TransSum: Translating Aspect and Sentiment Embeddings for Self-Supervised Opinion Summarization

Ke Wang, Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
{wangke17,wanxiaojun}@pku.edu.cn

Abstract

In this paper, we propose a novel self-supervised opinion summarization framework TransSum, which models opinion summaries as translations operating on the low-dimensional aspect and sentiment embedding spaces. Specifically, we propose two contrastive objectives to learn the crucial aspect and sentiment embeddings of reviews, by taking advantage of the intra- and inter-group invariances that have not been considered in previous studies. Furthermore, these embeddings can be used to reduce opinion redundancy and construct highly relevant reviews-summary pairs to train a supervised multi-input opinion summarization model. Experimental results on three different domains show that TransSum outperforms several strong baselines in generating informative, relevant and low-redundant summaries, unveiling the effectiveness of our approach.

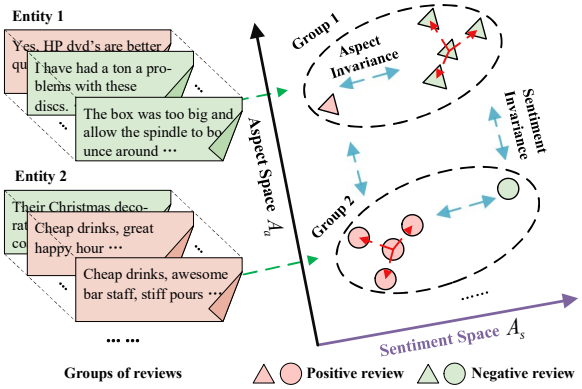

Figure 1: The proposed TransSum targets at learning corresponding aspect and sentiment embeddings for reviews (green arrows) through contrastive learning based on the aspect and sentiment invariances (blue arrows). These embeddings are used to construct reviews-summary pairs of high relevance (red arrows), so as to train a supervised multi-input opinion summarization model. Best view in color.

1 Introduction

Opinion summarization, which focuses on automatically generating summaries that reflect salient opinion information expressed in a group of documents (e.g., user reviews of a product in Figure 1), has been receiving great attention due to its usefulness and effectiveness for displaying massive opinion texts (Ku et al., 2006; Cheung et al., 2009; Chu and Liu, 2019). For example, a representative review summary of a product can not only replace large amounts of reviews for potential customers to read, but also provide more explanations than a simple overall sentiment rating, such as " What is the biggest complaint on the iPod 'screen' ?".

However, compared with supervised summarization in the domain of news articles, the annotated training data for opinion summarization is expensive to acquire. Due to the lack of gold-standard summaries for training, most existing works focus on unsupervised opinion summarization and treat it as a normal multi-document summarization task. They either struggle to reduce the opinion redundancy efficiently or output summaries lacking relevance to input reviews. Particularly, many previous studies focus on extractive approaches (Paul et al., 2010; Fabbrizio et al., 2014; Rossiello et al., 2017; Narayan et al., 2019), which copy texts from the input reviews but tend to be redundant and less informative (Chu and Liu, 2019). Some recently proposed abstractive methods are based on unsupervised representation learning, such as auto-encoder (Chu and Liu, 2019; Amplayo and Lapata, 2019; Brazinskas et al., 2020a) or variational auto-encoder (Brazinskas et al., 2020b; Angelidis et al., 2020), but mainly focus on the content transformation within each group of reviews. Other studies aim to create synthetic reviews-summary pairs to train a supervised multi-document summarization model (Amplayo and Lapata, 2019; Brazinskas et al., 2020b; Amplayo and Lapata, 2020; Amplayo

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 729–742
August 1–6, 2021.

et al., 2021), such as sampling a review from a corpus of product reviews and treating it as a summary of the remaining reviews, but such settings may lack rationality to guarantee the relevance of reviews and constructed pseudo-summaries.

In an effort to overcome these challenges, we propose a novel self-supervised framework for opinion summarization, TransSum, which consists of two main modules and does not require any gold summaries for training. (i) In the translation-based review modeling module, we expect to represent reviews with only their corresponding aspect and sentiment embeddings (as shown in Figure 1) with the purpose of reducing unnecessary information. We decompose each review into the aspect and sentiment embeddings through reconstruction and contrastive learning (van den Oord et al., 2018; He et al., 2020) based on two novel intra- and inter-group invariances: First, the real-world reviews in a group may discuss various opinions covering different aspects, but they are dependent with a specific entity (e.g., reviews about a specific product). Hence, the aspect information of the reviews in the same group should be closer than that of different groups (i.e, aspect invariance in Figure 1), that is, the distances between intra-group reviews should be less than the ones between inter-group reviews in the aspect embedding space. Second, the sentiment information of reviews with the same sentiment label should be closer than that of different sentiment labels (i.e, sentiment invariance in Figure 1), that is, the distances between reviews with the same sentiment should be less than the ones between reviews with different sentiments in the sentiment embedding space. (ii) In our multi-input opinion summarization module, we reduce opinion redundancy by combining similar embeddings, and use reviews with similar aspects (embeddings) to construct reviews-summary pairs of high relevance, which are used to train a supervised multi-input summarization model.

We conduct extensive experiments to show the superiority of our method. Experimental results on three different domains show that our method outperforms several strong baselines in generating informative, relevant, low-redundant and fluent summaries. We also perform ablation studies to analyze the effectiveness of the modules in our method.

In summary, our main contributions are:

- To the best of our knowledge, we are the first to generate opinion summaries from only the aspect and sentiment embeddings, which unlocks the critical bottleneck for unsupervised opinions modeling and takes a step forward towards more complex and controllable designs.
- We propose a novel self-supervised framework (TransSum) to generate opinion summaries without access to expensive annotations by disentangling reviews into aspect and sentiment embeddings and automatically constructing highly relevant reviews-summary pairs for model training.
- Experimental results on three domains show that our approach outperforms several strong baselines, especially in terms of relevance and non-redundancy.

2 TransSum

2.1 Overview

As aforementioned, a good opinion summary needs to cover major opinions/sentiments on different aspects of the entity (e.g., a movie, product, business) discussed in a group of reviews. Inspired by this observation, we propose a self-supervised framework (titled TransSum), aiming to generate opinion summaries without access to expensive annotations by interpreting them as translations operating on the aspect and sentiment embeddings.

