinion summary is built on the initial opinion features coming from an auxiliary opinion detection task, and such initial features already distinguish opinion words to some extent. Moreover, we propose a novel transformation network that helps strengthen the favorable correlations, e.g. between "restaurant' and "upscale", so that the produced opinion summary involves less noise.

Besides the opinion summary, another useful clue we explore is the aspect prediction history due to the inspiration of two observations: (1) In sequential labeling, the predictions at the previous time steps are useful clues for reducing the error space of the current prediction. For example, in the B-I-O tagging (refer to Section 2.1), if the previous prediction is "O", then the current prediction cannot be "I"; (2) It is observed that some sentences contain multiple aspect terms. For example, *"Apple is unmatched in product quality, aesthetics, craftmanship, and customer service"* has a coordinate structure of aspects. Under this structure, the previously predicted commonly-used aspect terms (e.g., "product quality") can guide the model to find the infrequent aspect terms (e.g., "craftmanship"). To capture the above clues, our model distills the information of the previous aspect detection for making a better prediction on the current state.

Concretely, we propose a framework for more accurate aspect term extraction by exploiting the opinion summary and the aspect detection history. Firstly, we employ two standard Long-Short Term Memory Networks (LSTMs) for building the initial aspect and opinion representations recording the sequential information. To encode the historical information into the initial aspect representations at each time step, we propose truncated history attention to distill useful features from the most recent aspect predictions and generate the history-aware aspect representations. We also design a selective transformation network to obtain the opinion summary at each time step. Specifically, we apply the aspect information to transform the initial opinion representations and apply attention over the transformed representations to generate the opinion summary. Experimental results show that our framework can outperform state-of-the-art methods.

2 The Proposed Model

2.1 The ATE Task

Given a sequence $X = \{x_1, ..., x_T\}$ of T words, the ATE task can be formulated as a token/word level sequence labeling problem to predict an aspect label sequence $Y = \{y_1, ..., y_T\}$, where each y_i comes from a finite label set $\mathcal{Y} = \{B, I, O\}$ which describes the possible aspect labels. As shown in the example below:

X	I	love	the	operation	system	and	preloaded	software
Y	O	O	O	B	I	O	B	I

B, I, and O denote beginning of, inside and outside of the aspect span respectively. Note that in commonly-used datasets such as [Pontiki *et al.*, 2016], the gold standard opinions are usually not annotated.

2.2 Model Description

As shown in Figure 1, our model contains two key components, namely **T**runcated **H**istory-**A**ttention (THA) and **S**elective **T**ransformation **N**etwork (STN), for capturing aspect detection history and opinion summary respectively. THA and STN are built on two LSTMs that generate the initial word representations for the primary ATE task and the auxiliary opinion detection task respectively. THA is designed to integrate the information of aspect detection history into the current aspect feature to generate a new history-aware aspect representation. STN first calculates a new opinion representation conditioned on the current aspect candidate. Then, we employ a bi-linear attention network to calculate the opinion summary as the weighted sum of the new opinion representations, according to their associations with the current aspect representation. Finally, the history-aware aspect representation and the opinion summary are concatenated as features for aspect prediction of the current time step.

Building Memory

As Recurrent Neural Networks can record the sequential information [Graves, 2012], we employ two vanilla LSTMs to build the initial token-level contextualized representations for sequence labeling of the ATE task and the auxiliary opinion word detection task respectively. For simplicity, let $\text{LSTM}^{\mathcal{T}}(x_t)$ denote an LSTM unit where $\mathcal{T} \in \{A, O\}$ is the task indicator. In the following sections, without specification, the symbols with superscript A and O are the notations used in the ATE task and the opinion detection task respectively. We use Bi-Directional LSTM to generate the initial token-level representations $h_t^{\mathcal{T}} \in \mathbb{R}^{2\text{dim}_h^{\mathcal{T}}}$ ($\text{dim}_h^{\mathcal{T}}$ is the dimension of hidden states):

$$h_t^{\mathcal{T}} = [\overrightarrow{\text{LSTM}}^{\mathcal{T}}(x_t); \overleftarrow{\text{LSTM}}^{\mathcal{T}}(x_t)], t \in [1, T]. \quad (1)$$

Capturing Aspect History

In principle, RNN can memorize the entire history of the predictions [Graves, 2012], but there is no mechanism to exploit the relation between previous predictions and the current prediction. As discussed above, such relation could be useful because of two reasons: (1) reducing the model's error space in predicting the current label by considering the definition of B-I-O schema, (2) improving the prediction accuracy for multiple aspects in one coordinate structure.

