Aspect-Based Sentiment Analysis by Aspect-Sentiment Joint Embedding

Anonymous EMNLP submission

Abstract

Aspect-based sentiment analysis of review texts is of great value for understanding user feedback in a fine-grained manner. It has in general two sub-tasks: (i) extracting aspects from each review, and (ii) classifying aspect-based reviews by sentiment polarity. In this paper, we propose a weakly-supervised approach for aspect-based sentiment analysis, which uses only a few keywords describing each aspect/sentiment without using any labeled examples. Existing methods are either designed only for one of the sub-tasks, neglecting the benefit of coupling both, or are based on topic models that may contain overlapping concepts. We propose to first learn ⟨sentiment, aspect⟩ joint topic embeddings in the word embedding space by imposing regularizations to encourage topic distinctiveness, and then use neural models to generalize the word-level discriminative information by pre-training the classifiers with embedding-based predictions and self-training them on unlabeled data. Our comprehensive performance analysis shows that our method generates quality joint topics and outperforms the baselines significantly (9.9 and 3.2 F1-score gain on average for aspect and sentiment classification respectively) on benchmark datasets[1].

1 Introduction

With the vast amount of reviews emerging on platforms like Amazon and Yelp, aspect-based sentiment analysis, which extracts opinions about certain facets of entities from text, becomes increasingly essential and benefits a wide range of downstream applications (Bauman et al., 2017; Nguyen et al., 2015).

Aspect-based sentiment analysis contains two sub-tasks: Aspect extraction and sentiment polarity classification. The former identifies the aspect covered in the review, whereas the latter decides its sentiment polarity.

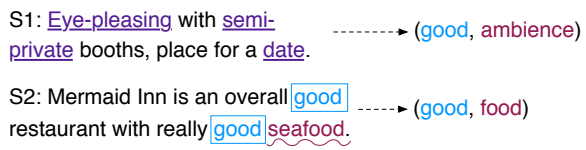

Figure 1: Two sample restaurant reviews. Pure aspect words are in red and wavy-underlined, and general opinion words are in blue and framed in boxes. Words implying both aspects and opinions (which we define as joint topics) are underlined and in purple.

We show two sample restaurant reviews in Fig. 1 together with their expected outputs—aspect and sentiment labels. Neural network models (Liu et al., 2015; Xu et al., 2018) have outperformed rule-based models (Hu and Liu, 2004; Zhuang et al., 2006), but they require large-scale fine-grained labeled data to train, which can be difficult to obtain. Some other studies leverage word embeddings to solve the aspect extraction problem in an unsupervised (He et al., 2017; Liao et al., 2019) or weakly-supervised setting (Angelidis and Lapata, 2018; Karamanolakis et al., 2019), *without* using any annotated documents. In this work, we also study the weakly-supervised setting, where only a few keywords are provided for each aspect and sentiment.

With a closer look at the two example reviews in Fig. 1, we observe that S2 includes a general opinion word "good" and a pure aspect word "seafood", which are separate hints for sentiment and aspect classification respectively. S1, on the other hand, does not address the target with plain and general words, but instead use more specific words like "semi-private" and "date" which are uniquely used when people feel good about the ambience instead of other aspects. Humans can interpret these unique and fine-grained terms as hints for a joint topic of ⟨good, ambience⟩, but this is hard for models that are solely trained for one sub-task. If a model

[1]Our model is released at https://github.com/emnlp2020-1675/joint-sentiment-and-aspect

can automatically learn the semantics of each joint topic of ⟨sentiment, aspect⟩, it will be able to identify representative terms of the joint topics such as "semi-private" which provide information for aspect and sentiment simultaneously, and will consequently benefit both aspect extraction and sentiment classification. Therefore, leveraging more fine-grained information by coupling the two subtasks will enhance both.

Several LDA-based methods consider learning joint topics (Zhao et al., 2010; Wang et al., 2015; Xu et al., 2012), but they rely on external resources such as part-of-speech (POS) tagging or opinion word lexicons. A recent LDA-based model (García-Pablos et al., 2018) uses pre-trained word embedding to bias the prior in topic models to jointly model aspect words and opinion words. Though working fairly well, topic models are generative models and do not enforce topic distinctiveness—topic-word distribution can largely overlap among different topics, allowing topics to resemble each other. Besides, topic models yield unstable results, causing large variance in classification results.

Our general idea is to learn a joint topic representation for each ⟨sentiment, aspect⟩ pair in the shared embedding space with words so that the surrounding words of topic embeddings nicely describe the semantics of a joint topic. This is accomplished by training topic embeddings and word embeddings on in-domain corpora and modeling the joint distribution of user-given keywords on all the joint topics. After learning the joint topic vectors, embedding-based predictions can be derived for any unlabeled review. However, these predictions are sub-optimal for sentiment analysis where word order plays an important role. To leverage the expressive power of neural models, we distill the knowledge from embedding-based predictions to convolutional neural networks (CNNs) (Krizhevsky et al., 2012) which perform compositional operations upon local sequences. A self-training process is then conducted to refine CNNs by using their high-confident predictions on unlabeled data.

