Opinions Summarization: Aspect Similarity Recognition Relaxes the Constraint of Predefined Aspects

Huy Tien Nguyen
Japan Advanced Institute of Science and Technology (JAIST), Japan
ntienhuy@jaist.ac.jp

Tung Le
Japan Advanced Institute of Science and Technology (JAIST), Japan
lttung@jaist.ac.jp

Minh Le Nguyen
Japan Advanced Institute of Science and Technology (JAIST), Japan
nguyenml@jaist.ac.jp

Abstract

Recently research in opinions summarization focuses on rating expressions by aspects and/or sentiments they carry. To extract aspects of an expression, most studies require a predefined list of aspects or at least the number of aspects. Instead of extracting aspects, we rate expressions by aspect similarity recognition (ASR), which evaluates whether two expressions share at least one aspect. This subtask relaxes the limitation of predefining aspects and makes our opinions summarization applicable in domain adaptation. For the ASR subtask, we propose an attention-cell LSTM model, which integrates attention signals into the LSTM gates. According to the experimental results, the attention-cell LSTM works efficiently for learning latent aspects between two sentences in both settings of in-domain and cross-domain. In addition, the proposed extractive summarization method using ASR shows significant improvements over baselines on the Opinosis corpus.

1 Introduction

Opinions Summarization is the collection of typical opinions mentioned in social media, blogs or forums on the web. This task helps customers to absorb better a large number of comments and reviews before making decisions as well as producers to keep track of what customers think about their products (Liu, 2012).

Due to the fast growth of data over the Internet, automatically opinions summarization has received a lot of attention in recent years. Most research focus on extractive summarization, where the most salient text units are identified and construct a summary. Ranking candidates for generic summarization usually bases on various handcrafted features such as sentence position and length (Radev et al., 2004), word frequency (Nenkova et al., 2006) or using neural networks for learning salient scores (Zhou et al., 2018).

In opinions summarization, however, this task is required to consider aspects and/or sentiments of text candidates for generating a concise and informative summary (Hu and Liu, 2006). The popular framework of this problem involves three subtasks (Hu and Liu, 2004): i) aspect discovery which extracts the properties of interested entities (e.g., battery life, design, customer service); ii) sentiment analysis which assigns sentiment polarity (positive and negative) towards the aspects extracted in the first step; and iii) summary generation which selects the most salient opinions to build a summary.

For the aspect discovery task, there are two main techniques: supervised and unsupervised learning. The former models the aspect extraction as a sequence labeling task. Due to predefining a list of aspect and heavily relying on annotated data, this approach suffers from domain adaptation problems. The latter uses a large amount of unlabeled data for abstracting aspects via the statistical topic modeling LDA (Blei et al., 2003) or the aspect-based autoencoder model (He et al., 2017a). However, these unsupervised techniques have limitations. First, we have to decide on a suitable number of aspects for each domain. Second, the existing methods require a sufficient amount of data while some domains may not have enough reviews, known as the cold-start problem (Moghaddam and Ester, 2013).

In extractive opinions summarization, most existing approaches use the aspects information for discarding potentially redundant units. For minimizing repeated information on the same aspect, we only need to identify whether two text

Proceedings of Recent Advances in Natural Language Processing, pages 487–496,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_058

No.	Sentence	Aspect Similarity
1	The pc runs so fast. I like its performance and price.	Yes
2	The food is cheap. The shop's location is good.	No
3	I love its pizza. I bought food from the restaurant.	Yes

Table 1: Some samples of Aspect Similarity Recognition

units have at least one aspect in common, which is called Aspect Similarity Recognition - ASR (Nguyen et al., 2018), rather than explicitly extracting aspects of each text unit. Table 1 shows some samples of the ASR task. Follow this observation, we propose an aspect-based summarization using ASR instead of aspect discovery. The advantage of ASR is to learn patterns and relations between two text units and not need to identify the aspects of each unit, therefore it is potential to cross-domain application. Our contributions in this work are as follows:

- We propose an attention-cell LSTM model (ACLSTM) for ASR which enhances the LSTM model via employing attention signals into the input gate and the memory cell. ACLSTM shows improvements compared to the conventional attention models for both settings of in-domain and cross-domain.