We propose a Truncated History-Attention (THA) component (the **THA** block in Figure 1) to explicitly model the aspect-aspect relation. Specifically, THA caches the most recent N^A hidden states. At the current prediction time step t, THA calculates the normalized importance score s_i^t of each cached state h_i^A ($i \in [t - N^A, t - 1]$) as follows:

$$a_i^t = \mathbf{v}^\top \tanh(\mathbf{W}_1 h_i^A + \mathbf{W}_2 h_t^A + \mathbf{W}_3 \tilde{h}_i^A), \quad (2)$$

$$s_i^t = \text{Softmax}(a_i^t). \quad (3)$$

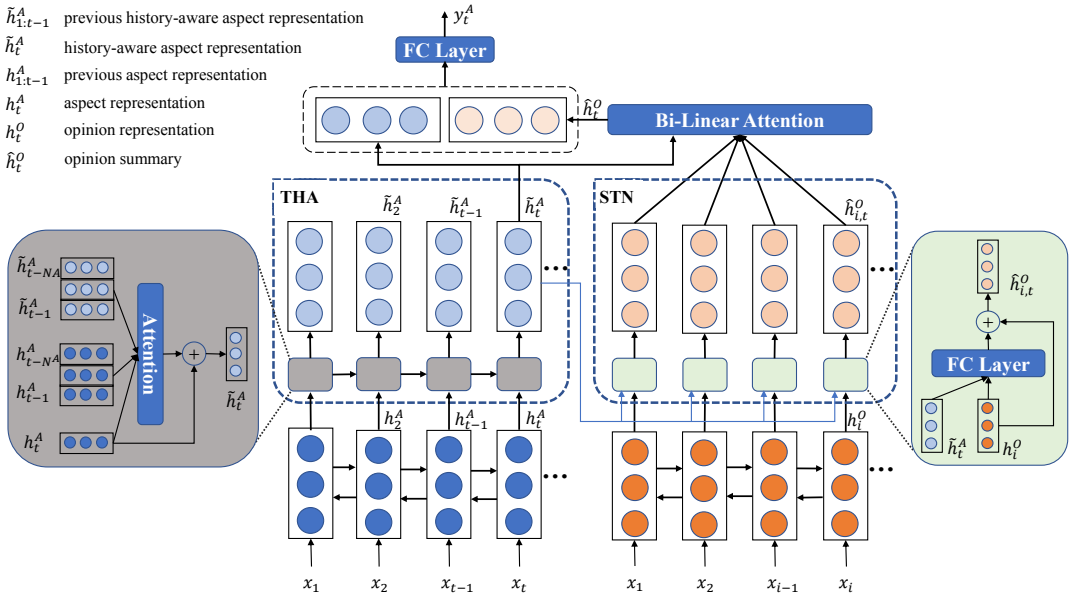


Figure 1: Framework architecture. The callouts on both sides describe how THA and STN work at each time step. Color printing is preferred.

$\tilde{h}_i^A$ denotes the previous history-aware aspect representation (refer to Eq. 5). $\mathbf{v} \in \mathbb{R}^{2\text{dim}_h^A}$ can be learned during training. $\mathbf{W}_{1,2,3} \in \mathbb{R}^{2\text{dim}_h^A \times 2\text{dim}_h^A}$ are parameters associated with previous aspect representations, current aspect representation and previous history-aware aspect representations respectively. Then, the aspect history $\hat{h}_t^A$ is obtained as follows:

$$\hat{h}_t^A = \sum_{i=t-N^A}^{t-1} s_i^t \times \tilde{h}_i^A. \tag{4}$$

To benefit from the previous aspect detection, we consolidate the hidden aspect representation with the distilled aspect history to generate features for the current prediction. Specifically, we adopt a way similar to the residual block [He *et al.*, 2016], which is shown to be useful in refining word-level features in Machine Translation [Wu *et al.*, 2016] and Part-Of-Speech tagging [Bjerva *et al.*, 2016], to calculate the history-aware aspect representations $\tilde{h}_t^A$ at the time step t:

$$\tilde{h}_t^A = h_t^A + \text{ReLU}(\hat{h}_t^A), \tag{5}$$

where ReLU is the relu activation function.

