Proceedings of Machine Learning Research 101:425–440, 2019 ACML 2019

An Attentive Memory Network Integrated with Aspect Dependency for Document-Level Multi-Aspect Sentiment Classification

Qingxuan Zhang QXZHANG@BIT.EDU.CN
Chongyang Shi * CY_SHI@BIT.EDU.CN
School of Computer, Beijing Institute of Technology, Beijing 100081, China

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

Document-level multi-aspect sentiment classification is one of the foundational tasks in natural language processing (NLP) and neural network methods have achieved great success in reviews sentiment classification. Most of recent works ignore the relation between different aspects and do not take into account the contexting dependent importance of sentences and aspect keywords. In this paper, we propose an attentive memory network for document-level multi-aspect sentiment classification. Unlike recent proposed models which average word embeddings of aspect keywords to represent aspect and utilize hierarchical architectures to encode review documents, we adopt attention-based memory networks to construct aspect and sentence memories. The recurrent attention operation is employed to capture long-distance dependency across sentences and obtain aspect-aware document representations over aspect and sentence memories. Then, incorporating the neighboring aspects related information into the final aspect rating predictions by using multi-hop attention memory networks. Experimental results on two real-world datasets TripAdvisor and BeerAdvocate show that our model achieves state-of-the-art performance.

Keywords: Memory network, Aspect dependency, Multi-aspect sentiment classification.

1. Introduction

Document-level multi-aspect sentiment classification (DMSC) is a fine-grained task in sentiment analysis [Liu (2012); Pang et al. (2008)], which aims to predict sentiment ratings from different aspects of products. With explosion of the Internet in this digital age, more and more people easily access the Internet and share their opinions on social media. Nowadays, users tend to give not only an overall rating of the product, but also score the different aspects of the product when they generate reviews on websites. The analysis of these aspect ratings could help companies better understand the pros and cons of the product or service. Therefore, it is very useful to perform document-level multi-aspect sentiment classification for analyzing reviews automatically.

Document-level sentiment classification is also a well-known task in sentiment analysis. Different from document-level sentiment classification, DMSC mainly focus on predicting different ratings for each aspect rather than an overall rating and the sentiment prediction of each aspect is actually a document-level sentiment classification task. Therefore, many

* Corresponding author

recent works on DMSC construct a multi-task learning framework [Caruana (1997); Luong et al. (2015)] with attention mechanism which is used to obtain different document representations for different aspects. Yin et al. (2017) propose a hierarchical iterative attention model to extract aspect-aware words and sentences. Li et al. (2018) introduce a hierarchical user aspect rating network to consider user preference and overall ratings jointly, and obtain state-of-the-art results.

Those recent works mentioned above on DMSC only use the mean word embedding of aspect keywords, but actually the frequency of each keyword of aspect is different in reviews and most sentences reflect the first and second highest frequency keywords of aspect. Therefore, we can model the importance of keywords of each aspect to obtain more effective aspect representations. In this work, we construct aspect memory with self-attention module, which contains the importance of each aspect keyword. Instead of modeling documents with hierarchical architecture, we encode the sentences with memory network and design a recurrent attention layer over aspect and sentence memories to obtain document representations. Otherwise, the recent proposed models on DMSC ignore the relation between different aspects. For example, in reviews "*The hotel is not worth the price, it took us two hours to be served.*", there are two aspects: "*price*" and "*service*". But the aspect "*service*" does not have enough context information to express sentiment, we are difficult to decide the polarity of aspect "*service*" unless considered with the aspect "*price*". In documents with multiple aspects, inspired by Hazarika et al. (2018), we suppose that the related information between aspects could be useful for sentiment predictions of aspects.

In this paper, we propose an attentive memory network to perform document-level multi-aspect sentiment classification. We model the aspects and the sentences as external memories by using attention-based memory networks. Then, a recurrent attention layer is adopted over these external memories to construct aspect-specific document representations. Finally, the multi-hop attention memory networks are introduced to mining related information from surrounding aspects and the final representations are fed to classifier for aspect rating predictions.

The contributions of this paper are as follows:

1. For modeling the keywords importance of aspects and capturing the internal structures of the sentences, we adopt attention-based memory network to construct aspect and sentence memories.

2. To capture long-distance dependency across sentences and obtain effective aspect-aware document representations, we propose a recurrent attention operation on aspect and sentence memories.

3. We propose a multi-hop attention memory network to incorporate the neighboring aspects related information into aspect rating predictions.

4. The experimental results demonstrate that our model achieves state-of-the-art performance.

The rest of the paper is organized as follows. After a summary of related work in Section 2, we describe the problem formalization of DMSC and our proposed model in Section 3. We provide experiments and evaluations in Section 4. Section 5 concludes this paper and discusses future directions.