We demonstrate the effectiveness of our model by conducting experiments on two benchmark datasets and show that our model outperforms all the baseline methods by a large margin. We also show that our model is able to describe joint topics with coherent term clusters.

Our contributions can be summarized as follows: (1) We propose a weakly-supervised method to enhance two sub-tasks of aspect-based sentiment analysis. Our method does *not* need any annotated data but only a few keywords for each aspect/sentiment. (2) We introduce an embedding learning objective that is able to capture the semantics of fine-grained joint topics of ⟨sentiment, aspect⟩ in the word embedding space. The embedding-based prediction is effectively leveraged by neural models to generalize on unlabeled data via self-training. (3) We demonstrate that our model generates high-quality joint topics and outperforms baselines significantly on two benchmark datasets.

2 Related Work

The problem of aspect-based sentiment analysis can be decomposed into two sub-tasks: aspect extraction and sentiment polarity classification. Most previous studies deal with them individually. There are various related efforts on aspect extraction (He et al., 2017), which can be followed by sentiment classification models (He et al., 2018). Other methods (García-Pablos et al., 2018) jointly solve these two sub-tasks by first separating target words from opinion words and then learning joint topic distributions over words. Below we first review relevant work on aspect extraction (Sec 2.1) and then turn to studies that jointly extract aspects and sentiment polarity (Sec 2.2).

2.1 Aspect Extraction

Early studies towards aspect extraction are mainly based on manually defined rules (Hu and Liu, 2004; Zhuang et al., 2006), which have been outperformed by supervised neural approaches that do not need labor-intensive feature engineering. While CNN (Xu et al., 2018) and RNN (Liu et al., 2015) based models have shown the powerful expressiveness of neural models, they can easily consume thousands of labeled documents thus suffer from the label scarcity bottleneck.

Various unsupervised approaches are proposed to model different aspects automatically. LDA-based methods (Brody and Elhadad, 2010; Chen et al., 2014) model each document as a mixture of aspects (topics) and output a ranking list or word distribution for each aspect. Recently, neural models have shown to extract more coherent topics. ABAE (He et al., 2017) uses an autoencoder to reconstruct sentences through aspect embedding and removes irrelevant words through attention mechanisms. LCC+GBC (Liao et al., 2019) incorpo-

rates sentence-level context along with local context to extract aspect words from the same category. CAt (Tulkens and van Cranenburgh, 2020) introduces a single head attention calculated by a Radio Basis Function (RBF) kernel to be the sentence summary. The unsupervised nature of these algorithms is hindered by the fact that the learned aspects often do not well align with user's interested aspects, and additional human effort is needed to map topics to certain aspects, not to mention some topics are irrelevant of interested aspects.

Several weakly-supervised methods address this problem by using a few keywords per aspect as supervision to guide the learning process. MATE (Angelidis and Lapata, 2018) extends ABAE by initializing aspect embedding using weighted average of keyword embeddings from each aspect. ISWD (Karamanolakis et al., 2019) co-trains a bag-of-word classifier and an embedding-based neural classifier to generalize the keyword supervision.

The above methods focus on the task of aspect extraction and thus do not take aspect-specific opinion words into consideration. However, the semantic meaning captured by a ⟨sentiment, aspect⟩ joint topic preserves more fine-grained information to imply the aspect of a sentence and thus can be used to improve the performance of aspect extraction.

2.2 Joint Extraction of Aspect and Sentiment

Most previous studies that jointly perform aspect and sentiment extraction are LDA-based methods. Zhao et al. (2010) include aspect-specific opinion models along with aspect models in the generative process. Wang et al. (2015) propose a restricted Boltzmann machine-based model that treats aspect and sentiment as heterogeneous hidden units. Xu et al. (2012) adapt LDA by introducing sentiment-related variables and integrating sentiment prior knowledge. All these methods rely on external resources such as part-of-speech (POS) tagging or opinion word lexicons. A more recent study that shares similar weakly-supervised setting with ours is W2VLDA (García-Pablos et al., 2018). They apply Brown clustering (Brown et al., 1992) to separate aspect-terms from opinion-terms and construct biased hyperparameters α and β by embedding similarity. In our results we show that our joint topics not only capture aspect-specific opinion words, but also other relevant terms like "vomit" for (bad, food) and "chaos" for (bad, ambience), which are critical for aspect-based sentiment analysis. Despite the effectiveness of topic models, they suffer from the drawback of not imposing discriminative constraints among topics—topic-word distribution can largely overlap among different topics, allowing redundant topics to appear and making it hard to classify them. We empirically show the advances of our method by capturing discriminative joint topic representations in the embedding space.

3 Problem Definition

Our weakly-supervised aspect-based sentiment analysis task is defined as follows. The input is a training corpus $\mathcal{D} = \{d_1, d_2, \ldots, d_{|\mathcal{D}|}\}$ of text reviews from a certain domain (e.g., restaurant or laptop) *without* any label for aspects or sentiment. A list of keywords l_a for each aspect topic (denoted as $a \in A$) and l_s for each sentiment polarity (denoted as $s \in S$) are provided by users as guidance. For each unseen review in the same domain, our model outputs a set of $\langle s, a \rangle$ labels.

