Proceedings of Machine Learning Research 95:630-645, 2018 ACML 2018

A Self-Attentive Hierarchical Model for Jointly Improving Text Summarization and Sentiment Classification

Hongli Wang WANGHLI8@MAIL2.SYSU.EDU.CN
Jiangtao Ren ISSRJT@MAIL.SYSU.EDU.CN
Schol of Data and Computer Science, Sun Yat-sen University, Guangdong, P.R.China 510006

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

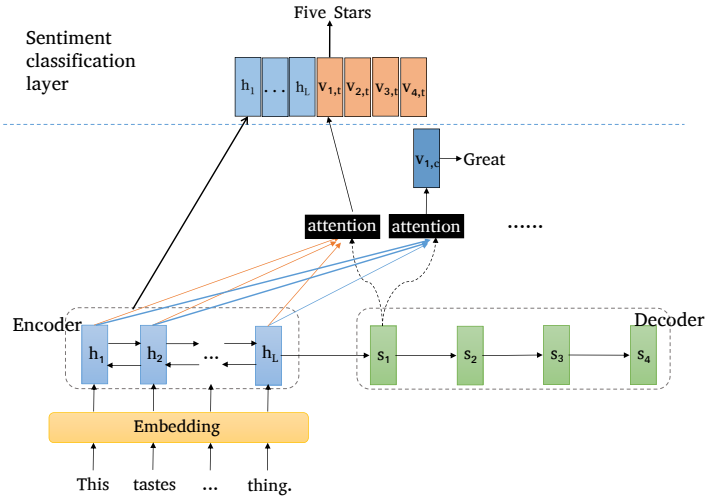
Text summarization and sentiment classification, in NLP, are two main tasks implemented on text analysis, focusing on extracting the major idea of a text at different levels. Based on the characteristics of both, sentiment classification can be regarded as a more abstractive summarization task. According to the scheme, a Self-Attentive Hierarchical model for jointly improving text Summarization and Sentiment Classification (SAHSSC) is proposed in this paper. This model jointly performs abstractive text summarization and sentiment classification within a hierarchical end-to-end neural framework, in which the sentiment classification layer on top of the summarization layer predicts the sentiment label in the light of the text and the generated summary. Furthermore, a self-attention layer is also proposed in the hierarchical framework, which is the bridge that connects the summarization layer and the sentiment classification layer and aims at capturing emotional information at text-level as well as summary-level. The proposed model can generate a more relevant summary and lead to a more accurate summary-aware sentiment prediction. Experimental results evaluated on SNAP amazon online review datasets show that our model outperforms the state-of-the-art baselines on both abstractive text summarization and sentiment classification by a considerable margin.

Keywords: Abstractive text summarization, Sentiment classification, Hierarchical end-to-end framework, Self-attention mechanism

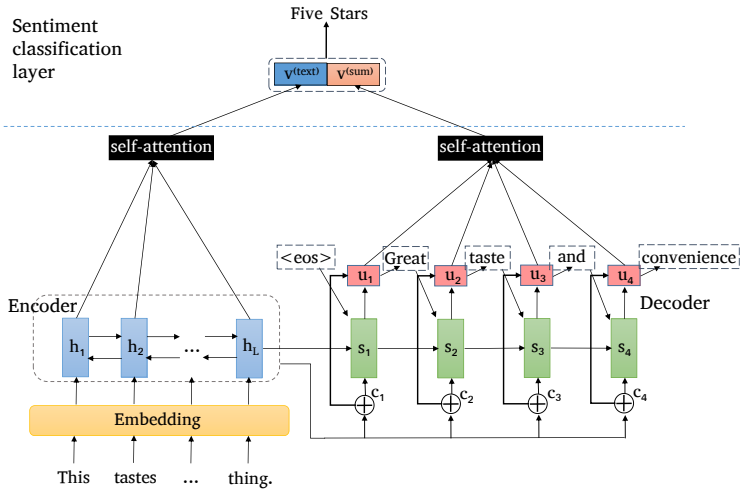
1. Introduction

Text summarization and sentiment classification are two of the most active and fundamental tasks in natural language processing (NLP), which are widely applied to analyze the textual materials in practical scenarios such as online news, website articles, and user reviews. Text summarization aims to create a representative summary with the major points of an original text. In general, there are two approaches to automatic summarization: extraction and abstraction. In this study, the abstraction-based summarization is mainly discussed and analyzed. Compared with the extraction-based summarization, which forms the summary by selecting a subset of existing words or phrases in the original text, the abstraction-based methods build an internal semantic representation and then apply natural language generation techniques to create a summary closer to a human-generated one. As for sentiment classification, generally known as sentiment analysis or opinion mining, it aims to extract and quantify affective states and subjective information of the writers from their texts by a series of sentiment labels. The summary expresses the major idea of an original text in

a shorter length whereas the sentiment label further summarizes the sentiment tendency of the text.



