A Variational Approach to Weakly Supervised Document-Level Multi-Aspect Sentiment Classification

Ziqian Zeng[1], Wenxuan Zhou[2], Xin Liu[1], and Yangqiu Song[1]
[1]Department of CSE, Hong Kong University of Science and Technology, HK
[2]Department of CS, University of Southern California, CA, USA
[1]{zzengae, xliucr, yqsong}@cse.ust.hk
[2]{zhouwenx}@usc.edu

Abstract

In this paper, we propose a variational approach to weakly supervised document-level multi-aspect sentiment classification. Instead of using user-generated ratings or annotations provided by domain experts, we use target-opinion word pairs as "supervision." These word pairs can be extracted by using dependency parsers and simple rules. Our objective is to predict an opinion word given a target word while our ultimate goal is to learn a sentiment polarity classifier to predict the sentiment polarity of each aspect given a document. By introducing a latent variable, i.e., the sentiment polarity, to the objective function, we can inject the sentiment polarity classifier to the objective via the variational lower bound. We can learn a sentiment polarity classifier by optimizing the lower bound. We show that our method can outperform weakly supervised baselines on TripAdvisor and BeerAdvocate datasets and can be comparable to the state-of-the-art supervised method with hundreds of labels per aspect.

1 Introduction

Document-level multi-aspect sentiment classification (DMSC) aims to predict the sentiment polarity of each aspect given a document which consists of several sentences describing one or more aspects (Wang et al., 2010, 2011; Yin et al., 2017). Solving the DMSC task is useful for providing both recommendations for users and suggestions for business owners on customer review platforms.

Aspect based sentiment classification (Tang et al., 2016a,b; Wang et al., 2016b; Chen et al., 2017; Ma et al., 2017; Wang et al., 2018) was usually done by supervised learning, where aspect-level annotations should be provided. Aspect-level annotations are not easy to obtain. Even when the platform provides the function to rate for different aspects, users are less likely to submit all of them. For example, about 37% of the aspect ratings are missing on TripAdvisor. If we can solve DMSC task without using aspect-level annotations, it can save human effort to annotate data or collect user-generated annotations on the platform.

Existing weakly supervised approaches (Wang et al., 2010, 2011) use overall polarities instead of aspect polarities as "supervision." Compared with the polarity of each aspect, it is relatively easy to obtain overall polarities. Specifically, they minimize the square loss between the overall polarity and the weighted sum of all aspect polarities. However, when users only care about a particular rare aspect, e.g., childcare services, these approaches cannot estimate parameters of the rare aspect incrementally. They have to re-collect documents which mentioned this rare aspect and estimate parameters of all aspects based on the new corpus. In addition, these approaches assume the document is a bag-of-words, which neglects the order of the words and fails to capture the similarity between words.

In this paper, we propose to use target-opinion word pairs as "supervision." Target-opinion word pairs can be helpful with our ultimate goal which is to learn a classifier to predict the sentiment polarity of each aspect given a document. For example, in a document "The bedroom is very spacious," if we can extract the target-opinion pair "bedroom-spacious," the sentiment polarity of the aspect *room* is likely to be *positive*. Hence, we propose to achieve the polarity classification goal by accomplishing another relevant objective: to predict an opinion word given a target word.

We can decompose the opinion word prediction objective into two sub-tasks. The first sub-task is to predict the sentiment polarity based on a document. For example, given a document "The bedroom is very spacious," it predicts the sentiment polarity of the aspect *room* to be *positive*. The sec-

Proceedings of NAACL-HLT 2019, pages 386–396
Minneapolis, Minnesota, June 2 - June 7, 2019.

ond sub-task is to predict the opinion word given a target word and a sentiment polarity predicted by the first sub-task. For example, knowing the fact that the sentiment polarity of the aspect *room* is *positive*, it predicts the opinion word associated with the target word "room" to be "spacious." By introducing a latent variable, i.e., the sentiment polarity of an aspect, to the opinion word prediction objective, we can inject the polarity classification goal (the first sub-task) into the objective via the variational lower bound which also incorporates the second sub-task. In this sense, our training objective is only based on the target-opinion word pairs which can be extracted by using dependency parsers and some manually designed rules. We consider our approach as weakly supervised learning because there is no direct supervision from polarity of each aspect.