2. Related Work

2.1. Multi-Aspect Sentiment Classification

Multi-aspect sentiment classification has already been studied extensively. Lu et al. (2011) propose Segmented Topic Model to model the whole document and build features, then use support vector regression to predict aspect ratings. McAuley et al. (2012) focus on learning which parts of a review correspond to each rated aspect and use multi-class SVM to predict ratings. Yin et al. (2017) model this task as a machine comprehension problem and adopt an iterative attention mechanism to build aspect-aware representations. Li et al. (2018) introduce a hierarchical user aspect rating network to consider user preference and overall ratings jointly, and obtain state-of-the-art results. Another related problem is called aspect-based sentiment classification (sentence-level task), which aims to identify the sentiment polarities of target aspects in their context. The method in Tang et al. (2015c) uses two dependent long short-term memory to model the semantic relatedness of a target with its context words in a sentence. [Ma et al. (2017); Wang et al. (2016)] employ LSTM or bidirectional LSTM to encode sentence information and add attention mechanism to obtain the aspect-specific representations for aspect-level sentiment classification. Xue and Li (2018) propose a model based on convolutional neural networks and gating mechanisms, which is more accurate and efficient for this task. Document-level sentiment classification (without aspect information) is also a related research field. Tang et al. (2015b) introduce a neural network to learn sentence representation with convolutional neural network or long short-term memory and generate document representation with gated recurrent neural network. Yang et al. (2016) propose a hierarchical attention network for document classification. [Chen et al. (2016); Wu et al. (2018)] apply hierarchical neural networks with user attention and product attention to generate document-level representations.

2.2. Memory Network

Memory network is a machine learning model first introduced by Weston et al. (2014), which has four basic components: *input(I), generalization(G), output(O) and response(R).* Then, Sukhbaatar et al. (2015) introduce end-to-end memory networks with a recurrent attention model over external memories, which is used to solve question answering and language modeling tasks. Nowadays, memory networks are widely used in the NLP field, such as question answering, sentiment classification, recommend system and so on. In question answering field, Miller et al. (2016) propose a key-value memory network to solve knowledge-base question answering task, which uses *(key, value)* vector pair as memories to represent documents. Kumar et al. (2016) introduce dynamic memory network with episodic memory module to process raw input-question-answer data for general question answering tasks. Then, Xiong et al. (2016) apply dynamic memory networks in visual and textual question answering for handling other forms of data such as pictures (non-text data). And in sentiment classification field, [Chen et al. (2017); Tang et al. (2016)] employ deep memory network with multiple attention layers to build text representations for aspect-level sentiment classification. Memory network has also achieved great success in recommend system. To address implicit collaborative filtering, Ebesu et al. (2018) use external memory and attention mechanism to learn user-item relations. Chen et al. (2018)

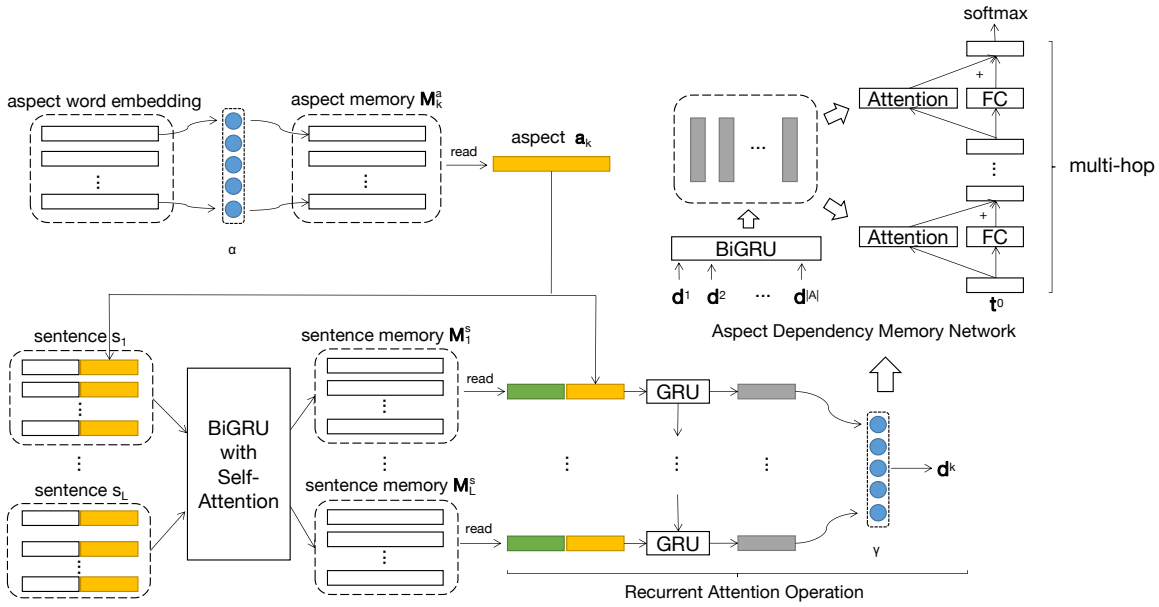


Figure 1: The overall architecture of our proposed Attentive Memory Network.

embed user's historical preferences with item-level and feature-level memory network for sequential recommendation. Huang et al. (2018) incorporate knowledge base information to enhance the representations of memory network, and combine RNN-based models and memory network to improve sequential recommendation performance.

