MTNA: A Neural Multi-task Model for Aspect Category Classification and Aspect Term Extraction On Restaurant Reviews

Wei Xue, Wubai Zhou, Tao Li and **Qing Wang**
School of Computing and Information Sciences
Florida International University
Miami, FL, USA
{wxue004, wzhou005, taoli, qwang028}@cs.fiu.edu

Abstract

Online reviews are valuable resources not only for consumers to make decisions before purchase, but also for providers to get feedbacks for their services or commodities. In Aspect Based Sentiment Analysis (ABSA), it is critical to identify aspect categories and extract aspect terms from the sentences of user-generated reviews. However, the two tasks are often treated independently, even though they are closely related. Intuitively, the learned knowledge of one task should inform the other learning task. In this paper, we propose a multi-task learning model based on neural networks to solve them together. We demonstrate the improved performance of our multi-task learning model over the models trained separately on three public dataset released by SemEval workshops.

1 Introduction

Aspect Based Sentiment Analysis (ABSA) (Liu and Zhang, 2012; Pontiki et al., 2016) task is proposed to better understand rapidly-growing online reviews than traditional opinion mining (Pang and Lee, 2008). ABSA aims to extract fine-grained insights such as named entities, aspects, and sentiment polarities. We focus on two subtasks in ABSA: aspect category classification (ACC) and aspect term extraction (ATE).

Given a predefined set of aspect categories, ACC aims to identify all the aspects discussed in a given sentence, while ATE is to recognize the word terms of target entities. For example, in restaurant reviews, suppose we have two aspects `Price` and `Food`. In the sentence *"The fish is carefully selected from all over the world and taste fresh and delicious."*, the aspect category is `Food`, and the aspect term is `fish`. There could be multiple aspect categories implied in one sentence; while in other sentences, there might be even no word corresponding to the given aspect category because of noisy aspect labels or fuzzy definition of the aspect. For example, the sentence *"I had a great experience."* expresses positive attitude towards the aspect `Restaurant`, but there is no corresponding word about it.

Recognizing the commonalities between ACC and ATE task can boost the performance of both of them. The aspect information of whole sentence can make it easier to differentiate the target terms from unrelated words; while recognized target terms are the hints for predicting aspect categories. Recently, neural networks have gained tremendous popularity and success in text classification (Kim, 2014; Kalchbrenner et al., 2014) and opinion mining (Irsoy and Cardie, 2014; Liu et al., 2015). In this paper, we consider ACC and ATE task together under a multi-task setting. We conduct experiments and analysis on SemEval datasets. Our model outperforms the conventional methods and competing deep learning models that tackle two problems separately.

2 Model

In this section, we specifically define the two tasks in ABSA: aspect category classification (ACC) and aspect term extraction (ATE), then present an end-to-end model MTNA (Multi-Task neural Networks for Aspect classification and extraction) that interleaves the two tasks.

We define ACC as a supervised classification task where the sentence should be labeled according to a subset of predefined aspect labels, and ATE as a sequential labeling task where the word tokens related to the given aspects should be

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 151–156,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