In other words, our model includes two classifiers: a sentiment polarity classifier and an opinion word classifier. In the sentiment polarity classifier, it predicts the sentiment polarity given a document. In the opinion word classifier, it predicts an opinion word based on a target word and a sentiment polarity. Compared with previous approaches (Wang et al., 2010, 2011), our approach can get rid of the assumption that the overall polarity should be observed and it is a weighted sum of all aspect polarities. Moreover, our approach can estimate parameters of a new aspect incrementally. In addition, our sentiment polarity classifier can be more flexible to capture dependencies among words beyond the bag-of-words representation if we use a deep neural network architecture to extract features to represent a document. We conducted experiments on two datasets, TripAdvisor (Wang et al., 2010) and BeerAdvocate (McAuley et al., 2012), to illustrate the effectiveness of our approach.

Our contributions are summarized as follows,

• We propose to solve DMSC task in a nearly unsupervised way.

• We propose to learn a classifier by injecting it into another relevant objective via the variational lower bound. This framework is flexible to incorporate different kinds of document representations and relevant objectives.

• We show promising results on two real datasets and we can produce comparable results to the supervised method with hundreds of labels per aspect.

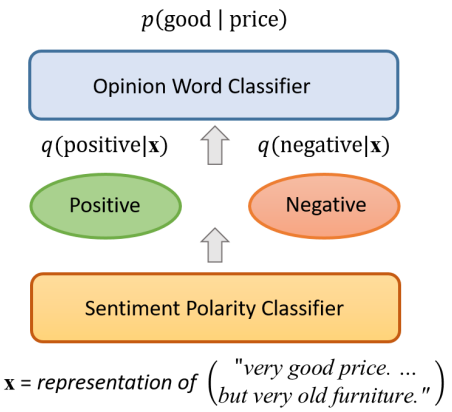

Figure 1: A sentiment polarity classifier and an opinion word classifier associated with the aspect *price*.

Code and data for this paper are available on https://github.com/HKUST-KnowComp/VWS-DMSC.

2 VWS-DMSC Approach

In this section, we describe our variational approach to weakly supervised DMSC (VWS-DMSC). In the next section, we present how we obtain target-opinion word pairs by using a rule-based extraction approach.

2.1 Overview

Our model consists of a sentiment polarity classifier and an opinion word classifier. Our task is document-level multi-aspect sentiment classification. For each aspect, we train a sentiment polarity classifier and an opinion word classifier. The input of the sentiment polarity classifier of each aspect is the same, i.e., a representation of a document. The target-opinion word pairs used in opinion word classifiers are different for different aspects.

Figure 1 shows the relation between two classifiers (on the aspect *price*). The input $\mathbf{x}$ of the sentiment polarity classifier is a representation of a document, e.g., bag-of-words or a representation learned by recurrent neural networks. The sentiment polarity classifier takes $\mathbf{x}$ as input and produces a distribution of sentiment polarity R_a of an aspect a, denoted as $q(R_a|\mathbf{x})$. If R_a only has two possible values, i.e., positive and negative, then outputs of the classifier are $q(positive|\mathbf{x})$ and $q(negative|\mathbf{x})$ respectively. The opinion word classifier takes a target word ("price") and a possible value of the sentiment polarity r_a as input, and estimates $p(\text{"good"}|r_a, \text{"price"})$. Our train-

ing objective is to maximize the log-likelihood of an opinion word given a target word, e.g., $p(\text{“good”}|\text{“price”})$. The likelihood is estimated based on the sentiment polarity classifier and the opinion word classifier.

2.2 Sentiment Polarity Classifier

The sentiment polarity classifier aims to estimate a distribution of sentiment polarity $q(R_a|\mathbf{x})$, where R_a is a discrete random variable representing the sentiment polarity and $\mathbf{x}$ is a feature representation of a document. We use a simple Softmax classifier here. We denote r_a as a possible value of the random variable R_a, representing a possible sentiment polarity. The model estimates the probability of class r_a as

$$q(R_a = r_a|\mathbf{x}) = \frac{\exp\left(\mathbf{w}_{r_a}^T\mathbf{x}\right)}{\sum_{r'_a}\exp\left(\mathbf{w}_{r'_a}^T\mathbf{x}\right)}\ , \quad (1)$$

where $\mathbf{w}_{r_a}$ is a vector associated with sentiment class r_a for aspect a.

