Aspect-Based Sentiment Classification with Attentive Neural Turing Machines

Qianren Mao [1,2], **Jianxin Li** [1,2], **Senzhang Wang** [3], **Yuanning Zhang** [1,2], **Hao Peng** [1,2], **Min He** [4] and **Lihong Wang** [4]

[1]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, China
[2]State Key Laboratory of Software Development Environment, Beihang University, China
[3]Nanjing University of Aeronautics and Astronautics
[4]National Computer Network Emergency Response Technical Team/Coordination Center of China
{maoqr, lijx, zhangyn, penghao}@act.buaa.edu.cn, szwang@nuaa.edu.cn, hemin@cert.org.cn, wlh@isc.org.cn

Abstract

Aspect-based sentiment classification aims to identify sentiment polarity expressed towards a given opinion target in a sentence. The sentiment polarity of the target is not only highly determined by sentiment semantic context but also correlated with the concerned opinion target. Existing works cannot effectively capture and store the inter-dependence between the opinion target and its context. To solve this issue, we propose a novel model of Attentive Neural Turing Machines (ANTM). Via interactive read-write operations between an external memory storage and a recurrent controller, ANTM can learn the dependable correlation of the opinion target to context and concentrate on crucial sentiment information. Specifically, ANTM separates the information of storage and computation, which extends the capabilities of the controller to learn and store sequential features. The read and write operations enable ANTM to adaptively keep track of the interactive attention history between memory content and controller state. Moreover, we append target entity embeddings into both input and output of the controller in order to augment the integration of target information. We evaluate our model on SemEval2014 dataset which contains reviews of Laptop and Restaurant domains and Twitter review dataset. Experimental results verify that our model achieves state-of-the-art performance on aspect-based sentiment classification.

1 Introduction

Sentiment analysis, known as opinion mining, has drawn increasing attention from researchers and industries due to its wide application in understanding people's attitude towards some topic or product reviews and so on. Aspect-based sentiment analysis (ABSA) is a fine-grained task in the field of text classification [Pontiki *et al.*, 2014; Peng *et al.*, 2018]. Several subtasks can be regarded as sentiment classification problems at sentence level, e.g., aspect level sentiment classification and aspect term level (opinion target level) sentiment classification. The goal of our paper is to infer the sentiment polarity (e.g., positive, neutral, negative) of the opinion target appearing in given comments. As shown in case 1 of the following example, we called it multiple-target-different-polarity sentence: $'The\ food\ is\ usually\ good\ but\ it\ is\ certainly\ not\ a\ relaxing\ place\ to\ go.'$. The opinion target collocates with frequently-used sentiment words in which the sentiment polarity of target $food$ corresponding to sentiment word $good$ is positive while the polarity of target $place$ corresponding to $isn't\ relaxing$ is negative.

Case 1: *The **food** is usually good but it certainly isn't a relaxing **place** to go.*

Case 2: *The only thing I can imagine is that Sony jumped on early specifications for **Vista** requirements from Microsoft and designed it to those inadequate requirements.*

In addition to the challenge of case 1 where the polarity could be opposite when different targets are considered, another challenge presented in case 2 is referred to as a long-sequential-distance sentence. Unlike other review expression cases in which sentiment words always follow forward or backward to target words in a nearby position, there is a long distance between target words and related sentiment words in case 2 which shows that the target word $Vista$ is far away from the corresponding sentiment word $inadequate$ and demonstrative pronoun it. Unfortunately, most recurrent neural networks are hard ro handle the sentence of case 2 due to their chain structure of non-linearities being prone to gradient vanishing.

Among previous works, approaches [Wang *et al.*, 2014; Tang *et al.*, 2016a; Wang *et al.*, 2016] focusing on multiple-target-different-polarity sentence have just simply concatenated target representation to hidden state of neural networks. However, these methods are deficient in modeling the inter-dependence between target and context where sentiment features have been separated by long-term dependencies. Inspired by memory augmented neural networks being successfully applied in Question & Answering (Q&A) task, MemNet, [Tang *et al.*, 2016b] and CEA [Yang *et al.*, 2018a] treat target entity or aspect as a query object, and to find sentiment clues in memory content. RAM model attempts to adopt multiple-attention mechanism to capture sentiment features separated by a long distance and performs well in target senti-

ment analysis. However, this model has overloaded the usage of memory representations which takes only a single memory to both represent a source sentence and track attention history. In addition, relying on number of attention layers makes models hard to achieve a stable performance, as being the same with MemNet.

