An Unsupervised Neural Attention Model for Aspect Extraction

Ruidan He[†‡], **Wee Sun Lee**[†], **Hwee Tou Ng**[†], and **Daniel Dahlmeier**[‡]
[†]Department of Computer Science, National University of Singapore
[‡]SAP Innovation Center Singapore
[†]{ruidanhe,leews,nght}@comp.nus.edu.sg
[‡]d.dahlmeier@sap.com

Abstract

Aspect extraction is an important and challenging task in aspect-based sentiment analysis. Existing works tend to apply variants of topic models on this task. While fairly successful, these methods usually do not produce highly coherent aspects. In this paper, we present a novel neural approach with the aim of discovering coherent aspects. The model improves coherence by exploiting the distribution of word co-occurrences through the use of neural word embeddings. Unlike topic models which typically assume independently generated words, word embedding models encourage words that appear in similar contexts to be located close to each other in the embedding space. In addition, we use an attention mechanism to de-emphasize irrelevant words during training, further improving the coherence of aspects. Experimental results on real-life datasets demonstrate that our approach discovers more meaningful and coherent aspects, and substantially outperforms baseline methods on several evaluation tasks.

1 Introduction

Aspect extraction is one of the key tasks in sentiment analysis. It aims to extract entity aspects on which opinions have been expressed (Hu and Liu, 2004; Liu, 2012). For example, in the sentence *"The beef was tender and melted in my mouth"*, the aspect term is *"beef"*. Two sub-tasks are performed in aspect extraction: (1) extracting all aspect terms (e.g., *"beef"*) from a review corpus, (2) clustering aspect terms with similar meaning into categories where each category represents a single aspect (e.g., cluster *"beef"*, *"pork"*, *"pasta"*, and *"tomato"* into one aspect *food*).

Previous works for aspect extraction can be categorized into three approaches: rule-based, supervised, and unsupervised. Rule-based methods usually do not group extracted aspect terms into categories. Supervised learning requires data annotation and suffers from domain adaptation problems. Unsupervised methods are adopted to avoid reliance on labeled data needed for supervised learning.

In recent years, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and its variants (Titov and McDonald, 2008; Brody and Elhadad, 2010; Zhao et al., 2010; Mukherjee and Liu, 2012) have become the dominant unsupervised approach for aspect extraction. LDA models the corpus as a mixture of topics (aspects), and topics as distributions over word types. While the mixture of aspects discovered by LDA-based models may describe a corpus fairly well, we find that the individual aspects inferred are of poor quality – aspects often consist of unrelated or loosely-related concepts. This may substantially reduce users' confidence in using such automated systems. There could be two primary reasons for the poor quality. Conventional LDA models do not directly encode word co-occurrence statistics which are the primary source of information to preserve topic coherence (Mimno et al., 2011). They implicitly capture such patterns by modeling word generation from the document level, assuming that each word is generated independently. Furthermore, LDA-based models need to estimate a distribution of topics for each document. Review documents tend to be short, thus making the estimation of topic distributions more difficult.

In this work, we present a novel neural approach to tackle the weaknesses of LDA-based methods. We start with neural word embeddings that al-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 388–397
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1036

ready map words that usually co-occur within the same context to nearby points in the embedding space (Mikolov et al., 2013). We then filter the word embeddings within a sentence using an attention mechanism (Bahdanau et al., 2015) and use the filtered words to construct aspect embeddings. The training process for aspect embeddings is analogous to autoencoders, where we use dimension reduction to extract the common factors among embedded sentences and reconstruct each sentence through a linear combination of aspect embeddings. The attention mechanism de-emphasizes words that are not part of any aspect, allowing the model to focus on aspect words. We call our proposed model *Attention-based Aspect Extraction* (ABAE).

In contrast to LDA-based models, our proposed method explicitly encodes word-occurrence statistics into word embeddings, uses dimension reduction to extract the most important aspects in the review corpus, and uses an attention mechanism to remove irrelevant words to further improve coherence of the aspects.

We have conducted extensive experiments on large review data sets. The results show that ABAE is effective in discovering meaningful and coherent aspects. It substantially outperforms baseline methods on multiple evaluation tasks. In addition, ABAE is intuitive and structurally simple. It can also easily scale to a large amount of training data. Therefore, it is a promising alternative to LDA-based methods proposed previously.