Capturing Opinion Summary

Previous works show that modeling aspect-opinion association is helpful to improve the accuracy of ATE, as exemplified in employing attention mechanism for calculating the opinion information [Wang *et al.*, 2017; Li and Lam, 2017]. MIN [Li and Lam, 2017] focuses on a few surrounding opinion representations and computes their importance scores according to the proximity and the opinion salience derived from a given opinion lexicon. However, it is unable to capture the long-range association between aspects and opinions. Besides, the association is not strong because only the distance information is modeled. Although CMLA [Wang *et al.*, 2017] can exploit global opinion information for aspect extraction, it may suffer from the noise brought in by attention-based feature aggregation. Taking the aspect term "fish" in *"Furthermore, while the **fish** is unquestionably fresh, **rolls** tend to be inexplicably bland."* as an example, it might be enough to tell "fish" is an aspect given the appearance of the strongly related opinion "fresh". However, CMLA employs conventional attention and does not have a mechanism to suppress the noise caused by other terms such as "rolls". Dependency parsing seems to be a good solution for finding the most related opinion and indeed it was utilized in [Wang *et al.*, 2016], but the parser is prone to generating mistakes when processing the informal online reviews, as discussed in [Li and Lam, 2017].

To make use of opinion information and suppress the possible noise, we propose a novel Selective Transformation Network (STN) (the **STN** block in Figure 1), and insert it before attending to global opinion features so that more important features with respect to a given aspect candidate will be highlighted. Specifically, STN first calculates a new opinion representation $\hat{h}_{i,t}^O$ given the current aspect feature $\tilde{h}_t^A$ as follows:

$$\hat{h}_{i,t}^O = h_i^O + \text{ReLU}(\mathbf{W}_4 \tilde{h}_t^A + \mathbf{W}_5 h_i^O), \tag{6}$$

where $\mathbf{W}_4$ and $\mathbf{W}_5 \in \mathbb{R}^{2\text{dim}_h^O \times 2\text{dim}_h^O}$ are parameters for history-aware aspect representations and opinion representations respectively. They map $\tilde{h}_t^A$ and h_i^O to the same subspace. Here the aspect feature $\tilde{h}_t^A$ acts as a "filter" to keep more important opinion features. Equation 6 also introduces a residual block to obtain a better opinion representation $\hat{h}_{i,t}^O$, which is conditioned on the current aspect feature $\tilde{h}_t^A$.

For distilling the global opinion summary, we introduce a bi-linear term to calculate the association score between $\tilde{h}_t^A$ and each $\hat{h}_{i,t}^O$:

$$w_{i,t} = \text{Softmax}(\tanh(\tilde{h}_t^A \mathbf{W}_{bi} \hat{h}_{i,t}^O + \mathbf{b}_{bi})), \tag{7}$$

where $\mathbf{W}_{bi}$ and $\mathbf{b}_{bi}$ are parameters of the Bi-Linear Attention layer. The improved opinion summary $\hat{h}_t^O$ at the time t is obtained via the weighted sum of the opinion representations:

$$\hat{h}_t^O = \sum_{i=1}^{T} w_{i,t} \times \hat{h}_{i,t}^O. \quad (8)$$

Finally, we concatenate the opinion summary $\hat{h}_t^O$ and the history-aware aspect representation $\tilde{h}_t^A$ and feed it into the top-most fully-connected (FC) layer for aspect prediction:

$$f_t^A = [\tilde{h}_t^A : \hat{h}_t^O], \quad (9)$$

$$P(y_t^A|x_t) = \text{Softmax}(\mathbf{W}_f^A f_t^A + \mathbf{b}_f^A). \quad (10)$$

Note that our framework actually performs a multi-task learning, i.e. predicting both aspects and opinions. We regard the initial token-level representations h_i^O as the features for opinion prediction:

$$P(y_i^O|x_i) = \text{Softmax}(\mathbf{W}_f^O h_i^O + \mathbf{b}_f^O). \quad (11)$$

$\mathbf{W}_f^{\mathcal{T}}$ and $\mathbf{b}_f^{\mathcal{T}}$ are parameters of the FC layers.