3. Method

In this section, we first introduce the problem formalization of DMSC. Then, we describe our proposed model in detail.

3.1. Formalization

We use $D = \{d_1,d_2,...,d_{|D|}\}$ to denote all review documents and $A = \{a_1,a_2,...,a_{|A|}\}$ to represent all aspects. Each instance d of D consists of l sentences $\{s_1,s_2,...,s_l\}$ and the i-th sentence $s_i = \{w_{i1},w_{i2},...,w_{in}\}$. The j-th aspect a_j in A contains a keyword set $\{w^a_{j1},w^a_{j2},...,w^a_{jm}\}$. DMSC aims to predict aspect ratings of these reviews according to their text information.

3.2. Attentive Memory Network for DMSC

In this section, we present our attentive memory network in detail. Fig. 1 shows the architecture of our model. Specifically, our model consists of three components: Aspect and Sentence Memory Building that encodes aspect and sentence information, Recurrent Attention Operation on Memory that generates the aspect-aware document representations and Aspect Dependency Memory Network that incorporates the surrounding aspects related information into final predictions.

3.2.1. Aspect and Sentence Memory Building

Given the i-th sentence $s_i = \{w_{i1},w_{i2},...,w_{in}\}$ in document d and aspect a_k that includes a set of words $\{w^a_{k1},w^a_{k2},...,w^a_{km}\}$, we first map each word in sentence and aspect into a low-dimensional real-value vector. All the word vectors are stacked in a word embedding matrix $E \in \mathbb{R}^{d\times|V|}$, where d is the dimension of word vector and $|V|$ is vocabulary size. To model the importance of each keyword of aspect a_k, we adopt self-attention module to learn the weights of the keywords automatically which are calculated as:

$$\alpha = softmax(W_{a2}\tanh(W_{a1}a_k)) \tag{1}$$

$$\mathbf{m}^a_{ki} = \alpha_i \mathbf{w}^a_{ki} \tag{2}$$

where $W_{a1} \in \mathbb{R}^{d\times d}$ and $W_{a2} \in \mathbb{R}^d$ are weight matrixes, the vector α contains the learned weights of aspect keywords. Then, all weighted keyword vectors $\{\mathbf{m}^a_{k1},\mathbf{m}^a_{k2},...,\mathbf{m}^a_{km}\}$ of aspect a_k are stacked and regarded as the aspect memory $\mathbf{M}^a_k \in \mathbb{R}^{d\times m}$.

To capture internal structures of sentences with given aspect a_k and make full use of aspect information, we first obtain the aspect embedding $\mathbf{a}_k$ which is read from aspect memory $\mathbf{M}^a_k$ as:

$$\mathbf{a}_k = \sum_i \mathbf{m}^a_{ki} \tag{3}$$

Then, we concatenate $\mathbf{a}_k$ to all words in the sentence $s_i = \{\mathbf{w}_{i1} \oplus \mathbf{a}_k,\mathbf{w}_{i2} \oplus \mathbf{a}_k,...,\mathbf{w}_{in} \oplus \mathbf{a}_k\}$ and a bi-direction recurrent neural network with self-attention Lin et al. (2017) is employed to encode sentence information. The sentence s_i is fed to a bidirectional GRU for gaining context dependency between words in a sentence:

$$\overrightarrow{\mathbf{h}}_{ij} = \overrightarrow{GRU}_s(\mathbf{w}_{ij}) \tag{4}$$

$$\overleftarrow{\mathbf{h}}_{ij} = \overleftarrow{GRU}_s(\mathbf{w}_{ij}) \tag{5}$$

And the final representation $\mathbf{h}_{ij} = [\overrightarrow{\mathbf{h}}_{ij}, \overleftarrow{\mathbf{h}}_{ij}]$. For simplicity, we note all the $\mathbf{h}_{ij}$ as $\mathbf{H}_i = (\mathbf{h}_{i1},\mathbf{h}_{i2},...,\mathbf{h}_{in})$. Then, the self-attention mechanism takes the whole GRU hidden states $\mathbf{H}_i$ as input and outputs a vector of weights β:

$$\beta = softmax(W_{s2}\tanh(W_{s1}\mathbf{H}_i)) \tag{6}$$

$$\mathbf{m}^s_{ij} = \beta_j \mathbf{h}_{ij} \tag{7}$$

where $W_{s1} \in \mathbb{R}^{2d_h\times 2d_h}$ and $W_{s2} \in \mathbb{R}^{2d_h}$ are weight matrixes, d_h is the hidden size of GRU and the sentence memory $\mathbf{M}^s_i$ stacks all weighted hidden states $\mathbf{m}^s_{ij}$ as $\mathbf{M}^s_i = \{\mathbf{m}^s_{i1},\mathbf{m}^s_{i2},...,\mathbf{m}^s_{in}\}$.