As noted in a recent theoretical model of importance in summarization (Peyrard, 2019), a good summary should meet three requirements: (i) minimum redundancy, (ii) maximum relevance with the input document(s), and (iii) maximum informativeness. Based on the observation that reviews are usually created to express users' sentiments on certain aspects of a specific entity (e.g., the price and battery of a PC), we reasonably define informativeness, the amount of new information contained in the opinion summary relative to the background knowledge, as the aspect and sentiment information. The purpose is to reduce unnecessary information in the opinion summary, such as personal information or other irrelevant details.

Specifically, TransSum consists of two main components: **(1)** A translation-based review modeling module that learns only aspect and sentiment embeddings from each review for opinion summarization, to keep only the key and useful information (requirement iii). The aspect and sentiment

embeddings of reviews are learned through reconstruction and two contrastive objectives, which take advantage of aspect and sentiment invariance of intra- and inter-group reviews (detailed in Sec 2.3). **(2)** A multi-input opinion summarization module that learns to generate the summary from the redundancy-reduced combination of the aspect and sentiment embeddings of input reviews (requirement i). It is trained by synthetic reviews-summary pairs of high relevance (requirement ii), which are constructed based on the assumption that reviews with the same aspect information (embeddings) are likely to express similar opinions (detailed in Sec 2.4).

2.2 Notations

More formally, let $\mathbb{D}$ denote a review corpus in a domain (e.g., Products' reviews), which consists of m groups of reviews. For each group $\boldsymbol{G}$, we assume that it contains n reviews $\{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_i \cdots, \boldsymbol{r}_n\}$ about a specific entity e (e.g., a product), n is not a fixed number. For each review $\boldsymbol{r}_i$ in $\boldsymbol{G}$, we define its number of tokens $\boldsymbol{r}_i$ as $|\boldsymbol{r}_i|$, that is, $\boldsymbol{r}_i = \{\boldsymbol{r}_i^{(1)}, \cdots, \boldsymbol{r}_i^{(|\boldsymbol{r}_i|)}\}$, and use $\boldsymbol{r}_{-i} = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_{i-1}, \boldsymbol{r}_{i+1}, \cdots, \boldsymbol{r}_n\}$ to represent the remaining $n-1$ reviews. Each review has a binary sentiment label x (i.g., positive or negative), which indicates the overall sentiment polarity of the review. The aspect and sentiment embeddings of $\boldsymbol{r}_i$ are denoted as $\boldsymbol{a}_i \in \mathbb{R}^{|\boldsymbol{r}_i| \times k}$ and $\boldsymbol{s}_i \in \mathbb{R}^{|\boldsymbol{r}_i| \times k}$. E and D are encoder and decoder respectively.

The goal of opinion summarization is to generate a summary $\boldsymbol{y}$ that covers opinions mentioned in the group of reviews, in other words, $\boldsymbol{y}$ can be considered *"a representative review"* that can replace the group of reviews $\{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_i \cdots, \boldsymbol{r}_n\}$ in terms of informativeness. Note that we cannot access gold-standard opinion summaries for each group of reviews, as the human-annotated summaries do not exist in most domains.

2.3 Translation-Based Review Modeling

The translation-based review modeling module aims to learn aspect and sentiment embeddings for reviews (the left block in Figure 2).

For each review $\boldsymbol{r}_i$ in the group $\boldsymbol{G}$, we encode it using a Transformer (Vaswani et al., 2017) encoder E, and the output encoding $\boldsymbol{h}_i \in \mathbb{R}^{|\boldsymbol{r}_i| \times k}$ is:

$$\boldsymbol{h}_i = E(\boldsymbol{r}_i) = (\boldsymbol{h}_i^{(1)}, \cdots, \boldsymbol{h}_i^{(|\boldsymbol{r}_i|)}), \quad (1)$$

where k is the embedding dimension. Inspired by Zhong et al. (2019), we initialize the token embeddings of E with the ones of the BERT-base model (Devlin et al., 2019). Then we use projection matrices $\boldsymbol{A}_a \in \mathbb{R}^{k \times k}$ and $\boldsymbol{A}_s \in \mathbb{R}^{k \times k}$ to project $\boldsymbol{h}_i$ to the aspect and sentiment spaces as $\boldsymbol{a}_i$ and $\boldsymbol{s}_i$ (the blue and red squares in Figure 2), respectively.

$$\boldsymbol{a}_i = \boldsymbol{h}_i \boldsymbol{A}_a = (\boldsymbol{a}_i^{(1)}, \cdots, \boldsymbol{a}_i^{(|\boldsymbol{r}_i|)}) \quad (2)$$

$$\boldsymbol{s}_i = \boldsymbol{h}_i \boldsymbol{A}_s = (\boldsymbol{s}_i^{(1)}, \cdots, \boldsymbol{s}_i^{(|\boldsymbol{r}_i|)}) \quad (3)$$

For later use, we further denote $\hat{\boldsymbol{a}}_i = \frac{1}{|\boldsymbol{r}_i|} \sum_{j=1}^{|\boldsymbol{r}_i|} \boldsymbol{a}_i^{(j)}$ ($\hat{\boldsymbol{a}}_i \in \mathbb{R}^k$) and $\hat{\boldsymbol{s}}_i = \frac{1}{|\boldsymbol{r}_i|} \sum_{j=1}^{|\boldsymbol{r}_i|} \boldsymbol{s}_i^{(j)}$ ($\hat{\boldsymbol{s}}_i \in \mathbb{R}^k$) to represent the mean vectors of the embeddings, repectively.

Translation-Based Reconstruction: We assume that each review is "an opinion summary" of the user's intention and attitude, and model the review as the translation from aspect and sentiment embeddings, that is, $\boldsymbol{c}_i = \boldsymbol{a}_i + \boldsymbol{s}_i$ (the yellow square in Figure 2):

$$\boldsymbol{c}_i = \{\boldsymbol{a}_i^{(1)} + \boldsymbol{s}_i^{(1)}, \cdots, \boldsymbol{a}_i^{(|\boldsymbol{r}_i|)} + \boldsymbol{s}_i^{(|\boldsymbol{r}_i|)}\}, \quad (4)$$

To maximize informativeness and reduce unnecessary information, we hope to reconstruct $\boldsymbol{r}_i$ from only the embeddings $\boldsymbol{c}_i$ with a decoder D. The reconstruction loss is:

$$\mathcal{L}_{rec} = \mathbb{E}_{\boldsymbol{r}_i \sim \mathbb{D}}[\ell(D(\boldsymbol{c}_i), \boldsymbol{r}_i)], \quad (5)$$

where ℓ is the cross entropy loss (de Boer et al., 2005) and D is a Transformer (Vaswani et al., 2017) decoder with cross-attention on $\boldsymbol{c}_i$. Following previous arts (Amplayo et al., 2021; ElSahar et al., 2020), we adopt label smoothing method (Szegedy et al., 2016) on $\boldsymbol{r}_i$ instead of computing with categorical distributions.

