Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised

Stefanos Angelidis and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
s.angelidis@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

We present a neural framework for opinion summarization from online product reviews which is knowledge-lean and only requires light supervision (e.g., in the form of product domain labels and user-provided ratings). Our method combines two weakly supervised components to identify salient opinions and form extractive summaries from multiple reviews: an aspect extractor trained under a multi-task objective, and a sentiment predictor based on multiple instance learning. We introduce an opinion summarization dataset that includes a training set of product reviews from six diverse domains and human-annotated development and test sets with gold standard aspect annotations, salience labels, and opinion summaries. Automatic evaluation shows significant improvements over baselines, and a large-scale study indicates that our opinion summaries are preferred by human judges according to multiple criteria.[1]

1 Introduction

Opinion summarization, i.e., the aggregation of user opinions as expressed in online reviews, blogs, internet forums, or social media, has drawn much attention in recent years due to its potential for various information access applications. For example, consumers have to wade through many product reviews in order to make an informed decision. The ability to summarize these reviews succinctly would allow customers to efficiently absorb large amounts of opinionated text and manufacturers to keep track of what customers think about their products (Liu, 2012).

The majority of work on opinion summarization is *entity-centric*, aiming to create summaries from text collections that are relevant to a particular entity of interest, e.g., product, person, company, and so on. A popular decomposition of the problem involves three subtasks (Hu and Liu, 2004, 2006): (1) *aspect extraction* which aims to find specific features pertaining to the entity of interest (e.g., battery life, sound quality, ease of use) and identify expressions that discuss them; (2) *sentiment prediction* which determines the sentiment orientation (positive or negative) on the aspects found in the first step, and (3) *summary generation* which presents the identified opinions to the user (see Figure 1 for an illustration of the task).

A number of techniques have been proposed for aspect discovery using part of speech tagging (Hu and Liu, 2004), syntactic parsing (Lu et al., 2009), clustering (Mei et al., 2007; Titov and McDonald, 2008b), data mining (Ku et al., 2006), and information extraction (Popescu and Etzioni, 2005). Various lexicon and rule-based methods (Hu and Liu, 2004; Ku et al., 2006; Blair-Goldensohn et al., 2008) have been adopted for sentiment prediction together with a few learning approaches (Lu et al., 2009; Pappas and Popescu-Belis, 2017; Angelidis and Lapata, 2018). As for the summaries, a common format involves a list of aspects and the number of positive and negative opinions for each (Hu and Liu, 2004). While this format gives an overall idea of people's opinion, reading the actual text might be necessary to gain a better understanding of specific details. Textual summaries are created following mostly extractive methods (but see Ganesan et al. 2010 for an abstractive approach), and various formats ranging from lists of words (Popescu and Etzioni, 2005), to phrases (Lu et al., 2009), and sentences (Mei et al., 2007; Blair-Goldensohn et al., 2008; Lerman et al., 2009; Wang and Ling, 2016).

In this paper, we present a neural framework for opinion extraction from product reviews. We follow the standard architecture for aspect-based summarization, while taking advantage of the success of neural network models in learning continuous features without recourse to preprocessing tools or linguistic annotations. Central to our system is the ability to accurately identify aspect-

[1] Our code and dataset are publicly available at https://github.com/stangelid/oposum.

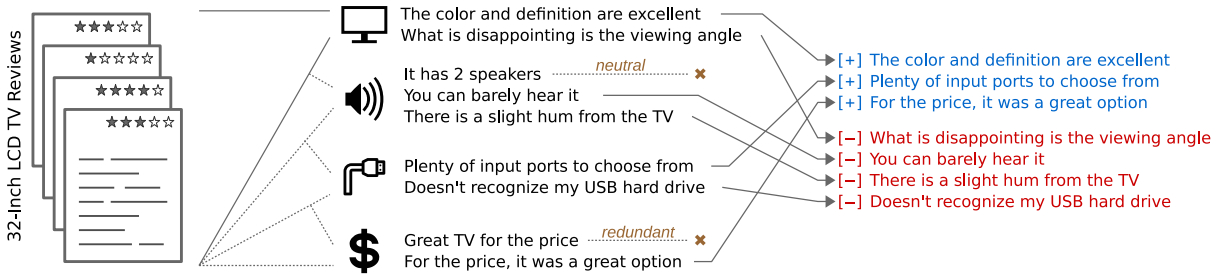

Figure 1: Aspect-based opinion summarization. Opinions on image quality, sound quality, connectivity, and price of an LCD television are extracted from a set of reviews. Their polarities are then used to sort them into positive and negative, while neutral or redundant comments are discarded.

specific opinions by using different sources of information freely available with product reviews (product domain labels, user ratings) and minimal domain knowledge (essentially a few aspect-denoting keywords). We incorporate these ideas into a recently proposed aspect discovery model (He et al., 2017) which we combine with a weakly supervised sentiment predictor (Angelidis and Lapata, 2018) to identify highly salient opinions. Our system outputs extractive summaries using a greedy algorithm to minimize redundancy. Our approach takes advantage of weak supervision signals only, requires minimal human intervention and no gold-standard salience labels or summaries for training.