- We introduce a novel aspect-based summarization using Aspect Similarity Recognition. According to the experiments, our method outperforms strong baselines on Opinosis corpus. We also evaluate our method in regard to domain adaptation.

The remainder of this paper is organized as follows: Section 2 reviews the related research, Section 3 describes the problem formulation, Section 4 and 5 respectively introduce the attention-cell LSTM for ASR and the proposed summarization using ASR, Section 6 discusses the experiments for ASR and summarization, and Section 7 concludes our work and future work.

2 Related Work

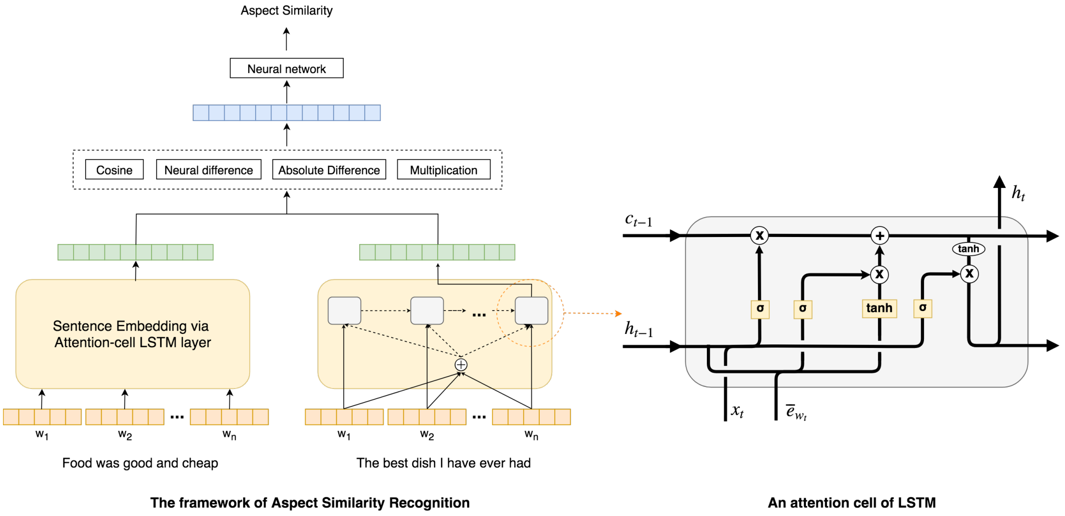
In the scope of this paper, we focus on discussing neural-based systems for generic and opinions summarization. For a comprehensive literature of non-neural techniques, we refer the reader to Liu and Zhang (2012).

For extractive generic summarization, Cao et al. (2015) rank sentences in a parsing tree via a recursive neural network. However, the model requires handcrafted features as input. Cheng and Lapata (2016) propose an end-to-end model for extracting words and sentences. In this system, a document is encoded via convolutional and recurrent layers, then an attention architecture is employed to extract sentences and words. Follow this work, Zhou et al. (2018) enhance the previous system by jointly learning to score and select sentences. By integrating sentence scoring and selecting into one phase, as the model selects a sentence, the sentence is scored according to the partial output summary and current extraction state.

To our knowledge, the first neural-based model of extractive opinions summarization is proposed by Kågebäck et al. (2014), which uses an unfolding recursive auto-encoder to learn phrase embeddings and measures similarity by Cosine and Euclidean distance. The limitation of this system is to purely rely on semantic similarity without taking into account the aspect information. Yang et al. (2017) use the unsupervised neural attention-based aspect autoencoder (ABAE) (He et al., 2017b) for presenting each aspect in an aspect embedding space. Then, the representative sentence for each aspect is selected via its distance with the centroid of that aspect. For summarization, however, ABAE is not efficient compared to K-mean in the aspects which occur more frequently in the dataset. Angelidis and Lapata (2018) introduce seed words of each domain to the autoencoder ABAE. This weakly-supervised model which is trained under multi-task objective outperforms the unsupervised model for aspect extraction. Different from the previous work in aspect-based opinions summarization, we apply aspect similarity recognition (ASR) instead of aspect extraction. ASR facilitates the problem of domain adaptation in summarization.