Contrastive Learning of Aspect and Sentiment Embeddings: We perform contrastive learning to learn the aspect and sentiment embeddings, based on the following two contrastive objectives: (i) the aspect embeddings of intra-group reviews should be "closer" to each other than the ones of inter-group reviews, and (ii) the sentiment embeddings of reviews with the same sentiment label should be "closer" to each other than the ones of reviews with different sentiment labels, even if they are in different groups.

More concretely, for the aspect embedding $\boldsymbol{a}_i$, we except to make its similarity with a "similar" sample $\boldsymbol{a}_i^+$ far greater than the one with a "dissimilar" sample $\boldsymbol{a}_i^-$, that is, $Sim(\hat{\boldsymbol{a}}_i, \hat{\boldsymbol{a}}_i{}^+) \gg$

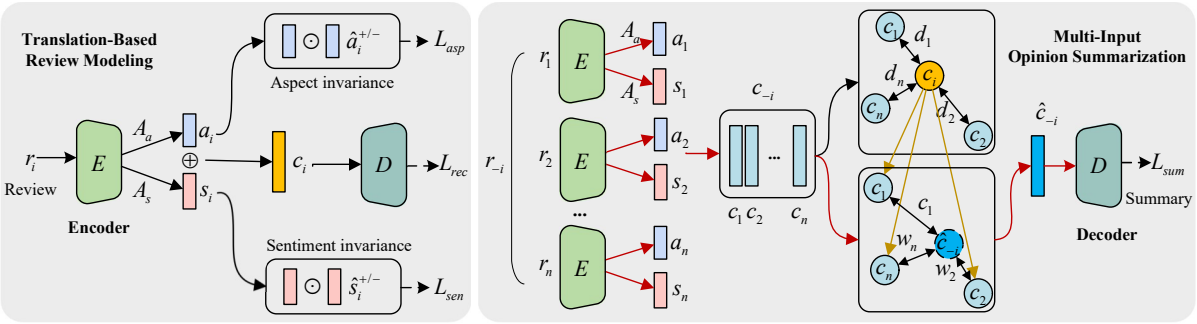

Figure 2: Architecture of TransSum, which consists of two main components: (1) a translation-based review modeling module that learns aspect and sentiment embeddings, and (2) a multi-input opinion summarization module that learns to generate summaries that are low-redundant and highly relevant to the input reviews. The encoder and decoder are shared, and the red arrows indicate the data flow in the inference phase. Best view in color.

$Sim(\hat{\boldsymbol{a}}_i, \hat{\boldsymbol{a}}_i^-)$. $\boldsymbol{a}_i^+$ is the aspect embedding of a review sampled from the same group, and $\boldsymbol{a}_i^-$ is the aspect embedding of a review sampled from other groups. In this work, we use the dot product between embeddings to measure similarity (i.e., Sim), which can be regarded as a measure of the angle between the two embeddings in the vector space. As a consequence, the aspect-based contrastive objective is:

$$\mathcal{L}_{asp} = -\mathbb{E}_{\boldsymbol{r}_i \sim \mathbb{D}}[\log \frac{\exp(Sim(\hat{\boldsymbol{a}}_i, \hat{\boldsymbol{a}}_i^+))}{\exp(Sim(\hat{\boldsymbol{a}}_i, \hat{\boldsymbol{a}}_i^+)) + \exp(Sim(\hat{\boldsymbol{a}}_i, \hat{\boldsymbol{a}}_i^-))}], \quad (6)$$

As for the sentiment embedding $\boldsymbol{s}_i$, the "similar" sample $\boldsymbol{s}_i^+$ is the sentiment embedding of a review sampled from different groups but with the same sentiment label, and the "dissimilar" sample $\boldsymbol{s}_i^-$ is the sentiment embedding of a review sampled from different groups and with a different sentiment label. Hence, the sentiment-based contrastive objective is defined as follows:

$$\mathcal{L}_{sen} = -\mathbb{E}_{\boldsymbol{r}_i \sim \mathbb{D}}[\log \frac{\exp(Sim(\hat{\boldsymbol{s}}_i, \hat{\boldsymbol{s}}_i^+))}{\exp(Sim(\hat{\boldsymbol{s}}_i, \hat{\boldsymbol{s}}_i^+)) + \exp(Sim(\hat{\boldsymbol{s}}_i, \hat{\boldsymbol{s}}_i^-))}]. \quad (7)$$

To the best of our knowledge, we are the first to go beyond the intra-group information modeling by further considering the inter-group level contrastive learning of aspect and sentiment embeddings.

To further enlarge the disagreements among the aspect/sentiment projection matrix and reduce the redundancy of parameters, we additionally add a regularization loss to encourage uniqueness:

$$\mathcal{L}_{reg} = ||\boldsymbol{A}_a^T \boldsymbol{A}_a - \boldsymbol{I}||^2 + ||\boldsymbol{A}_s^T \boldsymbol{A}_s - \boldsymbol{I}||^2, \quad (8)$$

where $\boldsymbol{I}$ is the identity matrix.

2.4 Multi-Input Opinion Summarization

After learning aspect and sentiment low-dimensional embeddings of reviews, we can construct reviews-summary pairs of high relevance based on the similarity of aspect embeddings, so as to train a supervised multi-input opinion summarization model (the right block in Figure 2).

Although real-world reviews in a group discuss various viewpoints covering different aspects under consideration, they are in fact focused on the same entity. In other words, they may repeat discussions about certain aspects many times, and may also include their own unique aspects. However, opinions on the same aspects are likely to be the same in real scenarios, e.g., knowing that most users complain about the high price of a product, the next price-focused review is likely to give a negative view. Based on this observation, we expect to reduce redundancy in similar aspects and use reviews with similar aspects to construct a high-quality data set whose reviews-summary data pairs are highly relevant.

High-Relevance Dataset Creation: We expect to find a subset of $\boldsymbol{r}_{-i}$ in which reviews are similar to $\boldsymbol{r}_i$ in the aspect embedding space, and use $\boldsymbol{r}_i$ as the target (pseudo) opinion summary of this subset of reviews. We have noticed that in real reviews, the majority of views on the same aspect are consistent with each other, so we believe most reviews-summary pairs created in this way can be used for training a model to capture and summarize the major opinions of the input reviews. More sophisticated ways for dataset creation will be left for further study.