To solve the above deficiencies, we propose Attentive Neural Turing Machines (ANTM) for aspect term level sentiment classification. We use a structured memory module to store past information with separation of neural network parameters, and utilize an interactive read-write operation to automatically search for sectional sentiment information from memory content. Specifically, our model contains an external memory which is stacked by word representations of the input sentence, and a recurrent controller to encode sentence feature representation. With an addressing operation, the external memory can be read and written, which helps to capture sentiment features of context related target words. Specifically, the read operation keeps tracking an interactive attention among context words to opinion target, while the write operation fixes contents of memory at each time. Finally, we concatenate the opinion target into each hidden vector to augment integration of target information before computing attention weights for the final sentiment classification.

We evaluate our approach on SemEval2014 dataset which contains reviews of Laptop and Restaurant domains and Twitter review dataset. Experimental results show that our model achieves substantial performance improvement over the two datasets. The prime contributions of our work can be summarized as follows.

- With appending opinion target information, our ANTM model is robust to resolve the problem of target-sensitive sentiment by an efficient way of interaction between external memory and neural network state.
- Our ANTM model separates the information of storage and computation, which can extends the capabilities of a recurrent neural network to learn and store sequential features, and helps improve semantic loss from long-term dependencies.
- Our ANTM model sets a new state-of-the-art performance on the task of aspect term/opinion target level sentiment classification.

2 Related Work

Recent research works of ABSA can be broadly categorized into neural network based methods and memory network based methods.

2.1 ABSA with Neural Networks

The ABSA is often interpreted as a multi-class classification problem in the literature. Traditional approaches usually first manually build a set of features and then run them through Support Vector Machine (SVM) classifiers [Jiang *et al.*, 2011; Han *et al.*, 2013; Kiritchenko *et al.*, 2014; Wagner *et al.*, 2014]. The feature-based models depend on the quality of laborious feature engineering work and are labor intensive. [Dong *et al.*, 2014; Nguyen and Shirai, 2015] construct a target dependent phrase dependency tree to identify the sentiment of the aspect/target in the sentence, and outperform recursive neural networks. More efficacious work tends to detect the polarity of aspect or aspect term words using conventional neural networks like long short-term memory (LSTM). These models aim to explore the potential correlation of aspect or aspect term words and sentiment polarity. TD-LSTM and TC-LSTM [Tang *et al.*, 2016a] take opinion target information into consideration, and achieve good performance in target-dependent classification. [Wang *et al.*, 2016] proposes AE-LSTM, AT-LSTM and ATAE-LSTM methods. These methods introduce attention mechanism to concentrate on different parts of the sentence when different aspects are taken as input, and the result shows that feeding the embeddings of aspect or aspect terms is important to capture the corresponding sentiment polarity especially for the case 1 problem mentioned before. Drawing on the experience of form for Q&A, some methods [Tang *et al.*, 2016b; Liu *et al.*, 2018] treat opinion target information as a query vector or interactive vector [Ma *et al.*, 2017; Fan *et al.*, 2018] to context, and achieve a very competitive performance. All the result above show that the attention mechanism and appending target information are both effective way to capture related sentiment information in response to the concerned opinion target.

2.2 ABSA with Memory Networks

The memory networks have initially been explored for the task of Q&A with End-To-End Memory Networks (MemN2N) [Sukhbaatar *et al.*, 2015] and Gated MemN2N [Liu and Perez, 2017], and the task of copy and associative recall with Neural Turing Machine (NTM) [Graves *et al.*, 2014]. Moreover, some deep learning methods with memory augmented neural networks have been used in sentiment classification tasks and holistically succeed gain success. [Tang *et al.*, 2016b] proposed a deep memory network with multi-hops/layers named MemNet for aspect level sentiment classification and achieved comparable performance with feature-based SVM system, and substantively outperformed standard LSTM architectures. Inspired by multi-hops memory from MemNet, [Yang *et al.*, 2018b] used multi-hops memory to learn abstractive sentiment-related representation for both entity and aspect and achieved a significant gain over several baselines. Unlike LSTMs used in sentiment classification, memory-augmented networks encouraged local changes in memory. This helps not only to find the structure in the training data, but also to generalize to sequences that are beyond the generalization power of LSTMs, such as longer sequences in algorithmic tasks.