4 Model

Figure 2 shows the workflow of our model. Our goal is to generate a set of ⟨sentiment, aspect⟩ predictions for each review.

We first learn an embedding space to explicitly represent the semantics of the topics (including both pure aspect/sentiment and joint ⟨sentiment, aspect⟩ ones) as embedding vectors, which are surrounded by the embeddings of the representative words of the topics. We also impose discriminative regularization on the embedding space to push different topics apart. To model the local sequential information which is crucial for sentiment analysis, we use CNN as the classifier by pre-training it on pseudo labels given by the cosine similarity between document embeddings and topic embeddings, and self-training it on unlabeled data to iteratively refine its parameters. Below we introduce the details of our model.

4.1 Joint-Topic Representation Learning

We learn the representations of words and topics on the in-domain corpus by following two principles: (1) distributional hypothesis (Sahlgren, 2008) and (2) topic distinctiveness. The first principle is achieved by an adaptation of the Skip-Gram model (Mikolov et al., 2013) through modeling both local and global contexts of words, and the second is achieved by a series of regularization

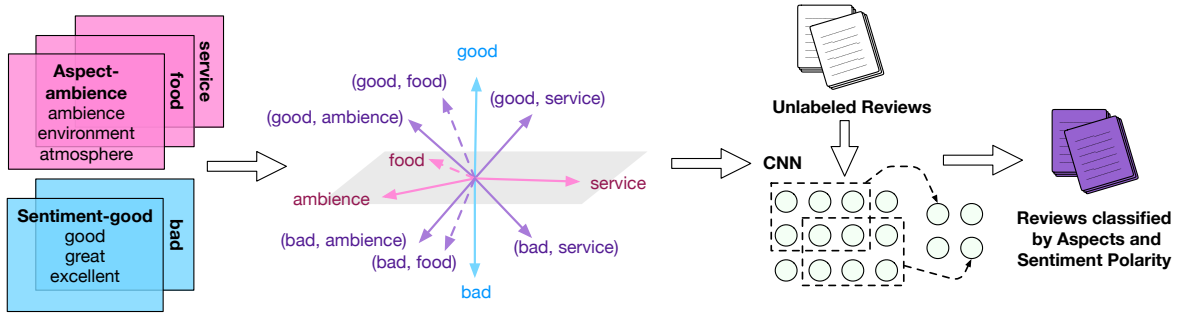

Figure 2: Overview of our model. We first leverage the in-domain training corpus and user-given keywords to learn joint topic representation in the word embedding space. The marginal probability of keywords belonging to an aspect/sentiment can be summed up by the joint distribtution over ⟨sentiment, aspect⟩ joint topics. Embedding-based prediction on unlabeled data are then leveraged by neural models for pre-training and self-training.

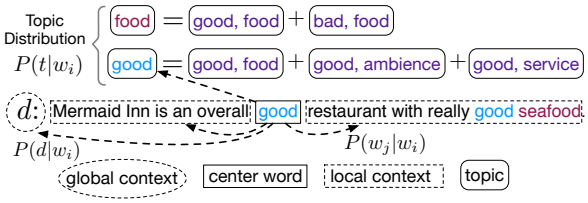

Figure 3: Embedding training.

objectives. Fig. 3 provides the overview of our embedding learning objectives with an example.

Modeling Local and Global Contexts. We learn word embeddings based on the assumption that words with similar contexts have similar meanings, and define contexts to be a combination of location contexts (Mikolov et al., 2013) and global contexts (Meng et al., 2019; Liao et al., 2019). The local context of a word w_i refers to other words whose distances are h words or less from w_i. To maximize the probability of seeing the local context of a word w_i, we use the following objective:

$$\mathcal{L}_l = -\sum_{w_i} \sum_{0<|j-i|\leq h} \log P(w_j|w_i), \quad (1)$$

where $P(w_j|w_i) \propto \exp(\boldsymbol{v}_j^\top \boldsymbol{u}_i)$, and $\boldsymbol{u}_i, \boldsymbol{v}_j$ are the center and context word embeddings.

The global context (Meng et al., 2019; Liao et al., 2019) of a word w_i refers to the document d where a word appears, based on the motivation that similar documents contain similar-meaning words. We use the following objective for global context:

$$\mathcal{L}_g = -\sum_{d\in\mathcal{D}} \sum_{w_i\in d} \log P(d|w_i), \quad (2)$$

where $P(d|w_i) \propto \exp(\boldsymbol{d}^\top \boldsymbol{u}_i)$.