3.2.2. Recurrent Attention Operation on Memory

After obtaining aspect memory $\mathbf{M}^a_k$ and sentence memory $\mathbf{M}^s_i$ which encodes aspect and sentence information, we notice that different sentences in documents d may have different influences with given aspect a_k. Therefore, we design a recurrent attention operation on aspect and sentence memory. First of all, we obtain the sentence embedding $\mathbf{s}_i$ which is read from sentence memory $\mathbf{M}^s_i$:

$$\mathbf{s}_i = \sum_j \mathbf{m}^s_{ij} \tag{8}$$

Then, we also concatenate $\mathbf{a}_k$ to all sentences in documents $d = \{\mathbf{s}_1 \oplus \mathbf{a}_k, \mathbf{s}_2 \oplus \mathbf{a}_k, ..., \mathbf{s}_l \oplus \mathbf{a}_k\}$ and the sentences are fed to a unidirectional GRU to capture long-distance dependency across sentences:

$$\mathbf{h}_i^d = GRU_d(\mathbf{s}_i) \tag{9}$$

After obtaining hidden representation $\mathbf{h}_i^d$, to measure the importance of the sentences, the attention mechanism is applied to obtain the aspect-aware document representation. The document representation $\mathbf{d}^k$ for aspect a_k is calculated as:

$$\gamma = softmax(W_{d2} \tanh(W_{d1}\mathbf{h}^d)) \tag{10}$$

$$\mathbf{d}^k = \sum_i \gamma_i \mathbf{h}_i^d \tag{11}$$

where $W_{d1} \in \mathbb{R}^{d_h \times d_h}$ and $W_{d2} \in \mathbb{R}^{d_h}$ are weight matrixes, γ_i measures the importance of the i-th sentence for aspect a_k and $\mathbf{d}^k$ is the aspect-aware document representation that summarizes all the information of sentences in a document.

3.2.3. Aspect Dependency Memory Network

We obtain the aspect-aware document representations $\{\mathbf{d}^1, \mathbf{d}^2, ..., \mathbf{d}^{|A|}\}$ for all aspects. In this work, we suppose that the related information from surrounding aspects is useful for sentiment predictions of aspects. Inspired by Tang et al. (2016), we propose multi-hop attention memory network for modeling aspect dependency. Different from the method in Tang et al. (2016), we use the bi-directional GRU for memory building rather than word embedding vectors as memory and replace linear operation with a fully connected layer. Before the aspect-aware document representations $\{\mathbf{d}^1, \mathbf{d}^2, ..., \mathbf{d}^{|A|}\}$ are fed to bi-directional GRU, inspired by Li et al. (2018), the document representations are concatenating with the overall rating $\mathbf{r}$ of document d which is helpful for aspect rating predictions.

The multi-hop attention memory networks have the following steps:

- **Input.** The bi-directional GRU is adopted to construct memory over aspect-aware document representations $\{\mathbf{d}^1, \mathbf{d}^2, ..., \mathbf{d}^{|A|}\}$ as following:

$$\overrightarrow{\mathbf{h}_i^m} = \overrightarrow{GRU}_m(\mathbf{d}^i) \tag{12}$$

$$\overleftarrow{\mathbf{h}_i^m} = \overleftarrow{GRU}_m(\mathbf{d}^i) \tag{13}$$

  And the memory representation $\mathbf{m}_i^d = [\overrightarrow{\mathbf{h}_i^m}, \overleftarrow{\mathbf{h}_i^m}]$, all representations are stacked as external memory $\mathbf{M}^d = \{\mathbf{m}_1^d, \mathbf{m}_2^d, ..., \mathbf{m}_{|A|}^d\}$. The target aspect-aware representation $\mathbf{m}_{tag}^d$ is transformed as initial state $\mathbf{t}_{tag}^{(0)}$.