3 Methodology

Problem Definition. Our task is concerned with aspect term/opinion target level sentiment analysis. Suppose the input sentence with k words, is $S = \{w_1, ..., w_i, ..., w_k\}$, the goal of our model is to predict the sentiment polarity of a given opinion target w_i. In practice, we choose the aspect term or the opinion target word as an opinion target in the task

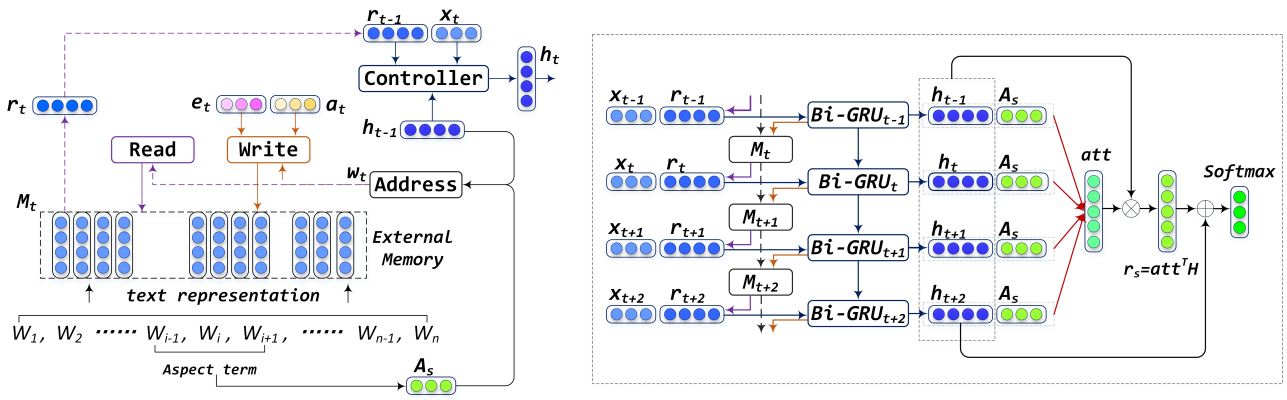


Figure 1: Illustration of our proposed Attentive Neural Turing Machines (ANTM) for opinion target sentiment classification.

of ABSA and sentiment analysis in Twitter respectively[1].

3.1 Attentive Neural Turing Machines

In this section, we will introduce our ANTM model in detail. The architecture of ANTM model for aspect term/opinion target level sentiment analysis is illustrated in Figure 1. The left of the figure is output in time t of the module of NTM memory M with text representation. The controller (navy blue) is fed on the current input x_t and the read vector r_{t-1} and the output vector h_{t-1}. The address operation with concatenating both hidden vector h_{t-1} and target A_s emits read (purple) and write (orange) heads to interact with the memory M. The right dashed box represents the whole model of ANTM, where the controller is a Bi-GRU. Particularly, the final representation computed by attention mechanism with concatenated the target vector to each output representation of the recurrent network.

The previous NTM model couples the neural networks to external memory via selective read and Write operations, and expands an ability to "deposit" and "process" information. This framework has been demonstrated to be effective in tasks of copying and sorting data sequences with extending the capabilities of neural networks. However, NTM may not be directly applicable to ABSA task due to a particular design for tasks of copying and sorting. Inspired by the vanilla NTM architecture, our ANTM model contains the same two basic components: a recurrent controller and an extended memory bank. Moreover, our model regards opinion target as a part of input to adaptively address important evidences from external memory where memory could be stacked by the word representations of a sentence. Moreover, with the attentive aspect participation where aspect information has been appended into each output hidden representations, our model efficiently concentrates on inter-dependence between sentiment semantic context words and different targets. Crucially, every component of our architecture is differentiable, making it easy to train with gradient descent in an end-to-end way.

[1]Note that the target could be a multi-word expression, e.g. "the Pope's visit to Palestine". We constitute the target embeddings as an average of the multiple words like [Sun *et al.*, 2015; Tang *et al.*, 2016b; Chen *et al.*, 2017] did.

3.2 The Extended Memory

The memory extends capabilities of neural networks by coupling them to the external memory resources since attentional processes to read from and write to the memory selectively. Thus, the external memory can be treated as a module storing addressable information, which makes sure of that the controller with neural networks captures more useful information. Our model stacks word embeddings of the input sentence into each memory slot as other memory networks did. Furthermore, we set other different external memory representations for experiments to evaluate the importance of memory to ANTM.

