Abstractive Opinion Tagging

Qintong Li∗
Shandong University
qintongli@mail.sdu.edu.cn

Piji Li†
Tencent AI Lab
piji@tencent.com

Xinyi Li‡
Peng Cheng Laboratory
lixinyimichael@gmail.com

Zhaochun Ren‡
Shandong University
zhaochun.ren@sdu.edu.cn

Zhumin Chen
Shandong University
chenzhumin@sdu.edu.cn

Maarten de Rijke
University of Amsterdam
m.derijke@uva.nl

ABSTRACT
In e-commerce, opinion tags refer to a ranked list of tags provided by the e-commerce platform that reflect characteristics of reviews of an item. To assist consumers to quickly grasp a large number of reviews about an item, opinion tags are increasingly being applied by e-commerce platforms. Current mechanisms for generating opinion tags rely on either manual labelling or heuristic methods, which is time-consuming and ineffective. In this paper, we propose the abstractive opinion tagging task, where systems have to automatically generate a ranked list of opinion tags that are based on, but need not occur in, a given set of user-generated reviews.

The abstractive opinion tagging task comes with three main challenges: (1) the noisy nature of reviews; (2) the formal nature of opinion tags vs. the colloquial language usage in reviews; and (3) the need to distinguish between different items with very similar aspects. To address these challenges, we propose an abstractive opinion tagging framework, named AOT-Net, to generate a ranked list of opinion tags given a large number of reviews. First, a sentence-level salience estimation component estimates each review’s salience score. Next, a review clustering and ranking component ranks reviews in two steps: first, reviews are grouped into clusters and ranked by cluster size; then, reviews within each cluster are ranked by their distance to the cluster center. Finally, given the ranked reviews, a rank-aware opinion tagging component incorporates an alignment feature and alignment loss to generate a ranked list of opinion tags. To facilitate the study of this task, we create and release a large-scale dataset, called eComTag, crawled from real-world e-commerce websites. Extensive experiments conducted on the eComTag dataset verify the effectiveness of the proposed AOT-Net in terms of various evaluation metrics.

1 INTRODUCTION
With the explosive growth of customer reviews in e-commerce scenarios, many online platforms, such as Amazon and Alibaba, provide opinion tags to enable potential buyers to make informed decisions without having to absorb large numbers of reviews. As shown in Figure 1, a sequence of opinion tags is a ranked list mined from a set of reviews that reflects different users’ preferences towards certain aspects of items. Many studies have focused on mining valuable information from reviews and shown promising results in various tasks, such as opinion summarization [1, 2, 5, 8, 24, 25, 40] and item description generation [13, 35]. So far, however, no study seems to have developed opinion tagging methods to generate opinion tags that reflect diverse opinions of item aspects in a concise manner. In this paper, we propose the task of abstractive opinion tagging, which aims to automatically generate a ranked list of opinion tags that stem from, but need not occur in, a given set of user-generated reviews. This is an abstractive rather than an extractive opinion tagging task as the opinion tags do not need to occur in the review.

To solve this new task, we face three challenges. First, in reviews of e-commerce items, the noisy nature of the reviews inevitably makes it hard to identify salient item-related features [16, 24]. According to human annotations (on the eComTag dataset described below), we find that almost 59% of review sentences are not item-related or do not contain opinions towards certain aspects. Second, different reviewers have different ways of expressing themselves [43], whereas the target opinion tags are usually in a more formal style. The colloquial language usage of user reviews [48]...
Figure 1: An example of a set of reviews and their corresponding opinion tags, where \( U_N \) denotes the index of reviews, whereas the number behind each opinion tag reflects the number of reviews belonging to the tag.

![Figure 1](image1.png)

Figure 2: Fraction of relevant reviews per opinion tag in the eComTag dataset.

![Figure 2](image2.png)

Figure 3: The average semantic similarity between opinion tags and opinion clusters for all samples in the eComTag dataset. Each color denotes an opinion cluster. The same opinion clusters share the same color across different opinion tags. The width of the color band denotes the degree of semantic similarity between an opinion tag and the corresponding opinion cluster. (Best viewed in color.)

![Figure 3](image3.png)

makes abstractive methods necessary to learn better review representations. Third, it is difficult to reflect the different perspectives on an item in an accurate and diverse manner. Many e-commerce platforms rank opinion tags to help customers distinguish between different items with very similar aspects. In Figure 2, we plot the number of relevant reviews on different ranked indexes of opinion tags in the eComTag dataset (described below), which suggests a natural ranking for the tags when presented to users.

To address the challenges listed above, we design an abstractive framework, named AOT-Net, which consists of three components: (1) a sentence-level salience estimation component that predicts a salience score for each review; (2) a review clustering and ranking component that first groups reviews into clusters, which we refer to as “opinion clusters,” and ranks reviews by opinion cluster size; this component then ranks reviews within each cluster by their distance to the cluster center; and (3) a rank-aware opinion tagging component that generates opinion tags with ranks.