2.3 Joint Training

All the components in the proposed framework are differentiable. Thus, our framework can be efficiently trained with gradient methods. We use the token-level cross-entropy error between the predicted distribution $P(y_t^{\mathcal{T}}|x_t)$ ($\mathcal{T} \in \{A, O\}$) and the gold distribution $P(y_t^{\mathcal{T},g}|x_t)$ as the loss function:

$$\mathcal{L}_{\mathcal{T}} = -\frac{1}{T}\sum_{t=1}^{T} P(y_t^{\mathcal{T},g}|x_t) \odot \log[P(y_t^{\mathcal{T}}|x_t)]. \quad (12)$$

Then, the losses from both tasks are combined to form the training objective of the entire model:

$$\mathcal{J}(\theta) = \mathcal{L}_A + \mathcal{L}_O, \quad (13)$$

where $\mathcal{L}_A$ and $\mathcal{L}_O$ represent the loss functions for aspect and opinion extractions respectively.

3 Experiment

3.1 Datasets

To evaluate the effectiveness of the proposed framework for the ATE task, we conduct experiments over four benchmark datasets from the SemEval ABSA challenge [Pontiki *et al.*, 2014; Pontiki *et al.*, 2015; Pontiki *et al.*, 2016]. Table 1 shows their statistics. D_1 (SemEval 2014) contains reviews of the laptop domain and those of D_2 (SemEval 2014), D_3 (SemEval 2015) and D_4 (SemEval 2016) are for the restaurant domain. In these datasets, aspect terms have been labeled by the task organizer.

Gold standard annotations for opinion words are not provided. Thus, we choose words with strong subjectivity from MPQA[3] to provide the distant supervision [Mintz *et al.*, 2009]. To compare with the best SemEval systems and the current state-of-the-art methods, we use the standard train-test split in SemEval challenge as shown in Table 1.

[3]http://mpqa.cs.pitt.edu/

		# Sentences	# Aspects	# Sentences with aspects
D_1	TRAIN	3045	2358	1484
	TEST	800	654	422
D_2	TRAIN	3041	1743	1020
	TEST	800	1134	194
D_3	TRAIN	1315	1192	832
	TEST	685	542	401
D_4	TRAIN	2000	1743	1233
	TEST	676	622	420

Table 1: Statistics of datasets.

3.2 Comparisons

We compare our framework with the following methods:

- **CRF-1**: Conditional Random Fields with basic feature templates[4].
- **CRF-2**: Conditional Random Fields with basic feature templates and word embeddings.
- **Semi-CRF**: First-order Semi-Markov Conditional Random Fields [Sarawagi *et al.*, 2004] and the feature templates in Cuong *et al.* [2014] are adopted.
- **LSTM**: Vanilla bi-directional LSTM with pre-trained word embeddings.
- **IHS_RD** [Chernyshevich, 2014], **DLIREC** [Toh and Wang, 2014], **EliXa** [San Vicente *et al.*, 2015], **NLANGP** [Toh and Su, 2016]: The winning systems in the ATE subtask in SemEval ABSA challenge [Pontiki *et al.*, 2014; Pontiki *et al.*, 2015; Pontiki *et al.*, 2016].
- **WDEmb** [Yin *et al.*, 2016]: Enhanced CRF with word embeddings, dependency path embeddings and linear context embeddings.
- **MIN** [Li and Lam, 2017]: MIN consists of three LSTMs. Two LSTMs are employed to model the memory interactions between ATE and opinion detection. The last one is a vanilla LSTM used to predict the subjectivity of the sentence as additional guidance.
- **RNCRF** [Wang *et al.*, 2016]: CRF with high-level representations learned from Dependency Tree based Recursive Neural Network.
- **CMLA** [Wang *et al.*, 2017]: CMLA is a multi-layer architecture where each layer consists of two coupled GRUs to model the relation between aspect terms and opinion words.