Since recent language pre-training models like ELMO [Peters *et al.*, 2018], GPT [Radford *et al.*, 2018] and BERT [Devlin *et al.*, 2018] have shown to be powerful for improving many natural language processing tasks on small-scale datasets, we argue that the pre-trained parameters from language models would potentially facilitate sentiment analysis with relative small-scale corpus. Thus we feed our text to the pre-trained BERT$_{Base}$ and BERT$_{Large}$[2], and then directly extract sequence output representations which contain richer contextual semantic information. Unlike existing BERT results that use a fine-tuning approach, we obtain the representations of corresponding sequence output as our external memory, named BERT$_B$ and BERT$_L$, respectively.

Specifically, all the word representations are encoded by column vectors in an embedding matrix $w \in \mathbb{R}^{d_w \times V}$, where d_w is the dimension of word representation and $|V|$ is vocabulary size. In order to keep full semantic information for sequential encoding, we do not separate text into two parts which only using context representations as an external memory as [Tang *et al.*, 2016b] does, but keep the full text representations $\{e_1..., e_i, ...e_k\}$ to stack into the slots of the external memory $\{M_t(1), ..., M_t(i), ..., M_t(k)\}$, where k is the number of memory slots and equivalent to the sentence length, the memory matrix is $M_t \in \mathbb{R}^{d \times k}$.

3.3 Target-Dependent Addressing Mechanisms

ANTM regards the concatenation of target vector and the output vector of the controller as a query vector to address im-

[2]https://github.com/google-research/bert

portant evidences from external memory M. We hope that sentiment information related to target can be imported to the controller input. Therefore, we combine content-based addressing mechanism to learn the interdependency weighting between target and context, which we call that operation the Target-Dependent Addressing Mechanisms (TDAM). As shown in Eq.1, $q_t \in \mathbb{R}^d$ is the query vector which concatenates the target vector and the last output vector produced by the controller. Finally, the TDAM addressing weights w_t^c is produced by a softmax function as follows.

$$w_t^c(i) = softmax\left(v_a^\top tanh\left(W_q\left[q_t; M_t(i)\right]\right)\right), \quad (1)$$

$$q_t = sigmoid\left(W_1[h_{t-1}; A_s] + b_1\right). \quad (2)$$

3.4 Writing Memory-State

There are two decomposed operations on writing to memory-state: an $erase$ followed by an add. The $erase$ is similar to the forget gates in LSTM/GRU, which determines how much information to be removed from memory cells. More specifically, the $erase$ vector specifies the values to be removed on each dimension in memory cells through normalized weights w_t^W emitted by a write head at time t. The memory vector $M_{t-1}(i)$ from the previous time-step is modified by the $erase$ vector $e_t \in \mathbb{R}^d$, where n is the number of words in a sentence $\forall i \in [1, n]$. Formally, the memory-state after $erase$ is given by:

$$\widetilde{M}_t(i) = M_{t-1}(i)\left[1 - w_t^W(i)\, e_t\right] \quad (3)$$

$$e_t = \sigma\left(W_t^e q_t\right), \quad (4)$$

where $e_t \in \mathbb{R}^d$ and the $w_t^W(i)$ specifies the weight associated with i^{th} memory cell in the same parametric form as in Eq.(1). Each write head also produces add vector $a_t \in \mathbb{R}^d$ which decides how much the current information should be written to the memory after the $erase$ operation:

$$M_t(i) = \widetilde{M}_t(i) + w_t^W(i)\, a_t \quad (5)$$

$$a_t = \sigma\left(W_t^a q_t\right). \quad (6)$$

The combined $erase$ and add operations of all write heads produce the final content of the memory at time t. The composite write operation is differentiable because both $erase$ and add are differentiable. $erase$ and add operations can adaptively modify and assign an importance score to each memory slot according to its semantic relatedness with opinion target.

3.5 Reading Memory-State

At each time t, let $M \in \mathbb{R}^{d \times k}$ be the memory contents and w_t^R be a vector of weights over the k slots emitted by read head at time t. The output vector is computed as a weighted sum of each piece of memory in $M_t(i)$, namely

$$r_t = \sum_{i=1}^{k} w_t^R(i)\, M_t(i), \quad (7)$$

where $r_t \in \mathbb{R}^k$, the elements $w_t(i) \in [0, 1]$ is the weight of M_t and $\sum_i w_t(i) = 1$. We use content-based addressing in determining w_t^R as the same parametric form as in Eq.(1).