In practice, we assign a weight w to each review in $\boldsymbol{r}_{-i}$, that is, assigning small values to low-

relevance reviews instead of looking for a subset of only high-relevance reviews (as shown by the yellow arrows in Figure 2). For each review $\boldsymbol{r}_j$ in $\boldsymbol{r}_{-i}$, we calculate the distance between it and $\boldsymbol{r}_i$ in the aspect embedding space as:

$$d_j = Sim(\hat{\boldsymbol{a}}_j, \hat{\boldsymbol{a}}_i), d_j \in \boldsymbol{d}_{-i}, \quad (9)$$

where $\boldsymbol{d}_{-i} = \{d_1, \cdots, d_{i-1}, d_{i+1}, \cdots, d_n\}$. Then we construct the reviews-summary pair $< \boldsymbol{r}_{-i}, \boldsymbol{r}_i >$ with the following weights $\boldsymbol{w}_{-i}$, which will be used later:

$$w_j = \frac{\exp(d_j)}{\sum_{d_z \in \boldsymbol{d}_{-i}} \exp(d_z)}, w_j \in \boldsymbol{w}_{-i}. \quad (10)$$

Note that some previous arts (Brazinskas et al., 2020b; ElSahar et al., 2020; Brazinskas et al., 2020a; Amplayo et al., 2021) adopted a leave-one-out self-supervision setting (Besag, 1975) similar to ours. But they did not take into account the relevance between each review and the pseudo summary, which can be considered as our special case with a uniform distribution $\boldsymbol{w}_{-i} = (\frac{1}{n-1}, \cdots, \frac{1}{n-1})$.

Embedding-Based Redundancy Reduction: Aside from creating a high-relevance synthetic dataset, we can use the learned embeddings to reduce redundancy. We regard the embedding differences of different reviews as their natural variation, and perform a weighted pooling operation to remove redundant information (similar embeddings). Therefore, we obtain $\hat{\boldsymbol{c}}_{-i}$ based on multiple inputs $\{w_1\boldsymbol{c}_1, \cdots, w_{i-1}\boldsymbol{c}_{i-1}, w_{i+1}\boldsymbol{c}_{i+1}, \cdots, w_n\boldsymbol{c}_n\}$. Note that $\boldsymbol{w}$ is a uniform distribution in the inference phase, that is, the weight of each input review is equal.

Finally, we generate the opinion summary of $\boldsymbol{r}_{-i}$ and the summarization loss $\mathcal{L}_{sum}$ is:

$$\mathcal{L}_{sum} = \mathbb{E}_{\boldsymbol{r}_i \sim \mathbb{D}}[\ell(D(\hat{\boldsymbol{c}}_{-i}), \boldsymbol{r}_i)], \quad (11)$$

where D is the decoder shared with the previous module and we also adopt label smoothing technique (Szegedy et al., 2016) on $\boldsymbol{r}_i$.

2.5 Training

Finally, we optimize the sum of the above losses:

$$\mathcal{L}_{final} = \mathcal{L}_{rec} + \mathcal{L}_{asp} + \mathcal{L}_{sen} + \mathcal{L}_{reg} + \mathcal{L}_{sum}. \quad (12)$$

We also explore non-equal weighting of the losses but do not find a meaningful difference in outcomes. We perform beam search decoding in the inference stage.

	Train	Dev	Test
Yelp	13,369 / 97.1	100 / 8.0	100 / 8.0
Amazon	192,742 / 24.9	28 / 8.0	32 / 8.0
RT	2,458 / 83.3	536 / 98.0	737 / 100.3

Table 1: Statistics of three datasets in different domains (i.e., businesses, products and movies). The format in the cells is "Number of entities / Average number of reviews per entity".

3 Experiments

Datasets: We conduct experiments on three opinion summarization benchmarks in different domains, including: (1) **Yelp** (Chu and Liu, 2019) which contains business customer reviews from the Yelp Dataset Challenge[1]; (2) **Amazon** (Brazinskas et al., 2020b) which includes a large corpus of product reviews for four Amazon categories (i.e., Electronics, Clothing, Shoes and Jewelry, Home and Kitchen, and Health and Personal Care)[2]; (3) **Rotten Tomatoes (RT)** (Wang and Ling, 2016) which has a large set of reviews for various movies written by critics[3]. The detailed statistics of the three datasets are shown in Table 1. For Yelp and Amazon, there are no gold standard summaries for large training corpora, but the small development and test sets have summaries written by Amazon Mechanical Turk (AMT) crowd-workers. In RT, each set of reviews has a gold-standard opinion summary written by an editor, but we do not use ground truth summaries for training due to the unsupervised setting. Note that all reviews have a binary sentiment label (e.g., positive or negative). For Yelp and Amazon which have 1–5 scale ratings, we mark reviews with scores below 3 as negative and the rest as positive. The implementation details of our method are shown in supplementary materials.

Compared Methods: We compare TransSum with several state-of-the-art unsupervised summary generation methods, and some of them can be essentially considered as special cases of our method.

For extractive systems where summaries are created by selecting a subset of salient sentences from the input reviews, they include: (1) **LexRank** (Erkan and Radev, 2004), a PageRank-like algorithm which selects the review closest to the centroid of a group as the summary; (2) **W2VCent**,

[1] https://github.com/sosuperic/MeanSum

[2] https://github.com/abrazinskas/Copycat-abstractive-opinion-summarizer

[3] https://web.eecs.umich.edu/∼wangluxy/data.html

a centroid-based multi-document summarization method that uses word embeddings (Mikolov et al., 2013) instead of TF-IDF to represent each sentence (Rossiello et al., 2017); and (3) **Multi-Lead-1** (See et al., 2017) which constructs the summary by selecting the leading sentences from each review of a group. Additionally, we also report the **upper bound** of extractive methods, i.e., the highest-scoring review in a group when computing ROUGE-L (Lin and Hovy, 2003) against reference summaries.