3 Problem Formulation

Every product e contains a set of reviews $R_e = \{r_i^e, ..., r_n^e\}$ expressing users' opinions on that product. A review r_i^e is viewed as a sequence of sentences $(s_1, ..., s_m)$. For each product e, our goal is to select the most salient sentences in reviews R_e for producing a summary. The proposed



The framework of Aspect Similarity Recognition

An attention cell of LSTM

Figure 1: The proposed framework for the aspect similarity task

approach is divided into subtasks as follows:

1. **Sentiment prediction** determines the overall polarity $p_s \in [-1, +1]$ a sentence carries, where $-1, +1$ respectively indicate maximally negative and positive. According to Angelidis and Lapata (2018), highly positive or negative opinions are more likely to contain informative text than neutral ones. In our system, we use the ensemble sentiment classifier proposed by Huy Tien and Minh Le (2017), which achieves strong performances at sentence level.

2. **Semantic textual similarity** measures the semantic similarity q_{ij} of two sentences i and j, which plays an important role in identifying the most informative sentences as well as redundant ones. We use the state-of-the-art multi-level comparison model (Tien et al., 2018) for this task.

3. **Aspect similarity Recognition (ASR)** predicts a probability r_{ij} that two sentences i and j shares at least one aspect. This subtask facilitates the elimination of redundant text in summarization, especially for domain adaptation.

4. **Summarization Generation** employs the three signals above for ranking sentences. A concise and informative summary of a product e is generated by selecting the most salient sentences from reviews R_e.

Section 4 describes in details the attention-cell LSTM for the ASR task and Section 5 explains how to combine the polarity, semantic and aspect similarity to produce a summary.

4 Attention Cell LSTM

According to Nguyen et al. (2018), recurrent neural networks efficiently capture aspect relationships. For dealing with the remaining difficulties of this task, the authors analyze the necessary of an attention mechanism. For that reason, we aim to emphasize salient words as encoding sentences over LSTM. A straightforward approach is to learn attention signals by self-attention and then apply these signals into inputs before feeding them into LSTM. In other words, these attention signals are applied to all gates of a LSTM cell. However, we assume emphasized input makes the cell forget more information on the previous state (the forget gate's function) while this state stores the most salient information by the support of attention signals. This conflict causes the inefficiency of integrating attention signals with LSTM. Therefore, we propose a novel LSTM cell which prevents the state from forgetting too much salient information as employing attention signals for encoding sentences. For the ASR task, the proposed attention-cell LSTM outperforms the conventional LSTM with/out using attention in both of settings: in-domain and cross-domain.

By representing a word w_i by a pre-trained word embedding e_{w_i}, we construct a sentence S of n words as a sequence of n word embeddings $S = [e_{w_1}, e_{w_2}, ..., e_{w_n}]$. Contextual information is incorporated in the word embeddings over the bidirectional GRU (Bahdanau et al., 2014) and then the self-attention signal a_i of w_i is learned as follows (from Yang et al. (2016)):

$$\overleftarrow{h_i} = \overleftarrow{GRU}(e_{w_i}) \tag{1}$$

$$\overrightarrow{h_i} = \overrightarrow{GRU}(e_{w_i}) \tag{2}$$

$$h_i = \overleftarrow{h_i} \oplus \overrightarrow{h_i} \tag{3}$$

$$u_i = tanh(W_a h_i + b_a) \tag{4}$$

$$a_i = \frac{exp(u_i^T u_a)}{\sum_i exp(u_i^T u_a)} \tag{5}$$

$$\bar{e}_{w_i} = e_{w_1} a_i \tag{6}$$

where $\oplus$ is concatenation operator, w_a, b_a, u_a are respectively a weight matrix, a bias, and a context vector. These parameters are randomly initialized and optimized during training.

A sentence s is transformed to a fix-length vector e_s by recursively applying a LSTM cell to each word embedding e_{w_t} and the previous step h_{t-1}. At each time step t, the LSTM unit with l-memory dimension defines six vectors in $\mathbb{R}^l$: input gate i_t, forget gate f_t, output gate o_t, tanh layer u_t, memory cell c_t and hidden state h_t (Tai et al., 2015). We modify the conventional LSTM cell to employ attention signals without the conflict of remembering and forgetting as follows:

$$i_t = \sigma(W_i \bar{e}_{w_t} + U_i h_{t-1} + b_i) \tag{7}$$

$$f_t = \sigma(W_f e_{w_t} + U_f h_{t-1} + b_f) \tag{8}$$

$$o_t = \sigma(W_o e_{w_t} + U_o h_{t-1} + b_o) \tag{9}$$

$$u_t = \tanh(W_u \bar{e}_{w_t} + U_u h_{t-1} + b_u) \tag{10}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot u_t \tag{11}$$

$$h_t = o_t \odot \tanh(c_t) \tag{12}$$

$$e_s = h_n \tag{13}$$

where $\sigma, \odot$ respectively denote a logistic sigmoid function and element-wise multiplication; W_i, U_i, b_i are respectively two weights matrices and a bias vector for input gate i. The denotation is similar to forget gate f, output gate o, tanh layer u, memory cell c and hidden state h. In the attention-cell LSTM, we introduce the attention signal a_t to only the input gate i_t and the tanh layer u_t, which are in charge of deciding what new information is going to be stored in the cell state. This approach allows the LSTM cell to employ attention for remembering salient information and avoid the unexpected effect of attention on the forget gate.

We visualize how the attention-cell LSTM manipulates attention signals in Figure 1. The four metrics are used to evaluate the relationship between two sentences e_{s_1} and e_{s_2} as follows (from Nguyen et al. (2018)):

Cosine similarity:

$$d_{cosine} = \frac{e_{s_1} \cdot e_{s_1}}{\|e_{s_1}\| \, \|e_{s_2}\|} \tag{14}$$

Multiplication vector & Absolute difference:

$$d_{mul} = e_{s_1} \odot e_{s_2} \tag{15}$$

$$d_{abs} = |e_{s_1} - e_{s_2}| \tag{16}$$

where $\odot$ is element-wise multiplication.

Neural difference:

$$x = e_{s_1} \oplus e_{s_2} \tag{17}$$

$$d_{neu} = W^{neu} x + b^{neu} \tag{18}$$

where W^{neu} and b^{neu} are respectively a weight matrix and a bias parameter.

As a result, we have a sentence-sentence similarity vector d^{sent} as follows:

$$d^{sent} = d_{cosine} \oplus d_{mul} \oplus d_{abs} \oplus d_{neu} \tag{19}$$

The sentence-sentence similarity vector is transferred into an aspect similarity label $\hat{y}$ through a two layers neural network as follows:

$$sim^{sent} = \sigma(W^{sent} d^{sent} + b^{sent}) \tag{20}$$

$$s\bar{i}m^{sent} = dropout(sim^{sent}) \tag{21}$$

$$\hat{y} = \sigma(W^y s\bar{i}m^{sent} + b^y) \tag{22}$$

where W^{sent}, W^y, b^{sent}, and b^y are weight matrices and bias parameters, respectively.

Dropout layer (Srivastava et al., 2014) is applied to our model. Dropout prevents networks from overfitting via randomly dropping out each hidden unit with a probability p on each presentation of each training case. We train this model under the cross entropy loss function and AdaDelta as the stochastic gradient descent (SGD) update rule. Details of Adadelta method can be found in Zeiler (2012).

5 Opinion Summarization

Given a product e, we aim to rank a set of sentences $D = \{s_i\}$ from the reviews talking about the product e. The procedure of scoring and selecting sentences for constructing an opinion summary K of the product e is as follows:

1. In the first step $t = 0$, we score each sentence $s_i \in D$ and select the most salient sentence $\hat{s}^0$ for the summary K:

$$asp_{s_i}^{t=0} = \frac{1}{|D|} \sum_{j \in D} r_{ij} \quad (23)$$

$$sim_{s_i}^{t=0} = \frac{1}{|D|} \sum_{j \in D} q_{ij} \quad (24)$$

$$sal_{s_i}^{t=0} = (1 + \alpha |p_{s_i}|) * asp_{s_i}^{t=0} * sim_{s_i}^{t=0} \quad (25)$$

$$\hat{s}^0 = \arg\max_{s_i \in D} \{sal_{s_i}^{t=0}\} \quad (26)$$

$$K^{t=1} = K \cup \{\hat{s}^0\} \quad (27)$$

$$D^{t=1} = D \setminus \{\hat{s}^0\} \quad (28)$$

At the step $t = 0$, the salient sal_{s_i} is computed by the semantic similarity sim_{s_i}, the aspect coverage sim_{s_i} and the polarity p_{s_i}. Different from the previous works, we also take into account the aspect coverage in which a sentence carrying more aspects has a higher salient score. In addition, the polarity of a sentence contributes to its ranking by a coefficient $\alpha \in [0, 1]$.