Dataset	Pos.	Neg.	Neu.	Tot.
Laptop-Train	994	870	464	2313
Laptop-Test	341	128	169	638
Restaurant-Train	2164	807	637	3608
Restaurant-Test	728	196	196	1120
Twitter-Train	1561	1560	3127	6248
Twitter-Test	173	173	346	692

Table 1: Statistics of the datasets. We have removed those sentences that present both positive and negative opinion towards a target.

3.6 Controller Network

The controller network receives inputs from an external environment and emits outputs to participate in training process of sentiment classification. This controller can be a feed-forward network or a recurrent neural network. We choose the typical recurrent neural network of Bi-GRU [Cho *et al.*, 2014] as the controller and to certify coupling the external memory could extend the capabilities of Bi-GRU. More formally, the controller with Bi-GRU of our model can be computed as follows:

$$h_t = GRU(h_{t-1}; r_{t-1}; x_t). \quad (8)$$

The controller interacts with the external memory via input and output vectors. On one hand, our model leverages r_t produced by the last memory-state at each time t. On the other hand, the output vector that the controller produced will act on the address operation as Eq.(2) shows. With different extents by address, read and write operations, our model can concentrate on different parts of sentiment words related to concerned target.

3.7 Attentive NTM

The frontal way of using aspect information is letting target representation play a part in address, read and write operations, which not only explicitly captures the importance of each sentiment semantic information, but also helps the external memory store these sentiment semantic clues as well. In order to utilize target information better, we append target representation into each hidden vector so as to enhance the final sentiment classification. The right part of Figure 1 shows the attentive operation using an attention mechanism with appending aspect information to two kinds of controller's output layer. Let H be a matrix consisting of output vectors $[h_1, h_2..., h_k]$ that the controller produced. The attention mechanism will produce an attention weight over H and the final representation r.

$$M_s = tanh\left(\left[W_h H; W_v V_a\right]\right), \quad (9)$$

$$att = softmax\left(w^T M_s\right), \quad (10)$$

$$r_s = Hatt^T, \quad (11)$$

where $M_s \in \mathbb{R}^{(d^h + d^t) \times k}$, and d^h is the dimension of output vector produced by the controller at time t. d^t is the dimension of opinion target vector. $V_a = [A_s; A_s; ...; A_s]$ represents the operation of concatenating target vector in sequence. w is a trained parameter vector and w^T is a transpose. att is a vector consisting of attention weights, and r_s is a weighted representation of sentence.

Methods	Laptop		Restaurant		Twitter	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
Feature-SVM	0.7049	-	0.8016	-	0.6340	0.6330
TD-LSTM	0.6810	-	0.7560	-	0.6662♮	0.6401♮
ATAE-LSTM	0.6870	-	0.7720	-	-	-
IAN	0.7210	-	0.7860	-	-	-
MemNet	0.7237	-	0.8032	-	0.6850♮	0.6691♮
RAM	0.7449	0.7135	0.8023	0.7080	0.6936	0.6730
ANTM+Glove42B	0.7491	0.7142	0.8143	0.7120	0.7011	0.6814
ANTM+BERT$_B$	0.7537	0.7189	0.8078	0.7154	0.7176	0.6921
ANTM+BERT$_L$	**0.7584**	**0.7249**	**0.8249**	**0.7210**	**0.7235**	**0.6945**

Table 2: Comparison of different methods on reviews form SemEval 2014 Task 4 and Twitter. The results with '♮' are retrieved from the papers of RAM. We take the experimental results of the Accuracy rate and Macro-F1 for comparison.

3.8 Sentiment Prediction

Inspired by [Rocktäschel *et al.*, 2016], we add $W_x h_N$ into the final sentence representation which is $h^* = tanh\left(W_r r_s + W_x h_N\right)$. Then we use a softmax classifier to predict distribution over potential labels $\widehat{y}$ of sentiment polarity for the concerned target. The classifier takes the representation h^* as input, which can be used to make a prediction directly or fed into a loss function.

$$loss = -\sum_{(s,a)\in S}\sum_{c\in C} p_c\left(s,a\right)\cdot log\left(\widehat{p}_c\left(s,a\right)\right) + \lambda\left\|\theta\right\|^2, \quad (12)$$

where W_s and b_s are the parameters for softmax layer. S means all training sentences, C is the collection of sentiment categories, (s,a) means a sentence-target pair. $\widehat{p}(s,a) = softmax\left(W_s h^* + b_s\right)$ is the probability of predicting sentiment distribution and $p_c\left(s,a\right)$ denotes the ground truth. We use backpropagation to calculate the gradients of all the parameters, and update them with stochastic gradient descent. λ is the L_2 regularization term. θ is the parameter set.