We also compare with six state-of-the-art abstractive models where summaries are generated from scratch, including: (1) **Opinosis** (Ganesan et al., 2010), a graph-based method that uses token-level redundancy to generate summaries; (2) **MeanSum** (Chu and Liu, 2019), an auto-encoder that generates summaries by reconstructing the mean of review encodings, which is in fact special cases of our method without contrastive transformations of aspect and sentiment embeddings and high-relevance dataset creation; (3) **OpinionDigest** (Suhara et al., 2020), a combination of an aspect-based sentiment analysis model and a phrase-to-review seq2seq model, which can be seen as using opinion phrases to model summaries rather than using the aspect and sentiment embeddings as we do; (4) **DenoiseSum** (Amplayo and Lapata, 2020), which create a synthetic dataset by treating a review and its noisy versions as the summary and pseudo-review input, instead of using the aspect similarity of real-world reviews like ours; (5) **CopyCat** (Brazinskas et al., 2020b), a hierarchical variational auto-encoder which learns a latent code of the summary and uses a leave-one-out self-supervision setting, and it can be regarded as a special case where TransSum does not consider the relevance of input reviews and the constructed summaries; and (6) **PlanSum** (Amplayo et al., 2021), which uses adversarial learning to learn the aspect and sentiment distributions of reviews, instead of the intra- and inter-group contrastive transformations we use. Note that we do not compare with methods using gold summaries, such as Brazinskas et al. (2020a).

3.1 Main Results

3.1.1 Automatic Evaluation

For automatic summary evaluation, we report the classical ROUGE (Lin and Hovy, 2003) scores on test sets. We report F-measure scores of ROUGE-1 (**R1**), ROUGE-2 (**R2**) and ROUGE-L (**RL**) in the experiments.

Table 2 contains the automatic evaluation results on three different datasets. From the results, we can see that: (1) Although extractive methods (e.g., LexRank, W2VCent and Multi-Lead-1) achieve comparable results, their upper bounds are affected by the data sets used. For example, the upper bound results of R2 and RL on Yelp are much lower than the other two, perhaps because most sentences on the Yelp dataset contain more redundant information. (2) Among abstractive models, OpinionDigest and CopyCat perform much better than Opinosis and MeanSum, showing the effectiveness of using opinion phrases or specific distributions to model opinion summaries. But our method surpasses them by a wide margin, indicating that the aspect and sentiment embeddings learned by contrastive learning are beneficial for modeling opinion summaries. (3) Impressively, we observe a large improvement brought by the creation of synthetic datasets (i.e., DenoiseSum, CopyCat and PlanSum), showing the usefulness of using reviews as pseudo-summaries. However, our method is superior to them, illustrating the importance of considering the relevance of the constructed reviews-summary pairs. (4) Overall, our model outperforms all baseline models on three datasets over all three metrics. It is also worth noting that TransSum even surpasses the upper bound of extractive methods on Yelp with an increase of 5.55, 2.3, and 2.2 points in ROUGE-1/2/L.

3.1.2 Human Evaluation

Further, we conduct a human evaluation to evaluate the quality of generated summaries more accurately.

We focus on five criteria: (1) the aspect-based informativeness indicator (**Aspect**) focuses on whether the summary covers common aspects discussed in the reviews, (2) the sentiment-based informativeness indicator (**Sentiment**) focuses on whether it agrees with their overall sentiment about different aspects. (3) the relevance indicator (**Relevance**) reflects whether the summary is relevant to the input reviews, (4) the non-redundancy indicator (**Non-Redundancy**) measures whether the summary contains unnecessary repetition, and (5) the fluency indicator (**Fluency**) shows whether the summary is well-formed and grammatical. We show the detailed questions in **supplementary materials**. We sampled 50, 32, and 50 review groups

Method	*Abstr.?*	Yelp			Amazon			RT		
		R1	**R2**	**RL**	**R1**	**R2**	**RL**	**R1**	**R2**	**RL**
Upper Bound (Extractive)	✗	31.07	6.11	18.11	33.98	7.88	21.60	30.94	10.75	24.95
LexRank(Erkan and Radev, 2004)	✗	25.50	2.64	13.37	28.74	5.47	16.75	14.88	1.94	10.50
W2VCent (Rossiello et al., 2017)	✗	24.61	2.85	13.81	28.73	4.97	17.45	13.93	2.10	10.81
Multi-Lead-1 (See et al., 2017)	✗	26.34	3.72	13.86	30.32	5.90	15.78	14.21	1.82	10.23
Opinosis (Ganesan et al., 2010)	✓	25.15	2.61	13.54	28.42	4.57	15.50	14.98	3.07	12.19
MeanSum (Chu and Liu, 2019)	✓	28.86	3.66	15.91	29.20	4.70	18.15	15.79	1.94	12.26
OpinionDigest(Suhara et al., 2020)	✓	29.30	5.77	18.56	-	-	-	-	-	-
DenoiseSum(Amplayo and Lapata, 2020)	✓	30.14	4.99	17.65	-	-	-	21.26	4.61	16.27
CopyCat(Brazinskas et al., 2020b)	✓	29.47	5.26	18.09	31.97	5.81	20.16	-	-	-
PlanSum(Amplayo et al., 2021)	✓	34.79	7.01	19.74	32.87	6.12	19.05	21.77	6.18	16.98
TransSum (Ours)	✓	**36.62***	**8.41***	**20.31***	**34.23***	**7.24***	**20.49***	**25.34***	**8.62***	**18.35***

Table 2: Automatic evaluation results on three datasets. We make the best results bold, and use "-" to indicate unreported results or unfound outputs. "$*$" means that the improvements over PlanSum are statistically significant with p-value ≤ 0.05 for t-test, and "*Abstr.?*" indicates whether the method is an abstractive approach.

	Method	Asp.	Sen.	Rel.	Non.	Flu.
Yelp	LexRank	-0.49	-0.31	-0.81	-0.60	-0.30
	PlanSum	-0.12	-0.23	-0.47	-0.13	-0.10
	TransSum	**0.26**	**0.13**	**0.57**	**0.24**	**0.02**
	Gold	0.35	0.41	0.71	0.49	0.38
Amazon	LexRank	-0.62	-0.47	-0.45	-0.53	-0.58
	PlanSum	0.08	-0.11	-0.06	-0.36	0.10
	TransSum	**0.28**	**0.19**	**0.15**	**0.38**	**0.17**
	Gold	0.42	0.39	0.36	0.51	0.31
RT	LexRank	-0.55	-0.32	-0.54	-0.73	-0.36
	PlanSum	-0.08	-0.14	-0.41	-0.33	-0.13
	TransSum	**0.26**	**0.17**	**0.42**	**0.45**	**0.20**
	Gold	0.37	0.29	0.53	0.61	0.29

Table 3: Human evaluations results in terms of the Best-Worst scaling. The kappa coefficient of judges is 0.72.