To clarify, our framework aims at extracting aspect terms where the opinion information is employed as auxiliary, while RNCRF and CMLA perform joint extraction of aspects and opinions. Nevertheless, the comparison between our framework and RNCRF/CMLA is still fair, because we do not use manually annotated opinions as used by RNCRF and CMLA, instead, we employ an existing opinion lexicon to provide weak opinion supervision.

3.3 Settings

We pre-processed each dataset by lowercasing all words and replace all punctuations with `PUNCT`. We use pre-trained

[4]http://sklearn-crfsuite.readthedocs.io/en/latest/

Models	D_1	D_2	D_3	D_4
CRF-1	72.77	79.72	62.67	66.96
CRF-2	74.01	82.33	67.54	69.56
Semi-CRF	68.75	79.60	62.69	66.35
LSTM	75.71	82.01	68.26	70.35
IHS_RD (D_1 **winner**)	74.55	79.62	-	-
DLIREC (D_2 **winner**)	73.78	84.01	-	-
EliXa (D_3 **winner**)	-	-	70.04	-
NLANGP (D_4 **winner**)	-	-	67.12	72.34
WDEmb	75.16	84.97	69.73	-
MIN	77.58	-	-	73.44
RNCRF	78.42	84.93	67.74♮	69.72*
CMLA	77.80	85.29	70.73	72.77*
OURS w/o **THA**	77.64	84.30	70.89	72.62
OURS w/o **STN**	77.45	83.88	70.09	72.18
OURS w/o **THA** & **STN**	76.95	83.48	69.77	71.87
OURS	**79.52**	**85.61**	**71.46**	**73.61**

Table 2: Experimental results (F_1 score, %). The first four methods are implemented by us, and other results without markers are copied from their papers. The results with '*' are reproduced by us with the released code by the authors. For RNCRF, the result with '♮' is copied from the paper of CMLA (they have the same authors). '-' indicates the results were not available in their papers.

GloVe 840B vectors[5] [Pennington *et al.*, 2014] to initialize the word embeddings and the dimension (i.e., $\dim_w$) is 300. For out-of-vocabulary words, we randomly sample their embeddings from the uniform distribution $\mathcal{U}(-0.25, 0.25)$ as done in [Kim, 2014]. All of the weight matrices except those in LSTMs are initialized from the uniform distribution $\mathcal{U}(-0.2, 0.2)$. For the initialization of the matrices in LSTMs, we adopt Glorot Uniform strategy [Glorot and Bengio, 2010]. Besides, all biases are initialized as 0's.

The model is trained with SGD. We apply dropout over the ultimate aspect/opinion features and the input word embeddings of LSTMs. The dropout rates are empirically set as 0.5. With 5-fold cross-validation on the training data of D_2, other hyper-parameters are set as follows: $dim_h^A = 100$, $dim_h^O = 30$; the number of cached historical aspect representations N^A is 5; the learning rate of SGD is 0.07.

3.4 Main Results

As shown in Table 2, the proposed framework consistently obtains the best scores on all of the four datasets. Compared with the winning systems of SemEval ABSA, our framework achieves 5.0%, 1.6%, 1.4%, 1.3% absolute gains on D_1, D_2, D_3 and D_4 respectively.

Our framework can outperform RNCRF, a state-of-the-art model based on dependency parsing, on all datasets. We also notice that RNCRF does not perform well on D_3 and D_4 (3.7% and 3.9% inferior than ours). We find that D_3 and D_4 contain many informal reviews, thus RNCRF's performance degradation is probably due to the errors from the dependency parser when processing such informal texts.

CMLA and MIN do not rely on dependency parsing, instead, they employ attention mechanism to distill opinion information to help aspect extraction. Our framework consistently performs better than them. The gains presumably come from two perspectives: (1) In our model, the opinion summary is exploited after performing the selective transformation conditioned on the current aspect features, thus the summary can to some extent avoid the noise due to directly applying conventional attention. (2) Our model can discover some uncommon aspects under the guidance of some commonly-used aspects in coordinate structures by the history attention.