2. In the next step t, the salient sentence $\hat{s}^t$ is selected as follows:

$$asp_{s_i}^{t} = \frac{1}{|D^t|} \sum_{j \in D^t} r_{ij} \quad (29)$$

$$sim_{s_i}^{t} = \frac{1}{|D^t|} \sum_{j \in D^t} q_{ij} \quad (30)$$

$$\bar{sal}_{s_i}^{t=0} = (1 + \alpha |p_{s_i}|) * asp_{s_i}^{t} * sim_{s_i}^{t} \quad (31)$$

To avoid the redundant information, we penalize each sentence s_i by the aspect similarity $acov_{s_i}^t$ and semantic similarity $scov_{s_i}^t$ of that sentence with the selected sentences, in which β is a coefficient:

$$acov_{s_i}^{t} = \frac{1}{|K^t|} \sum_{j \in K^t} r_{ij} \quad (32)$$

$$scov_{s_i}^{t} = \frac{1}{|K^t|} \sum_{j \in K^t} q_{ij} \quad (33)$$

$$sal_{s_i}^{t} = \bar{sal}_{s_i}^{t} - \beta * acov_{s_i}^{t} * scov_{s_i}^{t} \quad (34)$$

$$\hat{s}^t = \arg\max_{s_i \in D^t} \{sal_{s_i}^{t}\} \quad (35)$$

$$K^{t+1} = K^t \cup \{\hat{s}^t\} \quad (36)$$

$$D^{t+1} = D^t \setminus \{\hat{s}^t\} \quad (37)$$

3. We repeat step 2 until the number of selected sentences is reached or the most salient score at the current step t is lower than a threshold. To avoid missing topic words in a summary, in step 1 and 2, we only select sentences containing words belonging to the list of frequent words on that topic. According to our observation, the topic words are the most frequent.

6 Experiments & Results

6.1 Aspect Similarity Recognition

We evaluate the attention-cell LSTM on ASRcorpus (Nguyen et al., 2018), which contains sentences from the SemEval 2016 dataset with two domains: RESTAURANT and LAPTOP. Each sample is a pair of sentences annotated as aspect similarity ($label = 1$) or not aspect similarity ($label = 0$). Table 2 reports the statistic of ASRCorpus in details.

	RESTAURANT			LAPTOP		
	Train	Dev	Test	Train	Dev	Test
Sentences	1239	469	587	1657	382	573
Sentence pairs	458K	68K	98K	447K	26K	44K
Similarity	229K	34K	49K	223K	13K	22K
Not similarity	229K	34K	49K	223K	13K	22K
Vocabulary	3769			3649		

Table 2: Statistic of ASRCorpus

We compare our model to some strong baselines as well as the conventional recurrent networks using attention. We choose the optimal values of hyper-parameters in our model and baselines via a grid search on 30% of LAPTOP domain. Because the number of RESTAURANT's categories is smaller than LAPTOP's, the performance of RESTAURANT domain is better.

Table 3 reports the experimental results. By employing efficiently attention signals, the attention-cell LSTM outperforms the conventional recurrent models using attention. As we analysis in Section 4, applying attention to all gates of a LSTM cell causes the conflict of remembering and forgetting. This drawback makes the training of the LSTM-attention model inefficient. Consequently, the trained LSTM-attention model predicts the same label for all inputs.

We also evaluate how the models perform in cross-domain setting where the models are trained on one domain dataset and tested on the other.

These results also prove that these approaches are potential to cross-domain application. We observe that a set of salient words in each domain is different. Therefore, the support of attention signals in domain adaptation is not significant compared to the recurrent models without attention.

Method	RES	LAP	→RES	→LAP
Word Average	70.75	65.12	54.5	54.59
CNN	77.57	67.23	54.08	54.49
LSTM	79.4	70.21	59.1	57.59
BiLSTM	79.2	71.14	59.2	57.95
Attention	78.79	68	57.92	54.55
LSTM-attention	50	50	50	50
Attention-Cell LSTM	**80**	**72.73**	**59.77**	**58.1**
Attention-Cell BiLSTM	79.42	71.65	59.3	58

Table 3: The in-domain and cross-domain experimental results on the two domains: RESTAURANT and LAPTOP. ”→Y” denotes that models are tested on Y but trained on the other. Accuracy metric is used for evaluation. The results are statistically significant at $p < 0.05$ via the pairwise t-test.

