Making the Best Use of Review Summary for Sentiment Analysis

Sen Yang[♡△*], **Leyang Cui**[♡△◇*], **Jun Xie**[§], **Yue Zhang**[♡△†]
[♡]School of Engineering, Westlake University
[△]Institute of Advanced Technology, Westlake Institute for Advanced Study
◇ Zhejiang University § Tencent SPPD
senyang.stu@gmail.com cuileyang@westlake.edu.cn
stiffxie@tencent.com yue.zhang@wias.org.cn

Abstract

Sentiment analysis provides a useful overview of customer review contents. Many review websites allow a user to enter a summary in addition to a full review. Intuitively, summary information may give additional benefit for review sentiment analysis. In this paper, we conduct a study to exploit methods for better use of summary information. We start by finding out that the sentimental signal distribution of a review and that of its corresponding summary are in fact complementary to each other. We thus explore various architectures to better guide the interactions between the two and propose a hierarchically-refined review-centric attention model. Empirical results show that our review-centric model can make better use of user-written summaries for review sentiment analysis, and is also more effective compared to existing methods when the user summary is replaced with summary generated by an automatic summarization system.

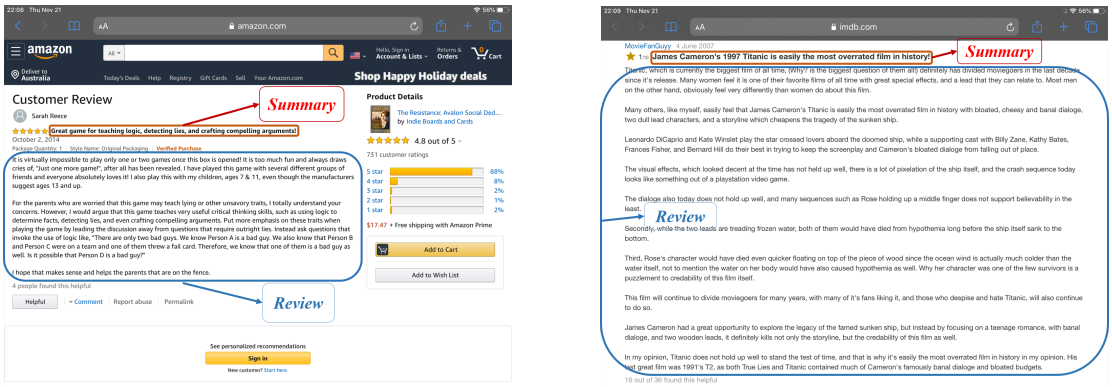
1 Introduction

Sentiment analysis (Pang et al., 2002; Kim and Hovy, 2004; Liu, 2012; Socher et al., 2013) is a fundamental task in natural language processing, which predicts the subjectivity and polarity of a given text. In practice, automatically extracting sentiment from user reviews has wide applications such as E-commerce and movie reviews (Manek et al., 2015; Guan et al., 2016; Kumari and Singh, 2016). In many review websites such as Amazon and IMDb, the user is allowed to give a summary in addition to the review, where summaries can contain more general information about the review. Figure 1 gives a few such examples. It is thus an interesting research question on how to make use of both review and summary information for better sentiment classification under such a scenario.

As shown in Figure 1, user-written summaries can be a brief version of reviews that is highly indicative of the user sentiment. Thus summaries can be used as additional training signals for sentiment classification. To this end, recent work (Ma et al., 2018; Wang and Ren, 2018) exploits multi-task learning. The model structure can be illustrated by Figure 2a. In particular, given a review input, a model is trained to simultaneously predict the sentiment and the summary. As a result, both summary and review features are integrated into the review encoder through back-propagation training.

While the above methods are highly effective, we find that the correlation between reviews and summaries can be subtle. As shown in Table 1, sometimes a summary does not directly convey sentiment as contained in the review itself. In other cases, the summary contains explicit sentiment, but the review does not. Empirically, we find in our experiments that the sentiment polarities as predicted from the reviews are consistent with those predicted from the summaries for only 73.9% of the test instances. Existing joint training methods take only the review as input at test time, and thus can be limited in its use of summary information. These facts suggest that it can be necessary to model deeper interaction between reviews and summaries for better sentiment classification.

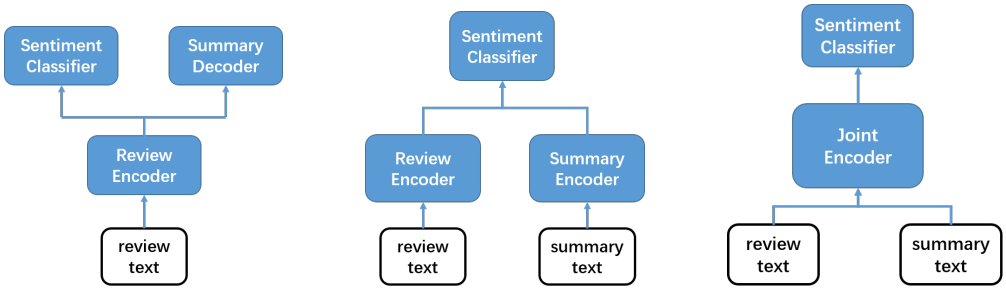
[*]Equal contribution.
[†]Corresponding author



(a) A review for *Avalon* (a card game) from Amazon

(b) A review for the movie *Titanic* from IMDb

Figure 1: Screenshots from two review websites, each containing a brief summary along with a review.