- **Multi-hop attention.** The multi-hop attention mechanism is used to select related information from other aspects over memory $\mathbf{M}^d$. At the h-th hop, the similarity between target $\mathbf{t}_{tag}^{(h-1)}$ and the memory slots in $\mathbf{M}^d$ is calculated as:

$$\alpha_{tag}^{(h)} = softmax(W_m \mathbf{M}^d + W_{tm}\mathbf{t}_{tag}^{(h-1)} + b_m) \tag{14}$$

where W_m, W_{tm} and b_m are trainable parameters shared in different hops. Here, $\alpha_{tagi}^{(h)}$ measures the relatedness between target aspect and aspect i. And the output $\mathbf{o}_{tag}^{(h)}$ of the h-th hop is obtained by summing the weighted memory slots in $\mathbf{M}^d$:

$$\mathbf{o}_{tag}^{(h)} = \sum_i \alpha_{tagi}^{(h)} \mathbf{m}_i^d \tag{15}$$

And the target representation $\mathbf{t}_{tag}^{(h-1)}$ is transformed with a fully connected layer:

$$\mathbf{t}_{tag}^{(h-1)} = \tanh(W_t \mathbf{t}_{tag}^{(h-1)} + b_t) \tag{16}$$

where $W_t \in \mathbb{R}^{2d_h \times 2d_h}$ and $b_t \in \mathbb{R}^{2d_h}$. Then, the target representation $\mathbf{t}_{tag}^{(h)}$ is undated at the end of the h-th hop:

$$\mathbf{t}_{tag}^{(h)} = \mathbf{t}_{tag}^{(h-1)} + \mathbf{o}_{tag}^{(h)} \tag{17}$$

3.3. Output and Model Training

We apply H hops in our model. After H hops, all the states $\{\mathbf{t}_{tag}^{(h)}\}_{1 \le tag \le |A|}$ are used for the final prediction over $|Y|$ classes which contain aspect-aware document information and related aspect information:

$$\hat{\mathbf{p}}(d, tag) = softmax(W_{tag}^c \mathbf{t}_{tag}^{(h)} + b_{tag}^c) \tag{18}$$

where $W_{tag}^c \in \mathbb{R}^{2d_h \times |Y|}$ and $b_{tag}^c \in \mathbb{R}^{|Y|}$ are parameters of softmax layer. In our model, cross-entropy error between gold sentiment distribution and our model's sentiment distribution is defined as loss function (L) for optimization when training:

$$L = -\sum_{d \in D} \sum_{tag=1}^{|A|} \sum_{i=1}^{|Y|} \mathbf{p}(d, tag) \cdot \log(\hat{\mathbf{p}}(d, tag)) \tag{19}$$

where $\mathbf{p}(d, tag)$ is a one-hot vector that the dimension of ground truth being 1 and others being 0.

4. Experiments

In this section, we conduct empirical experiments to verify the effectiveness of our model on DMSC. We first introduce the experimental settings, then we compare our model to the baseline methods to demonstrate its effectiveness on DMSC.

4.1. Dataset

We conduct our experiments on two real-word datasets namely TripAdvisor and BeerAdvocate. We tokenize the datasets by Stanford Corenlp[1] and the statistics of the datasets are described in Table 1. The used datasets are randomly split into training, development, and testing sets with 80/10/10%.

1. http://nlp.stanford.edu/software/corenlp.shtml

Table 1: The statistics of the TripAdvisor and BeerAdvocate datasets.

Dataset	#docs	#words/doc	#words/sent
TripAdvisor	29,391	251.7	18.0
BeerAdvocate	51,020	144.5	12.1

TripAdvisor reviews are obtained from Yin et al. (2017), which contain seven aspects (*value, room, location, cleanliness, check in/front desk, service, and business service*) and the aspect rating scale is 1-5.
BeerAdvocate reviews are obtained from Yin et al. (2017), which contain four aspects (*feel, look, smell, and taste*) and the aspect rating scale is 1-10.

4.2. Baseline Methods

We evaluate our model with several baseline methods:

- **Majority** is a basic baseline method, which assigns the majority sentiment label in the training set to each review document in the test set.
- **SVM** uses unigram, bigram text features and trains with LibLinear.
- **NBoW** uses the mean embedding of all words in a document as features which are fed into SVM classifier.
- **DAN** Iyyer et al. (2015) is a deep averaging network with several fully connected layers and uses the mean embedding of all words as input where applies word dropout to improve robustness.
- **CNN** Kim (2014) adopts convolution layer to extract words neighboring features over a sentence, then gets a fix-sized sentence representation by pooling layer.
- **LSTM** Tang et al. (2015a) uses Bi-LSTM to model the review document as a sequence and gets the forward and backward semantic information of each word, then averages the all semantic information as document representation to predict the sentiment polarity.
- **HAN** Yang et al. (2016) models the document in word and sentence level with a hierarchical structure and uses an attention mechanism to capture the important words and sentences.
- **DMSCMC** Yin et al. (2017) regards the task as machine comprehension problem and uses iterative attention modules to build up aspect-specific representations for reviews.
- **HARN** Li et al. (2018) incorporates aspect information and overall rating into a hierarchical network to obtain document representations, and obtains state-of-the-art results on document-level multi-aspect sentiment classification.