4 Experiments

4.1 Datasets and Hyperparameters Setting

We conduct experiments on two datasets. The first dataset comes from SemEval 2014 Task 4, and it contains two kinds of customers' reviews from the Laptop and Restaurant domains. The second dataset is the Tweet collection [Dong *et al.*, 2014], as table 1 shows. Each review provides text, aspect term/opinion target, and corresponding sentiment polarity. In our experiments, the dimension of the target and text word vectors are set to 300 in the case of considering the content of each memory slot being set by the word vector of Glove42B [Pennington *et al.*, 2014]. While, these vectors are set to 768-dimension or 1024-dimension when we treat the corresponding sequence output representations from BERT$_{Base}$ or BERT$_{Large}$ as our external memory respectively. We train our model with the L2-regularization weight of 0.001 and the initial learning rate of 0.01. We also set dropout of 0.5 to avoid over-fitting.

4.2 Comparison Methods

We choose the following methods as baselines.

Feature-based SVM [Kiritchenko *et al.*, 2014] uses SVMs on n-gram features, parse features and lexicon features.

TD-LSTM [Tang *et al.*, 2016a] extends LSTM to model the left and right target-dependent representations, then concatenates them for prediction.

ATAE-LSTM [Wang *et al.*, 2016] is developed based on AE-LSTM. This method appends the aspect embeddings into each input embeddings and hidden vector to strengthen the inter-dependence between context and target.

IAN [Ma *et al.*, 2017] uses an interactive attention to strengthen the inter-dependence between context and target.

MemNet [Tang *et al.*, 2016b] applies a deep memory network with three attention hops for aspect sentiment classification.

RAM [Chen *et al.*, 2017] builds a memory module to synthesize the word sequence features. Similar to MemNet, it adds a recurrent function and employs multiple attentions on memory to predict the opinion target sentiment.

We also list the variants of ANTM model, which are used to analyze the effects of the external memory with different word representations.

ANTM+Glove42B contains an external memory equipped by word embeddings of Glove42B and the controller network is equipped by the Bi-GRU network.

ANTM+BERT$_B$ contains an external memory which is stacked by sequence outputs of BERT$_{Base}$ and the controller network is equipped by the Bi-GRU network.

ANTM+BERT$_L$ contains an external memory which is stacked by sequence outputs of BERT$_{Large}$ and the controller network is equipped by the Bi-GRU network.

5 Results and Analysis

The results are shown in Table 2. We can see that ANTM+BERT$_L$ achieves the best performance among all the baselines. Compared with RAM, ANTM+BERT$_L$ improves the performance of accuracy rate about 1.35% and 2.26% on the Laptop and Restaurant category respectively. Among two recurrent neural network models, ATAE-LSTM with appending target embedding both in the input and output representations has an advantage in concentrating on different parts of sentence, so that it performs better than TD-LSTM. However, ATAE-LSTM still performs worse than our model, because it simply concatenates target embeddingss. Although feature-based SVM outperforms other two recurrent models of TD-LSTM and ATAE-LSTM in accuracy rate of Laptop and Restaurant categories, it is inferior to our model. In addition, memory networks of MemNet and RAM with multi-