To obtain deeper analysis, we inspect the attention-cell LSTM's performance on each class (e.g., “similarity” and “not similarity”) by precision, recall and F1 scores reported in Table 4. In both of the domains and settings, the model performs better on “not similarity” class than “similarity” class in terms of F1 score. According to the results in cross-domain setting, we could conclude that the models learn rules, patterns for identifying aspect similarity rather than remembering topic words and keywords in a particular domain.

Domain	*Class*	*Precision*	*Recall*	$F1$
RES	Not Similarity	0.76	0.88	0.81
	similarity	0.86	0.82	0.78
LAP	Not Similarity	0.68	0.87	0.76
	similarity	0.82	0.59	0.68
→RES	Not Similarity	0.58	0.68	0.63
	similarity	0.62	0.52	0.56
→LAP	Not Similarity	0.56	0.72	0.63
	similarity	0.61	0.44	0.51

Table 4: The attention-cell LSTM's performance on each class.

6.2 Opinion Summarization

The Opinosis dataset (Ganesan et al., 2010) includes user reviews of 51 different topics (e.g., hotel, car, product). Each topic includes between 50 and 575 sentences made by various authors and around 4 reference summaries created by human. The corpus is suited for opinion summarization as well as evaluating the ability of domain adaptation.

We use ROUGE to assess the agreement of generated summaries and gold summaries. Our experiments include ROUGE-1, ROUGE-2 and, ROUGE-SU4, which base on one-gram, bi-gram and skip-bigram co-occurrences respectively.

The model for each subtask in our summarization system is implemented as follows:

- Sentiment prediction: the ensemble classifier (Huy Tien and Minh Le, 2017) is trained on Stanford Sentiment Treebank (Socher et al., 2013) with the accuracy of 88.6%.

- Semantic textual similarity : the multi-level comparison model (Tien et al., 2018) is trained on STSbenchmark[1] with the accuracy of 82.45%.

- Aspect similarity recognition: the attention-cel LSTM is trained on the ASRcorpus of the both domains with the accuracy of 76.2%.

- Summary generation: we set $\alpha = 1.67$ and $\beta = 0.1$. The number of the most frequent words is three. These parameters are optimized over a set of 5 topics randomly selected from the Opinosis dataset. According to the analysis of (Ganesan et al., 2010), the size of a summary is two sentences.

For comparison, we use MEAD (Radev et al., 2000) and CW-Add$_{Euc}$ (Kågebäck et al., 2014) as baselines. MEAD is an extractive method based on cluster centroids which selects the salient sentences by a collection of the most important words. CW-Add$_{Euc}$ measures the Euclidean similarity between two sentences by their continuous vector space. In addition, we also report the contribution of using aspect and sentiment information in summarization. The results denoted OPT$_R$ and OPT$_F$ in Table 5 describe the upper bound score

[1] http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

Method	ROUGE-1			ROUGE-2			ROUGE-SU4		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
OPT$_R$	57.86	21.96	30.28	22.96	12.31	15.33	29.5	13.53	17.7
OPT$_F$	45.93	48.84	46.57	20.42	19.94	19.49	23.17	26.5	23.7
MEAD	**49.32**	9.16	15.15	**10.58**	1.84	3.08	**23.16**	1.02	1.89
CW-Add$_{Euc}$	29.12	22.75	24.88	5.12	3.6	4.1	10.54	7.59	8.35
The proposed summarizer									
Semantic	28.24	28.63	27.62	7.34	7.19	7	10.69	10.94	10.4
Semantic + Aspect	29.2	**29.19**	**28.24**	7.45	**7.29**	**7.12**	11.25	**11.26**	**10.78**
Aspect + Polarity	27.77	27.86	26.92	7.24	7.09	6.93	10.42	10.55	10.04
Semantic + Aspect + Polarity	28.56	28.31	27.5	7.06	6.84	6.71	10.92	10.83	10.4

Table 5: Performance comparison between the proposed methods and baselines.

of recall and F-score respectively. As the reference summaries of Opinosis are generated in abstractive approach by humans, our generated summaries cannot fully match with the reference summaries. For example, the maximum recall which an extractive method could achieve in ROUGE-1 is 57.86%.