We extend **DAN**, **CNN**, **LSTM**, **HAN** with the hierarchical architecture and multi-task framework as **MHDAN**, **MHCNN**, **MHLSTM**, **MHAN**. And our model is customized into three versions below:

- **AMN-wD** refers to attentive memory network without modeling aspect dependency.
- **AMN-wA** utilizes mean aspect keyword embeddings rather than using aspect memory.
- **AMN-BiGRU** uses BiGRU to model aspect dependency without using multi-hop attention memory networks.

4.3. Experimental Settings

In our experiments, the word embeddings are initialized with Glove[2] pre-trained vectors, whose embedding size is 200. For each aspect, we get the aspect keyword sets from Yin et al. (2017). The dimension of hidden state is set to 150. The model hyper-parameters are tuned based on the development set and trained using Adam with a learning rate of 0.001. To avoid model over-fitting, we use dropout with rate of 0.2.

We use *Accuracy* which measures DMSC performance and *Mean Squared Error (MSE)* to measure the divergences between predicted sentiment classes and ground truth classes as the evaluation metrics. The *Accuracy* and *MSE* metrics are defined as:

$$Accuracy = \frac{T}{N} \tag{20}$$

$$MSE = \frac{\sum_{i=1}^{N}(gt_i - pr_i)^2}{N} \tag{21}$$

where T means the number of predicted sentiment classes that are same with the ground truth sentiment classes, N is the number of the review documents for all aspects, gt_i represents the ground truth sentiment rating and pr_i represents the predicted sentiment rating.

4.4. Comparison Results with Baselines

We compare our model to the baseline methods on DMSC. The performance results over the two datasets are shown in Table 2. From Table 2, we have the following observations from the results:(1)Among the traditional models, the NBoW achieves the highest accuracy in both datasets which shows that the word embedding features are more effective than traditional n-gram features.(2)The neural network models outperform the traditional models which shows the superiority of neural network such as CNN extracts neighbored text features, LSTM models the text as a sequence and HAN encodes documents in hierarchical structures with attention mechanism.(3)And the MH-model is better than original model (such as MHLSTM is better than LSTM) that shows the multi-task framework and hierarchical structures are beneficial for DMSC.(4)DMSCMC considers the DMSC as a machine comprehension problem and regards the aspect as query to select the important information from reviews using an iterative attention module, the performance of which is outstanding

2. https://nlp.stanford.edu/projects/glove/

Table 2: Comparison of our model and other baseline methods on TripAdvisor and BeerAdvocate dataset for document-level multi-aspect sentiment classification.

Model	TripAdvisor		BeerAdvocate	
	Accuracy	MSE	Accuracy	MSE
Majority	23.89	2.549	24.41	4.545
SVM	35.26	1.963	25.79	3.270
NBoW	39.09	1.808	28.85	2.919
DAN	40.93	1.531	32.44	2.279
CNN	43.35	1.456	33.37	2.217
LSTM	44.02	1.470	34.78	2.097
HAN	44.68	1.301	36.03	1.920
MHDAN	42.47	1.549	32.54	2.376
MHCNN	43.79	1.398	35.33	1.976
MHLSTM	44.72	1.272	37.04	1.809
MHAN	44.94	1.210	36.82	1.813
DMSCMC	46.56	**1.083**	38.06	1.755
HARN	47.43	1.169	39.11	1.700
AMN	**48.66**	1.109	**40.19**	**1.686**

than neural network based models.(5)Different from DMSCMC, HARN utilizes hierarchical structures to encode documents and incorporates aspect information that is embedded as attention and overall ratings into the model, which results in HARN obtains state-of-the-art performance.(6)Our model encodes the aspects and sentences with memory network, and integrated with the related information from neighboring aspects by using multi-hop attention memory networks. Compared with DMSCMC and HARN, our proposed model achieves improvements of 1.2% and 1.0% in accuracy on TripAdvisor and BeerAdvocate.

To investigate the effects of aspect memory, modeling aspect dependency and our multi-hop attention memory networks, we also compare the all visions of our model to recent SOTA models on DMSC. The results are shown in Table 3. According to the results of Table 3, we can observe that:

- Compared to the model DMSCMC and HARN, almost all versions of our model have improvements on DMSC, which shows that utilizing the memory networks as our model's foundational structure is helpful for document-level multi-aspect sentiment classification task.

- According to the results of AMN-wD and AMN-wA, where AMN-wA obtains higher accuracy than AMN-wD, the related aspect information is more effective than modeling the keyword importance of aspects. And compared to the AMN-wD and AMN-wA, the result of AMN-BiGRU is better than both of them, indicating that modeling aspect dependency within aspects and constructing aspect memory are beneficial to final aspect rating predictions.

Table 3: Comparison of the variants of our model and SOTA methods on TripAdvisor and BeerAdvocate dataset.