CRF with basic feature template is not strong, therefore, we add CRF-2 as another baseline. As shown in Table 2, CRF-2 with word embeddings achieves much better results than CRF-1 on all datasets. WDEmb, which is also an enhanced CRF-based method using additional dependency context embeddings, obtains superior performances than CRF-2. Therefore, the above comparison shows that word embeddings are useful and the embeddings incorporating structure information can further improve the performance.

3.5 Ablation Study

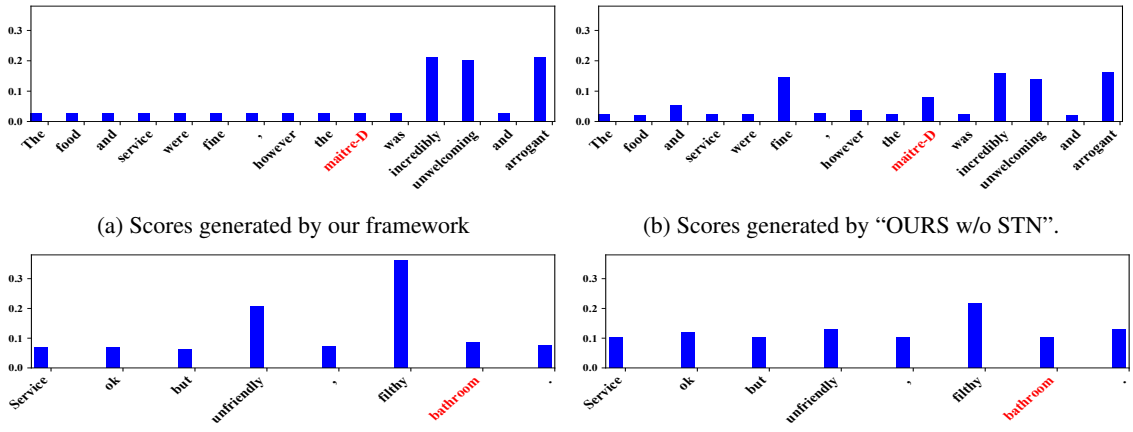
To further investigate the efficacy of the key components in our framework, namely, **THA** and **STN**, we perform ablation study as shown in the second block of Table 2. The results show that each of THA and STN is helpful for improving the performance, and the contribution of STN is slightly larger than THA. "OURS w/o THA & STN" only keeps the basic bi-linear attention. Although it performs not bad, it is still less competitive compared with the strongest baseline (i.e., CMLA), suggesting that only using attention mechanism to distill opinion summary is not enough. After inserting the STN component before the bi-linear attention, i.e. "OURS w/o THA", we get about 1% absolute gains on each dataset, and then the performance is comparable to CMLA. By adding THA, i.e. "OURS", the performance is further improved, and all state-of-the-art methods are surpassed.

3.6 Attention Visualization and Case Study

In Figure 2, we visualize the opinion attention scores of the words in two example sentences with the candidate aspects "maitre-D" and "bathroom". The scores in Figures 2a and 2c show that our full model captures the related opinion words very accurately with significantly larger scores, i.e. "incredibly", "unwelcoming" and "arrogant" for "maitre-D", and "unfriendly" and "filthy" for "bathroom". "OURS w/o STN" directly applies attention over the opinion hidden states h_i^O's, similar to what CMLA does. As shown in Figure 2b, it captures some unrelated opinion words (e.g. "fine") and even some non-opinionated words. As a result, it brings in some noise into the global opinion summary, and consequently the final prediction accuracy will be affected. This example demonstrates that the proposed STN works pretty well to help attend to more related opinion words given a particular aspect.

Some predictions of our model and those of LSTM and OURS w/o THA & STN are given in Table 3. The models incorporating attention-based opinion summary (i.e., OURS and OURS w/o THA & STN) can better determine if the commonly-used nouns are aspect terms or not (e.g. "device" in the first input), since they make decisions based on the global opinion information. Besides, they are able to extract some infrequent or even misspelled aspect terms (e.g.