AOT-Net works in such a way that the ranks of the opinion clusters are correlated with the ranks of opinion tags. To see why this is potentially useful, we consider the eComTag dataset (crawled from e-commerce sites and consisting of items, reviews, and opinion tags), group reviews into opinion clusters and rank clusters by cluster size. We then compute the semantic similarity between the opinion clusters and the opinion tags, obtained by averaging the semantic similarity between an opinion tag and reviews in an opinion cluster. Figure 3 visualizes the average semantic similarity between 10 opinion tags and 10 opinion clusters for all item samples in the eComTag dataset. Clearly, the opinion tags have broader chunks with the opinion clusters at the same rank, revealing an alignment between the ranked opinion tags and ranked opinion clusters. Furthermore, the opinion tags are semantically similar to the neighbors of the corresponding opinion clusters, which is reflected by the relatively wide color bands, e.g., the second opinion tag is semantically similar to the first opinion cluster and the third opinion cluster. The rank-aware opinion tagging component of AOT-Net integrates two alignment strategies: (1) an opinion tag and its corresponding opinion clusters are placed at similar ranks; and (2) an alignment loss explicitly encourages the model to focus on the aligned opinion clusters and ignore others.

To validate AOT-Net, we collect a new dataset, named eComTag, from Chinese e-commerce websites, containing reviews and opinion tags for 50,068 items. Our experiments based on the eComTag dataset show that AOT-Net is capable of significantly improving the generation performance on abstractive opinion tagging over state-of-the-art baselines.

Our contributions can be summarized as follows:

- We propose a new task, abstractive opinion tagging, to generate opinion tags based on large volumes of item reviews.
- We propose an abstractive framework AOT-Net to generate opinion tags based on reviews for a given item. AOT-Net has a sentence-level salience estimation component and a review clustering and ranking component to highlight salient reviews and rank reviews by clustering. We propose a rank-aware opinion tagging component with two alignment strategies to generate ranked opinion tags.
- We collect a large-scale dataset, namely eComTag, consisting of item reviews and opinion tags, to support research into abstractive opinion tagging. Experimental results conducted on the eComTag dataset demonstrate the effectiveness of the AOT-Net framework.

2 RELATED WORK

Related work comes in two categories: keyphrase generation and opinion summarization.

2.1 Keyphrase Generation

A lot of research has been conducted on generating keyphrases to summarize various types of text [11, 27, 28, 33, 36, 48, 49, 52]. Early approaches to keyphrase generation extract important phrases...
from the document as the results. Sequence tagging models have been applied to identify keyphrases [18, 30, 51]. Retrieval-based approaches utilize a two-step pipeline to extract and rank candidate keyphrases [21, 32, 34, 47]. Sun et al. [41] adopt an extractive graph-based approach, which applies a point network to generate a set of diverse keyphrases. Recently, abstractive approaches have also been explored. Chan et al. [10] propose a reinforcement learning approach for neural keyphrase generation that encourages a model to generate both sufficient and accurate keyphrases. Wang et al. [48] propose a topic-aware neural keyphrase generation method to identify topic words.

Unlike the work listed above, which only considers single document keyphrase generation, we consider opinion tagging from multiple documents, that is, from all of the reviews for a given item.

2.2 Opinion Summarization

Opinion summarization has become an emerging research topic in recent years. Early studies on opinion summarization focus on extracting salient sentences from text [3, 7, 15, 19, 29, 50]: Hu and Liu [19] identify item features mentioned in the reviews and then extract opinion sentences for the identified features. Unsupervised learning methods are utilized to extract a review summary by exploiting review helpfulness ratings [50]. To reduce redundancy, a greedy algorithm is also applied to form the final summaries [3]. Reflecting the most representative opinions from reviewers, many recent studies have shown that abstractive approaches are more appropriate for summarizing review text [6, 14, 17, 23, 44]: Gerani et al. [17] utilize a template filling strategy to generate a review summary; Wang and Ling [46] apply an encoder-decoder attention model to generate an abstractive summary for opinionated documents. The objective of the above summarization approaches is to generate coherent sentences to summarize opinions.

In contrast, we propose the abstractive opinion tagging task so as to generate opinion tags from a large number of user-generated reviews. In our scenario, opinion tags are more concise but without loss of essential information; they should help users comprehend reviews quickly and conveniently [25].

3 PROBLEM FORMULATION

Before detailing our proposed method, AOT-Net, we first formulate the abstractive opinion tagging problem. We use bold lowercase characters to denote vectors, and bold upper case characters to denote matrices. We write \( W \) and \( b \) for a projection matrix and a bias vector in a neural network layer, respectively. Suppose that there are \( M \) reviews for a given item. We denote each review \( X_i, 1 \leq i \leq M \) as a sequence of words, i.e., \( X_i = \{x_{i1}, \ldots, x_{iL_i}\} \), where \( L_i \) denotes the number of words in \( X_i \). In the same way, we assume that \( N \) opinion tags exist for each item. We denote each opinion tag \( Y_j, 1 \leq j \leq N \) as a sequence of words, i.e., \( Y_j = \{v_{j1}, \ldots, v_{jL_j}\} \), where \( L_j \) refers to the number of words in \( Y_j \). Given a set of reviews \( X = \{X_1, X_2, \ldots, X_M\} \), the task of abstractive opinion tagging is to generate a sequence of opinion tags \( Y = \{Y_1, Y_2, \ldots, Y_N\} \).