Model	TripAdvisor		BeerAdvocate	
	Accuracy	MSE	Accuracy	MSE
DMSCMC	46.56	**1.083**	38.06	1.755
HARN	47.43	1.169	39.11	1.700
AMN-wD	47.90	1.132	38.88	1.748
AMN-wA	48.17	1.120	39.33	1.752
AMN-BiGRU	48.29	1.122	39.47	1.750
AMN	**48.66**	1.109	**40.19**	**1.686**

- Compared to the AMN-BiGRU only using a BiGRU network to model aspect dependency, AMN achieves the better results, which validates the effectiveness of multi-hop attention memory networks on mining the surrounding aspects related information. And our AMN model integrated with all kinds of information obtains the best results compared with all versions of AMN.

4.5. Effect of Recurrent Attention Operation

In this section, we explain the effectiveness of recurrent attention operation over memories from two sides: the visualization of sentences attention weights and the influence of document length which refers to the number of sentences in a document.

4.5.1. The Attention Weights of Sentences

We sample a review document from TripAdvisor and visualize the aspect attention for case study. The visualization of sentence weights is shown in Fig. 2 (the second sentence is split into two lines due to its length). From Fig. 2, it can be seen that our model has the ability to choose the important sentence from the review for different aspects. For aspects "*value*" and "*room*", the second sentence "*the room is good, but the price is expensive.*" is obviously higher weights than the first and the third sentence because it contains the aspect-specific words such as *room*, *price* and their sentiment words. And for aspects "*stuff*" and "*service*", the third sentence has the highest weights due to the aspect-aware word *service* and sentiment word *normal*. It shows that the recurrent attention operation can capture the aspect-specific meanings in sentences and obtain the aspect-aware document representations.

4.5.2. Influence of Document Length

To investigate the effectiveness of recurrent attention operation over memories, we also explore the performance of our model on accuracy metric under various document lengths in TripAdvisior and BeerAdvocate. The performance of our model shows that the recurrent attention operation can capture the long-distance context dependency across sentences and

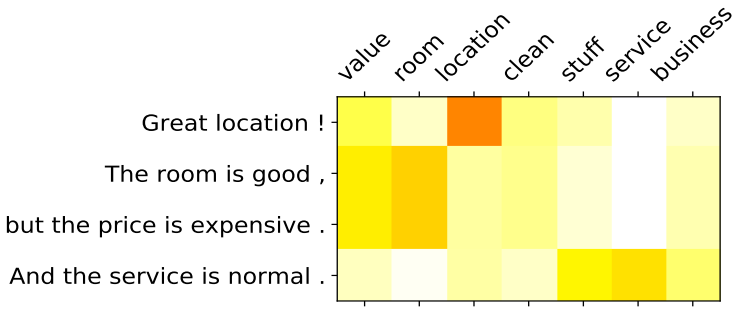
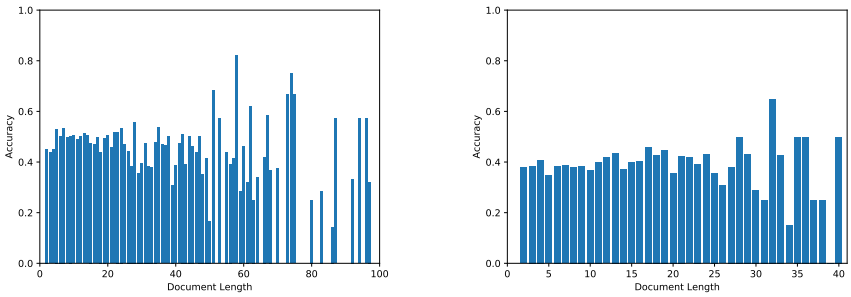


Figure 2: The attention visualization of sentences. Darker color means higher weight.



(*a*) TripAdvisor (*b*) BeerAdvocate

Figure 3: The accuracy results over different document lengths in TripAdvisior and BeerAdvocate.

the results in term of accuracy are shown in Fig. 3. The results in Fig. 3 have shown that our model can obtain good results when the document length is large (such as document length ¿ 80 in TripAdvisor and document length ¿ 30 in BeerAdvocate), illustrating that our model can gain the context information through long-length documents. On the other hand, the performance of our model is stable as the length of documents increases, indicating that our proposed model has good adaptability to both short and long reviews.

4.6. Aspect Dependency Memory Network Performance

In this section, to validate the performance of aspect dependency memory network, we try different hops with 1 to 5 of aspect dependency memory network and visualize the attention weights at different hops. The results of different hops of aspect dependency memory network are shown in Table 4. From Table 4, it can be observed that our model

Table 4: The results of different hops of aspect dependency memory network on TripAdvisor and BeerAdvocate datasset.