While MEAD selects long sentences (around 75 words) containing a lot of salient words to achieve a high score in recall but low in precision, our approach obtains a balance between these scores with quite shorter sentences (around 17 words).

Positive sentences
I purchased a 2007 Camry because of the looks of the redesigned model and because of the legendary Toyota quality and reliability.
The Concierge staff, exceptional and extremely helpful, right from suggestions on transportation excursion options to recommending an amazing restaurant.
When I checked in, I asked to be shown several rooms and the staff was happy to do so.
Negative sentences
My wife does say the vehicle is not as comfortable for long trips as other cars we've owned.
We had to go up a floor and into a service area to find ice.
The rude and poor service started from the concierge who was curt when I asked a question .

Table 6: Some sentences carrying the most polarity in the Opinosis dataset.

To analyze why sentiment signals cause negative impacts on the summarization generation, we inspect the most polarity sentences in the corpus. Some typical sentences are listed in Table 6. We observe that most of these sentences express individual experiences and too subjective to be selected for summarization. According to the Opinosis dataset, overstrong words (i.e., rude, extremely) and subjective words (i.e., my wife, I, we) are seldom present in a summary. These factors lead to an unexpected result of using polarity information in summarization although sentences carrying the most polarity are still informative.

Domain	*Class*	*Semantic + Aspect*
Tablet	More informative	33%
	Less informative	13%
	Equally informative	54%
Others	More informative	17%
	Less informative	8%
	Equally informative	72%

Table 7: Informative test for using Semantic with Aspect against without Aspect.

We expect that aspect signals support to generate an informative summary, which is a summary carrying salient information on various aspects. However, the ROUGE metric measures the number of matches between two pieces of text, so it is difficult to compare which one is more informative. Therefore, we execute an **informative test** to understand whether aspect signals help to generate a more informative summary. Given reference summaries and two summaries generated by the system with/out using aspect signals respectively, three persons are asked to select one of the three answers: which system's summary is more informative, or both of them are equally informative. The inter-rater agreement Cohen's Kappa score for each pair of assessors is higher than 0.74. The overall answer is concluded by the majority vote scheme. In case of receiving three different answers, that pair of summaries is assigned as equally informative. The result reported in Table 7 includes domain specification (15 samples in $Tablet$ and 36 samples in $Others$), which facilitates the evaluation of domain adaptation. As the ASR system is trained on the restaurant and laptop

Summary on the Comfort of Toyota Camry 2007	
Human	[1] The Camry offers interior comfort, while providing a quiet ride. Comfortable seating and easy to drive. [2] Overall very comfortable ride front and back. Nice and roomy. [3] Its very comfortable and a quiet ride with low levels of noise.
Semantic	The ride is quiet and comfortable. Very comfortable, quiet interior.
Semantic + Aspect	The ride is quiet and comfortable. Very comfortable ride and seating.
Summary on the location of Holiday Inn London	
Human	[1] Location is excellent, very close to the Glouchester Rd. Tube stop. [2] Excellent location. Near the tube station. [3] The location is excellent. The hotel is very convenient to shopping, sightseeing, and restaurants. It is located just minutes from the tube stations.
Semantic	Great location but don't bring the car! Great location great breakfast!
Semantic + Aspect	Great location but don't bring the car! Great location for the tube and bus!

Table 8: Human and system summaries for some products/services. For each topic, we list three summaries by human.

dataset, we consider tablet's topics in the Opinosis corpus as in-domain and others as out-of-domain. According to the informative test, the system with aspect dominates in both of the domains ($Tablet$ and $Others$). This result proves the contribution of aspect signals and the domain adaptation of the ASR system.