#	$\mathcal{L}_{rec}$	$\mathcal{L}_{asp}$	$\mathcal{L}_{sen}$	$\mathcal{L}_{reg}$	Yelp	Ama.	RT
1	✓				18.64	18.94	16.89
2	✓	✓		✓	19.91	19.61	17.46
3	✓		✓	✓	19.82	19.53	17.26
4		✓	✓	✓	19.96	20.04	17.93
5	✓	✓	✓		20.08	20.11	18.14
6	✓	✓	✓	✓	**20.31**	**20.49**	**18.35**

Table 4: Ablation study of different losses. $\mathcal{L}_{sum}$ is a basic loss, so all combinations in the table include it.

from the Yelp, Amazon, and RT test sets with human-annotated summaries, respectively. Then we employ five graduate students to evaluate each tuple containing summaries from LexRank (strong extractive baseline), PlanSum (strong abstractive baseline), TransSum (Ours) and the gold-standard summaries according to the criteria. Note that the order in which the summaries are presented to the judges is random. We use Best-Worst Scaling (Louviere et al., 2015), which has been shown to produce more reliable results than ranking scales (Kiritchenko and Mohammad, 2016). Specifically, each score is computed as the percentage of times it was selected as best minus the percentage of times it was selected as worst, and ranges from -1 (unanimously worst) to +1 (unanimously best).

The results are shown in Table 3. As shown, summaries generated by TransSum have better aspect- and sentiment-based informativeness, indicating that our model can effectively capture the salient opinion information. We find that extractive summaries tend to be more general or even irrelevant (e.g. LexRank on Yelp), but our model performs very well in terms of relevance. Our method also excels baselines in non-redundancy and fluency, showing that summaries generated TransSum are low-redundant and fluent. We show examples of generated summaries of our model and comparison systems in supplementary materials.

3.2 Ablation Study

3.2.1 Loss Effectiveness Analysis

We present in Table 4 various ablation studies on the three datasets, which assess the contribution of different losses. We report the ROUGE-L score on test sets.

Compared to the only reconstruction loss (i.e., row#1), the contrastive learning of aspect and sentiment embeddings (i.e., row#2 and row#3) can bring improvements of 1.27/0.67/0.57 and 1.18/0.59/0.37 points on three datasets, respectively. From row#5 and row#4, we observe that the reconstruction and

#	Model Variants	Yelp	Amazon	RT
1	TransSum	20.31	20.49	18.35
2	w/o BERT embeddings	20.05	20.14	18.05
3	w/o label smoothing	20.08	20.21	18.12
4	w/o beam search	20.02	20.08	17.99
5	w/o summary modeling	18.51	18.82	16.72
6	w/o weighted pooling	19.83	19.95	17.88

Table 5: Ablation study of different components. "w/o" means "without".

regularization losses are also useful for improving results. The last row shows that all our proposed losses in TransSum are helpful, especially $\mathcal{L}_{asp}$ and $\mathcal{L}_{sen}$, demonstrating the effectiveness of our model.

3.2.2 Module Effectiveness Analysis

To investigate the importance of the model's individual components, we perform ablations by removing the initialized BERT embeddings, label smoothing, beam search, the translation-based review modeling module, and weighted pooling operation (i.e., $\boldsymbol{w}_{-i}$ is a uniform distribution).

From the results in Table 5, all components play a role, yet the most significant drop (i.e., row#5) in ROUGE-L when the translation-based review modeling module is removed, demonstrating the great effectiveness of the aspect and sentiment embeddings learned through contrastive learning. Interestingly, even without learning aspect and sentiment embeddings, using high-relevant reviews-summary pairs created by only entangled representations (i.e., row#5) can also achieve competitive results. This further shows the importance of considering the relevance between reviews and pseudo-summaries.

4 Related Work

Unsupervised Opinion Summarization aims to automatically generate summaries for a group of opinions about a specific entity (e.g., user reviews of a product), and does not require any gold summaries (Ku et al., 2006; Kim et al., 2011; Chu and Liu, 2019). Most previous works focus on extractive approaches, which select a subset of salient sentences from the inputs based on topic-words (Paul et al., 2010; Fabbrizio et al., 2014), word-frequencies (Erkan and Radev, 2004; Nenkova and Vanderwende, 2005), word embeddings (Rossiello et al., 2017) or textual graphs (Radev et al., 2004). However, due to their shortcomings of copying text from the input (Banko and Vanderwende, 2004), studies of abstractive summarization methods have increased tremendously (Ganesan et al., 2010; Perez-Beltrachini et al., 2019; Zou et al., 2020; Mukherjee et al., 2020). Most of these abstractive works model the problem of opinion summarization as a normal multi-document summarization task, using an auto-encoder framework with attention (Chu and Liu, 2019; Amplayo and Lapata, 2019; Brazinskas et al., 2020a), variational distributions (Brazinskas et al., 2020b; Angelidis et al., 2020), or abstract meaning representations (Liu et al., 2015). Few of them pay attention to the opinion information, and model the opinion summary with opinion phrases (Suhara et al., 2020) or the aspect and sentiment distributions (Amplayo et al., 2021). To the best of our knowledge, we are the first to model opinion summaries with only aspect and sentiment embeddings, which are learned through two novel contrastive objectives based on the aspect and sentiment invariances.

Our work is also related to **contrastive learning**, which a popular unsupervised learning paradigm in the field of computer vision and speech, aiming to enlarge the embedding disagreements of different instances for representation learning (van den Oord et al., 2018; Ye et al., 2019; He et al., 2020). Although there have been studies using contrastive learning for summary evaluation (Wu et al., 2020), to our best knowledge, we are the first to use the contrastive transformation on natural textual samples to directly help summary generation, and open the door to research on modeling opinion summaries with aspect and sentiment embeddings.

5 Conclusion

In this paper, we propose a novel self-supervised framework TransSum, to generate opinion summaries with only the aspect and sentiment embeddings, which are beneficial for maximizing informativeness, reducing redundancy of repeated opinions in reviews, and creating synthetic datasets of highly relevant reviews-summary pairs for training. Extensive evaluation and ablation studies show our model outperforms competitive systems in generating informative, high-relevant, low-redundant and fluent summaries. We believe that the viewpoint from modeling opinion summaries with only aspect and sentiment embeddings proposed in this study may pave a new way to design more complex and controllable systems for unsupervised opinion summarization.

6 Acknowledgements

This work was supported by National Natural Science Foundation of China (61772036), Beijing Academy of Artificial Intelligence (BAAI) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We would like to appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised opinion summarization with content planning.