4 METHOD

4.1 Overview

Before providing the details of AOT-Net, our proposed method for abstractive opinion tagging, we first provide an overview in Figure 4. We divide AOT-Net into three main phases: (A) sentence-level salience estimation; (B) review clustering and ranking; and (C) rank-aware opinion tagging. For a set of reviews \( X \) about a given item, in phase (A), we derive a salience score \( z_i \) for each review \( X_i \) in \( X \) to estimate its item-aware salience information. In phase (B), reviews are first encoded into vector representations and weighted by corresponding salience scores. The weighted vector representations of reviews are clustered into \( K \) opinion clusters \( \{C_1, \ldots, C_K\} \) and ranked by cluster size. Reviews within each cluster are ranked by their distance to the cluster center. Then we flatten ranked reviews into word-level vector representations. In phase (C), we use review representations to generate ranked opinion tags via two alignment constraints, i.e., alignment features and alignment loss. We jointly learn all components in a multi-task learning framework.

4.2 A: Sentence-level Salience Estimation

The aim of the sentence-level salience estimating component is to compute a salience score for each review \( X_i \in X \). We design a sentence-level self-attention mechanism to highlight item-related reviews and reduce noise. First, the component reads each review sequence \( X_i = \{x_{i1}, \ldots, x_{iL_i}\} \) and uses a lookup table to convert each review word \( x_p \) to a word embedding vector \( x_p \in \mathbb{R}^d \). To incorporate the contextual information of the review text into the representation of each word, we feed each embedding vector \( x_p \) to a bi-directional Gated- Recurrent Unit (GRU) [12] to learn a hidden representation \( h_p \in \mathbb{R}^d \). More specifically, a bi-directional GRU consists of a forward GRU that reads the embedding sequence from \( x_1 \) to \( x_{L_i} \) and a backward GRU that reads from \( x_{L_i} \) to \( x_1 \):

\[
\overrightarrow{h}_p = \text{GRU}_f(x_p, \overrightarrow{h}_{p-1}), \quad (1)
\]

\[
\overleftarrow{h}_p = \text{GRU}_b(x_p, \overleftarrow{h}_{p+1}), \quad (2)
\]

where \( \overrightarrow{h}_p \in \mathbb{R}^{d/2} \) and \( \overleftarrow{h}_p \in \mathbb{R}^{d/2} \) denote the forward states of the forward GRU and backward GRU, respectively. We concatenate the last forward hidden state \( \overrightarrow{h}_{L_i} \) and last backward hidden state \( \overleftarrow{h}_1 \) to form the hidden representation for review \( X_i \), i.e., \( h_{X_i} = [\overrightarrow{h}_{L_i}; \overleftarrow{h}_1] \).

Next, we pass hidden representations of all reviews \( \{h_{X_i}\}_{i=1:M} \) to a self-attention layer to model more complex interactions among the reviews. We propose a salience context vector \( c_i \) for each review \( X_i \) to denote the shared information from other reviews:

\[
q_i = W_q h_{X_i}, \quad k_i = W_k h_{X_i}, \quad v_i = W_v h_{X_i}, \quad (3)
\]

\[
c_i = \sum_{f=1}^{M} \exp(q_i^T k_f) \exp(q_i^T k_f)^T v_f, \quad (4)
\]

where \( q_i, k_i, v_i \in \mathbb{R}^{d \times d} \) refer to query, key, and value vectors, respectively. These vectors are linearly transformed from hidden representation \( h_{X_i} \). Then we apply a residual connection from the review hidden representation \( h_{X_i} \) to the salience context
vector $c_i$ and feed it to a two-layer feed-forward network with a ReLU as the activation function:

$$h'_X = W_{s2}(ReLU(W_{s1}(h_X + c_i))).$$

(5)

Given the context-enhanced review hidden representation $h'_X$, we can derive the real-valued salience score $z_i$ for $X_i$:

$$z_i = \sigma(W_h h'_X + b_h),$$

(6)

where $W_h \in \mathbb{R}^d$ and $b_h \in \mathbb{R}$. $\sigma(\cdot)$ is the sigmoid activation function. The salience scores $\{s'_1, \ldots, s'_M\}$ serve as the salience weights of review representations for the later review clustering and ranking component.

In order to optimize the sentence-level salience estimating component, we manually label a binary salience label $z'_i \in \{0, 1\}$ for each review $X_i$, where 1 denotes “item-related” whereas 0 denotes “noisy”. Then, sentence-level salience estimation component is trained by minimizing the cross-entropy loss function:

$$L_{cla} = -\frac{1}{M} \sum_{i=1}^{M} z'_i \log(z_i) + (1 - z'_i) \log(1 - z_i).$$

(7)

### 4.3 B: Review Clustering and Ranking

We propose a review clustering and ranking component to learn the ranks of reviews by grouping reviews into ranked opinion clusters, which is the main prerequisite to accurately generate ranked opinion tags.