Document Representation The representation of a document $\mathbf{x}$ can be different using different feature extraction approaches. Traditional document representations of sentiment classification would be bag-of-words, n-gram, or averaged word embeddings. Recently, end-to-end recurrent neural network based models demonstrate a powerful capacity to extract features of a document. The state-of-the-art model in DMSC task is (Yin et al., 2017). We use it as the document representation in our model.

2.3 Opinion Word Classifier

The opinion word classifier aims to estimate the probability of an opinion word w_o given a target word w_t and a sentiment polarity r_a:

$$p(w_o|r_a, w_t) = \frac{\exp\left(\varphi(w_o, w_t, r_a)\right)}{\sum_{w'_o}\exp\left(\varphi(w'_o, w_t, r_a)\right)}\ , \quad (2)$$

where φ is a scoring function related to opinion word w_o, target word w_t, and sentiment polarity r_a. Here we use the dot product as the scoring function:

$$\varphi(w_o, w_t, r_a) = I\big((w_t, w_o) \in \mathcal{P}, w_t \in \mathcal{K}_a\big)\cdot\mathbf{c}_{r_a}^T\mathbf{w}_o\ , \quad (3)$$

where $\mathbf{w}_o$ is the word embedding of opinion word w_o, $\mathbf{c}_{r_a}$ is a vector associated with r_a, $\mathcal{P}$ is the set of pairs extracted from the document, $\mathcal{K}_a$ is the set of target words associated with aspect a, and $I(\cdot)$ is an indicator function where $I(true) = 1$ and $I(false) = 0$.

Given a target word w_t and a sentiment polarity r_a, we aim to maximize the probability of opinion words highly related to them. For example, opinion word “good” is usually related to target word “price” for aspect *value* with sentiment polarity *positive*, and opinion word “terrible” is usually related to target word “traffic” for aspect *location* with sentiment polarity *negative*.

2.4 Training Objective

The objective function is to maximize the log-likelihood of an opinion word w_o given a target word w_t. As we mentioned before, the objective function can be decomposed into two sub-tasks. The first one corresponds to the sentiment polarity classifier. The second one corresponds to the opinion word classifier. After introducing a latent variable, i.e., the sentiment polarity, to the objective function, we can derive a variational lower bound of the log-likelihood which can incorporate two classifiers:

$$\begin{aligned}
\mathcal{L} &= \log p(w_o|w_t) \\
&= \log \sum_{r_a} p(w_o, r_a|w_t) \\
&= \log \sum_{r_a} q(r_a|\mathbf{x})\Big[\frac{p(w_o, r_a|w_t)}{q(r_a|\mathbf{x})}\Big] \\
&\geq \sum_{r_a} q(r_a|\mathbf{x})\Big[\log\frac{p(w_o, r_a|w_t)}{q(r_a|\mathbf{x})}\Big] \\
&= \mathbb{E}_{q(R_a|\mathbf{x})}\left[\log p(w_o|r_a, w_t)p(r_a|w_t)\right] \\
&\quad + H(q(R_a|\mathbf{x})) \\
&= \mathbb{E}_{q(R_a|\mathbf{x})}\left[\log p(w_o|r_a, w_t)p(r_a)\right] \\
&\quad + H(q(R_a|\mathbf{x}))\ , \quad (4)
\end{aligned}$$

where $H(\cdot)$ refers to the Shannon entropy. By applying Jensen’s inequality, the log-likelihood is lower-bounded by Eq. (4). The equality holds if and only if the KL-divergence of two distributions, $q(R_a|\mathbf{x})$ and $p(R_a|w_t, w_o)$, equals to zero. Maximizing the variational lower bound is equivalent to minimizing the KL-divergence. Hence, we can learn a sentiment polarity classifier which can produce a similar distribution to the true posterior $p(R_a|w_t, w_o)$. Compared with $p(R_a|w_t, w_o)$, $q(R_a|\mathbf{x})$ is more flexible since it can take any kind of feature representations as input. We assume that

a target word w_t and a sentiment polarity r_a are independent since the polarity assignment is not influenced by the target word. We also assume that the sentiment polarity R_a follows a uniform distribution, which means $p(r_a)$ is a constant. We remove it in Eq. (4) to get a new objective function as follows:

$$\mathbb{E}_{q(R_a|\mathbf{x})}\left[\log p(w_o|r_a, w_t)\right] + H(q(R_a|\mathbf{x}))\ . \quad (5)$$