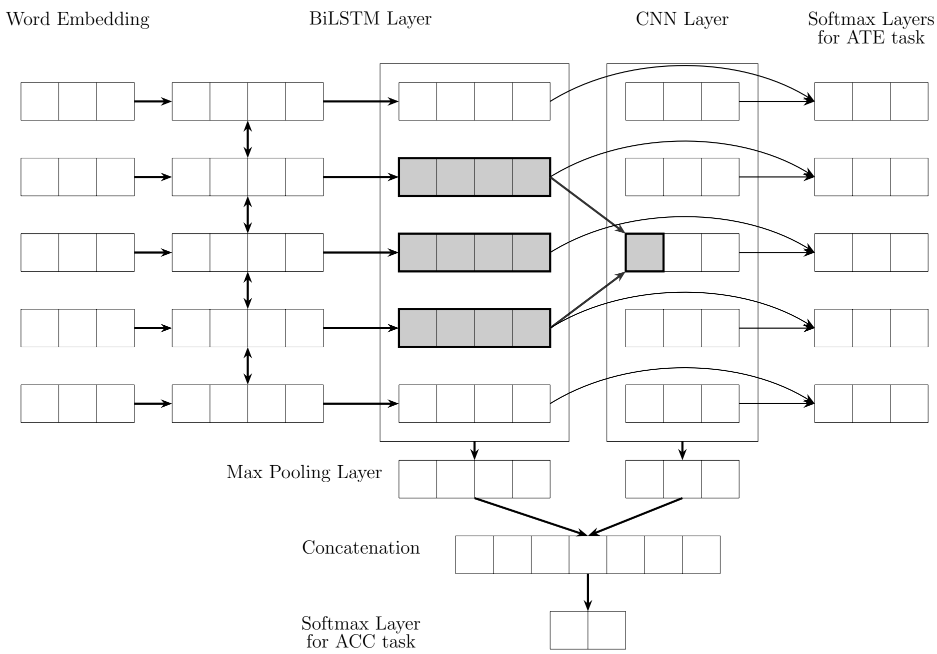

Figure 1: MTNA on a sequence of five words. The multi-task learning neural network combines BiLSTM and CNN layers together for ATE and ACC task respectively. One convolutional operation on BiLSTM layer is shown in the graph.

tagged according to a predefined tagging scheme, such as IOB (Inside, Outside, Beginning).

2.1 The Multi-task Learning Model

In this section, we describe our model MTNA.

Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) has memory cells and a group of adaptive gates to control the information flow of the network. It has good performance in named entity recognition task (NER) to simply stack embedding layer, Bi-directional LSTM (BiLSTM) layer and softmax layer together (Lample et al., 2016). ATE task can be viewed a special case of NER (Irsoy and Cardie, 2014). Convolutional Neural Networks (CNNs) have obtained good results in text classification, which usually consist of convolutional and pooling layers (Kim, 2014; Kalchbrenner et al., 2014; Toh and Su, 2016). They can be applied on ACC task immediately.

It should be noted that ACC task and ATE task are closely related. Aspect terms often implies the related aspect category. If the names of dishes appear in a sentence, it is easy to infer that this sentence is about the aspect `Food` and vice-versa. Multi-task learning can help the model of each task to focus its attention to relevant features, when the other task support the features with evidence (Ruder, 2017). Moreover, multi-task learning can obtain a common representation for all the tasks in the shared layers, which reduces noise in each task. We combine BiLSTM for ATE and CNN for ACC together in a multi-task framework. The parts for ACC task can utilize extra information learned in ATE task so that convolutional layers can focus on informative features. The tag prediction at each word in ATE task can also receive the distilled n-gram features of the surrounding words via convolutional operations.

The architecture of our model is shown in Figure 1. Specifically, a word embedding layer transforms indexed words to real valued vectors $\mathbf{x}_i$ with a pre-trained word embedding matrix (Mikolov et al., 2013; Pennington et al., 2014). Each sentence is represented by a matrix $\mathbf{S}$. A BiLSTM is applied on the outputs of word embedding layer $\mathbf{S}$, in which the two output vectors of the LSTMs are concatenated into a vector $\mathbf{h}_t$ for the t-th word. The represented features are further processed by a one-dimensional convolution layer with a set of kernels of different widths, so that the new feature maps $\mathbf{c}_t$ incorporate the information of words that are in the receptive field of the convolutions. For ATE task, we use softmax layer for each word in the given sentence to predict its tag. We further add skip connections between the LSTM layers to the softmax layers, since they are proved effective for training neural networks (He et al., 2016). To predict the aspect category of the sentence in ACC task, we use 1D max-over-time pooling layers (Collobert et al., 2011) which extracts maximum values from $\mathbf{h}_t$ and $\mathbf{c}_t$, a concatenation layer which joins the output vectors, and a softmax layer to output the probabilities of aspect categories. The final loss function of our model is a weighted sum of the loss functions of ACC task and ATE task. $L = L_{\text{acc}} + \lambda L_{\text{ate}}$, where λ is the weight parameter. L_{acc} is the cross-entropy loss function for ACC task; L_{ate} is the sentence-level log-likelihood for ATE task (Collobert et al., 2011; Lample et al., 2016).

3 Experiments

3.1 Datasets

For our experiments, we consider three data sets from SemEval workshops in recent years: SemEval 2014 Task 4 (SE14) (Pontiki et al., 2014), SemEval 2015 Task 12 (SE15) (Pontiki et al., 2015), and SemEval 2016 Task 5 (SE16) (Pontiki et al., 2016). We use the reviews in restaurant domain for all of them, and process SE14 into the same data format as the others. Each data set contains 2000 - 3000 sentences. For SE15 and SE16, an aspect label is a combination of an aspect and an attribute, like "Food#Price". There are 6 main aspects and total 12 configurations in SE15, SE16, while 5 aspects in SE14.