(*a*) Ma *et al.* (HSSC)



(*b*) this work (SAHSSC)

Figure 1: Comparison between the work proposed by Ma *et al.* and our proposed model.

There are lots of works in the research of text summarization and classification, but most of the existing models are simply built for one of them. In past years, there have been several systems of text analysis [Hole and Takalikar (2013); Mane et al. (2015)], which have been able to produce the summary and the sentiment label from the source content by lots of hand-crafted features. Some previous studies have extracted the important part of an original text to analyze the sentiment by summarization methods [Shetty and Bajaj (2015); Bhargava and Sharma (2017)], and one work trained text classification and the summarization jointly to improve the performance of summarization [Cao et al. (2017)].

However, these works only focus on either text summarization or sentiment classification. Unlike the previous works, Ma et al. (2018) have first jointly improved the two tasks within an end-to-end neural network-based framework, regarding the sentiment classification as a more abstract type of summarization.

As shown in Fig.1(*a*), the hierarchical model proposed by Ma et al. (2018) (namely `HSSC`) is composed of two recurrent neural networks (RNNs) – an encoder and a decoder, a MLP, as the classification layer, and two independent general attention mechanisms integrated on decoder for extracting different representation of the texts for word generation and classification. The model first encodes the complete source text, and then decodes one word at a time from the learned text representation for word generation. After generating the complete summary, the model predicts the sentiment label from a series of the learned text representations for sentiment classification. The hierarchical end-to-end framework is able to greatly combine the summarization layer and the sentiment classification layer, but it still exists insufficiency. It should be noted that the model still does not fully utilize the information of the generated summary, although it uses the hidden state of decoder RNN as extra information to guide the extraction of the text representations for sentiment classification by attention mechanism.

In this study, the generated summary from summarization is further explored, and a variant hierarchical framework is proposed towards jointly improving summarization and sentiment classification, which has a similar motivation but achieved differently, presented in Fig.1(*b*). Specifically, the proposed model generates the summary by an attention-based encoder-decoder layer, and then predicts the sentiment label based on an original text as well as the generated summary. Enhanced by such scheme, the model gains the information of sentiment from different levels of the text, which helps to make accurate and effective judgments on sentiment. In addition, the supervision of the generated summary by sentiment classification guides the summary decoder to generate the summary that has the same sentiment tendency as the original text.

To improve the model, the self-attention-based hierarchical framework instead of the attention-based scheme in Fig.1(*a*) is adopted, since the self-attention is a scalable attention mechanism and can also guide the attention of source content without extra information in some cases, like sentiment classification [Lin et al. (2017)]. Specifically, the self-attention layer between summarization layer and sentiment classification layer is employed to obtain the embedding representation of text and summary for sentiment classification in the hierarchical model. Compared with the attention mechanisms in Fig.1(*a*), which aim to extract the information of text for summarization and sentiment classification, the self-attention mechanisms in this study focus on the information of sentiment from the original text and the generated summary.

Overall, the contribution of this paper is to propose a *Self-Attentive Hierarchical model for jointly improving text Summarization and Sentiment Classification* (SAHSSC). Although our work is not the pioneer towards improving both the two tasks within an end-to-end neural framework, the proposed model has two improvements over the work proposed by Ma et al. (2018): (1) it further makes use of the generated summary from summarization in the joint task, enabling the hierarchical structure to build a close bond between the two tasks; (2) while creating the representations of text and summary, two different self-attention mechanisms are applied on encoder and decoder, which extract the information of sentiment

from text and the generated summary. In order to evaluate the performance of our model in comparison to the common state-of-the-art models, we experiment on Amazon online reviews datasets (SNAP). It shows that our model outperforms the current state-of-the-art models in multiple metrics on both abstractive summarization and sentiment classification.

The rest of the paper is organized as follows. In section 2, the proposed model is presented in details. Section 3 describes the experiments and the results. The related work is briefly described in section 4. In the end, the conclusions are drawn in section 5.

2. The Proposed Model

In this section, we first briefly give the problem formulation, and then introduce the proposed model in details. Finally, we present the overall loss function for training.

2.1. Problem Formulation

Given a review data pair (X, Y, l), where X, Y and l separately denote the original review text, the corresponding summary and the corresponding sentiment level, our model aims to map from the source text X to Y and l. Specifically, both the original content X and the summary Y are sequences of words:

$$X = \{x_1, x_2, ..., x_L\}$$

$$Y = \{y_1, y_2, ..., y_M\}$$

where L and M denote the number of words in the sequences X and Y, respectively. The label $l \in \{1, 2, ..., K\}$ denotes the level of sentiment of the original text X, from the lowest rating 1 to the highest rating K.