Our contributions in this work are three-fold: a novel neural framework for the identification and extraction of salient customer opinions that combines aspect and sentiment information and does not require unrealistic amounts of supervision; the introduction of an opinion summarization dataset which consists of Amazon reviews from six product domains, and includes development and test sets with gold standard aspect annotations, salience labels, and multi-document extractive summaries; a large-scale user study on the quality of the final summaries paired with automatic evaluations for each stage in the summarization pipeline (aspects, extraction accuracy, final summaries). Experimental results demonstrate that our approach outperforms strong baselines in terms of opinion extraction accuracy and similarity to gold standard summaries. Human evaluation further shows that our summaries are preferred over comparison systems across multiple criteria.

2 Related Work

It is outside the scope of this paper to provide a detailed treatment of the vast literature on opinion summarization and related tasks. For a comprehensive overview of non-neural methods we refer the interested reader to Kim et al. (2011) and Liu and Zhang (2012). We are not aware of previous studies which propose a neural-based system for end-to-end opinion summarization without direct supervision, although as we discuss below, recent efforts tackle various subtasks independently.

Aspect Extraction Several neural network models have been developed for the identification of aspects (e.g., words or phrases) expressed in opinions. This is commonly viewed as a supervised sequence labeling task; Liu et al. (2015) employ recurrent neural networks, whereas Yin et al. (2016) use dependency-based embeddings as features in a Conditional Random Field (CRF). Wang et al. (2016) combine a recursive neural network with CRFs to jointly model aspect and sentiment terms. He et al. (2017) propose an aspect-based autoencoder to discover fine-grained aspects without supervision, in a process similar to topic modeling. Their model outperforms LDA-style approaches and forms the basis of our aspect extractor.

Sentiment Prediction Fully-supervised approaches based on neural networks have achieved impressive results on fine-grained sentiment classification (Kim, 2014; Socher et al., 2013). More recently, *Multiple Instance Learning* (MIL) models have been proposed that use freely available review ratings to train segment-level predictors. Kotzias et al. (2015) and Pappas and Popescu-Belis (2017) train sentence-level predictors under a MIL objective, while our previous work (Angelidis and Lapata, 2018) introduced MILNET, a hierarchical model that is trained end-to-end on document labels and produces polarity-based opinion summaries of single reviews. Here, we use MILNET to predict the sentiment polarity of individual opinions.

Multi-document Summarization A few extractive neural models have been recently applied to generic multi-document summarization. Cao et al. (2015) train a recursive neural network using a ranking objective to identify salient sentences, while follow-up work (Cao et al., 2017) employs a multi-task objective to improve sentence extraction, an idea we adapted to our task. Yasunaga et al. (2017) propose a graph convolution network to represent sentence relations and estimate sentence salience. Our summarization method is tailored to the opinion extraction task, it identifies aspect-specific and salient units, while minimizing the redundancy of the final summary with a greedy selection algorithm (Cao et al., 2015; Yasunaga et al., 2017). Redundancy is also addressed in Ganesan et al. (2010) who propose a graph-based framework for abstractive summarization. Wang and Ling (2016) introduce an encoder-decoder neural method for extractive opinion summarization. Their approach requires direct supervision via gold-standard extractive summaries for training, in contrast to our weakly supervised formulation.

3 Problem Formulation

Let C denote a corpus of reviews on a set of products $E_{\mathrm{C}} = \{e_i\}_{i=1}^{|E_{\mathrm{C}}|}$ from a domain d_{C}, e.g., televisions or keyboards. For every product e, the corpus contains a set of reviews $R_e = \{r_i\}_{i=1}^{|R_e|}$ expressing customers' opinions. Each review r_i is accompanied by the author's overall rating y_i and is split into segments $(s_1, \ldots, s_m)$, where each segment s_j is in turn viewed as a sequence of words $(w_{j1}, \ldots, w_{jn})$. A segment can be a sentence, a phrase, or in our case an *Elementary Discourse Unit* (EDU; Mann and Thompson 1988) obtained from a *Rhetorical Structure Theory* (RST) parser (Feng and Hirst, 2012). EDUs roughly correspond to clauses and have been shown to facilitate performance in summarization (Li et al., 2016), document-level sentiment analysis (Bhatia et al., 2015), and single-document opinion extraction (Angelidis and Lapata, 2018).

A segment may discuss zero or more *aspects*, i.e., different product attributes. We use $A_{\mathrm{C}} = \{a_i\}_{i=1}^{K}$ to refer to the aspects pertaining to domain d_{C}. For example, *picture quality*, *sound quality*, and *connectivity* are all aspects of televisions. By convention, a *general* aspect is assigned to segments that do not discuss any specific aspects. Let $A_s \subseteq A_{\mathrm{C}}$ denote the set of aspects mentioned in segment s; $pol_s \in [-1, +1]$ marks the *polarity* a segment conveys, where -1 indicates maximally negative and $+1$ maximally positive sentiment. An opinion is represented by tuple $o_s = (s, A_s, pol_s)$, and $O_e = \{o_s\}_{s \in R_e}$ represents the set of all opinions expressed in R_e.

For each product e, our goal is to produce a summary of the most salient opinions expressed in reviews R_e, by selecting a small subset $S_e \subset O_e$. We expect segments that discuss specific product aspects to be better candidates for useful summaries. We hypothesize that *general* comments mostly describe customers' overall experience, which can also be inferred by their rating, whereas aspect-related comments provide specific reasons for their overall opinion. We also assume that segments conveying highly positive or negative sentiment are more likely to present informative opinions compared to neutral ones, a claim supported by previous work (Angelidis and Lapata, 2018).

We describe our novel approach to aspect extraction in Section 4 and detail how we combine aspect, sentiment, and redundancy information to produce opinion summaries in Section 5.