We use a standard transformer encoder [45] to convert each review $X_i$ into vector representations. Following Vaswani et al. [45], we first map each word $x \in X_i$ into its vectorized representation $x \in \mathbb{R}^{d_e}$ using a word embedding layer and a positional embedding layer, as shown in the following equation:

$$x = \text{Embed}(x) + \text{Pos}(x).$$

(8)

Then we use a transformer layer to encode global contextual information for words within $X_i$.

$$g = \text{LayerNorm}(x^{n-1} + \text{MHAtt}(x^{n-1})),$$

$$x^n = \text{LayerNorm}(g + \text{FFN}(g)).$$

(9)

(10)

where LayerNorm is the layer normalization proposed by Ba et al. [4]; MHAtt is the multi-head attention mechanism introduced by Vaswani et al. [45]; FFN is a two-layer feed-forward network with ReLU as hidden activation function; and $n$ is the number of transformer block layers. The word-level vector representations of review $X_i$ are $X_i = [x_1, \ldots, x_{L_i}] = [x^n_1, \ldots, x^n_{L_i}]$. To obtain the sentence representation $\tilde{X}_t$ of review $X_t$, we perform a hierarchical pooling [38] across its different words. The hierarchical pooling mechanism can preserve word order information and has demonstrated superior performance over average pooling or max-pooling on many semantic analysis tasks [38].

To highlight the item-aware reviews and ignore noisy reviews, the sentence representations of reviews are first weighted by corresponding salience scores, i.e., $\tilde{X}'_t = z'_t \tilde{X}_t$. Then we apply the $k$-means [31] algorithm on $\{X'_1, \ldots, X'_M\}$ to group corresponding $\{X'_1, \ldots, X'_M\}$ into $K$ opinion clusters. We rank opinion clusters from the largest (representing the highest number of reviews) to the smallest, denoted as $[C_1, \ldots, C_K]$. For each cluster, we rank reviews from the nearest (representing the distance between review and cluster center) to the farthest. Finally, we obtain a ranked list of reviews, represented as $[X'_1, \ldots, X'_L_1, \ldots, X'_K, \ldots, X'_K]$, where $X'_k,i$ is the $i$-th review in the $k$-th opinion cluster and $L_k$ is the number of reviews in the $k$-th opinion cluster.

We sequentially concatenate the vector representations of reviews in the ranked list and derive the final word-level representations, i.e., $X = [x_1, \ldots, x_{L_1}, \ldots, x_{K}, \ldots, x_{K,L_k}]$. Similarly, $x_{k,p}$ is the $p$-th word in the $k$-th opinion cluster and $q_k$ is the number of words in the $k$-th opinion cluster. Next, $X$ will serve as the memory bank for the later rank-aware opinion tagging component.

$k$ is manually assigned according to the number of reviews. If $M \leq 200$, $k = \lceil \frac{M}{200} \rceil$, otherwise $k = 20$.

In this paper, we only focus on the ranks of opinion clusters. We regard reviews or words in the same opinion cluster as equally important.
4.4 C: Rank-aware Opinion Tagging

Accurately generating opinion tags with ranks is challenging. Therefore, we propose a rank-aware opinion tagging component to generate ranked opinion tags.

In the training stage, we add a start token BOS at the beginning of each opinion tag, i.e., $Y' = [\text{BOS}; Y_j]$ where $[::]$ is the concatenation function. Then we concatenate all opinion tags into a sequence of words: $Y = [Y'_1; \cdots; Y'_N] = [y_{1,0}; \cdots; y_{1,L_y}; \cdots; y_{N,0}; \cdots; y_{N,L_y}]$, where $y_{j,q}$ is the $q$-th word of opinion tag $Y'_j$. $y_{j,0}$ is the BOS token of the $j$-th opinion tag. Our decoder, i.e., the rank-aware opinion tagging component, follows the transformer architecture [45].

From the data analysis in Figure 3, we know that the ranks of opinion tags have a strong correlation with the ranks of opinion clusters. The $j$-th opinion tag may pay attention to the $F$ most related opinion clusters and its surrounding neighbors simultaneously. Therefore, for each opinion tag $Y'_j$, we hypothesize that the model needs to focus on the $F$ most related opinion clusters. If an opinion cluster belongs to the $F$ focused opinion clusters, we call it a Focused Opinion Cluster (FOC). Otherwise, we call it an Outer Opinion Cluster (OOC). We design two alignment strategies between FOCs and opinion tags, i.e., incorporating alignment features and enforcing alignment loss, which help to improve the generation of ranked opinion tags.

**Alignment Feature.** Intuitively, the opinion tags and their FOCs are semantically similar in the vector space. To help the model capture the alignment between opinion tags and their FOCs, we incorporate alignment features $\text{Aln}()$ into the word-level representations of the opinion tags and opinion clusters. Formally, $\text{Aln}()$ is a function that maps an integer into a vector. Now, we explain how we will use it to represent the ranks of opinion tags and opinion clusters. First, we incorporate the alignment feature into the representations of opinion tags. The words in the $j$-th opinion tag have the same rank $j$ where $1 \leq j \leq N$. For each word $y_{j,q}$, the vectorized representation is the sum of the alignment feature $\text{Aln}(j)$, the word embedding $\text{Embed}(y_{j,q})$, and the positional embedding $\text{Pos}(y_{j,q})$:

$$y_{j,q} = \text{W}_r \text{Aln}(j) + \text{Embed}(y_{j,q}) + \text{Pos}(y_{j,q}),$$

(11)

where $\text{W}_r \in \mathbb{R}^{d_e \times d_e}$ is a trainable model parameter.