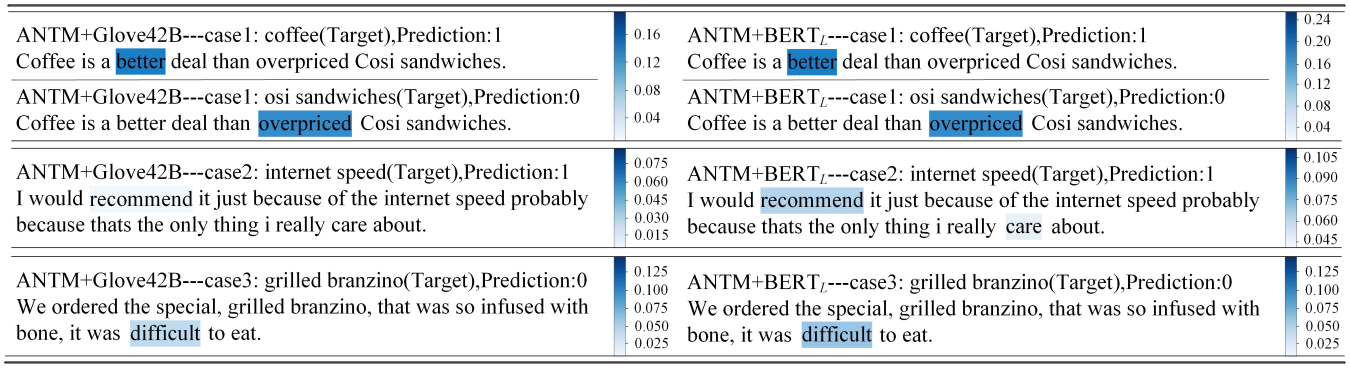
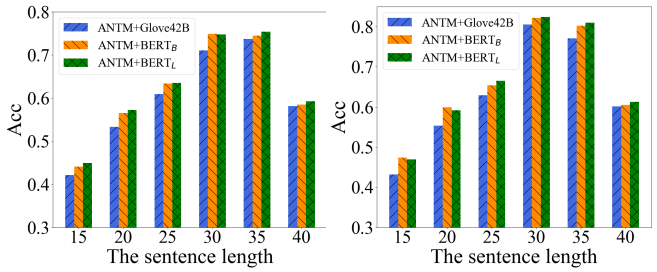


Figure 2: The attention visualizations on sentiment words. Deeper color implies larger attention weights. These cases contains the multiple-target-different-polarity sentence and the long-sequential-distance sentence.



(a) Laptop (b) Restaurant

Figure 3: Accuracy comparison of different methods with diverse sentence length on reviews from Restaurant and Laptop datasets.

attention mechanism have obtained comparative results compared with the feature-based SVM and augmented recurrent networks. MemNet regards aspect vector as a query vector to adaptively select important evidences from memory through multi-attention layers (hops). RAM model introduces an external memory and adopts non-linealy multiple attentions, and thus is more effective than MemNet.

Our models steadily outperforms the other two memory augmented networks. Particularly on Restaurant dataset, the improvement of ANTM+BERT$_L$ obtain more than 2% gain of Acc score compared with MemNet and RAM. Unlike MemNet and RAM, our model does not simply use the attention mechanism to match important sentiment semantic information, but computes the context of sentiment information as read vectors being feed into each cell of recurrent neural networks at each time step. Thus, our model not only encodes the inter-dependence between sentiment semantic context words and the target, but also extends the capabilities of neural networks.

Table 2 also shows that external memories with different word representations have subtly different effects on the last accuracy. ANTM+BERT$_L$ gets a slightly higher accuracy rate than others on three datasets. Moreover, ANTM+Glove42B outperforms other baseline models which verify ANTM can achieve distinguished performance. While, ANTM with memory of sequence output from the pre-trained representations of BERT is certainly better than ANTM with memory of the Glove42B word vectors in our task. ANTM+BERT$_L$ achieves 0.7584, 0.8249, 0.7235 Acc score over the three datasets and outperforms ANTM+Glove42B by 0.93%, 1.06%, 2.24%, respectively.

Figure 2 illustrates that our models have a robust ability to dispose of the multiple-target-different-polarity sentence. As the case 1 shows, both ANTM+Glove42B and ANTM++BERT$_L$ can find the corresponding sentiment clue of the concerned opinion target. Compared with the long-sequential-distance sentences of case 2 and 3, ANTM+BERT$_L$ performs better to capture more abstract sentiment semantic words with a slighter color of the attention weight of the sentiment verb "recommend" in case 2 and the semantic word "difficult" in case 3. These attention map cases validate the advantage of our method to keep track of attention weights especially in the long-sequential-distance text, in a way of enhancing the interaction between the target and its corresponding sentiment semantic context.

Figure 3 shows the classification accuracy of ANTM model with different input sentence length. Specifically, the highest accuracy of the Laptop dataset is achieved with sentence length of about 35, while Restaurant dataset gains the top accuracy with around 30. This is mainly because different datasets present different distributions of sentence length. 92.14% of sentences are less than 35 in length for the Laptop dataset with 2328 training data, while 91.38% of sentences are less than 30 in length for the Restaurant dataset with 3608 training data. From another point of view, this result proves that our model performs stably with different sentence length.