Regularizing Pure Aspect/Sentiment Topics. To endow the embedding space with discriminative power over the aspect/sentiment categories for better classification performance, we regularize the aspect topic embeddings $\boldsymbol{t}_a$ and sentiment topic embeddings $\boldsymbol{t}_s$ so that different topics are pushed apart. For example, the word "good" in Fig. 3 is a keyword for the sentiment topic *good*, and we aim to place $\boldsymbol{t}_{\text{good}}$ close to the word embedding of "good" in the embedding space while away from other topic embeddings (*i.e.*, $\boldsymbol{t}_{\text{bad}}$). To achieve this, we maximize the probability of using each topic keyword to predict its representing topic:

$$\mathcal{L}_{reg}^A = -\sum_{a\in A} \sum_{w_i\in l_a} \log P(t_a|w_i), \quad (3)$$

$$\mathcal{L}_{reg}^S = -\sum_{s\in S} \sum_{w_i\in l_s} \log P(t_s|w_i), \quad (4)$$

where l_a, l_s is the keyword list for aspect a, sentiment s, respectively; $P(t|w_i) \propto \exp(\boldsymbol{u}_i^\top \boldsymbol{t})$. Eqs. (3) and (4) empower the embedding space for classification purpose, that is, words can be "classified" into topics based on embedding similarity. For good initializations of $\boldsymbol{t}_a$ and $\boldsymbol{t}_s$, we use the average word embedding of user-provided keywords for each aspect and sentiment.

Regularizing Joint ⟨Sentiment, Aspect⟩ Topics. Now we examine the joint case, where $|S|\times|A|$ topics are regularized. We connect the learning of joint topic embeddings with pure aspect/sentiment topics by exploring the relationship between marginal distribution and joint distribution:

$$P(t_a|w_i) = \sum_{s\in S} P\left(t_{\langle s,a\rangle} \middle| w_i\right), \quad (5)$$

$$P(t_s|w_i) = \sum_{a\in A} P\left(t_{\langle s,a\rangle} \middle| w_i\right). \quad (6)$$

As an example, Fig. 3 shows that the marginal probability of the keyword "good" belonging to the sentiment topic "good" is equal to the probability sum of it belonging to ⟨good, food⟩, ⟨good, ambience⟩ and ⟨good, service⟩.

The objective for regularizing joint topics $\mathcal{L}_{joint}$ can be derived by replacing $P(t_a|w_i)$ in Eq. (3) with Eq. (5) and $P(t_s|w_i)$ in Eq. (4) with Eq. (6).

We also notice that general opinion words such as "good" (or pure aspect words such as "seafood") are equally irrelevant to the aspect (or sentiment) dimension, so we use a uniform distribution $\mathcal{U}$ to regularize their distribution over all the classes on the irrelevant dimension:

$$\mathcal{L}^A_{cross} = \sum_{s\in S}\sum_{w_i\in l_s} \mathrm{KL}\left(\mathcal{U}, P(t_a|w_i)\right), \quad (7)$$

$$\mathcal{L}^S_{cross} = \sum_{a\in A}\sum_{w_i\in l_a} \mathrm{KL}\left(\mathcal{U}, P(t_s|w_i)\right). \quad (8)$$

Putting the above objectives altogether, our final embedding learning objective is:

$$\mathcal{L} = \mathcal{L}_l + \lambda_g\mathcal{L}_g + \lambda_r(\mathcal{L}_{reg} + \mathcal{L}_{joint} + \mathcal{L}_{cross}), \quad (9)$$

where $\mathcal{L}_{reg} = \mathcal{L}^A_{reg} + \mathcal{L}^S_{reg}$, and the same for $\mathcal{L}_{joint}$ and $\mathcal{L}_{cross}$. For all the regularization terms, we treat them equally by using the same weight λ_r, which shows to be effective in practice.

4.2 Training CNNs for Classification

Word ordering information is crucial for sentiment analysis. For example, "Any movie is better than this one" and "this one is better than any movie" convey opposite sentiment polarities but have the exactly same words. The trained embedding space mainly captures word-level discriminative signals but is insufficient to model such sequential information. Therefore, we propose to train convolutional neural networks (CNNs) to generalize knowledge from the preliminary predictions given by the embedding space. Specifically, we first pre-train CNNs with soft predictions given by the cosine similarity between document embeddings and topic embeddings, and then adopt a self-training strategy to further refine the CNNs using their high-confident predictions on unlabeled documents.

Neural Model Pre-training. For each unlabeled review, we can (1) derive one distribution over the joint topics by calculating the cosine similarity between the document representation $\boldsymbol{d}$ and $\boldsymbol{t}_{\langle s,a\rangle}$, (2) derive separate distributions over sentiment and aspect variables using cosine similarity with marginal topics $\boldsymbol{t}_a$ and $\boldsymbol{t}_s$, or (3) combine (1) and (2) by adding the two sets of cosine scores. We find empirically that the last method achieves the best result, *i.e.*, the distribution of a test review d over the aspect and sentiment categories is computed as:

$$P(a|d) \propto \exp\left(T\cdot\left(\cos(\boldsymbol{t}_a,\boldsymbol{d}) + \frac{\sum_{s\in S}\cos(\boldsymbol{t}_{\langle s,a\rangle},\boldsymbol{d})}{|S|}\right)\right), \quad (10)$$

$$P(s|d) \propto \exp\left(T\cdot\left(\cos(\boldsymbol{t}_s,\boldsymbol{d}) + \frac{\sum_{a\in A}\cos(\boldsymbol{t}_{\langle s,a\rangle},\boldsymbol{d})}{|A|}\right)\right), \quad (11)$$

where $\boldsymbol{d}$ is obtained by averaging the word embeddings in d, and T is the temperature to control how greedy we want to learn from the embedding-based prediction.