2 Related Work

The problem of aspect extraction has been well studied in the past decade. Initially, methods were mainly based on manually defined rules. Hu and Liu (2004) proposed to extract different product features through finding frequent nouns and noun phrases. They also extracted opinion terms by finding the synonyms and antonyms of opinion seed words through WordNet. Following this, a number of methods have been proposed based on frequent item mining and dependency information to extract product aspects (Zhuang et al., 2006; Somasundaran and Wiebe, 2009; Qiu et al., 2011). These models heavily depend on predefined rules which work well only when the aspect terms are restricted to a small group of nouns.

Supervised learning approaches generally model aspect extraction as a standard sequence labeling problem. Jin and Ho (2009) and Li *et al.* (2010) proposed to use hidden Markov models (HMM) and conditional random fields (CRF), respectively with a set of manually-extracted features. More recently, different neural models (Yin et al., 2016; Wang et al., 2016) were proposed to automatically learn features for CRF-based aspect extraction. Rule-based models are usually not refined enough to categorize the extracted aspect terms. On the other hand, supervised learning requires large amounts of labeled data for training purposes.

Unsupervised approaches, especially topic models, have been proposed subsequently to avoid reliance on labeled data. Generally, the outputs of those models are word distributions or rankings for each aspect. Aspects are naturally obtained without separately performing extraction and categorization. Most existing works (Brody and Elhadad, 2010; Zhao et al., 2010; Mukherjee and Liu, 2012; Chen et al., 2014) are based on variants and extensions of LDA (Blei et al., 2003). Recently, Wang et al. (2015) proposed a restricted Boltzmann machine (RBM)-based model to simultaneously extract aspects and relevant sentiments of a given review sentence, treating aspects and sentiments as separate hidden variables in RBM. However, the RBM-based model proposed in (Wang et al., 2015) relies on a substantial amount of prior knowledge such as part-of-speech (POS) tagging and sentiment lexicons. A biterm topic model (BTM) that generates co-occurring word pairs was proposed in (Yan et al., 2013). We experimentally compare ABAE and BTM on multiple tasks in this paper.

Attention models (Mnih et al., 2014) have recently gained popularity in training neural networks and have been applied to various natural language processing tasks, including machine translation (Bahdanau et al., 2015; Luong et al., 2015), sentence summarization (Rush et al., 2015), sentiment classification (Chen et al., 2016; Tang et al., 2016), and question answering (Hermann et al., 2015). Rather than using all available information, attention mechanism aims to focus on the most pertinent information for a task. Unlike previous works, in this paper, we apply attention to an unsupervised neural model. Our experimental results demonstrate its effectiveness under an unsupervised setting for aspect extraction.

3 Model Description

We describe the Attention-based Aspect Extraction (ABAE) model in this section. The ultimate goal is to learn a set of aspect embeddings, where each aspect can be interpreted by looking at the nearest words (representative words) in the embedding space. We begin by associating each word w in our vocabulary with a feature vector $\mathbf{e}_w \in \mathbb{R}^d$. We use word embeddings for the feature vectors as word embeddings are designed to map words that often co-occur in a context to points that are close by in the embedding space (Mikolov et al., 2013). The feature vectors associated with the words correspond to the rows of a word embedding matrix $\mathbf{E} \in \mathbb{R}^{V \times d}$, where V is the vocabulary size. We want to learn embeddings of aspects, where aspects share the same embedding space with words. This requires an aspect embedding matrix $\mathbf{T} \in \mathbb{R}^{K \times d}$, where K, the number of aspects defined, is much smaller than V. The aspect embeddings are used to approximate the aspect words in the vocabulary, where the aspect words are filtered through an attention mechanism.

如何建立Asepct向量空间
Aspect向量矩阵怎么来的？？

Each input sample to ABAE is a list of indexes for words in a review sentence. Given such an input, two steps are performed as shown in Figure 1. First, we filter away non-aspect words by down-weighting them using an attention mechanism, and construct a sentence embedding $\mathbf{z}_s$ from weighted word embeddings. Then, we try to reconstruct the sentence embedding as a linear combination of aspect embeddings from $\mathbf{T}$. This process of dimension reduction and reconstruction, where ABAE aims to transform sentence embeddings of the filtered sentences ($\mathbf{z}_s$) into their reconstructions ($\mathbf{r}_s$) with the least possible amount of distortion, preserves most of the information of the aspect words in the K embedded aspects. We next describe the process in detail.