(a) Multi-task (b) Separate encoder (c) Joint encoder

Figure 2: Three model structures for incorporating summary into sentiment classification.

We conduct our investigation by treating both the review and the summary as inputs. In particular, we first compare the performance of sentiment classification using review only and using summary only, finding that the two sources of information are in fact complementary to each other. Second, as shown in Figure 2b, we investigate a simple method to integrate review and summary information by concatenating separately-learned representations. This method turns out to outperform models using review or summary inputs only. One limitation of this method, however, is that it does not capture the interaction between the review and summary information as thoroughly as the method shown in Figure 2a, in which the representation of a review contains summary knowledge also.

To address this issue, we further investigate a joint encoder structure between the review and the summary, which is demonstrated in Figure 2c. To this end, an intuitive method is co-attention (Xiong et al., 2017), which iteratively updates the representation of review and summary by consulting each other, as shown in Figure 3a. However, we find empirically that the review itself is relatively more indicative of the user sentiment compared to the summary. Given this observation, we further build a review-centric joint encoder. Different from the co-attention encoder, the review-centric model iteratively updates a review representation given a summary representation, but not vice versa.

We evaluate our proposed models on the SNAP (Stanford Network Analysis Project) Amazon review datasets (He and McAuley, 2016), which contain reviews and ratings together with user-written summaries if they exist. In scenarios where there is no user-written summary for a review, we use a pointer-generator network summarization model (See et al., 2017) to generate abstractive summaries. Empirical results show that our review-centric model outperforms a range of baselines, including multi-task, separate encoder and joint encoder methods. In addition, our review-centric model achieves new state-of-the-art results, giving 2.1% (with system-generated summary) and 4.8% (with gold summary) absolute improvements compared to the previous best method on the SNAP benchmark. To our knowl-

Rating: 5 stars

Review: *I can color right along with my grandchildren, without feeling intellectually compromised at the project. This book is* ***so amazing*** *that I have used some of the designs for stained glass windows. I* ***highly recommend*** *this for anyone who does not want to grow out of a favorite past time.*

Summary: *Now I don't have to grow up*

Rating: 5 stars

Review: *My son, 9, had outgrown his old helmet, so I bought this one. Less than three weeks later, he put it to the test. He is a daredevil who loves speed. Riding down a hill, that he isn't supposed to ride on, he lost control at 30 mph and landed on the side of his head, on the asphalt. He was knocked out momentarily, but passed the concussion screening at the ER. He is fine, other than some road rash. I hate to think how he might have fared with his old helmet, or no helmet. The ER doctor said that the spot he hit was about the worst place to hit for head injuries. Don't skimp on safety equipment, ever, especially for kids. I am ordering this exact same helmet as a replacement.*

Summary: ***Great buy!*** *Saved my son, thank you*

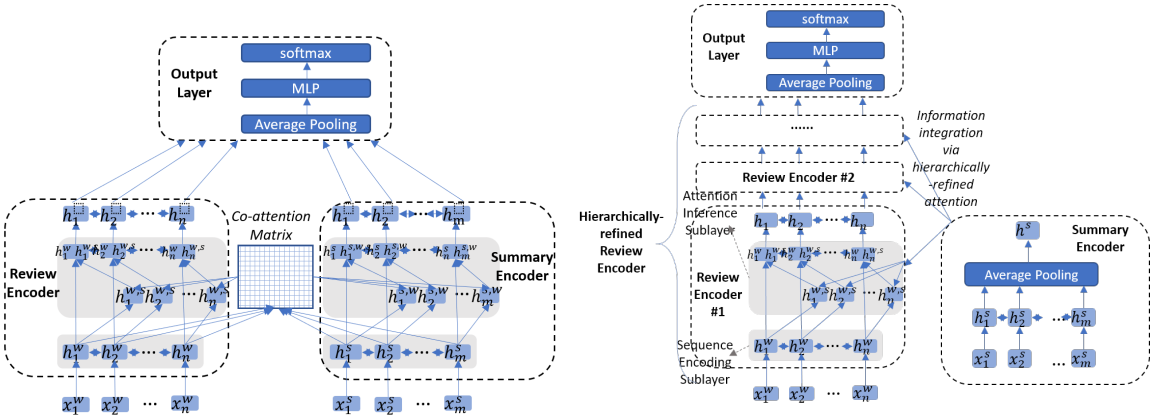
Table 1: Two examples of online reviews with summaries and ratings. Explicit sentiment phrases are in bold and underlined.

edge, we are the first to investigate the correlation between reviews and their summaries for expressing sentiment, and the first to empirically investigate different models making use of both reviews and summaries for better sentiment analysis. We release our code at `https://github.com/RingoS/sentiment-review-summary`.