4 Aspect Extraction

Our work builds on the aspect discovery model developed by He et al. (2017), which we extend to facilitate the accurate extraction of aspect-specific review segments in a more realistic setting. In this section, we first describe their approach, point out its shortcomings, and then present the extensions and modifications introduced in our *Multi-Seed Aspect Extractor* (MATE) model.

4.1 Aspect-Based Autoencoder

The *Aspect-Based Autoencoder* (ABAE; He et al. 2017) is an adaptation of the *Relationship Modeling Network* (Iyyer et al., 2016), originally designed to identify attributes of fictional book characters and their relationships. The model learns a segment-level aspect predictor without supervision by attempting to reconstruct the input segment's encoding as a linear combination of aspect embeddings. ABAE starts by pairing each word w with a pre-trained word embedding $\mathbf{v}_w \in \mathbb{R}^d$, thus constructing a word embedding dictionary $\mathbf{L} \in \mathbb{R}^{V \times d}$, where V is the size of the vocabulary. The model also keeps an aspect embedding dictionary $\mathbf{A} \in \mathbb{R}^{K \times d}$, where K is the number of aspects to be identified and i-th row $\mathbf{a}_i \in \mathbb{R}^d$ is a point in the word embedding space. Matrix $\mathbf{A}$ is initialized using the centroids from a

k-means clustering on the vocabulary's word embeddings.

The autoencoder, first produces a vector $\mathbf{v}_s$ for review segment $s = (w_1, \ldots, w_n)$ using an *attention encoder* that learns to attend on aspect words. A segment encoding is computed as the weighted average of word vectors:

$$\mathbf{v}_s = \sum_{i=1}^{n} c_i \mathbf{v}_{w_i} \quad (1)$$

$$c_i = \frac{exp(u_i)}{\sum_{j=1}^{n} exp(u_j)} \quad (2)$$

$$u_i = \mathbf{v}_{w_i}^{\mathsf{T}} \cdot \mathbf{M} \cdot \mathbf{v}'_s\,, \quad (3)$$

where c_i is the i-th word's attention weight, $\mathbf{v}'_s$ is a simple average of the segment's word embeddings and attention matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ is learned during training.

Vector $\mathbf{v}_s$ is fed into a softmax classifier to predict a probability distribution over K aspects:

$$\mathbf{p}_s^{asp} = \text{softmax}(\mathbf{W}\mathbf{v}_s + \mathbf{b})\,, \quad (4)$$

where $\mathbf{W} \in \mathbb{R}^{K \times d}$ and $\mathbf{b} \in \mathbb{R}^K$ are the classifier's weight and bias parameters. The segment's vector is then reconstructed as the weighted sum of aspect embeddings:

$$\mathbf{r}_s = \mathbf{A}^{\mathsf{T}} \mathbf{p}_s^{asp}\,. \quad (5)$$

The model is trained by minimizing a reconstruction loss $J_r(\theta)$ that uses randomly sampled segments $n_1, n_2, \ldots, n_{k_n}$ as negative examples:[2]

$$J_r(\theta) = \sum_{s \in \mathrm{C}} \sum_{i=1}^{k_n} \max(0, 1 - \mathbf{r}_s \mathbf{v}_s + \mathbf{r}_s \mathbf{v}_{n_i}) \quad (6)$$

ABAE is essentially a neural topic model; it discovers topics which will hopefully map to aspects, without any preconceptions about the aspects themselves, a feature shared with most previous LDA-style aspect extraction approaches (Titov and McDonald, 2008a; He et al., 2017; Mukherjee and Liu, 2012). These models will set the number of topics to be discovered to a much larger number (~ 15) than the actual aspects found in the data (~ 5). This requires a many-to-one mapping between discovered topics and genuine aspects which is performed manually.

[2] ABAE also uses a uniqueness regularization term that is not shown here and is not used in our Multi-Seed Aspect Extractor model.

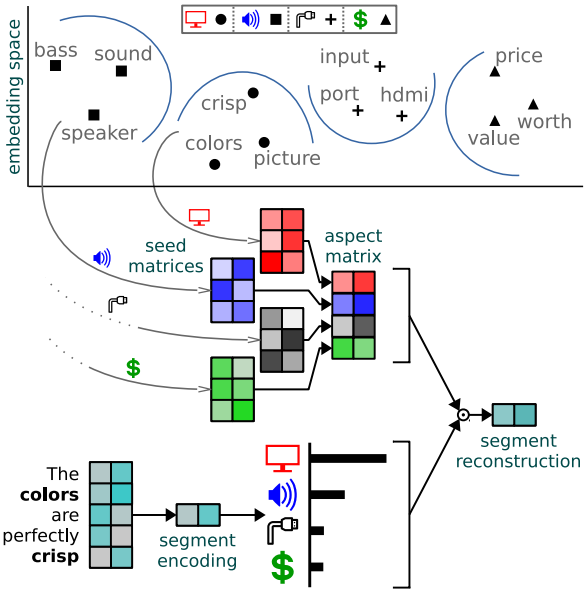

Figure 2: Multi-Seed Aspect Extractor (MATE).

4.2 Multi-Seed Aspect Extractor

Dynamic aspect extraction is advantageous since it assumes nothing more than a set of relevant reviews for a product and may discover unusual and interesting aspects (e.g., whether a plasma television has protective packaging). However, it suffers from the fact that the identified aspects are fine-grained, they have to be interpreted post-hoc, and manually mapped to coarse-grained ones.