3.1 Sentence Embedding with Attention Mechanism

We construct a vector representation $\mathbf{z}_s$ for each input sentence s in the first step. In general, we want the vector representation to capture the most relevant information with regards to the aspect (topic) of the sentence. We define the sentence embedding $\mathbf{z}_s$ as the weighted summation of word embeddings $\mathbf{e}_{w_i}$, $i = 1, ..., n$ corresponding to the word indexes in the sentence.

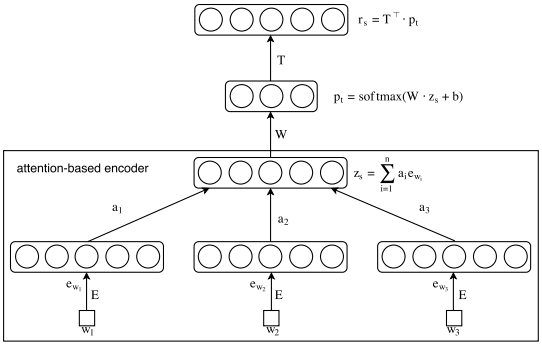

Figure 1: An example of the ABAE structure.

$$\mathbf{z}_s = \sum_{i=1}^{n} a_i \mathbf{e}_{w_i}. \quad (1)$$

For each word w_i in the sentence, we compute a positive weight a_i which can be interpreted as the probability that w_i is the right word to focus on in order to capture the main topic of the sentence. The weight a_i is computed by an attention model, which is conditioned on the embedding of the word $\mathbf{e}_{w_i}$ as well as the global context of the sentence:

$$a_i = \frac{\exp(d_i)}{\sum_{j=1}^{n} \exp(d_j)} \quad (2)$$

$$d_i = \mathbf{e}_{w_i}^{\top} \cdot \mathbf{M} \cdot \mathbf{y}_s \quad (3)$$

$$\mathbf{y}_s = \frac{1}{n} \sum_{i=1}^{n} \mathbf{e}_{w_i} \quad (4)$$

where $\mathbf{y}_s$ is simply the average of the word embeddings, which we believe captures the global context of the sentence. $\mathbf{M} \in \mathbb{R}^{d \times d}$ is a matrix mapping between the global context embedding $\mathbf{y}_s$ and the word embedding $\mathbf{e}_w$ and is learned as part of the training process. We can think of the attention mechanism as a two-step process. Given a sentence, we first construct its representation by averaging all the word representations. Then the weight of a word is assigned by considering two things. First, we filter the word through the transformation $\mathbf{M}$ which is able to capture the relevance of the word to the K aspects. Then we capture the relevance of the filtered word to the sentence by taking the inner product of the filtered word to the global context $\mathbf{y}_s$.

3.2 Sentence Reconstruction with Aspect Embeddings

We have obtained the sentence embedding. Now we describe how to compute the reconstruction of the sentence embedding. As shown in Figure 1, the reconstruction process consists of two steps of transitions, which is similar to an autoencoder. Intuitively, we can think of the reconstruction as a linear combination of aspect embeddings from $\mathbf{T}$:

$$\mathbf{r}_s = \mathbf{T}^\top \cdot \mathbf{p}_t \tag{5}$$

where $\mathbf{r}_s$ is the reconstructed vector representation, $\mathbf{p}_t$ is the weight vector over K aspect embeddings, where each weight represents the probability that the input sentence belongs to the related aspect. $\mathbf{p}_t$ can simply be obtained by reducing $\mathbf{z}_s$ from d dimensions to K dimensions and then applying a softmax non-linearity that yields normalized non-negative weights:

$$\mathbf{p}_t = softmax(\mathbf{W} \cdot \mathbf{z}_s + \mathbf{b}) \tag{6}$$

where $\mathbf{W}$, the weighted matrix parameter, and $\mathbf{b}$, the bias vector, are learned as part of the training process.