2 Related Work

Our work is partly related to previous work building well-designed matching models to capture the relationship between two texts. In reading comprehension, a matching model is required to capture the similarity among a given passage, a question and a candidate answer. Chen et al. (2016) adopted two GRUs to encode the passage and question, respectively, and a bilinear function to compute the similarity on each passage token. Xiong et al. (2017) make use of co-attention, which shares one single attention matrix between the passage and the question, calculating both passage-to-question and question-to-passage attention scores. For retrieval-based dialogue systems, models are required to calculate the matching score between a candidate response and a conversation context. In particular, *Sequential Matching Network* (Wu et al., 2017) captures matching information by constructing word-to-word and a sequence-to-sequence similarity matrices. *Deep Attention Matching Network* (Zhou et al., 2018) adopts self-attention and cross-attention modules to harvest intra-sentence relationship and inter-sentence relationship, respectively. To capture potential long-term label dependency in sequence labeling, Cui and Zhang (2019) use attention over label embeddings to refine the marginal label probabilities by calculating the similarity between a word sequence and a set of label embeddings. Compared with these methods, which model *matching* between two pieces of text, our work is different in that we consider how to effectively make use of the *complementary* property between a review and a corresponding summary for better review sentiment analysis.

Our work is related to previous work on sentiment analysis (Pang et al., 2002; Kim and Hovy, 2004; Liu, 2012), taking a whole review as input (Kim, 2014; Zhang et al., 2015; Yang et al., 2016; Johnson and Zhang, 2017) rather than specific aspects (Chen et al., 2017; Li et al., 2019). Different from previous work, we additionally consider user-generated or automatically-generated summaries as input. Our work is related to existing work on joint summarization and sentiment classification. Ma et al. (2018) propose a multi-view attention model for joint summarization and sentiment classification. Wang and Ren (2018) improve the model of Ma et al. (2018) by using additional attention on the generated text. Different from their work, we are not directly concerned about making better summaries. Instead, we make a broader discussion on how to make the best use of both review and summary for sentiment classification. Our work is also related to rationalizing sentiment predictions. Zhang et al. (2016) regard gold-standard rationales as additional input and used rationale-level attention for text classification. Bastings et al. (2019) propose an unsupervised latent model that selects a rationale and subsequently uses it for sentiment analysis. Our work is similar in that we can visualize the most salient words in sentiment classification. Different from the existing methods, our rationalization is based on the interaction between a review and a summary, with the latter guiding the visualization.



(a) Architecture of a co-attention joint encoder (b) Architecture of a review-centric joint encoder

Figure 3: Architectures of co-attention model and our review-centric model.

3 Task

3.1 Problem Formulation

The input to our task is a pair (X^w, X^s), where $X^w = x_1^w, x_2^w, ..., x_n^w$ is a review and $X^s = x_1^s, x_2^s, ..., x_m^s$ is a corresponding summary, the task is to predict the sentiment label $y \in [1, 5]$, where 1 denotes the most negative sentiment and 5 denotes the most positive sentiment. n and m denote the size of the review and summary in the number of words, respectively.

3.2 Research Questions

We aim to answer the following research questions empirically:

- RQ #1: What are the roles of and the correlation between a review and its summary for predicting the user rating;
- RQ #2: How to better leverage information from both the review and the summary for effective sentiment classification;

4 Method

All the methods that we investigate are based on a BiLSTM (Hochreiter and Schmidhuber, 1997) structure. We first discuss the basic BiLSTM to encode text (Sec 4.1), and then discuss two types of structures that use BiLSTM for separate encoding(Sec 4.2.1) and symmetric joint encoding (Sec 4.2.2), respectively. Finally, we discuss review-centric joint encoding (Sec 4.3) of the review and the summary.

4.1 Sequence Encoding

We use BiLSTM as the sequence encoder for all experiments. The input is a sequence of word representations $\mathbf{x} = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m = \{emb(x_1), ..., emb(x_m)\}$, where emb denotes a word embedding lookup table. Word representations are fed into a standard BiLSTM. We adopt a standard LSTM formulation, where a sequence of hidden states $\mathbf{h}_t$ are calculated from the sequence of $\mathbf{x}_t$ ($t \in [1, ..., m]$).

A forward left-to-right LSTM layer and a backward right-to-left LSTM yield a sequence of forward hidden states $\{\overrightarrow{\mathbf{h}}_1, ..., \overrightarrow{\mathbf{h}}_m\}$ and a sequence of backward hidden states $\{\overleftarrow{\mathbf{h}}_1, ..., \overleftarrow{\mathbf{h}}_m\}$, respectively. The two hidden states are concatenated to form a final representation:

$$\begin{aligned} \mathbf{h}_i &= [\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i] \\ \mathbf{H} &= \{\mathbf{h}_1, ..., \mathbf{h}_m\} \end{aligned} \tag{1}$$

This encoder structure serves as a basis for all the models. In particular, for the review-only and summary-only baselines, we use a single encoder as described above.