We propose a new weakly-supervised set-up for aspect extraction which requires little human involvement. For every aspect $a_i \in A_{\mathrm{C}}$, we assume there exists a small set of seed words $\{sw_j\}_{j=1}^{l}$ which are good descriptors of a_i. We can think of these *seeds* as query terms that someone would use to search for segments discussing a_i. They can be set manually by a domain expert or selected using a small number of aspect-annotated reviews. Figure 2 (top) depicts four television aspects (*image*, *sound*, *connectivity* and *price*) and three of their seeds in word embedding space. MATE replaces ABAE's aspect dictionary with multiple seed matrices $\{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K\}$. Every matrix $\mathbf{A}_i \in \mathbb{R}^{l \times d}$, contains one row per seed word and holds the seeds' word embeddings, as illustrated by the set of $[3 \times 2]$ matrices in Figure 2.

MATE still needs to produce an aspect matrix $\mathbf{A} \in \mathbb{R}^{K \times d}$, in order to reconstruct the input segment's embedding. We accomplish this by reducing each seed matrix to a single aspect embedding with the help of seed weight vectors $\mathbf{z}_i \in \mathbb{R}^l$ ($\sum_j z_{ij} = 1$), and concatenating the results, illus-

trated by the $[4 \times 2]$ aspect matrix in Figure 2:

$$\mathbf{a}_i = \mathbf{A}_i^\mathsf{T} \mathbf{z}_i \tag{7}$$

$$\mathbf{A} = [\mathbf{a}_1^\mathsf{T}; \ldots; \mathbf{a}_K^\mathsf{T}]\,. \tag{8}$$

The segment is reconstructed as in Equation (5). Weight vectors $\mathbf{z}_i$ can be uniform (for manually selected seeds), fixed, learned during training, or set dynamically for each input segment, based on the cosine distance of its encoding to each seed embedding. Our experiments showed that fixed weights, selected through a technique described below, result in most stable performance across domains. We only focus on this variant due to space restrictions (but provide more details in the supplementary material).

When a small number of aspect-annotated reviews are available, seeds and their fixed seed weights can be selected automatically. To obtain a ranked list of terms that are most characteristic for each aspect, we use a variant of the *clarity* scoring function which was first introduced in information retrieval (Cronen-Townsend et al., 2002). Clarity measures how much more likely it is to observe word w in the subset of segments that discuss aspect a, compared to the corpus as a whole:

$$\text{score}_a(w) = t_a(w) \log_2 \frac{t_a(w)}{t(w)}\,, \tag{9}$$

where $t_a(w)$ and $t(w)$ are the l_1-normalized *tf-idf* scores of w in the segments annotated with aspect a and in all annotated segments, respectively. Higher scores indicate higher term importance and truncating the ranked list of terms gives a fixed set of seed words, as well as their seed weights by normalizing the scores to add up to one. Table 1 shows the highest ranked terms obtained for every aspect in the *televisions* domain of our corpus (see Section 6 for a detailed description of our data).

4.3 Multi-Task Objective

MATE (and ABAE) relies on the attention encoder to identify and attend to each segment's aspect-signalling words. The reconstruction objective only provides a weak training signal, so we devise a multi-task extension to enhance the encoder's effectiveness without additional annotations.

We assume that aspect-relevant words not only provide a better basis for the model's aspect-based reconstruction, but are also good indicators of the product's domain. For example, the words *colors* and *crisp*, in the segment "*The colors are perfectly crisp*" should be sufficient to infer that the segment comes from a television review, whereas the words *keys* and *type* in the segment "*The keys feel great to type on*" are more representative of the keyboard domain. Additionally, all four words are characteristic of specific aspects.

Aspect	Top Terms
Image	picture color quality black bright
Sound	sound speaker quality bass loud
Connectivity	hdmi port computer input component
Price	price value money worth paid
Apps & Interface	netflix user file hulu apps
Ease of Use	easy remote setup user menu
Customer Service	paid support service week replace
Size & Look	size big bigger difference screen
General	tv bought hdtv happy problem

Table 1: Highest ranked words for the television corpus according to Equation (9).

Let $\mathrm{C}_{all} = \mathrm{C}_1 \cup \mathrm{C}_2 \cup \ldots$ denote the union of multiple review corpora, where C_1 is considered *in-domain* and the rest are considered *out-of-domain*. We use $d_s \in \{d_{\mathrm{C}_1}, d_{\mathrm{C}_2}, \ldots\}$ to denote the true domain of segment s and define a classifier that uses the vectors from our segment encoder as inputs:

$$\mathbf{p}_s^{dom} = \text{softmax}(\mathbf{W}_\mathrm{C} \mathbf{v}_s + \mathbf{b}_\mathrm{C})\,, \tag{10}$$

where $\mathbf{p}_s^{dom} = \langle p^{(d_{\mathrm{C}_1})}, p^{(d_{\mathrm{C}_2})}, \ldots \rangle$ is a probability distribution over product domains for segment s and $\mathbf{W}_\mathrm{C}$ and $\mathbf{b}_\mathrm{C}$ are the classifier's weight and bias parameters. We use the negative log likelihood of the domain prediction as the objective function, combined with the reconstruction loss of Equation (5) to obtain a multi-task objective:

$$J_{\mathrm{MT}}(\theta) = J_r(\theta) - \lambda \sum_{s \in \mathrm{C}_{all}} \log p^{(d_s)}\,, \tag{11}$$

where λ controls the influence of the classification loss. Note that the negative log-likelihood is summed over all segments in C_{all}, whereas $J_r(\theta)$ is only summed over the in-domain segments $s \in \mathrm{C}_1$. It is important not to use the out-of-domain segments for segment reconstruction, as they will confuse the aspect extractor due to the aspect mismatch between different domains.