3.3 Training Objective

ABAE is trained to minimize the reconstruction error. We adopted the contrastive max-margin objective function used in previous work (Weston et al., 2011; Socher et al., 2014; Iyyer et al., 2016). For each input sentence, we randomly sample m sentences from our training data as negative samples. We represent each negative sample as $\mathbf{n}_i$ which is computed by averaging its word embeddings. Our objective is to make the reconstructed embedding $\mathbf{r}_s$ similar to the target sentence embedding $\mathbf{z}_s$ while different from those negative samples. Therefore, the unregularized objective J is formulated as a hinge loss that maximize the inner product between $\mathbf{r}_s$ and $\mathbf{z}_s$ and simultaneously minimize the inner product between $\mathbf{r}_s$ and the negative samples:

$$J(\theta) = \sum_{s \in D} \sum_{i=1}^{m} \max(0, 1 - \mathbf{r}_s \mathbf{z}_s + \mathbf{r}_s \mathbf{n}_i) \tag{7}$$

where D represents the training data set and $\theta = \{\mathbf{E}, \mathbf{T}, \mathbf{M}, \mathbf{W}, \mathbf{b}\}$ represents the model parameters.

Domain	#Reviews	#Labeled sentences
Restaurant	52,574	3,400
Beer	1,586,259	9,245

Table 1: Dataset description.

3.4 Regularization Term

We hope to learn vector representations of the most representative aspects for a review dataset. However, the aspect embedding matrix $\mathbf{T}$ may suffer from redundancy problems during training. To ensure the diversity of the resulting aspect embeddings, we add a regularization term to the objective function J to encourage the uniqueness of each aspect embedding:

$$U(\theta) = \|\mathbf{T}_n \cdot \mathbf{T}_n^\top - \mathbf{I}\| \tag{8}$$

where $\mathbf{I}$ is the identity matrix, and $\mathbf{T}_n$ is $\mathbf{T}$ with each row normalized to have length 1. Any non-diagonal element $t_{ij} (i \neq j)$ in the matrix $\mathbf{T}_n \cdot \mathbf{T}_n^\top$ corresponds to the dot product of two different aspect embeddings. U reaches its minimum value when the dot product between any two different aspect embeddings is zero. Thus the regularization term encourages orthogonality among the rows of the aspect embedding matrix $\mathbf{T}$ and penalizes redundancy between different aspect vectors. Our final objective function L is obtained by adding J and U:

$$L(\theta) = J(\theta) + \lambda U(\theta) \tag{9}$$

where λ is a hyperparameter that controls the weight of the regularization term.

4 Experimental Setup

4.1 Datasets

We evaluate our method on two real-word datasets. The detailed statistics of the datasets are summarized in Table 1.

(1) **Citysearch corpus**: This is a restaurant review corpus widely used by previous works (Ganu et al., 2009; Brody and Elhadad, 2010; Zhao et al., 2010), which contains over 50,000 restaurant reviews from Citysearch New York. Ganu et al. (2009) also provided a subset of 3,400 sentences from the corpus with manually labeled aspects. These annotated sentences are used for evaluation of aspect identification. There are six manually defined aspect labels: *Food, Staff, Ambience, Price, Anecdotes,* and *Miscellaneous*.

(2) **BeerAdvocate**: This is a beer review corpus introduced in (McAuley et al., 2012), containing over 1.5 million reviews. A subset of 1,000 reviews, corresponding to 9,245 sentences, are annotated with five aspect labels: *Feel, Look, Smell, Taste,* and *Overall*.

4.2 Baseline Methods

To validate the performance of ABAE, we compare it against a number of baselines:

(1) **LocLDA** (Brody and Elhadad, 2010): This method uses a standard implementation of LDA. In order to prevent the inference of global topics and direct the model towards rateable aspects, each sentence is treated as a separate document.

(2) **k-means**: We initialize the aspect matrix T by using the k-means centroids of the word embeddings. To show the power of ABAE, we compare its performance with using the k-means centroids directly.

(3) **SAS** (Mukherjee and Liu, 2012): This is a hybrid topic model that jointly discovers both aspects and aspect-specific opinions. This model has been shown to be competitive among topic models in discovering meaningful aspects (Mukherjee and Liu, 2012; Wang et al., 2015).

(4) **BTM** (Yan et al., 2013): This is a biterm topic model that is specially designed for short texts such as texts from social media and review sites. The major advantage of BTM over conventional LDA models is that it alleviates the problem of data sparsity in short documents by directly modeling the generation of unordered word-pair co-occurrences (biterms) over the corpus. It has been shown to perform better than conventional LDA models in discovering coherent topics.