2.4.1 Approximation

The partition function of Eq. (2) requires the summation over all opinion words in the vocabulary. However, the size of opinion word vocabulary is large, so we use the negative sampling technique (Mikolov et al., 2013) to approximate Eq. (2). Specifically, we substitute $\log p(w_o|r_a, w_t)$ in the objective (5) with the following objective function:

$$\log\sigma\big(\varphi(w_o, w_t, r_a)\big) + \sum_{w'_o \in \mathcal{N}} \log\sigma\big(-\varphi(w'_o, w_t, r_a)\big)\ , \quad (6)$$

where w'_o is a negative sample of opinion words in the vocabulary, $\mathcal{N}$ is the set of negative samples and σ is the sigmoid function. Then our final objective function is rewritten as:

$$\mathbb{E}_{q(R_a|\mathbf{x})}\big[\log\sigma\big(\varphi(w_o, w_t, r_a)\big) + \sum_{w'_o \in \mathcal{N}} \log\sigma\big(-\varphi(w'_o, w_t, r_a)\big)\big] + \alpha H(q(R_a|\mathbf{x}))\ , \quad (7)$$

where α is a hyper-parameter which can adjust the expectation and entropy terms into the same scale (Marcheggiani and Titov, 2016).

3 Target Opinion Word Pairs Extraction

Target-opinion word pairs extraction is a well studied problem (Hu and Liu, 2004; Popescu and Etzioni, 2005; Bloom et al., 2007; Qiu et al., 2011). We designed five rules to extract potential target-opinion word pairs. Our method relies on Stanford Dependency Parser (Chen and Manning, 2014). We describe our rules as follows.

Rule 1: We extract pairs satisfying the grammatical relation *amod* (adjectival modifier) (De Marneffe and Manning, 2008). For example, in phrase "very good price," we extract "price" and "good" as a target-opinion pair.

Rule 2: We extract pairs satisfying the grammatical relation *nsubj* (nominal subject), and the head word is an adjective and the tail word is a noun. For example, in a sentence "The room is small," we can extract "room" and "small" as a target-opinion pair.

Dataset	TripAdvisor	BeerAdvocate
# docs	28,543	27,583
# target words	3,737	3,088
# opinion words	12,406	9,166
# pairs from R1	208,676	249,264
# pairs from R2	82,944	28,505
# pairs from R3	2,241	1,092
# pairs from R4	2,699	6,812
# pairs from R5	16,537	55,825

Table 1: Statistics of extracted target-opinion pairs .

Rule 3: Some verbs are also opinion words and they are informative. We extract pairs satisfying the grammatical relation *dobj* (direct object) when the head word is one of the following four words: "like", "dislike", "love", and "hate". For example, in the sentence "I like the smell," we can extract "smell" and "like" as a target-opinion pair.

Rule 4: We extract pairs satisfying the grammatical relation *xcomp* (open clausal complement), and the head word is one of the following word: "seem","look", "feel", "smell", and "taste". For example, in the sentence "This beer tastes spicy," we can extract "taste" and "spicy" as a target-opinion pair.

Rule 5: If the sentence contains some adjectives that can implicitly indicate aspects, we manually assign them to the corresponding aspects. According to (Lakkaraju et al., 2014), some adjectives serve both as target words and opinion words. For example, in the sentence "very tasty, and drinkable," the previous rules fail to extract any pair. But we know it contains a target-opinion pair, i.e., "taste-tasty." Most of these adjectives have the same root form with the aspects they indicated, e.g., "clean" (cleanliness), and "overpriced" (price). This kind of adjective can be extracted first and then we can obtain more similar adjectives using word similarities. For example, given "tasty," we could get "flavorful" by retrieving similar words.

Table 1 shows the statistics of the rule-based extraction on our two datasets. The first four rules can be applied to any dataset while the last one is domain dependent which requires human effort to identify these special adjectives. In practice, rule

5 can be removed to save human effort. The effect of removing rule 5 is shown in experiments.

After extracting potential target-opinion word pairs, we need to assign them to different aspects as supervision signals. We select some seed words to describe each aspect, and then calculate similarities between the extracted target (or opinion) word and seed words, and assign the pair to the aspect where one of its seed words has the highest similarity. The similarity we used is the cosine similarity between two word embeddings trained by word2vec (Mikolov et al., 2013). For example, suppose seed words {"room", "bed"} and {"business", "Internet"} are used to describe the aspect *room* and *business* respectively, and the candidate pair "pillow - soft" will be assigned to the aspect *room* if the similarity between "pillow" and "bed" is highest among all combinations.