4.2 Baselines

4.2.1 Separate Encoding

Two BiLSTMs are adopted to separately encode reviews and summaries. Both of the produced hidden state matrices are then delivered to two settings:

1) average-pooling baseline: the hidden state matrices are concatenated and then average-pooled to form a final representation for later prediction.

2) self-attention baseline: the hidden state matrices are separately processed using self-attention mechanism. Subsequently the two matrices are concatenated and average-pooled to produce the final representation for later prediction. Our self-attention module follows the implementation of Lin et al. (2017).

4.2.2 Symmetric Joint Encoding

On top of the sequence encoder, we separately adopt average pooling, self-attention (Lin et al., 2017), hard-attention (Xu et al., 2015; Shankar et al., 2018) and co-attention (Xiong et al., 2017) mechanisms as our joint baselines. In particular, co-attention can capture the interactions between review and summary by calculating the bidirectional symmetric attention flows with a shared attention weight matrix.

1) For joint encoder baselines using pooling and self-attention, only one BiLSTM is adopted, with concatenated review and summary texts as input.

2) The hard-attention baseline is trained using an additional extractive summarization objective. We implement our baseline following Xu et al. (2015) and Shankar et al. (2018). In particular, words in the review text that overlap with the corresponding summary are extracted in their original order to formulate a summary. The model calculates an additional loss between attention weights and extractive summary labels, so that the hard attention weights can be automatically produced during inference time.

3) As for co-attention baselines, we use two BiLSTMs to separately encode review and summary. The two hidden state matrices then interact with each other. The formulations are written as :

$$\begin{aligned} \mathbf{A} &= \mathbf{H}^w \mathbf{W}^w (\mathbf{H}^s \mathbf{W}^s)^\intercal \\ \mathbf{H}^w_{co\text{-}att} &= \mathbf{H}^w + \mathrm{softmax}(\frac{\mathbf{A}}{\sqrt{d}})\mathbf{H}^s \\ \mathbf{H}^s_{co\text{-}att} &= \mathbf{H}^s + \mathrm{softmax}(\frac{\mathbf{A}^\intercal}{\sqrt{d}})\mathbf{H}^w \end{aligned} \tag{2}$$

where $\mathbf{H}^w$ and $\mathbf{H}^s$ are the hidden states of reviews and summaries, respectively. $\mathbf{A} \in \mathbb{R}^{n \times m}$ is the co-attention matrix. n and m are the lengths of the review text and the summary text, respectively. d represents the hidden size of the BiLSTM. $\mathbf{H}^w_{co\text{-}att}$ and $\mathbf{H}^s_{co\text{-}att}$ are the co-attention representations of a review and its corresponding summary, respectively. They are then fed into subsequent layers for making predictions.

4.3 Review-centric Joint Encoding

This joint encoder model changes the review encoder over the baseline, while keeping the summary encoder. As shown in Figure 3b, the review encoder has a set of stacked layers, each consisting of a sequence encoding sublayer and an attention inference sublayer. The sequence encoding sublayers takes the same BiLSTM structure as the summary encoder, but with different model parameters. The attention inference sublayer integrates summary information into the review representation. By repeatedly consulting summary information, the review encoder obtains increasingly refined hidden states over layers.

Attention Inference Sublayer Formally, X^s and X^w are fed into sequence encoding layer, yielding $\mathbf{H}^w$ and $\mathbf{H}^s$, respectively. $\mathbf{h}^s$ is then obtained by average-pooling over $\mathbf{H}^s$. We model the dependencies between the original review and the summary with multi-head dot-product attention. Each head produces an attention vector $\boldsymbol{\alpha} \in \mathbb{R}^n$, which consists of a set of similarity scores between the hidden state of each token of the review text and the summary representation. The hidden states are calculated by

Domain	Size	#Review	#Summary
Toys & Games	168k	99.9	4.4
Sports & Outdoors	296k	87.2	4.2
Movies & TV	1,698k	161.6	4.8

Table 2: Data statistics. Size: number of samples, #Review: the average length of reviews, #Summary: the average length of summaries.

$$\begin{aligned} \boldsymbol{\alpha} &= \mathrm{softmax}(\frac{\mathbf{H}^w \mathbf{W}_i^Q (\mathbf{h}^s \mathbf{W}_i^K)^\intercal}{\sqrt{d_h/k}}) \\ \mathbf{head}_i &= \hat{\mathbf{A}}(\hat{\mathbf{H}}^s)^\intercal \mathbf{W}_i^V \\ \mathbf{H}^{w,s} &= \mathrm{concat}(\mathbf{head}_1, \ldots, \mathbf{head}_k), \end{aligned} \tag{3}$$

where superscripts w and s represent review and summary, respectively. $\hat{\mathbf{A}}$ and $\hat{\mathbf{H}}^s$ are the unsqueezed matrices of $\boldsymbol{\alpha}$ and $\mathbf{h}^s$. $\mathbf{W}_i^Q \in \mathbb{R}^{d_h \times \frac{d_h}{k}}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_h \times \frac{d_h}{k}}$ and $\mathbf{W}_i^V \in \mathbb{R}^{d_h \times \frac{d_h}{k}}$ are model parameters. Q, K and V represent *Query*, *Key* and *Value*, respectively. k is the number of parallel heads and $i \in [1, k]$ indicates which head is being processed.