4.3 Experimental Settings

Review corpora are preprocessed by removing punctuation symbols, stop words, and words appearing less than 10 times. For LocLDA, we use the open-source implementation GibbsLDA++[1] and for BTM, we use the implementation released by (Yan et al., 2013)[2]. We tune the hyperparameters of all topic model baselines on a held-out set with grid search using the topic coherence metric to be introduced later in Eq 10: for LocLDA, the Dirichlet priors $\alpha = 0.05$ and $\beta = 0.1$; for SAS and BTM, $\alpha = 50/K$ and $\beta = 0.1$. We run 1,000 iterations of Gibbs sampling for all topic models.

For the ABAE model, we initialize the word embedding matrix $\mathbf{E}$ with word vectors trained by word2vec with negative sampling on each dataset, setting the embedding size to 200, window size to 10, and negative sample size to 5. The parameters we use for training word embeddings are standard with no specific tuning to our data. We also initialize the aspect embedding matrix $\mathbf{T}$ with the centroids of clusters resulting from running k-means on word embeddings. Other parameters are initialized randomly. During the training process, we fix the word embedding matrix $\mathbf{E}$ and optimize other parameters using Adam (Kingma and Ba, 2014) with learning rate 0.001 for 15 epochs and batch size of 50. We set the number of negative samples per input sample m to 20, and the orthogonality penalty weight λ to 1 by tuning the hyperparameters on a held-out set with grid search. The results reported for all models are the average over 10 runs.

Following (Brody and Elhadad, 2010; Zhao et al., 2010), we set the number of aspects for the restaurant corpus to 14. We experimented with different number of aspects from 10 to 20 for the beer corpus. The results showed no major difference, so we also set it to 14. As in previous work (Brody and Elhadad, 2010; Zhao et al., 2010), we manually mapped each inferred aspect to one of the gold-standard aspects according to its top ranked representative words. In ABAE, representative words of an aspect can be found by looking at its nearest words in the embedding space using cosine as the similarity metric.

5 Evaluation and Results

We describe the evaluation tasks and report the experimental results in this section. We evaluate ABAE on two criteria:

- Is it able to find meaningful and semantically coherent aspects?
- Is it able to improve aspect identification performance on real-world review datasets?

5.1 Aspect Quality Evaluation

Table 2 presents all 14 aspects inferred by ABAE for the restaurant domain. Compared to gold-

[1] http://gibbslda.sourceforge.net
[2] http://code.google.com/p/btm/

Inferred Aspects	Representative Words	Gold Aspects
Main Dishes	beef, duck, pork, mahi, filet, veal	Food
Dessert	gelato, banana, caramel, cheesecake, pudding, vanilla	
Drink	bottle, selection, cocktail, beverage, pinot, sangria	
Ingredient	cucumber, scallion, smothered, stewed, chilli, cheddar	
General	cooking, homestyle, traditional, cuisine, authentic, freshness	
Physical Ambience	wall, lighting, ceiling, wood, lounge, floor	Ambience
Adjectives	intimate, comfy, spacious, modern, relaxing, chic	
Staff	waitstaff, server, staff, waitress, bartender, waiter	Staff
Service	unprofessional, response, condescending, aggressive, behavior, rudeness	
Price	charge, paid, bill, reservation, came, dollar	Price
Anecdotes	celebrate, anniversary, wife, fiance, recently, wedding	Anecdotes
Location	park, street, village, avenue, manhattan, brooklyn	Misc.
General	excellent, great, enjoyed, best, wonderful, fantastic	
Other	aged, reward, white, maison, mediocrity, principle	

Table 2: List of inferred aspects for restaurant reviews (left), with top representative words for each inferred aspect (middle), and the corresponding gold-standard aspect labels (right). Inferred aspect labels (left) were assigned manually.

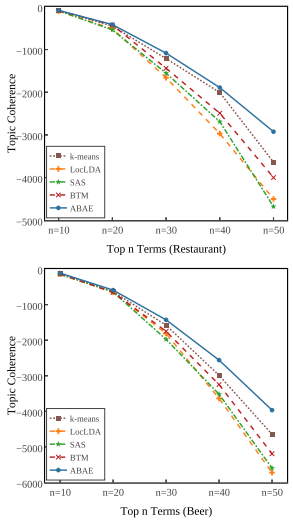

Figure 2: Average coherence score versus number of top n terms for the restaurant domain (top) and beer domain (bottom).

standard labels, the inferred aspects are more fine-grained. For example, it can distinguish main dishes from desserts, and drinks from food.