5 Opinion Summarization

We now move on to describe our opinion summarization framework which is based on the aspect extraction component discussed so far, a polarity prediction model, and a segment selection policy which identifies and discards redundant opinions.

Segment	Salience
1. The color and definition are perfect.	[+] 0.89
2. Set up was extremely easy,	[+] 0.79
3. Not worth $ 300.	[−] 0.75
4. The sound on this is horrendous.	[−] 0.52
5. The sound is TERRIBLE.	[−] 0.45
6. Nice and bright with good colors.	[+] 0.44

Table 2: Most salient opinions according to scores from Equation (12) for an LCD TV.

Domain	Products	Reviews	EDUs	Vocab
Laptop Cases	2,040 (10)	42,727 (100)	602K (1,262)	30,443
B/T Headsets	1,471 (10)	80,239 (100)	1.46M (1,344)	51,263
Boots	4,723 (10)	77,593 (100)	987K (1,198)	30,364
Keyboards	983 (10)	33,713 (100)	625K (1,396)	34,095
Televisions	1,894 (10)	56,510 (100)	1.47M (1,483)	59,051
Vacuums	1,184 (10)	68,266 (100)	1.50M (1,492)	46,259

Table 3: The OPOSUM corpus. Numbers in parentheses correspond to the human-annotated subset.

Opinion Polarity Aside from describing a product's aspects, segments also express polarity (i.e., positive or negative sentiment). We identify segment polarity with the recently proposed *Multiple Instance Learning Network* model (MILNET; Angelidis and Lapata 2018). Whilst trained on freely available document-level sentiment labels, i.e., customer ratings on a scale from 1 (negative) to 5 (positive), MILNET learns a segment-level sentiment predictor using a hierarchical, attention-based neural architecture.

Given review r consisting of segments $(s_1, \ldots, s_m)$, MILNET uses a CNN segment encoder to obtain segment vectors $(\mathbf{u}_1, \ldots, \mathbf{u}_m)$, each used as input to a segment-level sentiment classifier. For every vector $\mathbf{u}_i$, the classifier produces a sentiment prediction $\mathbf{p}_i^{stm} = \langle p_i^{(1)}, \ldots, p_i^{(M)} \rangle$, where $p_i^{(1)}$ and $p_i^{(M)}$ are probabilities assigned to the most negative and most positive sentiment class respectively. Resulting segment predictions $(\mathbf{p}_1^{stm}, \ldots, \mathbf{p}_m^{stm})$ are combined via a GRU-based attention mechanism to produce a document-level prediction $\mathbf{p}_r^{stm}$ and the model is trained end-to-end on the reviews' user ratings using negative log-likelihood.

The essential by-product of MILNET are segment-level sentiment predictions $\mathbf{p}_i^{stm}$, which are transformed into polarities pol_{s_i}, by projecting them onto the $[-1, +1]$ range using a uniformly spaced sentiment class weight vector.

Opinion Ranking Aspect predictions $\mathbf{p}_s^{asp} = \langle p_s^{(a_1)}, \ldots, p_s^{(a_K)} \rangle$ and polarities pol_s, form the opinion set $O_e = \{(s, A_s, pol_s)\}_{s \in R_e}$ for every product $e \in E_C$. For simplicity, we set the predicted aspect-set A_s to only include the aspect with the highest probability, although it is straightforward to allow for multiple aspects. We rank every opinion $o_s \in O_e$ according to its salience:

$$sal(o_s) = |pol_s| \cdot (\max_i p_s^{(a_i)} - p_s^{(\text{GEN})}), \quad (12)$$

where the quantity in parentheses is the probability difference between the most probable aspect and the *general* aspect. The salience score will be high for opinions that are very positive or very negative and are also likely to discuss a non-general aspect.

Opinion Selection The final step towards producing summaries is to discard potentially redundant opinions, something that is not taken into account by our salience scoring method. Table 2 shows a partial ranking of the most salient opinions found in the reviews for an LCD television. All segments provide useful information, but it is evident that segments 1 and 6 as well as 4 and 5 are paraphrases of the same opinions.

We follow previous work on multi-document summarization (Cao et al., 2015; Yasunaga et al., 2017) and use a greedy algorithm to eliminate redundancy. We start with the highest ranked opinion, and keep adding opinions to the final summary one by one, unless the cosine similarity between the candidate segment and any segment already included in the summary is lower than 0.5.

6 The OPOSUM Dataset

We created OPOSUM, a new dataset for the training and evaluation of **Opi**nio**n** **Sum**marization models which contains Amazon reviews from six product domains: *Laptop Bags*, *Bluetooth Headsets*, *Boots*, *Keyboards*, *Televisions*, and *Vacuums*. The six training collections were created by downsampling from the *Amazon Product Dataset*[3] introduced in McAuley et al. (2015) and contain reviews and their respective ratings. The reviews were segmented into EDUs using a publicly available RST parser (Feng and Hirst, 2012).