We train two CNN models separately for aspect and sentiment classification by learning from the two distributions in Eqs. (10) and (11). We leverage the knowledge distillation objective (Hinton et al., 2015) to minimize the cross entropy between the embedding-based prediction p_d and the output prediction q_d of the CNNs:

$$H(p_d, q_d) = -\sum_t P(t|d)\log Q(t|d). \quad (12)$$

Neural Model Refinement. The pre-trained CNNs only copy the knowledge from the embedding space. To generalize their current knowledge to the unlabeled corpus, we adopt a self-training technique to bootstrap the CNNs. The idea of self-training is to use the model's high-confident predictions on unlabeled samples to refine itself. Specifically, we compute a target score (Xie et al., 2016) for each unlabeled document based on the predictions of the current model by enhancing high-confident predictions via a squaring operation:

$$\text{target}(P(a|d)) = \frac{P(a|d)^2/f_a}{\sum_{a'\in A}P(a'|d)^2/f_{a'}},$$

where $f_a = \sum_{d\in\mathcal{D}}P(a|d)$. Since self-training updates the target scores at each epoch, the model is gradually refined by its most recent high-confident predictions. The self-training process is terminated when no more samples change label assignments after the target scores are updated. The resulting model can be used to classify any unseen reviews.

5 Evaluation

We conduct a series of quantitative and qualitative evaluation on benchmark datasets to demonstrate the effectiveness of our model.

Dataset	#Training reviews	#Test reviews
Restaurant	16,061	966
Laptop	14,683	749

Table 1: Dataset Statistics.

Dataset	Aspect	Keywords
Restaurant	Location	street block river avenue
	Drinks	beverage wines cocktail sake
	Food	spicy sushi pizza taste
	Ambience	atmosphere room seating environment
	Service	tips manager waitress servers
Laptop	Support	service warranty coverage replace
	OS	windows ios mac system
	Display	screen led monitor resolution
	Battery	life charge last power
	Company	hp toshiba dell lenovo
	Mouse	touch track button pad
	Software	programs apps itunes photoshop
	Keyboard	key space type keys

Table 2: Keywords of each aspect/sentiment.

5.1 Experimental Setup

Datasets: The two datasets are used for evaluation:

- **Restaurant**: For in-domain training corpus, we collect 16,061 unlabeled reviews from *Yelp Dataset Challenge*[2]. For evaluation, we use the benchmark dataset in the restaurant domain in SemEval-2016 (Pontiki et al., 2016), where each sentence is labeled with aspect and sentiment polarity. We remove sentences with multiple labels or with a *neutral* sentiment polarity to simplify the problem (otherwise a set of keywords can be added to describe it), leaving us with 5 aspects: *Location, Drinks, Food, Ambience*, and *Service*.
- **Laptop**: We leverage 14,683 unlabeled Amazon reviews under the laptop category collected by (He and McAuley, 2016) as in-domain training corpus. We also use the benchmark dataset in the laptop domain in SemEval-2016 for evaluation. There are altogether 21 aspects. We remove those rather rare ones and retain 8 aspects: *Support, OS, Display, Battery, Company, Mouse, Software,* and *Keyboard*. Detailed statistics of both datasets are listed in Table 1, and the aspects along with their keywords are in Table 2.

Preprocessing and Hyperparameter Setting. To preprocess the training corpus D, we use the word tokenizer provided by *NLTK*[3]. We also use a phrase mining tool, AutoPhrase (Shang et al., 2017), to extract meaningful phrases such as "great wine" and "numeric keypad" such that they can capture complicated semantics in a single text unit. In joint topic representation learning, these phrases have their own embeddings just as other words. In embedding training process, we set the hyperparameters as follows: embedding dimension = 100, local context window size $h = 5$, $\lambda_g = 2.5$, $\lambda_r = 1$, training epoch = 5. For neural model pre-training, we set $T = 20$. A CNN model is trained for each sub-task: aspect extraction and sentiment classification. Each model uses 20 feature maps for filters with window sizes of 2, 3, and 4. SGD is used with $1e-3$ as the learning rate in both pre-training and self-training and the batch size is set to 16.

[2]https://www.yelp.com/dataset/challenge
[3]https://www.nltk.org/

5.2 Quantitative Evaluation

We conduct quantitative evaluation on both aspect extraction and sentiment polarity classification.

Compared Methods. Our model is compared with several previous studies. A few of them are specially designed for aspect extraction but do not perform well on sentiment classification. So we only report their results on aspect extraction. For fair comparison, we use the same training corpus and test set for each baseline method. For weakly-supervised methods, they are fed with the same keyword list as ours.