5.1.1 Coherence Score

In order to objectively measure the quality of aspects, we use *coherence score* as a metric which has been shown to correlate well with human judgment (Mimno et al., 2011). Given an aspect z and a set of top N words of z, $S^z = \{w_1^z, ..., w_N^z\}$, the coherence score is calculated as follows:

$$C(z; S^z) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} log \frac{D_2(w_n^z, w_l^z) + 1}{D_1(w_l^z)} \quad (10)$$

where $D_1(w)$ is the document frequency of word w and $D_2(w_1, w_2)$ is the co-document frequency of words w_1 and w_2. A higher coherence score indicates a better aspect interpretability, i.e., more meaningful and semantically coherent.

Figure 2 shows the average *coherence score* of each model which is computed as $\frac{1}{K}\sum_{k=1}^{K} C(z_k; S^{z_k})$ on both the restaurant domain and beer domain. From the results, we make the following observations: (1) ABAE outperforms previous models for all ranked buckets. (2) BTM performs slightly better than LocLDA and SAS. This may be because BTM directly models the generation of biterms, while conventional LDA just implicitly captures such patterns by modeling word generation from the document level. (3) It is interesting to note that performing k-means on the word embeddings is sufficient to perform better than all topic model baselines, including BTM. This indicates that neural word embedding is a better model for capturing co-occurrence than LDA, even for BTM which specifically models the generation of co-occurring word pairs.

	k-means	LocLDA	SAS	BTM	ABAE
Restaurant	11	8	9	9	11
Beer	9	8	8	9	10

Table 3: Number of coherent aspects. K (number of aspects) = 14 for all models.

5.1.2 User Evaluation

As we want to discover a set of aspects that the human user finds agreeable, it is also necessary

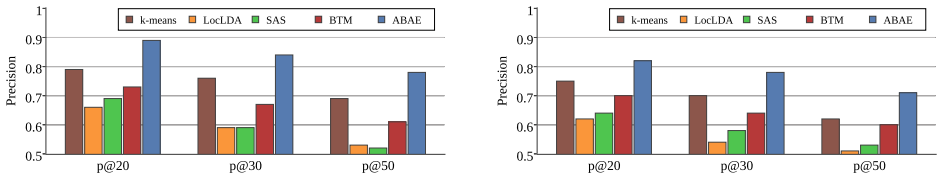

Figure 3: Average $p@n$ over all coherent aspects for the restaurant domain (left) and beer domain (right).

to carry out user evaluation directly. Following the experimental setting in (Chen et al., 2014), we recruited three human judges. Each aspect is labeled as coherent if the majority of judges assess that most of its top 50 terms coherently represent a product aspect. The numbers of coherent aspects discovered by each model are shown in Table 3. ABAE discovers the most number of coherent aspects compared with other models.

For a coherent aspect, each of its top terms is labeled as correct if and only if the majority of judges assess that it reflects the related aspect. We adopt *precision@n* (or $p@n$) to evaluate the results, which was also used in (Mukherjee and Liu, 2012; Chen et al., 2014). Figure 3 shows the average $p@n$ results over all coherent aspects for each domain. We can see that the user evaluation results correlate well with the coherence scores shown in Figure 2, where ABAE substantially outperforms all other models for all ranked buckets, especially for large values of n.

5.2 Aspect Identification

We evaluate the performance of sentence-level aspect identification on both domains using the annotated sentences shown in Table 1. The evaluation criterion is to judge how well the predictions match the true labels, measured by precision, recall, and F_1 scores. The results[4] are shown in Table 4 and Table 5.

Given a review sentence, ABAE first assigns an inferred aspect label which corresponds to the highest weight in $\mathbf{p}_t$ calculated as shown in Equation 6 . And we then assign the gold-standard label to the sentence according to the mapping between inferred aspects and gold-standard labels.

[3] k-means assigns a sentence an inferred aspect whose embedding is the closest to the averaged word embeddings of the sentence.