2.2. Summarization Layer

In the proposed model, the summarization layer is a standard Seq2Seq model with attention mechanism. The first idea of Seq2Seq was proposed to translate one sequence to another sequence through an encoder-decoder neural architecture in machine translation [Bahdanau et al. (2014)]. Recently, abstractive summary generation has been treated as sequence translation from an original text to a summary [Rush et al. (2015); See et al. (2017); Paulus et al. (2017)].

Formally, given a review text $X = \{x_1, x_2, ..., x_L\}$ represented into a sequence of word embeddings, the text encoder first reads the words in $\boldsymbol{x}$ and encodes them into a series of context vectors $H = (h_1, h_2, ..., h_L)$ though a bidirectional Long Short-term Memory Network (BiLSTM) in our model. The BiLSTM includes contextual information from past and future words into the vector representation h_t of a particular word vector x_t, as follows:

$$h_t = \overrightarrow{h_t} + \overleftarrow{h_t} \tag{1}$$

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(x_t, \overrightarrow{h}_{t-1}) \tag{2}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(x_t, \overleftarrow{h}_{t+1}) \tag{3}$$

where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are the hidden outputs of the forward LSTM and the backward passes of the BiLSTM respectively.

And then, the summary decoder sequentially generates a summary $Y = \{y_1, y_2, ..., y_M\}$ with context vectors as input, formally defined as follow:

$$p(Y|X) = \prod_{t=1}^{M} p(y_t|c_t, y_1, ..., y_{t-1}) \tag{4}$$

At t-th time step, the decoder RNN generates one word conditioned on the context vector c_t extracted by attention mechanism [Bahdanau et al. (2014)] and the decoder hidden state s_t. The generation probability of the t-th word can be calculated as:

$$p(y_t|X) = softmax(W_g u_t) \tag{5}$$

$$u_t = tanh(W_{st} s_t + W_{ct} c_t) \tag{6}$$

$$s_t = \overrightarrow{LSTM}(c_t, y_{t-1}, s_{t-1}) \tag{7}$$

where y_{t-1}, s_{t-1} are the last generated words and the hidden state of decoder LSTM at *t-1*-th time step, separately; W_g, W_{st} and W_{ct} are parameter matrices. And, given the context vectors H, the attention mechanism computes an attentive context vector c_t at t-th time step, which allows the decoder to get full information of the source text. c_t is computed as:

$$c_t = \sum_{i=1}^{N} \alpha_{ti} h_i \tag{8}$$

$$\alpha_{ti} = \frac{exp(e_{ti})}{\sum_{j=1}^{N} exp(e_{tj})} \tag{9}$$

$$e_{tj} = tanh(s_{t-1}^T W_t h_j) \tag{10}$$

where c_t is a weighted sum of context vector h_i in H, and the weight α_{ti} for the each h_i is sequentially computed by Equation 9 and 10; the W_t is a trainable parameter matrix.

2.3. Self-Attention Layer

Between the summarization layer and sentiment classification layer, the proposed self-attention layer consists of two independent self-attention mechanisms. The self-attention is a special case of the attention mechanism, which models the dependencies between tokens from the same sequence [Lin et al. (2017); Vaswani et al. (2017); Shen et al. (2018)]. Such an attention mechanism is usually used for sentence representation which abstracts sentence-level meanings. In our model, the self-attention layer aims to create the embedding representations of the original review text X and the summary Y generated from the summarization layer. Specifically, given the text context memory $H = (h_1, h_2, ..., h_L)$ carrying the semantics of the original review along all time steps from the encoder, self-attention first yields a weight matrix $A^{enc} = (a_1^{enc}, a_2^{enc}, ..., a_L^{enc})$, computed as:

$$A^{enc} = softmax(w_2^{enc} tanh(W_1^{enc} H^T)) \tag{11}$$

where w_1^{enc} is a parameter vector and W_2^{enc} is a parameter matrix. The *softmax()* is used to normalize the attention weights to sum up to 1.

Then, weighted by A^{enc}, we obtain the text vector representation $v^{(text)}$ by computing a weighted sum of H:

$$v^{(text)} = A^{enc}H \tag{12}$$

For the summary generated from decoder, the model collects the word representations of all time steps into the summary context memory $U = (u_1, u_2, ..., u_M)$. Similar to the text vector, the summary vector representation $v^{(sum)}$ is computed as:

$$v^{(sum)} = A^{dec}U \tag{13}$$

$$A^{dec} = softmax(w_2^{dec}tanh(W_1^{dec}U^T)) \tag{14}$$

In our proposed model, both the learned vector representations of text and summary are used for sentiment classification, but they provide the sentiment information from two different granularities for the sentiment prediction.