Model	TripAdvisor		BeerAdvocate	
	Accuracy	MSE	Accuracy	MSE
AMN-hop1	48.36	1.130	39.30	1.753
AMN-hop2	**48.66**	**1.109**	39.44	1.716
AMN-hop3	48.26	1.120	**40.19**	**1.686**
AMN-hop4	47.93	1.141	39.76	1.723
AMN-hop5	47.74	1.140	38.82	1.779

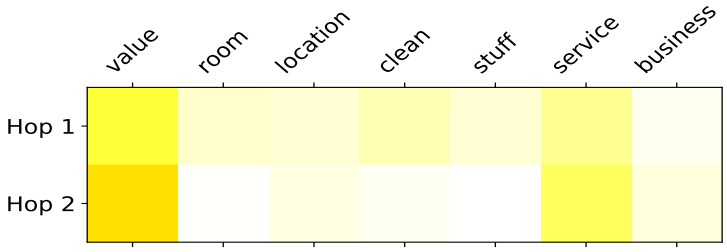


Figure 4: The attention weights of reviews "*The hotel is not worth the price, it took us two hours to be served.*" in aspect dependency memory network and the target aspect is "*service*". Darker color means higher weight.

generally works better with 2 or 3 hops. And our model with 1 hop performs not good as using more hops which shows that multi-hop attention mechanism can sufficiently capture the neighbored aspect related information than one-time attention mechanism. Otherwise, the performance is not getting better as the number of hop increases, such as the results of AMN-hop5 are worse than AMN-hop4. It may be because of that as the model's hop increases, the model becomes complex and less generalizable.

Then we visualize the attention weights of reviews "*The hotel is not worth the price, it took us two hours to be served.*" at different hops, where the target aspect is "*service*" and the visualization is shown in Fig. 4. The Fig. 4 shows that our aspect dependency memory network can sufficiently capture the related sentiment information from surrounding aspects. For target aspect "*service*", the weights of aspect "*value*" and aspect "*service*" document representations are higher than other aspect-aware representations at the same hop, which shows that our model can not only discover the document representation of the target

aspect, but also mine important information from the surrounding aspects. And the 2rd-hop attention weights of aspect "*value*" and aspect "*service*" are higher than the 1-st hop attention weights of themselves which illustrates once again that the multi-hop attention mechanism can effectively select the important information from other aspects than one-time attention mechanism.

5. Conclusion

In this paper, we propose an attentive memory network for document-level multi-aspect sentiment classification. To model the keywords importance of aspects and capture the internal structures of sentences, we construct aspect and sentence memories with attention-based memory networks. Then, the recurrent attention mechanism is adopted to capture long-distance context information across sentences and obtain aspect-aware document representations. For mining the neighboring aspect features, we introduce a multi-hop attention memory network. Extensive experiments show that our model outperforms state-of-the-art methods significantly.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2018YFB1003903), National Natural Science Foundation of China (No. 61502033, 61472034, 61772071, 61272361 and 61672098) and the Fundamental Research Funds for the Central Universities.

References

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659, 2016.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461, 2017.

Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. Sequential recommendation with user memory networks. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 108–116. ACM, 2018.

Travis Ebesu, Bin Shen, and Yi Fang. Collaborative memory network for recommendation systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 515–524. ACM, 2018.

Devamanyu Hazarika, Soujanya Poria, Prateek Vij, Gangeshwar Krishnamurthy, Erik Cambria, and Roger Zimmermann. Modeling inter-aspect dependencies for aspect-based sentiment analysis. In *Proceedings of the 2018 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 266–270, 2018.

Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. Improving sequential recommendation with knowledge-enhanced memory networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 505–514. ACM, 2018.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691, 2015.

Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387, 2016.

Junjie Li, Haitong Yang, and Chengqing Zong. Document-level multi-aspect sentiment classification by jointly modeling users, aspects, and overall ratings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 925–936, 2018.

Zhouhan Lin, Minwei Feng, Cicero Nogueira Dos Santos, Yu Mo, and Yoshua Bengio. A structured self-attentive sentence embedding. 2017.

Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th international conference on data mining workshops*, pages 81–88. IEEE, 2011.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*, 2015.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*, 2017.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE, 2012.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*, 2016.

Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*, 2015a.

Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015b.

Duyu Tang, Bing Qin, and Ting Liu. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1014–1023, 2015c.

Duyu Tang, Bing Qin, and Ting Liu. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900*, 2016.

Yequan Wang, Minlie Huang, Li Zhao, et al. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615, 2016.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

Zhen Wu, Xin-Yu Dai, Cunyan Yin, Shujian Huang, and Jiajun Chen. Improving review representations with user attention and product attention for sentiment classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406, 2016.

Wei Xue and Tao Li. Aspect based sentiment analysis with gated convolutional networks. *arXiv preprint arXiv:1805.07043*, 2018.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

Yichun Yin, Yangqiu Song, and Ming Zhang. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2054, 2017.