To obtain a better view of the advantages and disadvantages in our system, we show some generated summaries against reference summaries in Table 8. In extractive methods, the most salient sentences are selected from different reviewers, so it is possible to have repeated information in a summary. For instance in the case #1, the first sentence mentions *quiet* and *comfortable ride* while the second one contains *ride* and *seating*. Although these sentences still have different opinions (i.e., *quiet* vs *seating*), the repeat of *comfortable ride* downgrades the generated summary's quality. For improvement, we suggest a post-processing for a more concise summary by filtering redundant information. As the proposed aspect-based system ranks a sentence by not only semantic cover but also aspect cover, it selects the more salient opinions for summarization. For instance, although both of the systems extract different features (e.g., interior vs seating, breakfast vs tube and bus), the opinions (i.e., seating, tube and bus) chosen by the system with aspect support are more suited to the reference summaries.

In each topic, although the reference summaries and generated summary share most of the meaning, they deliver information in different ways and words. This fact makes the quality evaluation of generated summaries difficult. In addition to the ROUGE metric, we conducted the informative test for quality evaluation. However, for a large corpus or multiple systems comparison, this test requires a huge amount of human effort. Therefore, it is a high demand to have a reliable metric for summaries evaluation without human involvement.

7 Conclusion

In this work, we introduced a novel aspect-based opinions summarization framework using aspect similarity recognition. This subtask relaxes the constraint of predefined aspects in conventional aspect categorization tasks. For ASR tasks, we proposed an attention-cell LSTM to integrate efficiently attention signals into LSTM. This approach outperforms the baselines on both settings of in-domain and cross-domain. For summarization, we evaluated our system on the Opinosis corpus. In addition to ROUGE metric, an informative test with human involvement was implemented to show the domain adaptation ability of our system and how informative our generated summaries are. In the corpus, we observe that sentences carrying the most polarity are not suited to summarization. Therefore, employing sentiment for summarization needs deeper analysis. Due to the ASR task's advantage, we believe that it has a high demand in some fundamental tasks of natural language processing such as information retrieval, and sentence comparison.

Acknowledgements

This work was supported by JST CREST Grant, Japan. We are also immensely grateful to the reviewers for their comments.

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *EMNLP*.

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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, AAAI'15, pages 2153–2159.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 484–494.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pages 340–348.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017a. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017b. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, KDD '04, pages 168–177. https://doi.org/10.1145/1014052.1014073.

Minqing Hu and Bing Liu. 2006. Opinion extraction and summarization on the web. In *Proceedings of the National Conference on Artificial Intelligence*. volume 2.

Nguyen Huy Tien and Nguyen Minh Le. 2017. An ensemble method with sentiment features and clustering support. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, pages 644–653. http://aclweb.org/anthology/I17-1065.

Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2014. Extractive summarization using continuous vector space models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*. Association for Computational Linguistics, pages 31–39.

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

Bing Liu and Lei Zhang. 2012. *A Survey of Opinion Mining and Sentiment Analysis*, Springer US, Boston, MA, pages 415–463.

Samaneh Moghaddam and Martin Ester. 2013. The flda model for aspect-based opinion mining: Addressing the cold start problem. In *Proceedings of the 22Nd International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '13, pages 909–918. https://doi.org/10.1145/2488388.2488467.

Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '06, pages 573–580. https://doi.org/10.1145/1148170.1148269.

Huy Tien Nguyen, Quan-Hoang Vo, and Minh-Le Nguyen. 2018. A deep learning study of aspect similarity recognition. In *Proceedings of the 10th International Conference on Knowledge and Systems Engineering*.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004. Mead - a platform for multidocument multilingual text summarization. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Richard Socher, Alex Perelygin, Jean Y 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 conference on empirical methods in natural language processing (EMNLP)*. Citeseer, volume 1631, page 1642.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15:1929–1958.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. 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)*. Association for Computational Linguistics, pages 1556–1566.

Huy Nguyen Tien, Minh Nguyen Le, Yamasaki Tomohiro, and Izuha Tatsuya. 2018. Sentence modeling via multiple word embeddings and multi-level comparison for semantic textual similarity. *CoRR* abs/1805.07882. http://arxiv.org/abs/1805.07882.

Yinfei Yang, Cen Chen, Minghui Qiu, and Forrest Bao. 2017. Aspect extraction from product reviews using category hierarchy information. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, pages 675–680.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. 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*. Association for Computational Linguistics, pages 1480–1489. https://doi.org/10.18653/v1/N16-1174.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR* abs/1212.5701. http://arxiv.org/abs/1212.5701.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 654–663.