[5]https://nlp.stanford.edu/projects/glove/



(a) Scores generated by our framework

(b) Scores generated by "OURS w/o STN".

(c) Scores generated by our framework

(d) Scores generated by "OURS w/o STN".

Figure 2: Opinion attention scores (i.e. $w_{i,t}$ in Equation 7) with respect to "maitre-D" and "bathroom".

Input sentences	Output of LSTM	Output of OURS w/o THA & STN	Output of OURS
1. *the device speaks about it self*	*device*	**NONE**	**NONE**
2. *Great survice !*	**NONE**	*survice*	*survice*
3. *Apple is unmatched in product quality, aesthetics, craftmanship, and custormer service*	*quality, aesthetics, custormer service*	*quality, customer service*	*product quality, aesthetics, craftmanship, custormer service*
4. *I am pleased with the fast log on, speedy WiFi connection and the long battery life*	*WiFi connection, battery life*	*log, WiFi connection, battery life*	*log on, WiFi connection, battery life*
5. *Also, I personally wasn't a fan of the portobello and asparagus mole*	*asparagus mole*	*asparagus mole*	*portobello and asparagus mole*

Table 3: Case analysis. In the input sentences, the gold standard aspect terms are underlined and in red.

"survice" in the second input) based on the indicative clues provided by opinion words. For the last three cases, having aspects in coordinate structures (i.e. the third and the fourth) or long aspects (i.e. the fifth), our model can give precise predictions owing to the previous detection clues captured by THA. Without using these clues, the baseline models fail.

4 Related Work

Some initial works [Hu and Liu, 2004] developed a bootstrapping framework for tackling Aspect Term Extraction (ATE) based on the observation that opinion words are usually located around the aspects. [Popescu and Etzioni, 2005] and [Qiu *et al.*, 2011] performed co-extraction of aspect terms and opinion words based on sophisticated syntactic patterns. However, relying on syntactic patterns suffers from parsing errors when processing informal online reviews. To avoid this drawback, [Liu *et al.*, 2012; Liu *et al.*, 2013] employed word-based translation models. Specifically, these models formulated the ATE task as a monolingual word alignment process and aspect-opinion relation is captured by alignment links rather than word dependencies. The ATE task can also be formulated as a token-level sequence labeling problem. The winning systems [Chernyshevich, 2014; San Vicente *et al.*, 2015; Toh and Su, 2016] of SemEval ABSA challenges employed traditional sequence models, such as Conditional Random Fields (CRFs) and Maximum Entropy (ME), to detect aspects. Besides heavy feature engineering, they also ignored the consideration of opinions.

Recently, neural network based models, such as LSTM-based [Liu *et al.*, 2015] and CNN-based [Poria *et al.*, 2016] methods, become the mainstream approach. Later on, some neural models jointly extracting aspect and opinion were proposed. [Wang *et al.*, 2016] performs the two task in a single Tree-Based Recursive Neural Network. Their network structure depends on dependency parsing, which is prone to error on informal reviews. CMLA [Wang *et al.*, 2017] consists of multiple attention layers on top of standard GRUs to extract the aspects and opinion words. Similarly, MIN [Li and Lam, 2017] employs multiple LSTMs to interactively perform aspect term extraction and opinion word extraction in a multi-task learning framework. Our framework is different from them in two perspectives: (1) It filters the opinion summary by incorporating the aspect features at each time step into the original opinion representations; (2) It exploits history information of aspect detection to capture the coordinate structures and previous aspect features.

5 Concluding Discussions

For more accurate aspect term extraction, we explored two important types of information, namely aspect detection history, and opinion summary. We design two components, i.e. truncated history attention, and selective transformation network. Experimental results show that our model dominates those joint extraction works such as RNCRF and CMLA on the performance of ATE. It suggests that the joint extraction sacrifices the accuracy of aspect prediction, although the ground-truth opinion words were annotated by these authors.

Moreover, one should notice that those joint extraction methods do not care about the correspondence between the extracted aspect terms and opinion words. Therefore, the necessity of such joint extraction should be obelized, given the experimental findings in this paper.