For the target $j$-th opinion tag, the ranks of FOCs are set to $j$ as well, while the ranks of OOCs are set to 0. Then we enhance the word vectors in X with alignment features to capture the alignment between reviews and opinion tags:

$$\text{aln}_{x_{j,p}} = \begin{cases} \text{Aln}(j), & x_{j,p} \in \text{FOCs} \\ \text{Aln}(0), & x_{j,p} \in \text{OOCs}. \end{cases}$$

(12)

Next, the alignment features $[\text{aln}_{x_{j,1}}, \ldots; \text{aln}_{x_{j,N}}]$ are added into the review representations $[x_{1,1}; \ldots; x_{K,\text{len}x}]$ to obtain the alignment-enhanced representations $R = [r_{1,1}; \ldots; r_{K,\text{len}x}]$ for the words in opinion clusters:

$$r_{j,p} = \text{W}_r \text{aln}_{x_{j,p}} + x_{j,p},$$

(13)

where $\text{W}_r \in \mathbb{R}^{d_e}$ is a model parameter. Figure 5 summarizes the construction.

At each decoding step $q$ of the $j$-th opinion tag, the decoder reads the embeddings of the last prediction $y_{j,q-1}$:

$$y_{j,q} = \text{transformer}\_\text{decoder}(y_{j,q-1}),$$

(14)

where $y_{j,q} \in \mathbb{R}^{d_e}$ is the target word representation. Next, we introduce our decoder in detail.

To capture semantic and alignment information from the opinion clusters, a multi-head cross-attention MHAtt [45] is applied to compute the attention score $[\sigma_1, \ldots, \sigma_{k,\text{len}k}]$ between the last prediction $y_{j,q-1}$ and $[r_{1,1}; \ldots; r_{K,\text{len}k}]$:

$$\begin{align*}
\text{u}^z_{j,q-1} &= \text{W}_d y_{j,q-1} \\
\text{k}^z_{j,p} &= \text{W}_k r_{j,p} \\
\alpha^z_{j,p} &= \frac{\exp(\text{u}^z_{j,q-1}^T \text{k}^z_{j,p})}{\sum_{k=1}^K \sum_{p'=1}^{|V|} \exp(\text{u}^z_{j,q-1}^T \text{k}^z_{p'})}.
\end{align*}$$

(15-17)

where $\text{u}^z_{j,q-1} \in \mathbb{R}^{d_h}, \text{k}^z_{j,p} \in \mathbb{R}^{d_k}$ are query and key vectors that are linearly transformed from $y_{j,q-1}$ and $r_{j,p}$ as in [45]; $z \in \{1, \ldots, n_h\}$ indicates the $z$-th head among $n_h$ heads; $d_h = d_k = n_h$ is the dimension of each head.

The attention scores $[\sigma^z_1, \ldots, \sigma^z_{k,\text{len}k}]$ are then used to compute an aggregated vector $c_{j,q}$ for target word $y_{j,q}$:

$$c_{j,q} = \sum_{i=1}^{\text{len}k} \sum_{j=1}^{n_c} \sigma^z_{i,j} r_{i,j}; \cdots; \sum_{i=1}^{\text{len}k} \sum_{j=1}^{n_c} \sigma^z_{i,j} r_{i,j},$$

(18)

Then we feed the last word representation $y_{j,q-1}$ and vector $c_{j,q}$ to a two-layer feed-forward network with a ReLU as the activation function and a highway layer normalization on top:

$$\begin{align*}
s_{j,q-1} &= \text{LayerNorm}(y_{j,q-1}^n + c_{j,q}) \\
y_{j,q} &= \text{LayerNorm}(s_{j,q-1} + \text{FFN}(s_{j,q-1})),
\end{align*}$$

(19-20)

where $y_{j,q} \in \mathbb{R}^{d_e}$ is the target word representation. After that, we use $y_{j,q}$ to compute a probability distribution over the words in a predefined vocabulary $V$, as shown in the following equation:

$$p_{y,q}(y_{j,q} \mid [y_{1,0}, \ldots, y_{j,q-1}, X]) = \text{softmax}(\text{W}_p y_{j,q} + b_p),$$

(21)

where $\text{W}_p \in \mathbb{R}^{|V| \times d_e}, b_p \in \mathbb{R}^{|V|}$ are trainable parameters. To enable our model to generate out-of-vocabulary (OOV) words, we adopt the copy mechanism [37] to predict OOV words by directly copying words from the opinion clusters. We first compute a soft gate $p_{gen} \in [0, 1]$ between generating a word from the predefined vocabulary $V$ and copying a word from the input reviews $X$:

$$p_{gen} = \sigma(\text{W}_g y_{p,q} + b_g),$$

(22)
where \( W_b \in \mathbb{R}^{d} \) and \( b_y \in \mathbb{R} \) are trainable parameters. Finally, we can derive the final next-word probability distribution \( P(y_{i,q}) \):
\[
\hat{a}_{i,p} = \frac{\sum_{j} \alpha_{i,p} y_{j,q}}{n_h},
\]
\[
P(y_{i,q}) = \log \frac{p_{gen} p_{v}(y_{i,q})} {1 - p_{gen}} \sum_{i,p} \hat{a}_{i,p}. \tag{24}
\]

where we use \( P(y_{i,q}) \) to denote \( P(y_{i,q} \mid [y_{1,0}, \ldots, y_{i,q-1}], X) \) for brevity. We use the negative log-likelihood of the ground-truth words \( y_{i,q} \) as the generation loss function:
\[
L_{gen} = - \sum_{j=1}^{N} \sum_{q=1}^{L_{y,j}} \log P(y_{j,q}^* \mid [y_{1,0}^*, \ldots, y_{i,q-1}^*], X). \tag{25}
\]