To evaluate our methods and facilitate research, we produced a human-annotated subset of the dataset. For each domain, we uniformly sampled (across ratings) 10 different products with 10 reviews each, amounting to a total of 600 reviews, to be used only for development (300) and testing (300). We obtained EDU-level aspect annotations, salience labels and gold standard opinion

[3] http://jmcauley.ucsd.edu/data/amazon/

Aspect Extraction (F1)	L. Bags	B/T H/S	Boots	Keyb/s	TVs	Vac/s	**AVG**
Majority	37.9	39.8	37.1	43.2	41.7	41.6	40.2
ABAE	38.1	37.6	35.2	38.6	39.5	38.1	37.9
ABAE$_{init}$	41.6	48.5	41.2	41.3	45.7	40.6	43.2
MATE	46.2	52.2	45.6	43.5	48.8	42.3	46.4
MATE+MT	**48.6**	**54.5**	**46.4**	**45.3**	**51.8**	**47.7**	**49.1**

Salience (MAP/P@5)	L. Bags	B/T H/S	Boots	Keyb/s	TVs	Vac/s	**AVG**
MILNET	21.8 / 40.0	19.8 / 36.7	17.0 / 39.3	14.1 / 28.0	14.3 / 36.0	14.6 / 31.3	16.9 / 35.2
ABAE$_{init}$	19.9 / 48.5	27.5 / 49.7	13.8 / 28.1	19.0 / 44.9	16.8 / 42.4	16.1 / 34.0	18.8 / 41.3
MATE	23.0 / 57.1	30.9 / 50.7	15.4 / 31.9	21.0 / 43.1	18.7 / 44.7	19.9 / 44.0	21.5 / 45.2
MATE+MT	26.3 / 60.8	37.5 / 66.7	17.3 / 33.6	20.9 / 44.9	23.6 / 48.0	22.4 / 43.9	24.7 / 49.6
MILNET+ABAE$_{init}$	27.1 / 56.0	33.5 / 66.5	19.3 / 34.8	22.4 / 51.7	19.0 / 43.7	20.8 / 43.5	23.7 / 49.4
MILNET+MATE	28.2 / 54.7	36.0 / 66.5	21.7 / 39.3	24.0 / 52.0	20.8 / 46.1	23.5 / 49.3	25.7 / 51.3
MILNET+MATE+MT	**32.1 / 69.2**	**40.0 / 74.7**	**23.3 / 40.4**	**24.8 / 56.4**	**23.8 / 52.8**	**26.0 / 53.1**	**28.3 / 57.8**

Table 4: Experimental results for the identification of aspect segments (top) and the retrieval of salient segments (bottom) on OPOSUM's six product domains and overall (AVG).

summaries, as described below. Statistics are provided in Table 3 and in supplementary material.

Aspects For every domain, we pre-selected nine representative aspects, including the *general* aspect. We presented the EDU-segmented reviews to three annotators and asked them to select the aspects discussed in each segment (multiple aspects were allowed). Final labels were obtained using a majority vote among annotators. Inter-annotator agreement across domains and annotated segments using Cohen's Kappa coefficient was $K = 0.61$ ($N = 8{,}175$, $k = 3$).

Opinion Summaries We produced opinion summaries for the 60 products in our benchmark using a two-stage procedure. First, all reviews for a product were shown to three annotators. Each annotator read the reviews one-by-one and selected the subset of segments they thought best captured the most important and useful comments, without taking redundancy into account. This phase produced binary *salience* labels against which we can judge the ability of a system to identify important opinions. Again, using the Kappa coefficient, agreement among annotators was $K = 0.51$ ($N = 8{,}175$, $k = 3$).[4] In the second stage, annotators were shown the salient segments they identified (for every product) and asked to create a final extractive summary by choosing opinions based on their popularity, fluency and clarity, while avoiding redundancy and staying under a budget of 100 words. We used ROUGE (Lin and Hovy, 2003) as a proxy to inter-annotator agreement. For every product, we treated one reference summary as system output and computed how it agrees with the rest. ROUGE scores are reported in Table 5 (last row).

[4]While this may seem moderate, Radev et al. (2003) show that inter-annotator agreement for extractive summarization is usually lower ($K < 0.30$).

7 Experiments

In this section, we discuss implementation details and present our experimental setup and results. We evaluate model performance on three subtasks: aspect identification, salient opinion extraction, and summary generation.

Implementation Details Reviews were lemmatized and stop words were removed. We initialized MATE using 200-dimensional word embeddings trained on each product domain using skip-gram (Mikolov et al., 2013) with default parameters. We used 30 seed words per aspect, obtained via Equation (9). Word embeddings $\mathbf{L}$, seed matrices $\{\mathbf{A}_i\}_{i=1}^{K}$ and seed weight vectors $\{\mathbf{z}_i\}_{i=1}^{K}$ were fixed throughout training. We used the Adam optimizer (Kingma and Ba, 2014) with learning rate 10^{-4} and mini-batch size 50, and trained for 10 epochs. We used 20 negative examples per input for the reconstruction loss and, when used, the multi-tasking coefficient λ was set to 10. Seed words and hyperparameters were selected on the development set and we report results on the test set, averaged over 5 runs.