2.4. Sentiment Classification Layer

The sentiment classification layer is a feed-forward MLP network. Given the text vector representation $v^{(text)}$ and the summary vector representation $v^{(sum)}$, the classification layer computes the probability distribution of the sentiment labels, as follow:

$$p(l|X) = softmax(W_{vt}v^{(text)} + W_{vs}v^{(sum)}) \tag{15}$$

where W_{vt} and W_{vs} are trainable parameter matrices. The logistics layer makes the final prediction with the top probability of the sentiment label.

2.5. Overall Loss Function

The proposed model is trained by minimizing a joint loss for summarization and sentiment classification, as following:

$$L = L_s + \lambda L_c \tag{16}$$

where

$$L_s = -\sum_t y_t \log p(y_t|X) \tag{17}$$

$$L_c = -l \log p(l|X) \tag{18}$$

Here, L_s and L_c separately denote the loss of summarization and that of sentiment classification, which both are the categorical cross entropy. And λ is a hyper-parameter to balance the two losses, $\lambda = 0.5$ is set in this work.

3. Experiments

To demonstrate the effectiveness of the proposed method, we conducted extensive experiments. In this section, we first introduce our experimental settings. Then, we report and analyze the experimental results on text summarization and sentiment classification compared with several popular baselines and the state-of-the-art model. Finally, we provide the further analysis by ablation study and visual interpretation of the proposed model.

3.1. Experimental Settings

3.1.1. Datasets

In the experiment, we evaluate on **SNAP Amazon Reviews Dataset** originally provided by He and McAuley [He and McAuley (2016)], a part of **S**tanford **N**etwork **A**nalysis **P**roject **(SNAP)**[1]. The dataset contains product reviews and metadata from Amazon[2], including 142.8 million reviews spanning May 1996 - July 2014. Raw data includes product, reviews content, user information, rating and summaries. In this work, we form the benchmark datasets with three subsets of **Toys & Games**, **Movies & TV** and **Gourmet_Foods** in the Amazon Reviews Datasets, and pair each review content with the corresponding summary and sentiment label from the three raw datasets. The statistics of three benchmark datasets used in our experiments is shown as Table 1.

Dataset	Total Size	# Review	# Summary	Sentiment
Toys & Games	≈167k	99.8	4.4	{1,2,3,4,5}
Gourmet_Foods	≈151k	93.0	4.5	{1,2,3,4,5}
Movies & TV	≈1,697k	161.1	4.9	{1,2,3,4,5}

Table 1: Statistics of the datasets. # denotes the average length.

3.1.2. Evaluation Metric

For abstractive summarization, the ROUGE (Recall-Oriented Understudy for Gisty Evaluation) metrics are used for the automatic evaluation of the generated summaries [Lin and Hovy (2003)]. ROUGE is based on the comparison of n-grams between the produced summary and reference summaries, as:

$$ROUGE_N = \frac{\sum_{S\in\{reference\ summaries\}} \sum_{gram_n\in S} Countmatch(gram_n)}{\sum_{S\in\{reference\ summaries\}} \sum_{gram_n\in S} (gram_n)} \tag{19}$$

Following previous work [Chopra et al. (2016); Rush et al. (2015); See et al. (2017)], our summarization evaluation is based on three variants of ROUGE[3], namely, ROUGE-1 (unigram), ROUGE-2 (bigram), ROUGE-L (longest-common substring) in the reported result.

For evaluation metric on sentiment classification, we use the category accuracy for predefined sentiment labels, which is the accuracy of five-class sentiment.

3.1.3. Implementation Details

In the experiments, we use a vocabulary of 50k words for both the original texts and summaries, and replace the OOV words with ***<unk>***. For *Toys & Games*, *Gourmet_Foods* and *Movies & TV* dataset, the word embedding dimension and the hidden size of our model are respectively set to 256, 256, 512. The word embedding is random initialized and learned from training. We conduct mini-batch training with batch size of 64 and randomly shuffle

1. http://snap.stanford.edu/data/web-Amazon.html
2. https://www.amazon.com
3. We obtain the ROUGE scores using the pyrouge package at https://pypi.org/project/pyrouge/.

the training data at every epoch in the training. And, we use the Adam optimization with the initial learning rate $lr = 0.0003$, momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$ to minimize the training loss [Kingma and Ba (2014)]. Following Rush et al. (2015), we split the learning rate lr by half if the validation loss doesn't improve for an epoch, and train the model for total 20 epochs. Moreover, at training time, we use dropout with different dropout rates p of 0.2, 0.2, 0.0 for Toys, Foods and Movies datasets to avoid overfitting [Gal and Ghahramani (2016)], and clip the gradients with a maximum gradient norm of 10.0 [Pascanu et al. (2013)].