- **CosSim**: The topic representation is averaged by the embedding of seed words trained by Word2Vec. Cosine similarity is computed between a test sample and the topics to classify the sentence.
- **ABAE** (He et al., 2017): An attention-based model to unsupervisedly extract aspects. An autoencoder is trained to reconstruct sentences through aspect embeddings. The learned topics need to be manually mapped to aspects.
- **CAt** (Tulkens and van Cranenburgh, 2020): A recent method for unsupervised aspect extraction. A single head attention is calculated by a Radio Basis Function kernel to be the sentence summary.
- **MATE** (Angelidis and Lapata, 2018): A weakly-supervised aspect extraction method that extends ABAE by initializing the aspect embedding using weighted average of keyword embedding.
- **ISWD** (Karamanolakis et al., 2019): A weakly-supervised aspect extraction method that co-

Methods	Restaurant				Laptop			
	Accuracy	Precision	Recall	macro-F1	Accuracy	Precision	Recall	macro-F1
CosSim	64.20	54.55	47.82	49.85	55.87	60.55	54.37	50.83
ABAE(He et al., 2017)	66.98	50.41	53.51	50.10	58.34	59.00	57.25	55.61
CAt(Tulkens and van Cranenburgh, 2020)	66.25	57.34	56.51	50.51	63.02	69.76	68.13	64.12
MATE(Angelidis and Lapata, 2018)	67.34	56.13	51.27	51.77	66.23	64.18	65.50	64.74
ISWD(Karamanolakis et al., 2019)	69.56	58.77	54.19	52.80	67.24	68.47	69.33	67.85
W2VLDA(García-Pablos et al., 2018)	71.05	61.98	61.46	54.30	70.42	73.18	71.38	69.38
BERT(Devlin et al., 2019)	75.98	62.30	**80.26**	60.83	74.29	77.91	74.22	72.35
Ours w/o joint	80.64	72.18	68.75	68.28	76.70	77.65	77.81	77.24
Ours w/o self train	84.47	69.00	77.58	70.79	77.84	78.63	78.93	78.09
Ours	**84.99**	**74.09**	77.50	**74.85**	**77.97**	**78.67**	**79.19**	**78.28**

Table 3: Quantitative evaluation on aspect identification (%).

Methods	Restaurant				Laptop			
	Accuracy	Precision	Recall	macro-F1	Accuracy	Precision	Recall	macro-F1
CosSim	70.19	68.73	63.57	63.87	65.87	70.55	64.37	60.83
W2VLDA	77.36	79.69	73.02	70.30	74.50	75.36	75.65	74.48
BERT	79.50	77.72	77.99	77.85	70.68	75.10	71.26	73.67
Ours w/o joint	82.71	81.64	80.69	81.16	74.77	74.81	75.31	74.65
Ours w/o self train	82.61	83.19	78.82	80.15	75.70	76.74	76.97	75.69
Ours	**83.64**	**83.45**	**80.76**	**81.74**	**76.23**	**76.84**	**77.25**	**76.20**

Table 4: Quantitative evaluation on sentiment polarity classification (%).

trains a bag-of-seed-words classifier and an embedding-based classifier.

- **W2VLDA** (García-Pablos et al., 2018): A state-of-the-art topic modeling based method that leverages keywords for each aspect/sentiment to jointly do aspect/sentiment classification.
- **BERT** (Devlin et al., 2019): A recent proposed deep language model. We utilize the pre-trained BERT (12-layer, 768 dimension, uncased) and implement a simple weakly-supervised method that fine-tunes the model by providing pseudo labels for sentences containing keywords from a given aspect/sentiment.
- **Ours w/o joint**: An ablation of our proposed model. Neural model is pre-trained on separate topic embedding for each sentiment and aspect.
- **Ours w/o self train**: An ablation of our proposed model without self-training process.

Aspect Extraction. We report the results of aspect extraction of our model and all the baselines in Table 3. We use four metrics for evaluation: Accuracy, Precision, Recall and macro-F1 score. We observe that weakly-supervised methods tend to have a better performance than unsupervised ones, suggesting that using keywords to enrich the semantics of labels is a promising direction to increase classification performance. As shown in Table 3, our model, even without self-training, outperforms baseline methods on most of the metrics by a large margin, indicating that our model obtains substantial benefits from learning the semantics of fine-grained joint topics, and self-training boosts the performance further. We observe that our model can deal with cases where the target of the sentence is not explicitly mentioned. For example, our model correctly labels "It's to die for!" as (good, food). Although nothing mentioned is related to food, "to die for" appears in other sentences addressing the tastiness of food, thus is captured by the joint topic of (good, food). Our model performs lower than BERT on Recall on the **Restaurant** dataset. This is because BERT is a very deep neural language model pre-trained on the Wikipedia corpus (∼2,500 million words) containing much more commonsense knowledge. We examine the errors made by our method and discover that it can be confused by very general descriptions. Our model marks "Fine dining restaurant quality." as (good, ambience), while the true label is (good, food). In fact, we think this sentence might mean both food and ambience are good. More case studies comparing our full model and our model w/o joint embedding are shown in Appendix A.