Aspect Extraction We trained aspect models on the collections of Table 3 and evaluated their predictions against the human-annotated portion of each corpus. Our MATE model and its multi-task counterpart (MATE+MT) were compared against a majority baseline and two ABAE variants: vanilla ABAE, where aspect matrix $\mathbf{A}$ is initialized using k-means centroids and fine-tuned during training; and ABAE$_{init}$, where rows of $\mathbf{A}$

are fixed to the centroids of respective seed embeddings. This allows us to examine the benefits of our multi-seed aspect representation. Table 4 (top) reports the results using micro-averaged F1. Our models outperform both variants of ABAE across domains. ABAE$_{init}$ improves upon the vanilla model, affirming that informed aspect initialization can facilitate the task. The richer multi-seed representation of MATE, however, helps our model achieve a 3.2% increase in F1. Further improvements are gained by the multi-task model, which boosts performance by 2.7%.

Opinion Salience We are also interested in our system's ability to identify salient opinions in reviews. The first phase of our opinion extraction annotation provides us with binary salience labels, which we use as gold standard to evaluate system opinion rankings. For every product e, we score each segment $s \in R_e$ using Equation (12) and evaluate the obtained rankings via Mean Average Precision (MAP) and Precision at the 5th retrieved segment (P@5).[5] Polarity scores were produced via MILNET; we obtained aspect probabilities from ABAE$_{init}$, MATE, and MATE+MT. We also experimented with a variant that only uses MILNET's polarities and, additionally, with variants that ignore polarities and only use aspect probabilities.

Results are shown in Table 4 (bottom). The combined use of polarity and aspect information improves the retrieval of salient opinions across domains, as all model variants that use our salience formula of Equation (12) outperform the MILNET- and aspect-only baselines. When comparing between aspect-based alternatives, we observe that the extraction accuracy correlates with the quality of aspect prediction. In particular, ranking using MILNET+MATE+MT gives best results, with a 2.6% increase in MAP against MILNET+MATE and 4.6% against MILNET+ABAE$_{init}$. The trend persists even when MILNET polarities are ignored, although the quality of rankings is worse in this case.

Opinion Summaries We now turn to the summarization task itself, where we compare our best performing model (MILNET+MATE+MT), with and without a redundancy filter (RD), against the following methods: a baseline that selects segments *randomly*; a *Lead* baseline that only selects the leading segments from each review; *SumBasic*, a generic frequency-based extractive summarizer (Nenkova and Vanderwende, 2005); *LexRank*, a generic graph-based extractive summarizer (Erkan and Radev, 2004); *Opinosis*, a graph-based abstractive summarizer that is designed for opinion summarization (Ganesan et al., 2010). All extractive methods operate on the EDU level with a 100-word budget. For Opinosis, we tested an aspect-agnostic variant that takes every review segment for a product as input, and a variant that uses MATE's groupings of segments to produce and concatenate aspect-specific summaries.

[5] A system's salience ranking is individually compared against labels from each annotator and we report the average.

Summarization	ROUGE-1	ROUGE-2	ROUGE-L
Random	35.1	11.3	34.3
Lead	35.5	15.2	34.8
SumBasic	34.0	11.2	32.6
LexRank	37.7	14.1	36.6
Opinosis	36.8	14.3	35.7
Opinosis+MATE+MT	38.7	15.8	37.4
MILNET+MATE+MT	43.5	21.7	42.8
MILNET+MATE+MT+RD	**44.1**	**21.8**	**43.3**
Inter-annotator Agreement	54.7	36.6	53.9

Table 5: Summarization results on OPOSUM.

	Inform.	Polarity	Coherence	Redund.
Gold	2.04	**8.70**	**10.93**	**6.11**
This work	**9.26**	3.15	1.11	2.96
Opinosis	-12.78	-10.00	-9.08	-9.45
Lead	1.48	-1.85	-2.96	0.37

Table 6: *Best-Worst Scaling* human evaluation.

Table 5 presents ROUGE-1, ROUGE-2 and ROUGE-L F1 scores, averaged across domains. Our model (MILNET+MATE+MT) significantly outperforms all comparison systems ($p < 0.05$; paired bootstrap resampling; Koehn 2004), whilst using a redundancy filter slightly improves performance. Assisting Opinosis with aspect predictions is beneficial, however, it remains significantly inferior to our model (see the supplementary material for additional results).

We also performed a large-scale user study. For every product in the OPOSUM test set, participants were asked to compare summaries produced by: a (randomly selected) human annotator, our best performing model (MILNET+MATE+MT+RD), Opinosis, and the Lead baseline. The study was conducted on the Crowdflower platform using *Best-Worst Scaling* (BWS; Louviere and Woodworth 1991; Louviere et al. 2015), a less labour-intensive alternative to paired comparisons that has been shown to produce more reliable results than rating scales (Kiritchenko and Mohammad, 2017). We arranged every 4-tuple of competing summaries into four triplets. Every triplet was