3.1.4. Baselines

For the comprehensive comparison based on the evaluation metrics, we first choose several popular baselines on abstractive summarization or sentiment classification, which are comparable to our model.

For abstractive text summarization, following the previous work [Hu et al. (2015)], the baselines are as follows:

- Sequence-to-Sequence model (S2S): It uses an uni-LSTM layer to map the source content to a vector, and then uses another uni-LSTM layer to decode the target summary [Sutskever et al. (2014)].
- Attention-based Sequence-to-Sequence model (S2S-att): The standard Sequence-to-Sequence model with global attention mechanism [Bahdanau et al. (2014)].

For sentiment classification, we compare our model with three strong classifiers, as follows:

- BiLSTM: The BiLSTM model uses a bidirectional LSTM and max pooling across all the LSTM hidden outputs to get the embedding vector of source context, then uses a 1-layer MLP to output the classification result.
- CNN: The CNN model uses the same scheme as BiLSTM model, but substituting BiLSTM with one layer of 1-D convolutional network.
- Self-attention-based BiLSTM model (BiLSTM-SA): The BiLSTM model integrated with the self-attention mechanism.

To further analyze the performance of our model on both summarization and sentiment classification, we apply the current state-of-the-art model proposed by Ma et al. (2018) and its joint baseline model on the same datasets, as follows:

- a joint model of Attention-based Sequence-to-Sequence model and BiLSTM model (S2S-att+BiLSTM): S2S-att and BiLSTM share the same bi-directional LSTM encoder, and the S2S-att produces the summary with a uni-directional LSTM decoder, while the BiLSTM predicts the sentiment label with an MLP.
- HSSC: The structure of HSSC has been presented in Section 1.

In the experiments of above baseline models, based on the performance of the validation sets, we tune their hyper-parameters to yield the best performance.

3.2. Experimental Result

Model	Toys & Games			
	ROUGE-1	ROUGE-2	ROUGE-L	Accuracy
S2S	16.01	3.50	15.76	-
S2S-att	16.96	4.04	16.65	-
CNN	-	-	-	70.3
BiLSTM	-	-	-	70.7
BiLSTM-SA	-	-	-	71.9
S2S-att+BiLSTM	16.82	4.29	16.72	70.8
HSSC	17.78	4.74	17.64	72.0
SAHSSC (this work)	**18.88**	**5.12**	**18.75**	**72.5**

Model	Gourmet_Foods			
	ROUGE-1	ROUGE-2	ROUGE-L	Accuracy
S2S	14.55	3.41	14.39	-
S2S-att	15.02	3.74	14.83	
CNN	-	-	-	71.1
BiLSTM	-	-	-	70.9
BiLSTM-SA	-	-	-	71.8
S2S-att+BiLSTM	15.12	3.81	14.87	71.3
HSSC	15.63	4.14	15.22	72.0
SAHSSC (this work)	**16.23**	**4.53**	**16.03**	**72.4**

Model	Movies & TV			
	ROUGE-1	ROUGE-2	ROUGE-L	Accuracy
S2S	11.53	2.96	11.39	-
S2S-att	12.47	3.32	12.13	-
CNN	-	-	-	67.1
BiLSTM	-	-	-	67.8
BiLSTM-SA	-	-	-	68.8
S2S-att+BiLSTM	12.51	3.57	12.36	68.1
HSSC	13.67	4.52	13.47	68.7
SAHSSC (this work)	**14.34**	**4.88**	**13.87**	**69.2**

Table 2: Experimental results evaluated on the test sets of *Toys & Games*, *Gourmet_Foods* and *Movies & TV* dataset on abstractive text summarization and sentiment classification on three types of ROUGE metric and accuracy of 5-class sentiment.

The experimental results reported on ROUGE score for summarization and accuracy for classification applied on the three test sets is presented in Table 2. Note that the ROUGE scores on the SNAP datasets are lower than that on the other standard datasets for summarization, such as DUC and LCSTS [Hu et al. (2015)]. The reason for the difference lies in the source of datasets: the common datasets for summarization are generally derived from the formal news and official reports, but the SNAP datasets consists of a large number of online user reviews, in which most of the texts are informal and full of noise.

For abstractive text summarization, the results show that the joint models including S2S-att+BiLSTM, HSSC and the proposed model (SAHSSC) can achieve better performance on ROUGE scores than S2S and S2S-att, which indicates that the supervision of the sentiment labels improves the representation of the original text. In addition, the proposed model in this study still outperforms S2S-att+BiLSTM and HSSC on a considerable margin, which

indicates that our model is able to generate better summaries by learning to map from the generated summary to sentiment label. On the whole, our model (SAHSSC) achieves the best performance over the competitive state-of-the-art baselines in terms of abstractive summarization on the three datasets.