### 4.5 Multi-task Training Objective

We adopt a multi-task learning framework to jointly minimize the salience classification loss, alignment loss, and generation loss. The objective function is:
\[
\mathcal{L} = \lambda_1 \mathcal{L}_{cla} + \lambda_2 \mathcal{L}_{aln} + \lambda_3 \mathcal{L}_{gen}, \tag{27}
\]
where \( \lambda_1, \lambda_2, \lambda_3 \) are hyper-parameters that control the weights of these three losses. We set \( \lambda_1 = \lambda_2 = \lambda_3 = 1 \). Thus, each component of our joint model can be trained end-to-end.

### 5 EXPERIMENTAL SETUP

We set up experiments to compare AOT-Net against a number of relevant baselines. We are interested in the overall performance of AOT-Net and in understanding the effectiveness of the salience estimation and ranking alignment.

#### 5.1 Experiments

We report on five experiments. First, we compare AOT-Net against a number of baselines to assess its overall performance. Then we conduct ablation studies to analyze the influence of different components in AOT-Net as follows: (i) w/o SSE is AOT-Net without the sentence-level salience estimating component (SSE). (ii) w/o RCR is AOT-Net without the review clustering and ranking component (RCR). (iii) w/o AF is AOT-Net without alignment feature (AF). (iv) w/o AL is AOT-Net without alignment loss (AL). To further explore the effectiveness of the sentence-level self-attention mechanism in SSE, we consider AOT-RNN, the method only considers BiGRU in salience score prediction; whereas we write AOT-Embed for the method that employs MLP to replace BiGRU in SSE. Fourth, we analyze the performance of AOT-Net for different sizes of FOCs. Lastly, we provide a case study about abstractive opinion tagging.

### 5.2 Baselines

We compare AOT-Net with the following methods: (i) TF-IDF is an extractive approach that selects the important words as summaries based on term frequency and inverse document frequency; (ii) TextRank [34] is an unsupervised algorithm based on weighted-graphs; (iii) RNN is a sequence to sequence model with attention implemented by bi-directional GRU layer [12]; (iv) PG-Net [37] is a classical opinion summarization model based on the encoder-decoder framework with attention and copy mechanisms; and (v) the Transformer [45] is a Transformer-based encoder-decoder model with a copy mechanism, which is a strong baseline widely-adopted in opinion summarization.

### 5.3 The eComTag Dataset

Since there is no available opinion tagging dataset, we build a new one, named eComTag from several Chinese e-commerce websites. We collect nearly 112k items, sampling from different domains, including Food (7.62%), Books (10.57%), Sports (3.96%), Clothes (3.16%), Medical (1.62%), Electronics (29.51%), Cosmetic (37.43%), Furniture (0.33%), and Entertainment (8.23%). For each domain, there are a set of reviews and a list of opinion tags. Since reviewers may comment on multiple aspects, e.g., “The dim sum tasted extremely fresh, and the price was quite reasonable!”, we split each review into sentences by punctuation. We use a sentence to denote a review in our paper. Then, we remove samples where the number of opinion tags is smaller than 4 or the number of reviews is fewer than 50. Finally, we construct eComTag with 40,162 item samples. Users may write reviews arbitrarily, which results in many meaningless expressions, such as “Love, love, LOVE this space!”, and “come with my boyfriend.” To teach our model to distinguish these noisy sentences, we annotate each review with a binary salience label via human judgment. If a review is item-related, we label the review as 1, otherwise 0. The salience labels are the supervision signals for the sentence-level salience estimating component.

Finally, each item sample consists of a set of reviews, a set of corresponding salience labels, and a sequence of opinion tags. For text preprocessing, we tokenize texts using the Jieba toolkit\(^6\) and maintain a 50k vocabulary. In eComTag, about 30% of samples have more than 1024 words in reviews. We randomly split the dataset into training/validation/test sets with 8:1:1 ratio. The statistics are shown in Table 1. Specially, we define the present tag (Pr) as the exact tag that appears in reviews and absent tag (Ab) as the tag unseen in reviews. The proportion of absent tags is close to 75%, which further proves the necessity to apply abstractive methods on opinion tagging.

### 5.4 Evaluation Metrics

We employ two information retrieval metrics to evaluate the opinion tag generation: the macro F@k score and the normalized discounted cumulative gain (NDCG@k) score. Both are widely used to measure word overlap \([9, 41]\). To measure diversity of the generated opinion tags, we adopt the Distinct-2 score \([22]\) and a macro Unique-N score. We compute Unique-N as follows, Unique-N = \( \sum_{i=1}^{T} N_i / T \), where \( N_i \) is the number of distinct opinion tags in the \( i \)-th sample.

\(^6\)https://github.com/fxsjy/jieba
Table 1: Statistics of the eComTag dataset. “Pr” and “Ab” denote the proportion of present tags and absent tags respectively. "MTN" and "MTL" indicate max tag number and max tag length for a sample respectively.