Following Vaswani et al. (2017), we adopt a residual connection around each attention inference layer:

$$\mathbf{H} = \mathrm{LayerNorm}(\mathbf{H}^w + \mathbf{H}^{w,s}) \tag{4}$$

$\mathbf{H}$ is then fed to the subsequent sequence encoding layer as input, if any.

According to the equations of standard LSTM and Equation 3, tokens of the original review that are the most relevant to the summary are particularly focused on by consulting summary representation. The hidden states $\mathbf{H}^{w,s}$ are thus a representation matrix of the review text that encompasses key features of summary representation. Multi-head attention ensures that multi-faced semantic dependency features can be captured, which is beneficial for scenarios where multiple key points exist in one review.

4.4 Output Layer

Global average pooling is applied on $\mathbf{H}$, followed by a classifier layer:

$$\begin{aligned} \mathbf{h}^{avg} &= \text{avg-pooling}(\mathbf{h}_1, ..., \mathbf{h}_n) \\ \mathbf{p} &= \mathrm{softmax}(\mathbf{W}\mathbf{h}^{avg} + \mathbf{b}) \\ \hat{y} &= \operatorname{argmax} \mathbf{p}, \end{aligned} \tag{5}$$

where $\hat{y}$ is the predicted sentiment label; $\mathbf{W}$ and $\mathbf{b}$ are parameters to be learned.

Training Given a dataset $D = \{(X_t^w, X_t^s, y_t)\}|_{t=1}^{|T|}$, our models can be trained by minimizing

$$L = -\sum_{t=1}^{|T|} \log(\mathbf{p}^{[y_t]}) \tag{6}$$

where $\mathbf{p}^{[y_t]}$ denotes the value of the label in $\mathbf{p}$ that corresponds to y_t.

5 Experiments

The SNAP Amazon Review Dataset[1] (McAuley and Leskovec, 2013) consists of around 34 million Amazon reviews in different domains, such as books, games, sports and movies. Each review mainly consists of a product ID, a piece of user information, a plain text review, a user-written summary and an

[1] http://snap.stanford.edu/data/web-Amazon.html

Model	T & G	S & O	M & T	Average
Multi-task				
HSSC	71.9	73.2	68.9	71.3
SAHSSC	72.5	–	69.2	70.9
Summary Only				
Pooling	73.0	69.5	68.1	70.2
Self-attention	71.9	70.4	68.9	70.4
Review Only				
Pooling	73.3	71.2	71.7	72.1
Self-attention	73.5	71.8	72.3	72.5
Separate Encoder				
Pooling	74.4	73.9	73.8	74.0
Self-attention	75.8	73.1	73.7	74.2
Joint Encoder				
Hard-attention	73.4	72.1	73.9	73.1
Pooling	75.4	73.4	73.2	74.0
Self-attention	75.7	74.3	74.1	74.7
Co-attention	76.1	74.2	74.3	74.9
Review-centric Model	**76.6**	**76.1**	**75.9**	**76.2**

Table 3: Results using gold summary as input.

Model	T & G	S & O	M & T	Average
Separate Encoder				
Pooling	71.8	72.2	72.5	72.2
Self-attention	73.1	72.5	72.6	72.7
Joint Encoder				
Pooling	73.8	72.0	72.0	72.6
Self-attention	73.9	71.6	72.4	72.6
Co-attention	73.8	72.2	72.7	72.9
Review-centric Model	**74.8**	**72.6**	**72.8**	**73.4**

Table 4: Results using system-generated summary as input.

Model	T & G	S & O	M & T	Average
Co-attention (*review*)	75.1	74.8	74.7	74.9
Co-attention (*summary*)	74.1	74.2	73.4	73.9
Co-attention (*concat*)	76.1	74.2	74.3	74.9
Summary-centric	73.8	74.5	74.9	74.4
Review-centric	76.6	76.1	75.9	76.2

Table 5: Result comparison among different interacting schemes.

overall sentiment rating which ranges from 1 to 5. For fair comparison with previous work, we adopt the same domains and partition used by Ma et al. (2018) and Wang and Ren (2018), which includes three datasets (Toys & Games, Sports & Outdoors and Movies & TV). The statistics of our adopted dataset are shown in Table 2. For each dataset, the first 1000 samples are taken as the development set, the next 1000 samples as the test set, and the rest as the training set.