As for sentiment classification, BiLSTM-SA surprisingly performs better compared with two standard neural-network-based classifiers, CNN and BiLSTM, and even surpasses the performance of S2S-att+BiLSTM, which shows that the self-attention mechanism enables the model to learn a better text embedding for sentiment classification. With regard to the joint models, the better results than the three classifiers can be attributed to more labeled data and better representation of the original text. Moreover, the proposed model gets better accuracy than S2S-att+BiLSTM and HSSC, which demonstrates that the information of summary is beneficial to predict the sentiment label, and the model also obtains more effective information of the summary for sentiment prediction. In general, due to making use of the information of summary by self-attention mechanism, the proposed model in this study achieves the best performance over the popular baselines and the current state-of-the-art joint model in terms of sentiment classification on the three datasets.

3.3. Ablation study

Model	**Toys & Games**			
	ROUGE-1	ROUGE-2	ROUGE-L	Accuracy
w/o text-attn	18.05	4.53	17.91	67.3
w/o summary-attn	17.14	4.19	16.98	72.2
SAHSSC(full model)	18.88	5.12	18.75	72.5
Model	**Gourmet_Foods**			
	ROUGE-1	ROUGE-2	ROUGE-L	Accuracy
w/o text-attn	15.77	3.98	15.43	68.1
w/o summary-attn	15.26	3.76	14.93	71.9
SAHSSC(full model)	16.23	4.53	16.03	72.4
Model	**Movies & TV**			
	ROUGE-1	ROUGE-2	ROUGE-L	Accuracy
w/o text-attn	13.75	4.23	13.56	66.8
w/o summary-attn	12.53	3.47	12.46	68.7
SAHSSC(full model)	14.34	4.88	13.87	69.2

Table 3: Comparison between partial models and full model of ablation study. ROUGE and Accuracy respectively evaluate the performance of summarization and classification.

For the purpose of further testing the effectiveness of the approach in this study, the effect of each component of the proposed self-attentive architecture has been investigated, which enables the model to combine the summarization layer with the sentiment classification layer. The expanded comparisons have been also made between the partial model without text-level self-attention or summary-level self-attention and the full model, as shown in Table 3.

Without text-level attention module, the performances of both abstractive summarization and sentiment classification drop off, especially the sentiment classification. This is because that the text encoder can be directly guided by sentiment classification layer to learn the representation of the text. Moreover, the model without summary-level attention module reports the relatively poor performance on abstractive text summarization as the sentiment classification layer enables the summary decoder to generate the better summaries by directly back-propagating its gradient. Compared with the two partial models, the full model yields significant improvements on both summarization and classification.

3.4. Visualization

As an interpretation of the learned self-attentive text embedding and summary embedding from the self-attention layer, we plot heatmaps for some reviews of *Toys & Games*, *Gourmet_Foods* and *Movies & TV* datasets, as shown in Figure 2. Some examples of 1-star and 5-star from three test sets in SNAP amazon reviews datasets are randomly selected. In the heatmap, the attention score with red colors of different transparency is marked, and the deeper color denotes the higher attention score. As you can see, the model can capture the informative words or phrases that strongly indicate the sentiment and opinion in the original text by self-attention, such as *"really disappointed", "horrible", "love", "a fun game", "five stars"* and *"yuck"*. However, the text-level attention also focuses on lots of useless and even disturbed information of the original text, such as *"no major crashes", "works", "regular scene", "ended up", "ibs reaction" and "forgettable warrior"*. Moreover, the distraction problem of attention mechanism may become more serious in the longer review text. As far as summary is concerned, the generated summary can express the main idea of the original text with a relatively short sentence and the summary-level attention can further emphasize the key factor on sentiment and opinion of the original text, such as *"fun", "awful", "good" and "terrible"*.

4. Related Work

Inspired by the recent success of neural machine translation (NMT) [Bahdanau et al. (2014)], Rush et al. (2015) first proposed an encoder-decoder model for abstraction-based text summarization, in which an attentive convolutional encoder compresses texts and a feedforward neural network language model generates summaries. Chopra et al. (2016) introduced a recurrent neural networks (RNN) decoder to generate abstractive summaries. To handle out-of-vocabulary (OOV) problem, Nallapati et al. (2016) modelled rare/unseen words by a generator-pointer model so that the decoder is able to generate words in source texts, and See et al. (2017) further incorporated the pointer-generator model with the coverage mechanism. Paulus et al. (2017) combined the supervised word prediction with reinforcement learning (RL) for abstractive summarization.