Product domain: Televisions **Product name:** Sony BRAVIA 46-Inch HDTV	
Human	Plenty of ports and settings. Easy hookups to audio and satellite sources. The sound is good and strong. This TV looks very good. and the price is even better. The on-screen menu/options is quite nice. and the internet apps work as expected. The picture is clear and sharp. which is TOO SLOW to stream HD video... The software and apps built into this TV. are difficult to use and setup. Their service is handled off shore making. communication a bit difficult. :(
LexRank	Get a Roku or Netflix box. I watch cable, Netflix, Hulu Plus, YouTube videos and computer movie files on it. Sound is good much better. DO NOT BUY! this SONY Bravia ' Smart ' TV... and avoid the Sony apps at all costs. Because of these two issues, I returned the Sony TV. Also you can change the display and sound settings on each port. However, the streaming speed for netflix is just down right terrible. Most of the time I just quit. Since I do not own the cable box, So, I have the cable.
Opinosis	The picture and not bright at all even compared to my 6-year old sony lcd tv. It will not work with an hdmi. Connection because of a conflict with comcast's dhcp. Being generous because I usuallly like the design and attention to detail of sony products). I am very disappointed with this tv for two reasons: picture brightness and channel menu. Numbers of options available in the on-line area of the tv are numerous and extremely useful. Wow look at the color, look at the sharpness of the picture, amazing and the amazing.
This work	Plenty of ports and settings and have been extremely happy with it. The sound is good and strong. The picture is beautiful. And the internet apps work as expected. And the price is even better. Unbelieveable picture and the setup is so easy. Wow look at the color, look at the sharpness of the picture. The Yahoo! widgets do not work. And avoid the Sony apps at all costs. Communication a bit difficult. :(

Figure 3: Human and system summaries for a product in the *Televisions* domain.

shown to three crowdworkers, who were asked to decide which summary was *best* and which one was *worst* according to four criteria: *Informativeness* (How much useful information about the product does the summary provide?), *Polarity* (How well does the summary highlight positive and negative opinions?), *Coherence* (How coherent and easy to read is the summary?) *Redundancy* (How successfully does the summary avoid redundant opinions?).

For every criterion, a system's score is computed as the percentage of times it was selected as best minus the percentage of times it was selected as worst (Orme, 2009). The scores range from -100 (unanimously worst) to +100 (unanimously best) and are shown in Table 6. Participants favored our model over comparison systems across all criteria (all differences are statistically significant at $p < 0.05$ using post-hoc HD Tukey tests). Human summaries are generally preferred over our model, however the difference is significant only in terms of coherence ($p < 0.05$).

Finally, Figure 3 shows example summaries for a product from our televisions domain, produced by one of our annotators and by 3 comparison systems (LexRank, Opinosis and our MILNET+MATE+MT+RD). The human summary is primarily focused on aspect-relevant opinions, a characteristic that is also captured to a large extent by our method. There is substantial overlap between extracted segments, although our redundancy filter fails to identify a few highly similar opinions (e.g., those relating to the picture quality). The LexRank summary is inferior as it only identifies a few useful opinions, and instead selects many general or non-opinionated comments. Lastly, the abstractive summary of Opinosis does a good job of capturing opinions about specific aspects but lacks in fluency, as it produces grammatical errors. For additional system outputs, see supplementary material.

8 Conclusions

We presented a weakly supervised neural framework for aspect-based opinion summarization. Our method combined a seeded aspect extractor that is trained under a multi-task objective without direct supervision, and a multiple instance learning sentiment predictor, to identify and extract useful comments in product reviews. We evaluated our weakly supervised models on a new opinion summarization corpus across three subtasks, namely aspect identification, salient opinion extraction, and summary generation. Our approach delivered significant improvements over strong baselines in each of the subtasks, while a large-scale judgment elicitation study showed that crowdworkers favor our summarizer over competitive extractive and abstractive systems.

In the future, we plan to develop a more integrated approach where aspects and sentiment orientation are jointly identified, and work with additional languages and domains. We would also like to develop methods for abstractive opinion summarization using weak supervision signals.

Acknowledgments We gratefully acknowledge the financial support of the European Research Council (award number 681760).

References

Stefanos Angelidis and Mirella Lapata. 2018. Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *Proceedings of the WWW Workshop on NLP Challenges in the Information Explosion Era (NLPIX)*, Beijing, China.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2017. Improving multi-document summarization via text classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3053–3059, San Francisco, California, USA.

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

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 299–306, New York, NY, USA.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Wei Vanessa Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 60–68, Jeju Island, Korea.

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

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 388–397, Vancouver, Canada.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceeding of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, Washington, USA.

Minqing Hu and Bing Liu. 2006. Opinion extraction and summarization on the web. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1621–1624, Boston, Massachusettes, USA.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, San Diego, California.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 465–470, Vancouver, Canada.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain.

Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 597–606, Sydney, Australia.

Lun-Wei Ku, Yun-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Syposium on Computational Approaches to Analysing Weblogs*, pages 100–107, Palo Alto, California, USA.

Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. 2009. Sentiment summarization: Evaluating and learning user preferences. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 514–522. Association for Computational Linguistics.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles, California, USA.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 71–78. Association for Computational Linguistics.

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

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. *Mining Text Data, Springer*, pages 415–463.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443, Lisbon, Portugal.

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

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. *University of Alberta: Working Paper*.

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web*, pages 131–140, Madrid, Spain.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52, Santiago, Chile.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, pages 171–180, Banff, Alberta, Canada.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, Lake Tahoe, California, USA.

Arjun Mukherjee and Bing Liu. 2012. Modeling review comments. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 320–329, Jeju Island, Korea.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. Technical report.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and HB. Technical report.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research*, 58:591–626.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346, Vancouver, British Columbia, Canada.

Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 375–382. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA.

Ivan Titov and Ryan McDonald. 2008a. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 308–316, Columbus, Ohio, USA.

Ivan Titov and Ryan McDonald. 2008b. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, pages 111–120, Beijing, China.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57. Association for Computational Linguistics.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626, Austin, Texas, USA.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462, Vancouver, Canada.

Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2979–2985, New York, NY, USA.

Ying Zhao, George Karypis, and Usama Fayyad. 2005. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168.