Reinald Kim Amplayo and Mirella Lapata. 2019. Informative and controllable opinion summarization. *CoRR*, abs/1909.02322.

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1934–1945. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2020. Extractive opinion summarization in quantized transformer spaces. *CoRR*, abs/2012.04443.

Michele Banko and Lucy Vanderwende. 2004. Using n-grams to understand the nature of summaries. In *Proceedings of HLT-NAACL 2004: Short Papers, Boston, Massachusetts, USA, May 2-7, 2004*. The Association for Computational Linguistics.

Julian Besag. 1975. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195.

Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. 2005. A tutorial on the cross-entropy method. *Ann. Oper. Res.*, 134(1):19–67.

Arthur Brazinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4119–4135. Association for Computational Linguistics.

Arthur Brazinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5151–5169. Association for Computational Linguistics.

Jackie Chi Kit Cheung, Giuseppe Carenini, and Raymond T. Ng. 2009. Optimization-based content selection for opinion summarization. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pages 7–14, Suntec, Singapore. Association for Computational Linguistics.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Hady ElSahar, Maximin Coavoux, Matthias Gallé, and Jos Rozen. 2020. Self-supervised and controlled multi-document opinion summarization. *CoRR*, abs/2004.14754.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Giuseppe Di Fabbrizio, Amanda Stent, and Robert J. Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. In *INLG 2014 - Proceedings of the Eighth International Natural Language Generation Conference, Including Proceedings of the INLG and SIGDIAL 2014 Joint Session, 19-21 June 2014, Philadelphia, PA, USA*, pages 54–63. The Association for Computer Linguistics.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 340–348. Tsinghua University Press.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. IEEE.

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and ChengXiang Zhai. 2011. Comprehensive review of opinion summarization.

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*.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. 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 811–817. The Association for Computational Linguistics.

Lun-Wei Ku, Yu-Ting Liang, Hsin-Hsi Chen, et al. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 100107, pages 1–167.

Chin-Yew Lin and Eduard H. Hovy. 2003. 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*. The Association for Computational Linguistics.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman M. Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1077–1086. The Association for Computational Linguistics.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Rajdeep Mukherjee, Hari Chandana Peruri, Uppada Vishnu, Pawan Goyal, Sourangshu Bhattacharya, and Niloy Ganguly. 2020. Read what you need: Controllable aspect-based opinion summarization of tourist reviews. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1825–1828. ACM.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2019. What is this article about? extreme summarization with topic-aware convolutional neural networks. *J. Artif. Intell. Res.*, 66:243–278.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, 101.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Michael J. Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 66–76. ACL.

Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. 2019. Generating summaries with topic templates and structured convolutional decoders. 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 5107–5116. Association for Computational Linguistics.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. 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 1059–1073. Association for Computational Linguistics.

Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manag.*, 40(6):919–938.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the Workshop on Summarization and Summary Evaluation Across Source Types and Genres, MultiLing@EACL 2017, Valencia, Spain, April 3, 2017*, pages 12–21. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. Opiniondigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5789–5798. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

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, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. 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 47–57. The Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3612–3621. Association for Computational Linguistics.

Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6210–6219. Computer Vision Foundation / IEEE.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. 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 1049–1058. Association for Computational Linguistics.

Yanyan Zou, Xingxing Zhang, Wei Lu, Furu Wei, and Ming Zhou. 2020. Pre-training for abstractive document summarization by reinstating source text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3646–3660. Association for Computational Linguistics.

A Implementation Details

We implement our method on top of the Transformer-base (Vaswani et al., 2017) implemented in Fairseq (Ott et al., 2019). The token embeddings of BERT-base (Devlin et al., 2019) used for initialization are provided by Transformers (Wolf et al., 2020). The dimension k is 768 and the number of the attention heads is 4. Both the encoder and decoder have 6 layers, and the maximum sequence length l is set to 200. The beam size of the beam search is set to 5. We set the dropout rate to 0.1 and Adam (Kingma and Ba, 2015) with learning rate to 1e-5, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We implement our model based on PyTorch and use four TITAN X graphic cards for learning.

B Human Evaluation Questions

- **Aspect-based informativeness**: The summary sentences should cover common aspects discussed in the group of reviews.
- **Sentiment-based informativeness**: The summary sentences should agree with the overall sentiment about different aspects in the group of reviews.
- **Relevance**: The summary sentences should relevant to the input reviews.
- **Non-Redundancy**: The summary sentences should not contains unnecessary repetition.
- **Fluency**: The summary sentences should be grammatically correct, easy to read and understand.

Input Reviews	
1. i got the roast duck won ton noodles . the noodles were good and firm and the wontons were 100 % shrimp which was very good . the roast duck and roast suckling pig was bland . but not bad for the prices .	
2. lost on the shuffle . lucky i had extra time . ordered soup noodles with beef brisket and tendon . normally very fast as everything is already cooked . they had to ask me about my order 3 times after i asked them to check on it after waiting for 25 mins . when the food finally came , taste was good and portion was pretty big . beef tendon and brisket noodle ... the tendon was sooo soft and gooey .... mmm . but i had to deduct a point for the service mishap . and it wasn ' t busy yet ...	
3. always fantastic food with great prices . i went every weekend for a month in the summer . the owners are always friendly . if you are going later in the evening or late , don ' t order the tea (milk tea) or coffee . they boil it all day and by then its completely gross . but other wise , i ' ve never had a bad experience here .	
4. food portions were small and nothing special . bonus its its open late .	
5. - solid chinese eats - it gives me a good feeling when a restaurant is full of people . and this one normally is . (especially those of the same ethnic background as the cuisine) - if ever i have a craving for congee or roasted pork on rice , i ' m here . - oh and it ' s mad cheap - which is a nice bonus . i dig healthy competition .	
6. the price point is a little higher than the places i frequent in richmond hill / markham and the selection is smaller , but if i ever craze decent , solid , authentic chinese food when i ' m downtown , i come here !	
7. initially went to chinatown to eat beef brisket noodle soup at kings noodle but they were closed on wednesdays . walked down dundas and found this spot , decided to try it and as really surprised . the noodles were tasted good , much more generous portions compared to kings noodle and they were the same price . would recommend this place !	
8. this is an awesome place you can go for chinatown area . nice service , delicious food , and what you need more ?	
Gold	service can be a little slow here . the noodles are really good . i think it ' s a bit expensive though for what you get . there are other places that are cheaper but i don ' t know how they taste , so i can only comment on here . it ' s definitely worth checking out though . i had to wait a bit for my food but still pretty good experience .
LexRank	food portions were small and nothing special . bonus its its open late . walked down dundas and found this spot , decided to try it and as really surprised . the noodles were good and firm and the wontons were 100 % shrimp which was very good . the roast duck and roast suckling pig was bland . but not bad for the prices .
PlanSum	i've been to this place several times and i have never had a bad experience. the food is always good and the service is good. i love the fact that they are open late, so if you're looking for a quick lunch or dinner, this is the place to go.
Ours	the noodles are good but the price is a little expensive . the staff is always helpful and friendly . an awesome chinese eats you can go in the chinatown area . come by yourself !