The recent advancement of neural architecture makes LSTM and CNN popular on sentiment classification of text analysis. Kim (2014) first found that CNN achieved excellent results on sentiment analysis, and Tang et al. (2015) introduced neural network approach to learn continuous document representation for sentiment classification with convolutional neural networks (CNNs) or long short-term memory networks (LSTMs). Another study carried out by Zhang et al. (2015) explored the effectiveness of character-level convolutional

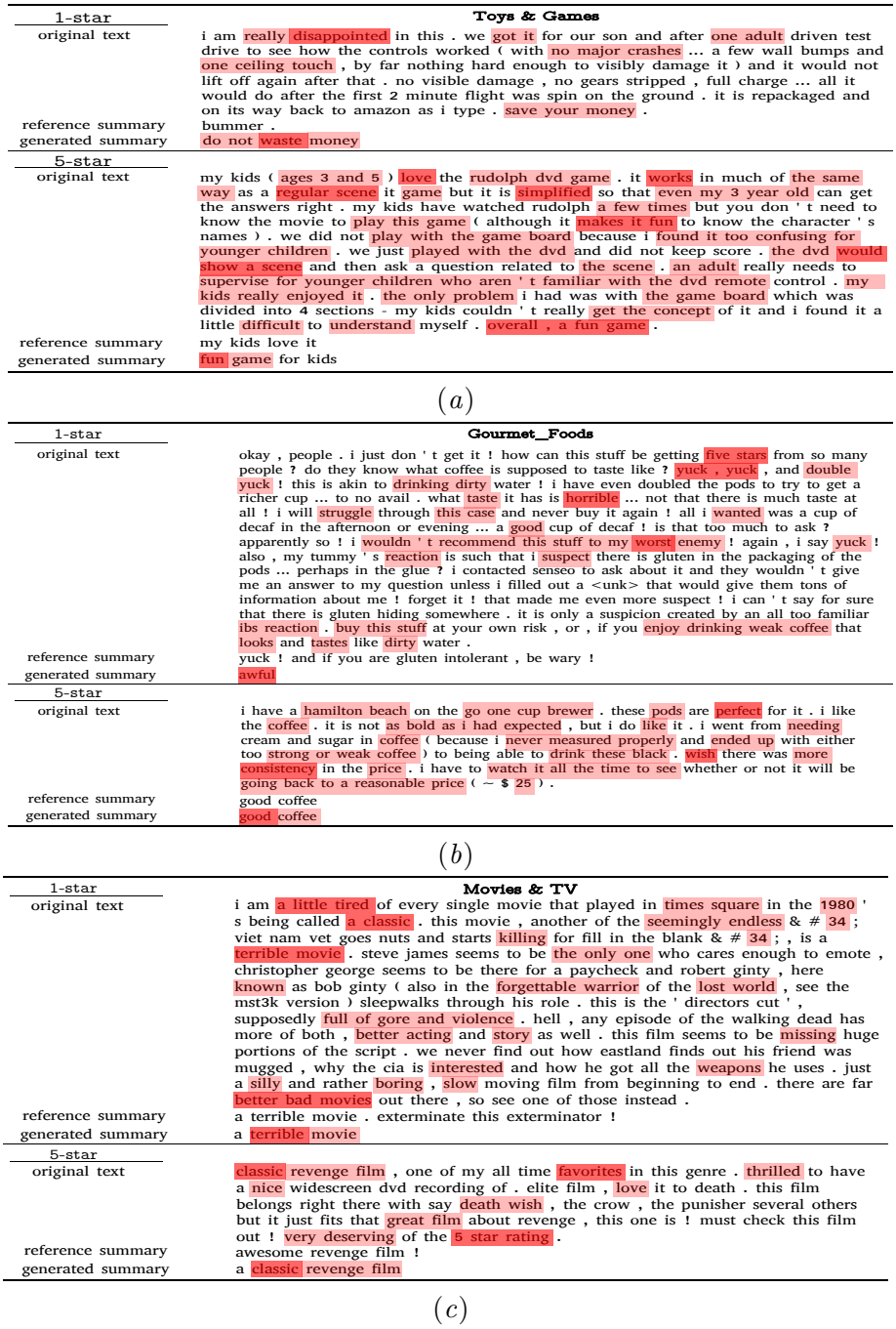


Figure 2: Heatmap of the attention scores of *Toys & Games*, *Gourmet_Foods* and *Movies & TV* test sets.

networks (CNNs) for text classification. For the purpose of cross-language sentiment classification without machine translation strategies, Becker et al. (2017) proposed an efficient

deep neural model constituted by CNNs or LSTMs. In a standard sentiment classification model, there are several LSTM or CNN layers to generate a sentence embedding and a multi-layer perceptron (MLP) to predict the sentiment label from the embedding. Lin et al. (2017) further enhanced the LSTM layer by integrating with self-attention mechanism to create the sentence embedding.