4 Experiment

In this section, we report average sentiment classification accuracies over all aspects on binary DMSC task.

4.1 Datasets

We evaluate our model on TripAdvisor (Wang et al., 2010) and BeerAdvocate (McAuley et al., 2012; Lei et al., 2016; Yin et al., 2017) datasets, which contain seven aspects (value, room, location, cleanliness, check in/front desk, service, and business) and four aspects (feel, look, smell, and taste) respectively. We run the same preprocessing steps as (Yin et al., 2017). Both datasets are split into train/development/test sets with proportions 8:1:1. All methods can use development set to tune their hyper-parameters. Ratings of TripAdvisor and BeerAdvocate datasets are on scales of 1 to 5 and 0 to 5 respectively. But in BeerAdvocate, 0 star is rare, so we treat the scale as 1 to 5. We convert original scales to binary scales as follows: 1 and 2 stars are treated as negative, 3 is ignored, and 4 and 5 stars are treated as positive. In BeerAdvocate, most reviews have positive polarities, so to avoid the unbalanced issue, we perform data selection according to overall polarities. After data selection, the number of reviews with negative overall polarities and that with positive overall polarities are equal.

4.2 Compared Methods

To demonstrate the effectiveness of our method, we compare our model with following baselines:

Majority uses the majority of sentiment polarities in training sets as predictions.

Lexicon means using an opinion lexicon to assign sentiment polarity to an aspect (Read and Carroll, 2009; Pablos et al., 2015). We combine two popular opinion lexicons used by Hu and Liu (2004) and Wilson et al. (2005) to get a new one. If an opinion word from extracted pairs is in positive (negative) lexicon, it votes for positive (negative). When the opinion word is with a negation word, its polarity will be flipped. Then, the polarity of an aspect is determined by using majority voting among all opinion words associated with the aspect. When the number of positive and negative words is equal, we adopt two different ways to resolve it. For **Lexicon-R**, it randomly assigns a polarity. For **Lexicon-O**, it uses the overall polarity as the prediction. Since overall polarities can also be missing, for both Lexicon-R and Lexicon-O, we randomly assign a polarity in uncertain cases and report both mean and std based on five trials of random assignments.

Assign-O means directly using the overall polarity of a review in the development/test sets as the prediction for each aspect.

LRR assumes the overall polarity is a weighted sum of the polarity of each aspect (Wang et al., 2010). LRR can be regarded as the only existing weakly supervised baseline where both algorithm and source code are available.

BoW-DMSC-A is a simple softmax classifier using all annotated training data where the input is a bag-of-words feature vector of a document.

N-DMSC-A is the state-of-the-art neural network based model (Yin et al., 2017) (**N-DMSC**) in DMSC task using all annotated training data, which serves an upper bound to our method.

N-DMSC-O is to use overall polarities as "supervision" to train an N-DMSC and apply it to the classification task of each aspect at the test time.

N-DMSC-{50,100,200,500,1000} is the N-DMSC algorithm using partial data. In order to see our method is comparable to supervised methods using how many labeled data, we use $\{50, 100, 200, 500, 1000\}$ annotations of each aspect to train N-DMSC and compare them to our method. In addition to annotated data for training, there are extra 20% annotated data for validation.

Dataset	TripAdvisor				BeerAdvocate			
	DEV		TEST		DEV		TEST	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Majority	0.6286	–	0.6242	–	0.6739	–	0.6726	–
Lexicon-R	0.5914	0.0021	0.5973	0.0018	0.5895	0.0020	0.5881	0.0025
Lexicon-O	0.7153	0.0012	0.7153	0.0015	0.6510	0.0023	0.6510	0.0021
Assign-O	0.7135	0.0016	0.7043	0.0020	0.6652	0.0028	0.6570	0.0034
N-DMSC-O	0.7091	–	0.7064	–	0.6386	–	0.6493	–
LRR	0.6915	0.0045	0.6947	0.0024	0.5976	0.0110	0.5941	0.0113
VWS-DMSC (Our)	0.7577	0.0016	0.7561	0.0012	0.7502	0.0058	0.7538	0.0066
N-DMSC-50	0.7255	0.0231	0.7270	0.0204	0.7381	0.0143	0.7442	0.0157
N-DMSC-100	0.7482	0.0083	0.7487	0.0069	0.7443	0.0126	0.7493	0.0145
N-DMSC-200	0.7531	0.0040	0.7550	0.0043	0.7555	0.0096	0.7596	0.0092
N-DMSC-500	0.7604	0.0028	0.7616	0.0040	0.7657	0.0066	0.7713	0.0070
N-DMSC-1000	0.7631	0.0054	0.7638	0.0042	0.7708	0.0066	0.7787	0.0053
N-DMSC-A	0.8281	–	0.8334	–	0.8576	–	0.8635	–
BoW-DMSC-A	0.8027	–	0.8029	–	0.8069	–	0.8089	–