3.2 Experiment Setup

Following the experiment settings used by most competitors (Toh and Su, 2016; Khalil and El-Beltagy, 2016; Machacek, 2016) in SemEval 2016, we convert the multi-label aspect classification into multiple one-vs-all binary classifications. F1-score is used to measure the performance of each model for ACC task, and another F1 measure adapted for ATE task.

For MTNA model, we use the pre-trained word embedding GloVe (Pennington et al., 2014) of 200 dimensions to initialize the embedding layer. The word vectors that are out of GloVe vocabulary are randomly initialized between -0.1 and 0.1. During the training process, the embedding vectors are fine-tuned. We choose three kinds of convolution kernels which have the width of 3, 4, 5. Each of them has 100 kernels (Kim, 2014). We use tanh function as the nonlinear active function in convolution layers based on the results of cross validation. We train the model with Adadelta (Zeiler, 2012). For each binary classifier, a 5-fold cross validations is used to tune other hyper-parameters: mini-batch size from $\{10, 20, 50\}$, dropout rate from $\{0.1, 0.2, 0.5\}$, the dimension of LSTM cells from $\{100, 200, 500\}$, and the weight λ in the loss function from $\{0.1, 1, 10\}$.

3.3 Compared methods

Top models in SemEval. For ACC task, NRC-Can (Kiritchenko et al., 2014) and NLANGP (Toh and Su, 2015) are top models in 2014 and 2015 respectively, both of which use SVM. NLANG (Toh and Su, 2016) adopts CNN-like neural network in 2016. For ATE task, CRF (Toh and Wang, 2014; Toh and Su, 2015, 2016) is the best model on all of three data sets.

BiLSTM-CRF. To assess whether CNN can improve the performance of ATE, we use a standard Bi-directional LSTM with CRF layer (Lample et al., 2016) as the baseline to tag words.

MTNA-s. To evaluate to what extent that ATE loss function can improve the performance of the ACC task, we compare MTNA with its variance MTNA-s, the loss function of which does not include that of ATE task. However, this model keeps LSTM layer as a feature extractor before the convolution layers as MTNA does.

4 Results and Analysis

The comparison results of all methods on three datasets are shown in Table 1.

On ACC task, MTNA outperforms over other compared methods, which are proposed for a single task and cannot utilize the information from the other task. On ATE task, there are small improvement compared with conditional random field. It empirically proves that multi-task learning can benefit both tasks. MTNA has higher F1-scores compared with BiLSTM-CRF. The results confirm the effectiveness of additional convolution features for the ATE task.

MTNA-s, a smaller model without layers for ATE task, also performs better than CNN. It proves that LSTM can provide the feature engineering which captures the long-distance dependency (Zhang et al., 2016). On the aspects other than `Restaurant`, MTNA-s has slightly lower scores than MTNA, which again demonstrates the effectiveness of multi-task learning.

	SE14		SE15		SE16	
	ACC	ATE	ACC	ATE	ACC	ATE
Top models	88.57	**84.01**	62.68	67.11	73.03	72.34
BiLSTM-CRF	-	83.24	-	66.82	-	71.87
MTNA-s	87.95	-	64.32	-	75.69	-
MTNA	**88.91**	83.65	**65.97**	**67.73**	**76.42**	**72.95**

Table 1: Comparison results in F1 scores on three datasets.

Model	Aspect Category Classification			AspectTerm Extraction		
	Food	Restaurant	Service	Food	Restaurant	Service
CNN	86.29	65.27	84.02	-	-	-
Bi-LSTM-CRF	-	-	-	73.96	54.34	87.55
MTNA-s	86.41	67.89	84.93	-	-	-
MTNA	87.33	66.07	86.09	74.67	56.59	88.70
	Ambience	Drinks	Location	Ambinece	Drinks	Location
CNN	81.55	67.36	69.25	-	-	-
Bi-LSTM-CRF	-	-	-	76.23	71.38	56.77
MTNA-s	81.08	69.23	70.06	-	-	-
MTNA	83.18	68.75	71.43	77.79	72.21	60.16

Table 2: F1 scores of models on SE16 across six aspects

To access the performance of methods across different aspects, we combine all sentences labeled by the same aspect regardless of any attribute, then conduct experiments as before. We re-implement CNN model, which is used in NLANG 2016. The results are as shown in Table 2. ACC task on the aspect `Restaurant` is more difficult than the task on other aspects. Both CNN and MTNA have lower F1-scores on this aspect. The reason is that some sentences have restaurant names as target terms. However, there are around 40.1% sentences with `Restaurant` label that do not have annotated words in the training dataset, 41.2% in test dataset. Meanwhile, all methods have better results in ATE task on the aspect `Service` than on the other aspects, because target word tokens do not have much variety.