There are some studies concerning with both text summarization and sentiment classification. Hole and Takalikar (2013) and Mane et al. (2015) jointed the text summarization and the sentiment classification into a text analysis system as two independent function modules. Shetty and Bajaj (2015) and Bhargava and Sharma (2017) analyzed the sentiment by summarization to extract important parts of the text. However, the above systems only train the summarization part and the sentiment classification part independently, and require rich hand-crafted features. Cao et al. (2017) proposed a model to improve the performance of text summarization by jointly training the text classification. Unlike above, Ma et al. (2018) first attempted to jointly improve text summarization and sentiment classification within a end-to-end neural framework. Because of the inspiration from Ma *et al.*, a novel self-attentive hierarchical model is proposed in this paper. As mentioned before in Section One, this study differs from the previous studies to further make use of the generated summary, and to enhance the hierarchical framework with self-attention mechanisms.

5. Conclusion

In this paper, the proposed model jointly performs text summarization and sentiment classification within a self-attentive hierarchical neural framework. Compared with the state-of-the-art joint model and other popular baselines, our model achieves better performance on both the summarization and sentiment classification in the extensive experiments on the Amazon reviews datasets, which shows that our work is a better method to jointly improve the two taks. For future research, we believe that jointly learning multiple tasks within an end-to-end neural network framework is an interesting and effective direction, and there would be more such datasets for further application.

Acknowledgments

This research is partially supported by National Natural Science Foundation of China (No. U1711263).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL http://arxiv.org/abs/1409.0473.

Willian Becker, Jonatas Wehrmann, Henry E. L. Cagnini, and Rodrigo C. Barros. An efficient deep neural architecture for multilingual sentiment analysis in twitter. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017.*, pages 246–251, 2017. URL https://aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15404.

Rupal Bhargava and Yashvardhan Sharma. Msats: Multilingual sentiment analysis via text summarization. In *Cloud Computing, Data Science & Engineering-Confluence, 2017 7th International Conference on*, pages 71–76. IEEE, 2017.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. Improving multi-document summarization via text classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3053–3059, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14525.

Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 93–98, 2016. URL http://aclweb.org/anthology/N/N16/N16-1012.pdf.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1019–1027, 2016. URL http://papers.nips.cc/paper/6241-a-theoretically-grounded-application-of-dropout-in-recurrent-neural-networks.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 507–517, 2016. doi: 10.1145/2872427.2883037. URL http://doi.acm.org/10.1145/2872427.2883037.

Vikrant Hole and Mukta Takalikar. Real time tweet summarization and sentiment analysis of game tournament. *International Journal of Science and Research*, 4(9):1774–1780, 2013.

Baotian Hu, Qingcai Chen, and Fangze Zhu. LCSTS: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1967–1972, 2015. URL http://aclweb.org/anthology/D/D15/D15-1229.pdf.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751, 2014. URL http://aclweb.org/anthology/D/D14/D14-1181.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

Chin-Yew Lin and Eduard H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Human Language Technology Conference of the North American*

Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003, 2003. URL http://aclweb.org/anthology/N/N03/N03-1020.pdf.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017. URL http://arxiv.org/abs/1703.03130.

Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4251–4257, 2018. doi: 10.24963/ijcai.2018/591. URL https://doi.org/10.24963/ijcai.2018/591.

Vinod L Mane, Suja S Panicker, and Vidya B Patil. Summarization and sentiment analysis from user health posts. In *Pervasive Computing (ICPC), 2015 International Conference on*, pages 1–4. IEEE, 2015.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290, 2016. URL http://aclweb.org/anthology/K/K16/K16-1028.pdf.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1310–1318, 2013. URL http://jmlr.org/proceedings/papers/v28/pascanu13.html.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304, 2017. URL http://arxiv.org/abs/1705.04304.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389, 2015. URL http://aclweb.org/anthology/D/D15/D15-1044.pdf.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083, 2017. doi: 10.18653/v1/P17-1099. URL https://doi.org/10.18653/v1/P17-1099.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans,*

Louisiana, USA, February 2-7, 2018, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16126.

Ashmita Shetty and Ruhi Bajaj. Auto text summarization with categorization and sentiment analysis. *International Journal of Computer Applications*, 130(7):57–60, 2015.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014. URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.

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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1422–1432, 2015. URL http://aclweb.org/anthology/D/D15/D15-1167.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010, 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657, 2015. URL http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.