Table 2: Averaged accuracies on DMSC of unsupervised, weakly supervised, and supervised methods on TripAdvisor and BeerAdvocate. Methods involve randomness also report standard deviation.

Since the sampled labeled data may vary for different trials, we perform five trials of random sampling and report both mean and std of the results.

For our method, denoted as **VWS-DMSC**, the document representation we used is obtained from N-DMSC (Yin et al., 2017). They proposed a novel hierarchical iterative attention model in which documents and pseudo aspect related questions are interleaved at both word and sentence-level to learn an aspect-aware document representation. The pseudo aspect related questions are represented by aspect related keywords. In order to benefit from their aspect-aware representation scheme, we train an N-DMSC to extract the document representation using only overall polarities. In the iterative attention module, we use the pseudo aspect related keywords of all aspects released by Yin et al. (2017). One can also use document-to-document autoencoders (Li et al., 2015) to generate the document representation. In this way, our method can get rid of using overall polarities to generate the document representation. Hence, unlike LRR, it is not necessary for our method to use overall polarities. Here, to have a fair comparison with LRR, we use the overall polarities to generate document representation. For our method, we do not know which state is positive and which one is negative at training time, so the Hungarian algorithm (Kuhn, 1955) is used to resolve the assignment problem at the test time.

4.3 Results and Analysis

We show all results in Table 2, which consists of three blocks, namely, unsupervised, weakly supervised, and supervised methods.

For unsupervised methods, our method can outperform majority on both datasets consistently. But other weakly supervised methods cannot outperform majority on BeerAdvocate dataset, which shows these baselines cannot handle unbalanced data well since BeerAdvocate is more unbalanced than TripAdvisor. Our method outperforms Lexicon-R and Lexicon-O, which shows that predicting an opinion word based on a target word may be a better way to use target-opinion pairs, compared with performing a lexicon lookup using opinion words from extract pairs. Good performance of Lexicon-O and Assign-O demonstrates the usefulness of overall polarities in development/test sets. N-DMSC-O trained with the overall polarities cannot outperform Assign-O since N-DMSC-O can only see overall polarities in training set while Assign-O can see overall polarities for both development and test sets and does not involve learning and generalization.

For weakly supervised methods, LRR is the only open-source baseline in the literature on weakly supervised DMSC, and our method outperforms LRR by **6**% and **16**% on TripAdvisor and BeerAdvocate datasets. N-DMSC-O can also be considered as a weakly supervised method be-

Dataset	TripAdvisor		BeerAdvocate	
Rule	DEV	TEST	DEV	TEST
R1	0.7215	0.7174	0.7220	0.7216
R2	0.7172	0.7180	0.6864	0.6936
R3	0.6263	0.6187	0.6731	0.6725
R4	0.6248	0.6279	0.6724	0.6717
R5	0.5902	0.5856	0.7095	0.7066
- R1	0.7538	0.7481	0.7458	0.7474
- R2	0.7342	0.7368	0.7504	0.7529
- R3	0.7418	0.7397	0.7565	0.7558
- R4	0.7424	0.7368	0.7518	0.7507
- R5	0.7448	0.7440	0.7550	0.7548
All	0.7577	0.7561	0.7502	0.7538

Table 3: Averaged accuracies on DMSC. "R1 – R5" means only using a rule while "-R1 – -R5" means removing a rule from all the rules.

cause it only uses overall polarities as "supervision," and we still outperform it significantly. It is interesting that LRR is worse than N-DMSC-O. We guess that assuming that the overall polarity is a weighted sum of all aspect polarities may not be a good strategy to train each aspect's polarity or the document representation learned by N-DMSC is better than the bag-of-words representation.