6 Conclusion

In this paper, we proposed an model of Attentive Neural Machines (ANTM) for aspect term/opinion target level sentiment analysis. The motivation of ANTM is to deploy an external memory to separate storage information from computation in this way to extend the capabilities of neural networks. Experimental results show that ANTM performs superior performance compared with these baselines, especially can boost the efficiency of dispose of the long-sequential-distance text. ANTM also proves the extended memory with drawing support from sequence output of pre-trained BERT model performs better in the small-scale corpuses. Potential future plan is to demonstrate the stability and superiority of ANTM applying to longer sequences for other sentiment analysis tasks.

Acknowledgements

Jianxin Li is the corresponding author. This work is supported by the National Natural Science Foundation of China (NSFC) (No.61772151, No.61872022 and No.61421003) and SKLSDE-2018ZX16. The co-author Senzhang Wang is supported by NSFC(No.61602237). The co-author Min He is supported by National Key R&D Program of China(No.2017YFB0803305). We also thank our anonymous reviewers for their constructive comments.

References

[Chen *et al.*, 2017] Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. Recurrent attention network on memory for aspect sentiment analysis. In *EMNLP*, 2017.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[Dong *et al.*, 2014] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *ACL(Volume 2: Short Papers)*, 2014.

[Fan *et al.*, 2018] Feifan Fan, Yansong Feng, and Dongyan Zhao. Multi-grained attention network for aspect-level sentiment classification. In *EMNLP*, 2018.

[Graves *et al.*, 2014] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. 2014.

[Han *et al.*, 2013] Qi Han, Junfei Guo, and Hinrich Schuetze. Codex: Combining an SVM classifier and character n-gram language models for sentiment analysis on twitter text. In *SemEval@NAACL-HLT*, pages 520–524, 2013.

[Jiang *et al.*, 2011] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *ACL*, 2011.

[Kiritchenko *et al.*, 2014] Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *SemEval 2014*, 2014.

[Liu and Perez, 2017] Fei Liu and Julien Perez. Gated end-to-end memory networks. In *ACL*, 2017.

[Liu *et al.*, 2018] Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. Content attention model for aspect based sentiment analysis. In *WWW*, 2018.

[Ma *et al.*, 2017] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. Interactive attention networks for aspect-level sentiment classification. In *AAAI*, 2017.

[Nguyen and Shirai, 2015] Thien Hai Nguyen and Kiyoaki Shirai. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *EMNLP*, 2015.

[Peng *et al.*, 2018] Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *WWW*, 2018.

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

[Peters *et al.*, 2018] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018.

[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. *SemEval-2014*, 2014.

[Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *Technical report, OpenAI*, 2018.

[Rocktäschel *et al.*, 2016] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. Reasoning about entailment with neural attention. In *ICLR*, 2016.

[Sukhbaatar *et al.*, 2015] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, 2015.

[Sun *et al.*, 2015] Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. Modeling mention, context and entity with neural networks for entity disambiguation. In *IJCAI*, 2015.

[Tang *et al.*, 2016a] Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. Effective lstms for target-dependent sentiment classification. In *COLING*, 2016.

[Tang *et al.*, 2016b] Duyu Tang, Bing Qin, and Ting Liu. Aspect level sentiment classification with deep memory network. In *EMNLP*, 2016.

[Wagner *et al.*, 2014] Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. Dcu: Aspect-based polarity classification for semeval task 4. In *SemEval 2014*, 2014.

[Wang *et al.*, 2014] Senzhang Wang, Xia Hu, Philip S. Yu, and Zhoujun Li. Mmrate: inferring multi-aspect diffusion networks with multi-pattern cascades. In *KDD*, 2014.

[Wang *et al.*, 2016] Yequan Wang, Minlie Huang, Li Zhao, et al. Attention-based lstm for aspect-level sentiment classification. In *EMNLP*, 2016.

[Yang *et al.*, 2018a] Jun Yang, Runqi Yang, Chongjun Wang, and Junyuan Xie. Multi-entity aspect-based sentiment analysis with context, entity and aspect memory. In *AAAI*, 2018.

[Yang *et al.*, 2018b] Jun Yang, Runqi Yang, Chongjun Wang, and Junyuan Xie. Multi-entity aspect-based sentiment analysis with context, entity and aspect memory. In *AAAI*, 2018.