[4] Note that the values of P/R/F_1 reported are the average over 10 runs (except some values taken from published results in Table 4). Thus the F_1 values cannot be computed directly from corresponding P/R values

Aspect	Method	Precision	Recall	F_1
Food	LocLDA	0.898	0.648	0.753
	ME-LDA	0.874	0.787	0.828
	SAS	0.867	0.772	0.817
	BTM	0.933	0.745	0.816
	SERBM	0.891	**0.854**	**0.872**
	k-means[3]	0.931	0.647	0.755
	ABAE	**0.953**	0.741	0.828
Staff	LocLDA	0.804	0.585	0.677
	ME-LDA	0.779	0.540	0.638
	SAS	0.774	0.556	0.647
	BTM	**0.828**	0.579	0.677
	SERBM	0.819	0.582	0.680
	k-means	0.789	0.685	0.659
	ABAE	0.802	**0.728**	**0.757**
Ambience	LocLDA	0.603	0.677	0.638
	ME-LDA	0.773	0.558	0.648
	SAS	0.780	0.542	0.640
	BTM	0.813	0.599	0.685
	SERBM	0.805	0.592	0.682
	k-means	0.730	0.637	0.677
	ABAE	**0.815**	**0.698**	**0.740**

Table 4: Aspect identification results on the restaurant domain. The results of LocLDA and ME-LDA are taken from (Zhao et al., 2010); the results of SAS and SERBM are taken from (Wang et al., 2015).

For the restaurant domain, we follow the experimental settings of previous work (Brody and Elhadad, 2010; Zhao et al., 2010; Wang et al., 2015) to make our results comparable. To do that, (1) we only used the single-label sentences for evaluation to avoid ambiguity (about 83% of labeled sentences have a single label), and (2) we only evaluated on three major aspects, namely *Food*, *Staff*, and *Ambience*. The other aspects do not show clear patterns in either word usage or writing style, which makes these aspects very hard for even humans to identify. Besides the baseline models, we also compare the results with other published models, including MaxEnt-LDA (ME-LDA) (Zhao et al., 2010) and SERBM (Wang et al., 2015). SERBM has reported state-of-the-art results for aspect identification on the restaurant corpus to date. However, SERBM relies on a substantial amount of prior knowledge.

Aspect	Method	Precision	Recall	F_1
Feel	k-means	0.720	0.815	0.737
	LocLDA	**0.938**	0.537	0.675
	SAS	0.783	0.695	0.730
	BTM	0.892	0.687	0.772
	ABAE	0.815	**0.824**	**0.816**
Taste	k-means	0.533	0.413	0.456
	LocLDA	0.399	**0.655**	0.487
	SAS	0.543	0.496	0.505
	BTM	0.616	0.467	**0.527**
	ABAE	**0.637**	0.358	0.456
Smell	k-means	**0.844**	0.295	0.422
	LocLDA	0.560	0.488	0.489
	SAS	0.336	0.673	0.404
	BTM	0.541	0.549	0.527
	ABAE	0.483	**0.744**	**0.575**
Taste+Smell	k-means	0.697	0.828	0.740
	LocLDA	0.651	**0.873**	0.735
	SAS	0.804	0.759	0.769
	BTM	0.885	0.760	0.815
	ABAE	**0.897**	0.853	**0.866**
Look	k-means	0.915	0.696	0.765
	LocLDA	0.963	0.676	0.774
	SAS	0.958	0.705	0.806
	BTM	0.953	0.854	0.872
	ABAE	**0.969**	**0.882**	**0.905**
Overall	k-means	0.693	0.648	0.639
	LocLDA	0.558	0.690	0.603
	SAS	0.618	0.664	0.619
	BTM	**0.699**	0.715	0.700
	ABAE	0.654	**0.828**	**0.725**

Table 5: Aspect identification results on the beer domain.

We make the following observations from Table 4: (1) ABAE outperforms all other models on F_1 score for aspects *Staff* and *Ambience*. (2) The F_1 score of ABAE for *Food* is worse than SERBM while its precision is very high. We analyzed the errors and found that most of the sentences we failed to recognize as *Food* are general descriptions without specific food words appearing. For example, the true label for the sentence "*The food is prepared quickly and efficiently.*" is *Food*. ABAE assigns *Staff* to it as the highly focused words according to the attention mechanism are *quickly* and *efficiently* which are more related to *Staff*. In fact, although this sentence contains the word *food*, we think it is a rather general description of service. (3) ABAE substantially outperforms k-means for this task although both methods perform well for extracting coherent aspects as shown in Figure 2 and Figure 3. This shows the power brought by the attention mechanism, which is able to capture the main topic of a sentence by only focusing on aspect-related words.