5.1 Experimental Settings

We use GloVe (Pennington et al., 2014) 300-dimensional embeddings as pretrained word vectors. The LSTM hidden size is set to 256. We use Adam (Kingma and Ba, 2015) to optimize all models, with an learning rate of $3e-4$, momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. The dropout rate α and number of attention heads M are separately set depending on the size of each dataset and using development experiments, which are $\alpha = 0.5$ and $M = 1$ for Toys & Games, $\alpha = 0.2$ and $M = 1$ for Sports & Outdoors and $\alpha = 0.0$ and $M = 2$ for Movies & TV. We adopt two layers for our review-centric model.

In addition to conducting experiments with user-written summaries, we additionally perform experiments by replacing the user-written summary with a system-generated summary for two reasons. First, we want to know the extent to which our method can be generalized to settings of traditional sentiment classification, where the input consists of only one piece of text. This is the setting adopted by most previous research. Second, two of our baselines, namely HSSC (Ma et al., 2018) and SAHSSC (Wang and Ren, 2018), adopt this setting and use summary information via multi-task learning. For generating summaries, we separately adopt a pointer-generator network (PG-Net) with coverage mechanism (See et al., 2017) trained on the training set.

5.2 Results

Our main results are shown in Tables 3 and 4. It can be seen from Table 3 that the baseline using review only outperforms that using summary only, which indicates that the review is more informative than the summary. In addition, the *Separate Encoder* models outperform both the *Summary Only* and the *Review Only* models, which indicates that additional summary input is beneficial to sentiment analysis. Finally, the *Joint Encoder* models generally outperform the *Separate Encoder* models, which suggests that modeling interactions between review and summary is superior to separate encoder structure. In particular, hard-attention receives more supervision information compared with soft-attention, by using supervision signals from *extractive* summaries. However, it underperforms the soft-attention model, which indicates that the most salient words for making sentiment classification may not strictly overlap with *extractive* summaries. Among soft-attention methods, co-attention achieves better performance

compared to self-attention, which may result from the fact that co-attention allows mutual interactions between review and summary.

In Table 3, all architectures using system-generated summary as additional input outperform *Review Only* models, demonstrating that even imperfect summary can still give additional benefit for sentiment prediction. However, models using system-generated summary perform significantly worse than those using gold summary, verifying the importance of high quality summaries.

In both tables, the review-centric model outperforms all the baseline models on all the three datasets. In particular, the review-centric model gives 1.3% and 0.5% improvements compared with the best baseline (co-attention) with both gold summary and system-generated summary, respectively. Our *Joint Encoder* models also outperform HSSC (Ma et al., 2018) and SAHSSC (Wang and Ren, 2018). In particular, these two multi-task models use summary information in training, thereby enhancing a review-only sentiment classifier. Their performance is competitive compared to the review only models. By further using user-written summaries directly, both *Separate Encoder* and *Joint Encoder* models outperform these models. It is worth noting that in Table 4, our methods still outperform the baselines with the same input settings, showing the effect of joint encoding.

5.3 Discussion

In this section, we aim to answer the research questions raised in Section 3.2.

5.3.1 RQ #1

We first explore the correlation between reviews and summaries with regard to carrying the user sentiment. In particular, we empirically compare the predictions of two simplest conditions, including using review only (abbreviated as *review-only*) and using gold summary only (abbreviated as *summary-only*), based on BiLSTM+pooling, on the Toys & Games dataset. For the purpose of exploring correlation, we focus on a special part of the test set, named as *conflicting-set*, on which *review-only* and *summary-only* have conflicting predictions with each other. We assume that a review in *conflicting-set* contains a different sentiment rating from that of its corresponding summary. *Conflicting-set* takes 26.1% of the whole test set, which suggests that such *conflicting* samples are frequently seen in the dataset. Additionally, we define the complement of *conflicting set* as *non-conflicting-set*. We also define *union-set*, which is a subset of *conflicting-set* and is composed of the samples for which at least one of the two models (*review-only* and *summary-only*) has correct predictions. The experimental results are shown in Figure 4.

Correlation As shown in Figure 4, the co-attention model gives a low accuracy of 50.1% on *conflicting-set*, which is much lower compared to its performance on *non-conflicting-set* (85.1%). This wide gap suggests that conventional models have difficulty handling conflicting situations. It can also be seen from Figure 4 that both *review-only* and *summary-only* obtain poor performance on *conflicting-set* (41.8% and 35.2%, respectively). However, the sum accuracy of the two models, which forms the third bar on both sides of Figure 4, takes $41.8\% + 35.2\% = 77.0\%$ of *conflicting-set*, which suggests that review and summary information are highly complementary to each other under conflicting situations.