5 Related Work

LSTM (Hochreiter and Schmidhuber, 1997) has been applied on target extraction (Irsoy and Cardie, 2014; Liu et al., 2015). In the workshop of SemEval-2016, this sequential neural network is used to extract features for the subsequent CRF prediction (Toh and Su, 2016). In a multi-layer attention model (Wang et al., 2017), several attention subnetworks (Bahdanau et al., 2014) are used to extract aspect terms and opinion terms together without considering ACC task.

As a special case of text classification, ACC task is often treated as a supervised classification task. CNN (LeCun et al., 1998) has been used for sentiment classification (Kim, 2014; Kalchbrenner et al., 2014) and aspect classification (Toh and Su, 2016).

Collobert et al. (Collobert et al., 2011) proposed a multi-task learning system using deep learning methods for various natural language processing tasks. However, the system with window approach cannot be jointly trained with that using sentence window approach. Moreover, only embedding layer (lookup table) and linear layer are shared among tasks, which limited the utilization of shared information. On NER task, the predictions of this model depend only on the information of the current word rather than the surrounding context. The most relevant model is Dependency Sensitive Convolutional Neural Networks (DSCNN) (Zhang et al., 2016). The goal of DSCNN is solely for text classification, but our model is designed for multi-task learning of ACC and ATE.

6 Conclusion

We introduce two important tasks, e.g., aspect category classification and aspect term extraction in aspect based sentiment analysis. We propose a multi-task learning model based on recurrent neural networks and convolutional neural networks to solve the two tasks at the same time. Finally, the comparative experiments demonstrate the effectiveness of our model across three public datasets. We can utilize other linguistic information, such as POS tags and the distributional representation learned from character level convolutional neural network in the future work.

Acknowledgment

The work was supported in part by the National Science Foundation under Grant Nos. IIS-1213026 and CNS-1461926; and a FIU Dissertation Year Fellowship.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arxiv* .

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research* 12:2493–2537.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation* 9(8):1735–1780.

Ozan Irsoy and Claire Cardie. 2014. Opinion Mining with Deep Recurrent Neural Networks. In *EMNLP*. pages 720–728.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *ACL*. pages 655–665.

Talaat Khalil and Samhaa R El-Beltagy. 2016. NileTMRG at SemEval-2016 Task 5 - Deep Convolutional Neural Networks for Aspect Category and Sentiment Extraction. *SemEval@NAACL-HLT* pages 271–276.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*. pages 1746–1751.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif M. Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *SemEval@COLING*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 437–442.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *NAACL-HLT*. pages 260–270.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *IEEE* 86(11):2278–2324.

Bing Liu and Lei Zhang. 2012. A Survey of Opinion Mining and Sentiment Analysis. *Mining Text Data* (Chapter 13):415–463.

Pengfei Liu, Shafiq R Joty, and Helen M Meng. 2015. Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings. In *EMNLP*. pages 1433–1443.

Jakub Machacek. 2016. BUTknot at SemEval-2016 Task 5 - Supervised Machine Learning with Term Substitution Approach in Aspect Category Detection. *SemEval@NAACL-HLT* pages 301–305.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. pages 3111–3119.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieva* 2:1–135.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*. pages 1532–1543.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *SemEval 2015*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 486–495.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *SemEval@COLING*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 27–35.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *SemEval@NAACL-HLT*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 19–30.

Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. *arxiv* .

Zhiqiang Toh and Jian Su. 2015. NLANGP: Supervised Machine Learning System for Aspect Category Classification and Opinion Target Extraction. In *SemEval@NAACL-HLT*. pages 496–501.

Zhiqiang Toh and Jian Su. 2016. NLANGP at SemEval-2016 Task 5: Improving Aspect Based Sentiment Analysis using Neural Network Features. In *SemEval@NAACL-HLT*. pages 282–288.

Zhiqiang Toh and Wenting Wang. 2014. DLIREC: Aspect Term Extraction and Term Polarity Classification System. In *SemEval@COLING*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 235–240.

Weya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled Multi-Layer Attentions for Co-Extraction of Aspect and Opinion Terms. In *AAAI*. pages 3316–3322.

Matthew D Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *arxiv* .

Rui Zhang, Honglak Lee, and Dragomir Radev. 2016. Dependency Sensitive Convolutional Neural Networks for Modeling Sentences and Documents. In *NAACL-HLT*. pages 1512–1521.