References

[Bjerva *et al.*, 2016] Johannes Bjerva, Barbara Plank, and Johan Bos. Semantic tagging with deep residual networks. *arXiv preprint arXiv:1609.07053*, 2016.

[Chernyshevich, 2014] Maryna Chernyshevich. Ihs r&d belarus: Cross-domain extraction of product features using crf. In *Proc. of SemEval*, pages 309–313, 2014.

[Cuong *et al.*, 2014] Nguyen Viet Cuong, Nan Ye, Wee Sun Lee, and Hai Leong Chieu. Conditional random field with high-order dependencies for sequence labeling and segmentation. *JMLR*, 15(1):981–1009, 2014.

[Glorot and Bengio, 2010] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of AISTATS*, volume 9, pages 249–256, 2010.

[Graves, 2012] Alex Graves. Supervised sequence labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pages 5–13. Springer, 2012.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of CVPR*, pages 770–778, 2016.

[Hu and Liu, 2004] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proc. of KDD*, pages 168–177, 2004.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. In *Proc. of EMNLP*, pages 1746–1751, 2014.

[Li and Lam, 2017] Xin Li and Wai Lam. Deep multi-task learning for aspect term extraction with memory interaction. In *Proc. of EMNLP*, pages 2886–2892, 2017.

[Liu *et al.*, 2012] Kang Liu, Liheng Xu, and Jun Zhao. Opinion target extraction using word-based translation model. In *Proc. of EMNLP*, pages 1346–1356, 2012.

[Liu *et al.*, 2013] Kang Liu, Heng Li Xu, Yang Liu, and Jun Zhao. Opinion target extraction using partially-supervised word alignment model. In *Proc. of IJCAI*, pages 2134–2140, 2013.

[Liu *et al.*, 2015] Pengfei Liu, Shafiq Joty, and Helen Meng. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proc. of EMNLP*, pages 1433–1443, 2015.

[Liu, 2012] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[Luong *et al.*, 2015] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proc. of EMNLP*, pages 1412–1421, 2015.

[Manek *et al.*, 2016] Asha S Manek, P Deepa Shenoy, M Chandra Mohan, and KR Venugopal. Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and svm classifier. *WWWJ*, 20:135–154, 2016.

[Mintz *et al.*, 2009] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proc. of ACL*, pages 1003–1011, 2009.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proc. of EMNLP*, pages 1532–1543, 2014.

[Pontiki *et al.*, 2014] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proc. of SemEval*, pages 27–35, 2014.

[Pontiki *et al.*, 2015] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proc. of SemEval*, pages 486–495, 2015.

[Pontiki *et al.*, 2016] Maria Pontiki, Dimitris Galanis, and et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proc. of SemEval*, pages 19–30, 2016.

[Popescu and Etzioni, 2005] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proc. of EMNLP*, pages 339–346, 2005.

[Poria *et al.*, 2016] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49, 2016.

[Qiu *et al.*, 2011] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27, 2011.

[San Vicente *et al.*, 2015] Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. Elixa: A modular and flexible absa platform. In *Proc. of SemEval*, pages 748–752, 2015.

[Sarawagi *et al.*, 2004] Sunita Sarawagi, William W Cohen, et al. Semi-markov conditional random fields for information extraction. In *Proc. of NIPS*, pages 1185–1192, 2004.

[Toh and Su, 2016] Zhiqiang Toh and Jian Su. Nlangp at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In *Proc. of SemEval*, pages 282–288, 2016.

[Toh and Wang, 2014] Zhiqiang Toh and Wenting Wang. Dlirec: Aspect term extraction and term polarity classification system. In *Proc. of SemEval*, pages 235–240, 2014.

[Wang *et al.*, 2016] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proc. of EMNLP*, pages 616–626, 2016.

[Wang *et al.*, 2017] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proc. of AAAI*, pages 3316–3322, 2017.

[Wu *et al.*, 2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. of ICML*, pages 2048–2057, 2015.

[Yin *et al.*, 2016] Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. Unsupervised word and dependency path embeddings for aspect term extraction. In *Proc. of IJCAI*, pages 2979–2985, 2016.