5.3.2 RQ #2

Interacting Scheme As shown in Tables 3 and 4, our review-centric model gives better results compared to the co-attention method. The only difference between these two methods is the interacting scheme between the review and the summary. We thus conduct experiments to further explore the influences of different interacting schemes. In particular, we train co-attention models using three types of top-layer representations, including using review representations only, using summary representations only and using the concatenation of both of the above. The experiment results are shown in Table 5. It can be seen that co-attention (*review*) and co-attention (*concat*) give rather similar performance, which indicates the relative importance of review representation. Moreover, the review-centric model outperforms all co-attention methods, which indicates that our review-centric attention is better than symmetric attention (e.g. co-attention).

In addition, we empirically compare our model with a model with the same structure but reverse inputs, named as the summary-centric model. As shown in Table 5, our review-centric model outperforms the

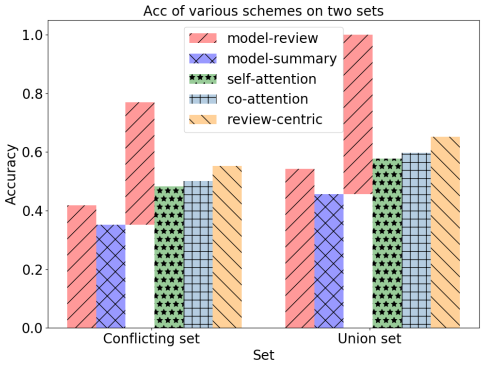


Figure 4: Analysis on *conflicting-set*. We stack the accuracy of *review-only* and *summary-only* to form the third bar, which is *union-set*.

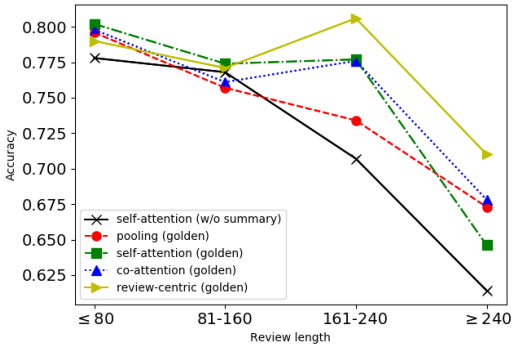


Figure 5: Accuracy against the review length with gold summary input.

- Summary fun for the whole new game in all ages ! ! ! fun ! ! !
- Review I bought this hoping to encourage my 9 and 10 year olds to 1) enjoy games and 2) practice spelling. WHO KNEW we 'd all fall in love. My mom comes over every afternoon and forces us all to play ! It **is** **quite** **fun** and anyone can win. I love it and buy it for friends. Highly recommend this game. So **much** **easier** **and** carefree than keeping score and strategy in Scrabble .

(a) Attention heatmap with system-generated summary

- Summary Favorite Game to Teach to Newbies
- Review I play a lot of Board Games. I play so many that I have a collection of games that are very fun , but very hard to Ingenious is the is a simple game of placing tiles on a board and then counting the number of matching symbols to score points. It **'s** **easy** **to** **teach** and easy to learn. And it 's immensely is a deep game in this simple idea of placing tiles on a board and making the best chains of It 's so easy that kids can play and do well and that adults can play and try to use strategy and still come in second to the of the games in my **collection** are **hard** to explain. There 's games with rules that change each round about who goes first , or there are games that have special rules about what you can and ca n't do on your turn. Ingenious is not one of these , it is a game that is Simple and Complex at the same time. It 's a lot of **fun** and it 's really easy to teach and still is one of the **most** enjoyable games I 've ever played.I **would** **recommend** **this** **to** anyone that is looking for a good board game that they can play over and over **again**

(b) Attention heatmap with gold summary

Figure 6: Visualization of self-attention and hierarchically-refined attention, with system-generated summary (a) and gold summary (b). (1) BiLSTM+self-attention: dot line / blue color; (2) the first layer of our review-centric model: straight line / pink color; (3) the second layer of our review-centric model: dash line / yellow color. Deeper color indicates higher attention weight.

summary-centric model by a large margin, which suggests that focusing on the review side is better than focusing on the summary side for predicting sentiment ratings.

5.4 Analysis

Intersection with Union-set We find that, on *conflicting-set*, 92.1% of the self-attention baseline's correct predictions, 91.0% of the co-attention baseline's correct predictions and 91.0% of the review-centric model's correct predictions come from *union-set*. The line of high ratios suggests that explicit sentiment indication in at least one piece of text between the review and the summary is necessary for making a correct prediction on *conflicting-set*.

In addition, our review-centric model slightly underperforms the co-attention model on *non-conflicting-set* (84.2% comparing with 85.1%). However, it still outperforms the co-attention model by 0.5% on the whole Toys & Games test set, which results from the fact that the the former outperforms the latter by a large margin of 5.1% on *conflicting-set*, and more specifically, 5.5% on *union-set*. The review-centric model's superior performance on *union-set* verifies its strength on making better use of the complementary correlation between the review and the summary. It also suggests that the two models hold different inductive biases when encoding reviews and summaries.