Sentiment Polarity Classification. We compare our model against baseline methods on sentiment classification and show the results in Table 4. Since some methods are designed for aspect extraction and do not perform well on sentiment classification, we do not report their results. As shown in

	Ambience	Service	Food	Support	Keyboard	Battery
Good	cozy, intimate, comfortable, loungy, great music	professional, polite, knowledgable, informative, helpful	huge portion, flavourful, super fresh, husband loves, authentic italian	accidental damage protection, accidental damage warranty, generous, guarantee, commitment	tactile feedback, tactile feel, classic, nicely spaced, chiclet style	lasts long, charges quickly, high performance, lasting, great power
Bad	cramped, unbearable, uncomfortable, dreary, chaos	inattentive, ignoring, extremely rude, condescending, inexperienced	microwaved, flavorless, vomit, frozen food, undercooked	completely useless, denied, refused, blamed, apologize	large hands, shallow, cramped, wrong key, typos	completely dead, drained, discharge, unplugged, torture

Table 5: Keywords retrieved by joint topic representations.

the table, our model outperforms all the baselines on both datasets. We also observe that methods only leveraging local contexts (CosSim and BERT) do not perform well compared to methods that leverage both global and local contexts (Ours and W2VLDA) on **Laptop** dataset. Since "good" and "bad" are a pair of antonyms, they can have very similar collocations, so models purely capturing local contexts do not distinguish them well.

5.3 Qualitative Evaluation

Representative terms for joint topics. To evaluate the quality of the joint topic representation, we retrieve their representative terms by ranking the embedding cosine similarity between words and each joint topic vector. For brevity, we randomly sample 3 aspects from each dataset and pair them up with the two sentiment polarity to form 12 joint topics. We list their top terms in Table 5. Results show that the representative terms form coherent and meaningful topics, and they are not restricted to be adjectives, such as "vomit" in (bad, food) and "commitment" in (good, support). Another interesting observation is that "cramped" appears in both (bad, ambience) in restaurant domain and (bad, keyboard) in laptop domain, suggesting that our model captures different meanings of words based on in-domain corpus.

Joint Topic Representation Visualization. To understand how the joint topics are distributed in the embedding space, along with the aspect and sentiment topics, we use PCA (Jolliffe, 2011) for dimension reduction to visualize topic embedding trained on the **Restaurant** corpus in Fig. 4. An interesting observation is that, some aspect topics (e.g., *ambience*) lie approximately in the middle of their joint topics ("*good, ambience*" and "*bad, ambience*"), showing that our embedding learning objective understands the joint topics as decomposition of their "marginal" topics, which fits with our goal to learn fine-grained topics.

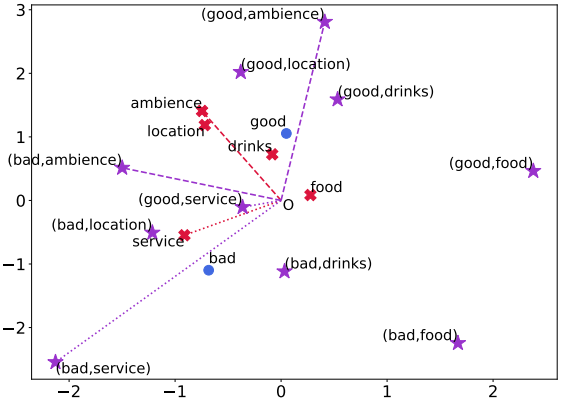

Figure 4: Visualization of joint topics (purple stars), aspect topics (red crosses) and sentiment topics (blue dots) in the embedding space.

6 Conclusion

In this paper we propose to enhance weakly-supervised aspect-based sentiment analysis by learning the representation of ⟨sentiment, aspect⟩ joint topic in the embedding space to capture more fine-grained information. We introduce an embedding learning objective that leverages user-given keywords for each aspect/sentiment and models their distribution over the joint topics. The embedding-based predictions are then used for pre-training neural models, which are further refined via self-training on unlabeled corpus. Experiment results show that our method learns high-quality joint topics and outperforms previous studies substantially.

In the future, we plan to adapt our methods to more general applications that are not restricted to the field of sentiment analysis, such as doing multiple-dimension classification (e.g., topic, location) on general text corpus.

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment

prediction and they are both weakly supervised. In *EMNLP*.

Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Samuel Brody and Noémie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *HLT-NAACL*.

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Comput. Linguistics*, 18:467–479.

Zhiyuan Chen, Arjun Mukherjee, and Bing Liu. 2014. Aspect extraction with automated prior knowledge learning. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Aitor García-Pablos, Montse Cuadros, and German Rigau. 2018. W2vlda: Almost unsupervised system for aspect based sentiment analysis. *Expert Syst. Appl.*, 91:127–137.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *ACL*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Exploiting document knowledge for aspect-level sentiment classification. *ArXiv*, abs/1806.04346.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *ArXiv*, abs/1602.01585.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD '04*.

Ian T. Jolliffe. 2011. Principal component analysis. In *International Encyclopedia of Statistical Science*.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019. Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. In *EMNLP/IJCNLP*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Ming Liao, Jing Li, Haisong Zhang, Lingzhi Wang, Xixin Wu, and Kam-Fai Wong. 2019. Coupling global and local context for unsupervised aspect extraction. In *EMNLP/IJCNLP*.