<table>
<thead>
<tr>
<th>Data</th>
<th>Sample</th>
<th>Pr</th>
<th>Ab</th>
<th>MTN</th>
<th>MTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>40,162</td>
<td>24.1%</td>
<td>75.9%</td>
<td>19</td>
<td>40</td>
</tr>
<tr>
<td>Validation</td>
<td>4,953</td>
<td>24.3%</td>
<td>75.7%</td>
<td>20</td>
<td>39</td>
</tr>
<tr>
<td>Test</td>
<td>4,953</td>
<td>24.1%</td>
<td>75.9%</td>
<td>14</td>
<td>32</td>
</tr>
</tbody>
</table>

Moreover, we design two metrics, Exact Rank Match (ERM) and Fuzzy Rank Match (FRM), to evaluate the rank accuracy of opinion tags. ERM is defined as the one-to-one exact match proportion between true tags and predicted tags. Inspired by the Embedding Score [26], FRM first maps predicted tags and the corresponding true tags into the same vector space, and then computes the average cosine similarity between their vector representations.

5.5 Implementation Details

We adopt the Adam [20] optimizer with settings $\{\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-6}, lr = 10^{-4}\}$ and we vary the learning rate following [45]. We add dropout [39] with keeping rate 0.8 and label smoothing [42] with smoothing factor 0.1 to avoid over-fitting. We use the Tencent AI Lab Chinese Embeddings\(^7\) for initialization of the word embedding layers. The rest of the parameters are randomly initialized. The dimensions of the alignment feature, word embedding layers, and positional embedding layers are set to 200. We set the batch size to 16 and use the validation loss for early stopping. When inference, we set the maximum decoding step as 50. We use a bidirectional GRU [12] with 2 layers to implement the sentence-level salience estimating component. All RNN-based models have 256 hidden units. All transformer-based models have 300 hidden units; the feed-forward hidden size is set to 50 for all layers. We set the $F$ in Section 4.4 to 3 ($3$ is the number of Focused Opinion Clusters at each decoding step) to ensure the target tag token have enough relevant reviews to reference and avoid introducing too much interference information simultaneously. AOT-Net was trained on a single Tesla V100 GPU and is implemented using PyTorch. All hyperparameters and models are selected on the validation set and the results are reported on the test set.

6 RESULTS AND ANALYSIS

Overall evaluation results on generating opinion tags are listed in Table 2. We find that all the abstractive models significantly outperform all the traditional extractive baselines. Thus we conclude that informal and colloquial nature of user-generated reviews make item-related features indiscernible using the unsupervised extraction methods. As expected, we also find that PG-Net significantly outperforms RNN, which implies that the copying mechanism is useful for opinion summarization. We can see AOT-Net significantly outperforms baseline Transformer in terms of all metrics and achieves the best performance for most metrics. The results of our ablation studies are shown in the lower part of the Table 2. We observe that after removing the SSE component, performances of AOT-Net in terms of most metrics drops obviously. If we do not rank and group reviews into opinion clusters before decoding (i.e., w/o RCR), although the diversity metric has a slight increase, the rank accuracy of AOT-Net decreases as we anticipated. We also find that after removing alignment feature or alignment loss mechanisms in the decoder, the performance of both retrieval and rank accuracy metrics (i.e., $F_1$ and FRM) degrades. We will conduct a detailed analysis of the individual components in the following sections.

6.1 Salience Estimation Analysis

As shown in Table 2, AOT-Net achieves a 17.1% and 2.72% increase over “w/o SSE” in terms of Micro-Distinct-2 and Macro-Distinct-2, respectively. Similar improvements can be observed for other metrics. This demonstrates that the predicted salience scores help AOT-Net to focus on more valuable reviews. To verify the effectiveness of SSE with more details, in Figure 6 we list the accuracy scores of AOT-Embed, AOT-RNN, and AOT-Net for sentence-level salience estimation. We find that AOT-Net outperforms both AOT-Embed and AOT-RNN, which verifies the effectiveness of BiGRU and self-attention mechanisms. AOT-RNN achieves 7.2% increase over AOT-Embed in terms of accuracy for all reviews, which verifies the advantage of BiGRU in representing review. In terms of accuracy, we find that AOT-Net gives a 0.4% and 5.1% increase over AOT-RNN for all reviews and item-related reviews, respectively. This indicates that AOT-Net benefits from self-attention mechanisms, which capture the shared information to distinguish item-related reviews.

![Figure 6: Accuracy values for sentence-level salience estimation.](image)

6.2 Number of FOCs

To evaluate the effect of the number of FOCs on the performance of rank-aware opinion tagging, we examine the performance of AOT-Net with different values of $F$ (see Section 4.4) in terms of ERM, FRM, and Distinct-2, respectively. As shown in Table 3, $F = 1$ significantly decreases the model rank and diversity performance. This suggests that only focusing on a single opinion cluster ignores many related reviews. We also find that when $F = 5$, the performance of AOT-Net slightly decreases; AOT-Net achieves the best performance in terms of all metrics when $F = 3$. Hence, we infer that $F$ is a trade-off between focusing on relevant reviews and removing irrelevant noise.

6.3 Case Study

Figure 7 shows an example illustrating the 5 highest attention weights $\alpha_{i,p}$ during the generation of opinion tags. We see that the

\(^7\)https://ai.tencent.com/ailab/nlp/embedding.html
Table 2: Evaluation results on the eComTag dataset. Results in bold are leading results in terms of the corresponding metric.