Review Length Figure 5 shows the accuracy of the average-pooling model, the self-attention model, the co-attention model and the review-centric model against review length. As the review length increases, the performance of all models decreases. BiLSTM+self-attention does not outperform BiLSTM+pooling on long text. Our review-centric method gives better results compared to all baseline models for long reviews, demonstrating that the review-centric model is effective for producing more abstract representations. The superior performance may result from the hierarchical review-centric attention mechanism, which maintains the most salient information while ignoring redundant information of the source review text. The review-centric model can thus be more robust when the review has noisy sentimental words or phrases, which are commonly seen in long reviews (e.g., the example in Figure 6b).

Case Study Our models have a natural advantage of interpretability thanks to the use of the attention inference sublayer. We visualize the hierarchically-refined review-centric attention of two sample cases from the test set of Toys & Games, and also self-attention distribution for fair comparison. To make the visualizations clear and to avoid confusion, we choose to visualize the most salient parts, by rescaling all attention weights into an interval of $[0, 100]$ and adopting 50 as the threshold for attention visualization (only attention weights ≥ 50 are visualized).

Figure 6a shows an example with system-generated summary that has 5 stars as the gold rating score. The summary text is "*fun for the whole new game in all ages ! ! ! fun ! ! !*", which suggests that the game is 1) interesting (from word "*fun*") and 2) not difficult to learn (from phrase "*all ages*"). It can be seen that both the self-attention model and the first layer of our review-centric model attend to the strongly positive phrase "*quite fun*", which is relevant to the word "*fun*" in the summary. In comparison, the second layer attends to the phrase "*much easier*", which is relevant to the phrase "*in all ages*" in the summary. This verifies our review-centric model's effectiveness of leveraging abstractive summary information.

Figure 6b illustrates a 5-star-rating example with a gold summary. The summary text is "*Favorite Game to Teach to Newbies*". As shown in the heatmap, self-attention attends only to general sentimental words such as "*hard*", "*fun*", "*immensely*" and "*most*", which deviates from the main idea of the document text. In comparison, the first layer of our review-centric model attends to phrases like "*easy to teach*", which is a perfect match of the phrase "*teach to newbies*" in the summary. This shows that the shallow attention inference sublayer can learn direct similarity matching information under the supervision of summarization. In addition, the second layer of our review-centric model attends to phrases including "*would recommend this to anyone*", which links to "*easy to teach*" and "*Teach to Newbies*", showing that the deep attention inference sublayer of our model can learn underlying connections between the review and the summary.

6 Conclusion

We empirically analyzed the correlation between reviews and summaries for customer review sentiment analysis, found that they are complementary to each other for carrying user sentiment. We investigated a range of joint encoder models for better modeling the interactions between reviews and summaries and proposed a novel review-centric method, which hold different inductive bias to capture the complementary correlation. Empirical results verified the effectiveness of joint encoding for review and summary among strong baselines and existing work, showing that a review-centric model outperforms a symmetric co-attention model.

Acknowledgments

We thank all anonymous reviewers for their constructive comments. We also would like to acknowledge funding support from the Westlake University and Bright Dream Robotics Joint Institute for Intelligent Robotics and a research grant from Tencent.

References

Joost Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2963–2977.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany, August. Association for Computational Linguistics.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. 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, Copenhagen, Denmark, September. Association for Computational Linguistics.

Leyang Cui and Yue Zhang. 2019. Hierarchically-refined label attention network for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4106–4119, Hong Kong, China, November. Association for Computational Linguistics.

Ziyu Guan, Long Chen, Wei Zhao, Yi Zheng, Shulong Tan, and Deng Cai. 2016. Weakly-supervised deep learning for customer review sentiment classification. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 3719–3725.

Ruining He and Julian McAuley. 2016. 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 '16, pages 507–517, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570, Vancouver, Canada, July. Association for Computational Linguistics.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, page 1367–es, USA. Association for Computational Linguistics.

Yoon Kim. 2014. 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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

N. Kumari and S. N. Singh. 2016. Sentiment analysis on e-commerce application by using opinion mining. In *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, pages 320–325.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6714–6721.

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

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. 2018. 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.

Asha S. Manek, P. Deepa Shenoy, M. Chandra Mohan, and K. R. Venugopal. 2015. Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and svm classifier. *World Wide Web*, 20:135–154.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 165–172, New York, NY, USA. ACM.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, July 6-7, 2002*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. 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.

Shiv Shankar, Siddhant Garg, and Sunita Sarawagi. 2018. Surprisingly easy hard-attention for sequence to sequence learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 640–645, Brussels, Belgium, October-November. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. 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 5998–6008.

Hongli Wang and Jiangtao Ren. 2018. A self-attentive hierarchical model for jointly improving text summarization and sentiment classification. In *Proceedings of The 10th Asian Conference on Machine Learning, ACML 2018, Beijing, China, November 14-16, 2018.*, pages 630–645.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada, July. Association for Computational Linguistics.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2048–2057. JMLR.org.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. 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 1480–1489.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. 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.

Ye Zhang, Iain James Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 795–804.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia, July. Association for Computational Linguistics.