Pengfei Liu, Shafiq R. Joty, and Helen M. Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *EMNLP*.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance M. Kaplan, and Jiawei Han. 2019. Spherical text embedding. In *NIPS*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. 2015. Sentiment analysis on social media for stock movement prediction. *Expert Syst. Appl.*, 42:9603–9611.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, and et. al. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Magnus Sahlgren. 2008. The distributional hypothesis. *The Italian Journal of Linguistics*, 20:33–54.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2017. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30:1825–1837.

Stéphan Tulkens and Andreas van Cranenburgh. 2020. Embarrassingly simple unsupervised aspect extraction. *ArXiv*, abs/2004.13580.

Linlin Wang, Kang Liu, Zhu Cao, Jun Zhao, and Gerard de Melo. 2015. Sentiment-aspect extraction based on restricted boltzmann machines. In *ACL*.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *ACL*.

Xueke Xu, Songbo Tan, Yue Liu, Xueqi Cheng, and Zheng Lin. 2012. Towards jointly extracting aspects and aspect-specific sentiment knowledge. In *CIKM '12*.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *EMNLP*.

Li Zhuang, Feng Jing, and Xiaoyan Zhu. 2006. Movie review mining and summarization. In *CIKM '06*.

A Case studies.

We list several test samples along with their ground truth and model predictions in Table 6 and Table 7. Some conflicting cases between ours and the ground truth are rather ambiguous. For example, the ground truth of the second example in Table 6 is (good, location), but we still think given that the review mentions "the outdoor atmosphere" and uses terms like "sitting on the sidewalk" and "cool evening", it is more relevant to ambience than location, as is output by our full model. The gold aspect label for the second and the third reviews in Table 7 are both "company", but apparently these two sentences are talking about two different aspects: the product itself and the service of the company. Though the output of our model, "os" and "support" for these two sentences may not be the most precise prediction, at least our model treats them as two different aspects. For other difficult sentences, human may still have a more precise judgement with much more common knowledge then our model trained on a corpus with only thousands of reviews.

Review	Ground Truth	Output of Full Model	Output of Model w/o joint embedding
The wait staff is very freindly, they make it feel like you're eating in a freindly little european town.	(good, service)	(good, ambience)	(good, location)
The outdoor atmosphere of sitting on the sidewalk watching the world go by 50 feet away on 6th avenue on a cool evening was wonderful.	(good, location)	(good, ambience)	(good, ambience)
It's simply the best meal in NYC.	(good, food)	(good, food)	(good, location)
You can get a table without a reservation if you get there early I they don't make you by bottles.	(good, service)	(good, service)	(bad, service)
The sauce tasted more like Chinese fast food than decent Korean.	(bad, food)	(good, food)	(bad, food)
My wife had barely###touched that mess of a dish.	(bad, food)	(bad, food)	(good, food)
This is undoubtedly my favorite modern Japanese brasserie (that don't serve sushi), and in my opinion, one of the most romantic restaurants in the city!	(good, ambience)	(good, food)	(good, location)
We took advanatage of the half price sushi deal on saturday so it was well worth it.	(good, food)	(good, food)	(bad, food)
Somehow working the italian charm with constant mille grazie does not constitute proper service.	(bad, service)	(good, service)	(bad, service)
If you don't like it, I don't know what to tell you.	(good, food)	(bad, food)	(bad, service)

Table 6: Comparison of predictions on sample **Restaurant** reviews between our full model and model pre-trained w/o joint topic embedding.

Review	Ground Truth	Output of Full Model	Output of Model w/o joint embedding
NO junkware!!	(good, software)	(good, software)	(good, display)
I definitely will buy a Mac again if and when this computer ever fails.	(good, company)	(good, os)	(good, os)
I don't have the inclination or time to devote to a companies tech support, search functions, or hold times.....dropped the HP and never looked back.	(bad, company)	(bad, support)	(bad, company)
I find myself adjusting it regularly.	(bad, display)	(bad, display)	(bad, mouse)
I thought learning the Mac OS would be hard, but it is easily picked up if you are familiar with a PC.	(good, os)	(good, os)	(bad, os)
They told me to reprogram the computer, and I didn't need to do that, and I lost pictures of my oldests first 2 years of her life.	(bad, support)	(bad, support)	(good, support)
It was nearly impossible to find a laptop that had A) a side keyboard.	(good, keyboard)	(bad, keyboard)	(good, keyboard)
But, hey, it's an Apple.	(good, company)	(bad, company)	(good, company)
I'm no power###user, but I have had no learning###curve with the MAC and I don't do anything geeky enough forcing me to learn the OS.	(good, os)	(good, os)	(bad, os)
The battery lasted 12 months, then pffft.....gone.	(bad, battery)	(bad, battery)	(good, battery)

Table 7: Comparison of predictions on sample **Laptop** reviews between our full model and model pre-trained w/o joint topic embedding.