<table>
<thead>
<tr>
<th>Models</th>
<th>F1@5</th>
<th>F1@10</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>ERM</th>
<th>FRM</th>
<th>Micro</th>
<th>Macro</th>
<th>Unique-N</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>0.039</td>
<td>0.038</td>
<td>0.018</td>
<td>0.019</td>
<td>0.16</td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TextRank</td>
<td>0.0019</td>
<td>0.0018</td>
<td>0.0091</td>
<td>0.0097</td>
<td>0.06</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>0.2895</td>
<td>0.2753</td>
<td>0.7383</td>
<td>0.7701</td>
<td>0.17</td>
<td>0.44</td>
<td>0.60</td>
<td>62.19</td>
<td>7.447</td>
</tr>
<tr>
<td>PG-Net</td>
<td>0.3138</td>
<td>0.2896</td>
<td>0.7600</td>
<td>0.8009</td>
<td>0.19</td>
<td>0.44</td>
<td>1.33</td>
<td>65.78</td>
<td>6.798</td>
</tr>
<tr>
<td>Transformer</td>
<td>0.2833</td>
<td>0.2756</td>
<td>0.6916</td>
<td>0.7483</td>
<td>0.25</td>
<td>0.59</td>
<td>1.57</td>
<td>89.23</td>
<td>8.851</td>
</tr>
<tr>
<td>AOT-Net</td>
<td>0.3529</td>
<td>0.3492</td>
<td>0.7473</td>
<td>0.8045</td>
<td>0.31</td>
<td>0.64</td>
<td>1.30</td>
<td>94.35</td>
<td>8.953</td>
</tr>
<tr>
<td>w/o SSE</td>
<td>0.2930</td>
<td>0.2822</td>
<td>0.7022</td>
<td>0.7563</td>
<td>0.25</td>
<td>0.61</td>
<td>1.11</td>
<td>91.85</td>
<td>8.957</td>
</tr>
<tr>
<td>w/o RCR</td>
<td>0.3434</td>
<td>0.3370</td>
<td>0.7353</td>
<td>0.7913</td>
<td>0.31</td>
<td>0.63</td>
<td>1.77</td>
<td>92.94</td>
<td>8.935</td>
</tr>
<tr>
<td>w/o AF</td>
<td>0.3141</td>
<td>0.3056</td>
<td>0.7194</td>
<td>0.7768</td>
<td>0.28</td>
<td>0.62</td>
<td>1.21</td>
<td>93.23</td>
<td>8.997</td>
</tr>
<tr>
<td>w/o AL</td>
<td>0.3406</td>
<td>0.3336</td>
<td>0.7322</td>
<td>0.7857</td>
<td>0.30</td>
<td>0.64</td>
<td>1.30</td>
<td>93.11</td>
<td>8.926</td>
</tr>
</tbody>
</table>

Table 3: Performance on different numbers (F) of FOCs.

<table>
<thead>
<tr>
<th>Models</th>
<th>ERM</th>
<th>FRM</th>
<th>Macro</th>
<th>Distinct-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>F=1</td>
<td>0.13</td>
<td>0.58</td>
<td></td>
<td>62.36</td>
</tr>
<tr>
<td>F=3</td>
<td>0.31</td>
<td>0.64</td>
<td></td>
<td>94.35</td>
</tr>
<tr>
<td>F=5</td>
<td>0.16</td>
<td>0.54</td>
<td></td>
<td>73.96</td>
</tr>
</tbody>
</table>

Thus, we conclude that the alignment feature and alignment loss in AOT-Net are helpful to capture the ranks of the opinion tags.

7 CONCLUSION AND FUTURE WORK

In this paper, we have proposed the abstractive opinion tagging task, which aims to automatically rank a ranked list of opinion tags from a large number of reviews. We have proposed a rank-aware abstractive opinion tagging framework (AOT-Net) that includes a sentence-level salience estimating component, a review clustering and ranking component, and a rank-aware opinion tagging component. To validate the effectiveness of AOT-Net, we conducted extensive experiments on a newly collected real-world dataset, eComTag. Experiments show that AOT-Net achieves state-of-the-art performance on the abstractive opinion tagging task. AOT-Net has two main advantages over previous work. On the one hand, it generates more concise opinion tags; on the other hand, the ranked lists of generated opinion tags help users distinguish products with very similar aspects. Our work provides a plausible solution to greatly reduce human annotation costs for online e-commerce opinion tagging. Although we focused mostly on e-commerce portals, our methods are also broadly applicable to other settings with opinionated content, such as microblogs.

Limitations of our work include low efficiency and coarse-grained salience estimation. As to our future work, we will adopt the deep clustering network instead of k-means algorithm. Also, pre-trained language models could provide more power to enhance our sentence salience estimation. Few-shot learning for handling unbalanced review distributions among different domains could be another direction. It will be also interesting to explore user interactions with AOT-Net to generate personalized opinion tags in the future.

CODE AND DATA

The source code and dataset used in this paper are available at https://github.com/qtli/AOT.

ACKNOWLEDGMENTS

We thank our reviewers for valuable feedback. This work was supported by the National Key R&D Program of China with grant No. 2020YFB1406704, the Natural Science Foundation of China.
REFERENCES