For supervised block methods, BoW-DMSC-A and N-DMSC-A are both supervised methods using all annotated data, which can be seen as the upper bound of our algorithm. N-DMSC-A outperforms BoW-DMSC-A, which shows that the document representation based on neural network is better than the bag-of-words representation. Hence, we use the neural networks based document representation as input of the sentiment polarity classifier. Our results are comparable to N-DMSC-200 on TripAdvisor and N-DMSC-100 on BeerAdvocate.

4.4 Ablation Study

To evaluate effects of extracted rules, we performed an ablation study. We run our algorithm VWS-DMS with each rule kept or removed over two datasets. If no pairs extracted for one aspect in training set, the accuracy of this aspect will be 0.5, which is a random guess. From the Table 3 we can see that, the rule R1 is the most effective rule for both datasets. Rules R3/R4/R5 are less effective on their own. However, as a whole, they can still improve the overall performance. When considering removing each of rules, we found that our algorithm is quite robust, which indicates missing one of the rules may not hurt the performance much. Hence, if human labor is a major concern, rule 5 can be discarded. We found that sometimes removing one rule may even result in better accuracy (e.g., "-R3" for BeerAdvocate dataset). This means this rule may introduce some noises into the objective function. However, "-R3" can result in worse accuracy for TripAdvisor, which means it is still complementary to the other rules for this dataset.

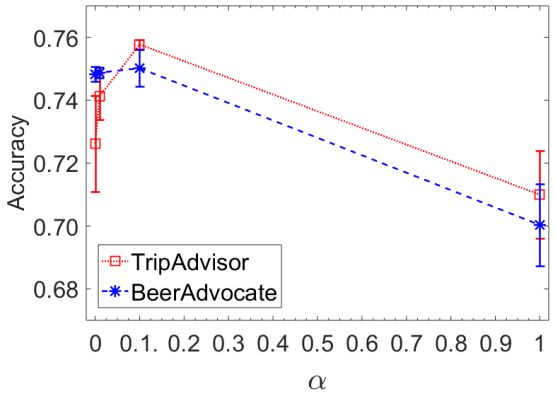

Figure 2: Parameter sensitivity analysis.

4.5 Parameter Sensitivity

We also conduct parameter sensitivity analysis of our approach. The parameter α in Equation (7) adjusts the expectation and entropy terms on the same scale. We test $\alpha = \{0, 0.01, 0.1, 1\}$ for both of the datasets. As we can see from Figure 2, $\alpha = 0.1$ is a good choice for both datasets.

4.6 Implementation Details

We implemented our models using TensorFlow (Abadi et al., 2016). For N-DMSC and LRR, we used code released by Yin et al. (2017) and Wang et al. (2010) respectively and followed their preprocessing steps and optimal settings.

Parameters are updated by using ADADELTA (Zeiler, 2012), an adaptive learning rate method. To avoid overfitting, we impose weight decay and drop out on both classifiers. The regularization coefficient and drop out rate are set to 10^{-3} and 0.3 respectively. The number of negative samples and α in our model are set to 10 and 0.1 respectively. For each document and each aspect, multiple target-opinion pairs are extracted. The opinion word classifier associated with an aspect will predict five target-opinion pairs at a time. These five target-opinion pairs are selected with bias. The

probability of a pair being selected is proportional to the frequency of the opinion word to the power of -0.25. In this way, opinion words with low frequency are more likely to be selected compared to the uniform sampling. In order to initialize both classifiers better, the word embeddings are retrofitted (Faruqui et al., 2015) using PPDB (Ganitkevitch et al., 2013) semantic lexicons.

5 Related Work

In this section, we review the related work on document-level multi-aspect sentiment classification, target-opinion word pairs extraction, and variational methods.