Table 6: Examples of opinion summaries generated by multiple systems on the Yelp dataset.

Input Reviews	
1. the only thing i would like to see is an aux cord when i do n't want to charge my phone , but it 's not a huge deal . the sound is great , and worth the money . the remote works with your phone , and that 's precisely what i wanted .	
2. you need to buy an adaptor for ipod nano 's so it was disappointing when my son opened it up on christmas and could not use it for his ipod nano . it does not state that anywhere on the box or when i ordered it .	
3. love , love , love the ability to save multiple preset radio stations , and the sound is clear , crisp ... amazing ! it almost makes waking up a pleasure . another feature i never thought i wanted , but really appreciate , is the ability to set the brightness of the clock readout . brilliant !	
4. the sound of the radio is of real quality . i also like having the two separate alarms and the alarm is not obnoxious yet still wakes us up . my wife charges her iphone on it regularly and works out well . we like the sony so much i got one for my son and his wife for a christmas present .	
5. i was looking quite awhile to locate a decent sounding radio / ipod player which would also charge my ipod . this is perfect for our family . it 's a lot smaller than i thought , which is good . and when we update to an iphone 5 , there is a $ 5 adapter to get so we can still use this radio . perfect !	
6. as always , sony has a 'winner ' in this combined am / fm radio and docking station . great sound , looks good and wife is very pleased as she put it in her craft work area . finding the combo of am / fm was n't easy either.lots of fm only units . this is a great product .	
7. while i like the dream machine i do n't know why there 's so much static . it 's nearly impossible to get a couple of my favorite radio stations without constant static in the background . my other radio does n't do that . i 've even tried different locations for it . that 's a big disappointment and shortcoming of the product .	
8. my husband really like this speaker ... love it ! its so easy to operate by setting the alarm. i like the way when you put your iphone 4s to the dock its charge at the same time while you are you using it ... ! great product	
Gold	this fm/am radio , iphone docking station and alarm clock is a perfect combination ! the sound is amazing , the alarm clock is not annoying , and the design looks great ! it would be nice to have a place to use an aux cord and certain apple products require a $ 5 adapter to use the docking station but other than that , this product is fantastic !
LexRank	while i like the dream machine i do n't know why there 's so much static . great sound , looks good and wife is very pleased as she put it in her craft work area . the sound of the radio is of real quality .
PlanSum	this fm/am radio , iphone docking station and alarm clock is a perfect combination ! the sound is amazing , the alarm clock is not annoying , and the design looks great ! it would be nice to have a place to use an aux cord and certain apple products require a $ 5 adapter to use the docking station but other than that , this product is fantastic ! this is a great product . it has amazing sound quality , and the dual-band feature is nice . i love the fact that it charges my device while it 's docked . not only is this thing functional , but it also looks great and does n't take up a lot of space . overall , i would advise buying this if your needs call for an awesome speaker that doubles as a charging station . this is exactly what i 've been looking for ! it has great sound quality and it 's really easy to dock my iphone in it (it charges my phone at the same time ! bonus) the alarm is easy to use and does it 's job . some iphones and ipods will need an adapter , so make sure to check that out first .
Ours	i love this radio for its sound, which is amazingly clear and quite great . overall , this is a great product , and good for my family. i would advise buying this .

Table 7: Examples of opinion summaries generated by multiple systems on the Amazon dataset.

Input Reviews	
1. The best reason to see The Haunting is the sheer sumptuousness of its creepy-crawly set designs.	
2. Has an unseen enchantment, so aptly sets spinning like a huge magnificent gyroscope on a string	
3. In The Haunting, the moviemakers succeed in something very difficult: creating a haunted house with real personality and terror.	
4. All the stops are pulled out to provide a state-of-the-art, slam-bang movie experience.	
5. Looking terrified and screaming is really all that's required in David Self's inane script.	
6. Director Jan de Bont, known for the razzle he put into the exciting movie Speed and the subsequent dud Twister, proves himself unable to break away from depending on dazzle to substitute for substance.	
7. To my surprise, I find myself recommending The Haunting on the basis of its locations, its sets, its art direction, its sound design, and the overall splendor of its visuals.	
8. It's all hokum from beginning to end.	
9. A flat, draggy exercise in cheesy special effects and grandiose art direction palming itself off as a horror film.	
10. More hokey than haunting.	
11. When The Haunting finally limps to its conclusion, you may feel like booing the screen.	
12. This all-flash, no-substance–and no scare–thriller is a textbook example of soulless, money-burning Hollywood hype products.	
13. One of the most misguided big-screen diversions to come along in some time, considering the clear potential it has.	
14. so thoroughly misguided and muddled Dreamworks, the studio foisting this bomb on the public, ought to hire a special corps of ushers to hand out sympathy cards to patrons as they exit the theater.	
15. This is as far from the Poverty Row gasps of The Blair Witch Project as you can get, and more fun.	
16. I wouldn't waste more than the price of a video rental on this one.	
17. High-tech remake is dumb and overblown.	
18. All logic is deadened by the obnoxious special effects!	
19. Once the screaming begins, so will your laughing	
20. Glossy but lackluster.	
21. It's just a conglomeration of cheap fright tactics and a booming bass track meant to get you to jump out of your seat.	
22. An exercise in missed opportunities and bad filmmaking!	
23. The Haunting is a muddled mess that defies any rationality.	
24. The only thing scary about the new version is realizing that someone keeps giving director Jan De Bont money to make movies.	
25. The characters are on the dramatic equivalent of Death Row.	
.....	
Gold	sophisticated visual effects fail to offset awkward performances and an uneven script .
LexRank	the characters are on the dramatic equivalent of death row .
PlanSum	the haunting is a very good movie, but it's a lot of fun, and the filmmakers have been raised by the original.
Ours	unfortunately , this is one haunting with the obnoxious special effects that are bloated and wretchedly overdone !

Table 8: Examples of opinion summaries generated by multiple systems on the Rotten Tomatoes dataset.