For the beer domain, in addition to the five gold-standard aspect labels, we also combined *Taste* and *Smell* to form a single aspect – *Taste+Smell*. This is because these two aspects are very similar

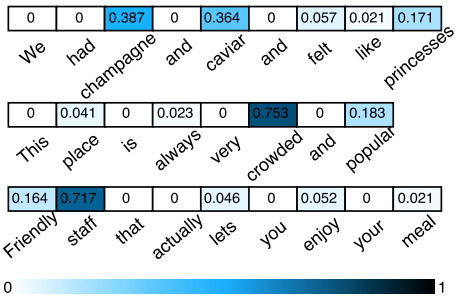

Figure 4: Visualization of the attention layer.

and many words can be used to describe both aspects. For example, the words *spicy*, *bitter*, *fresh*, *sweet*, etc. are top ranked representative words in both aspects, which makes it very hard even for humans to distinguish them. Since *Taste* and *Smell* are highly correlated and difficult to separate in real life, a natural way to evaluate is to treat them as a single aspect.

We can see from Table 5 that due to the issue described above, all models perform poorly on *Taste* and *Smell*. ABAE outperforms previous models in F_1 scores on all aspects except for *Taste*. The results demonstrate the capability of ABAE in identifying separable aspects.

Aspect	Method	Precision	Recall	F_1
Food	ABAE$^-$	0.898	0.739	0.791
	ABAE	**0.953**	**0.741**	**0.828**
Staff	ABAE$^-$	0.784	0.669	0.693
	ABAE	**0.802**	**0.728**	**0.757**
Ambience	ABAE$^-$	0.782	0.660	0.703
	ABAE	**0.815**	**0.698**	**0.740**

Table 6: Comparison between ABAE and ABAE$^-$ on aspect identification on the restaurant domain.

5.3 Validating the Effectiveness of Attention Model

Figure 4 shows the weights of words assigned by the attention model for some example sentences. As we can see, the weights learned by the model correspond very strongly with human intuition. In order to evaluate how attention model affects the overall performance of ABAE, we conduct experiments to compare ABAE and ABAE$^-$ on aspect identification, where ABAE$^-$ denotes the model in which the attention layer is switched off and sentence embedding is calculated by averaging its word embeddings: $z_s = \frac{1}{n}\sum_{i=1}^{n} e_{w_i}$. The results on the restaurant domain are shown in Table 6. ABAE achieves substantially higher precision and recall on all aspects compared with

ABAE$^-$, which demonstrates the effectiveness of the attention mechanism.

6 Conclusion

We have presented ABAE, a simple yet effective neural attention model for aspect extraction. In contrast to LDA models, ABAE explicitly captures word co-occurrence patterns and overcomes the problem of data sparsity present in review corpora. Our experimental results demonstrated that ABAE not only learns substantially higher quality aspects, but also more effectively captures the aspects of reviews than previous methods. To the best of our knowledge, we are the first to propose an unsupervised neural approach for aspect extraction. ABAE is intuitive and structurally simple, and also scales up well. All these benefits make it a promising alternative to LDA-based methods in practice.

Acknowledgements

This research is partially funded by the Economic Development Board and the National Research Foundation of Singapore.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Zhiyuan Chen, Arjun Mukherjee, and Bing Liu. 2014. Aspect extraction with automated prior knowledge learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *Proceedings of the 12th International Workshop on the Web and Databases*.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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

Wei Jin and Hung Hay Ho. 2009. A novel lexicalized HMM-based learning framework for web opinion mining. In *Proceedings of the 26th International Conference on Machine Learning*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 2nd International Conference on Learning Representations*.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. 2010. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

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

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *Proceedings of the 12th IEEE International Conference on Data Mining*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics* 37:9–27.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International World Wide Web Conference*.

Linlin Wang, Kang Liu, Zhu Cao, Jun Zhao, and Gerard de Melo. 2015. Sentiment-aspect extraction based on restricted Boltzmann machines. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

Wenya Wang, Sinno J. 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*.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd International World Wide Web Conference*.

Yichun Yin, Furu Wei, Li Dong, Kaiming 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*.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*.