Document-level Multi-Aspect Sentiment Classification. Wang et al. (2010) proposed a LRR model to solve this problem. LRR assumes the overall polarity is a weighted sum of all aspect polarities which are represented by word frequency features. LRR needs to use aspect keywords to perform sentence segmentation to generate the representation of each aspect. To address the limitation of using aspect keywords, LARAM (Wang et al., 2011) assumes that the text content describing a particular aspect is generated by sampling words from a topic model corresponding to the latent aspect. Both LRR and LARAM can only access to overall polarities in the training data, but not gold standards of aspect polarities. Meng et al. (2018) proposed a weakly supervised text classification method which can take label surface names, class-related keywords, or a few labeled documents as supervision. Ramesh et al. (2015) developed a weakly supervised joint model to identify aspects and the corresponding sentiment polarities in online courses. They treat aspect (sentiment) related seed words as weak supervision. In the DMSC task which is a fine-grained text classification task, the label surface names or keywords for some aspects would be very similar. Given that the inputs are the same and the supervisions are similar, weakly supervised models cannot distinguish them. So we do not consider them as our baselines. Yin et al. (2017) modeled this problem as a machine comprehension problem under a multi-task learning framework. It also needs aspect keywords to generate aspect-aware document representations. Moreover, it can access gold standards of aspect polarities and achieved state-of-the-art performance on this task. Hence, it can serve as an upper bound. Some sentence-level aspect based sentiment classification methods (Wang et al., 2016b, 2018) can be directly applied to the DMSC task, because they can solve aspect category sentiment classification task. For example, given a sentence "the restaurant is expensive," the aspect category sentiment classification task aims to classify the polarity of the aspect category "price" to be *negative*. The aspect categories are predefined which are the same as the DMSC task. Some of them (Tang et al., 2016a,b; Chen et al., 2017; Ma et al., 2017) cannot because they are originally designed for aspect term sentiment classification task. For example, given a sentence "I loved their fajitas," the aspect term sentiment classification task aims to classify the polarity of the aspect term "fajitas" to be *positive*. The aspect terms appearing in the sentence should be provided as inputs.

Target Opinion Word Pairs Extraction. There are two kinds of methods, namely, rule based methods and learning based methods to solve this task. Rule based methods extract target-opinion word pairs by mining the dependency tree paths between target words and opinion words. Learning based methods treat this task as a sequence labeling problem, mapping each word to one of the following categories: target, opinion, and other.

(Hu and Liu, 2004) is one of earliest rule based methods to extract target-opinion pairs. An opinion word is restricted to be an adjective. Target words are extracted first, and then an opinion word is linked to its nearest target word to form a pair. Popescu and Etzioni (2005) and Bloom et al. (2007) manually designed dependency tree path templates to extract target-opinion pairs. If the path between a target word candidate and an opinion word candidate belongs to the set of path templates, the pair will be extracted. Qiu et al. (2011) identified dependency paths that link opinion words and targets via a bootstrapping process. This method only needs an initial opinion lexicon to start the bootstrapping process. Zhuang et al. (2006) adopted a supervised learning algorithm to learn valid dependency tree path templates, but it requires target-opinion pairs annotations.

Learning based methods require lots of target-opinion pairs annotations. They trained conditional random fields (CRF) (Lafferty et al., 2001) based models (Jakob and Gurevych, 2010; Yang and Cardie, 2012; Wang et al., 2016a) or deep neural networks (Liu et al., 2015; Wang et al., 2017; Li

and Lam, 2017) to predict the label (target, opinion or other) of each word. Jakob and Gurevych (2010) and Li et al. (2012) extracted target-opinion pairs without using using any labeled data in the domain of interest, but it needs lots of labeled data in another related domain.

In this paper, we only use very simple rules to extract target-opinion pairs to validate the effectiveness of our approach. If better pairs can be extracted, we can further improve our results.

Variational Methods. Variational autoencoders (Kingma and Welling, 2014; Rezende et al., 2014) (VAEs) use a neural network to parameterize a probability distribution. VAEs consists of an encoder which parameterizes posterior probabilities and a decoder which parameterizes the reconstruction likelihood given a latent variable. VAEs inspire many interesting works (Titov and Khoddam, 2015; Marcheggiani and Titov, 2016; Šuster et al., 2016; Zhang et al., 2018; Chen et al., 2018) which are slightly different from VAEs. Their encoders produce a discrete distribution while the encoder in VAEs yields a continuous latent variable. Titov and Khoddam (2015) aimed to solve semantic role labeling problem. The encoder is essentially a semantic role labeling model which predicts roles given a rich set of syntactic and lexical features. The decoder reconstructs argument fillers given predicted roles. Marcheggiani and Titov (2016) aimed to solve unsupervised open domain relation discovery. The encoder is a feature-rich relation extractor, which predicts a semantic relation between two entities. The decoder reconstructs entities relying on the predicted relation. Šuster et al. (2016) tried to learn multi-sense word embeddings. The encoder uses bilingual context to choose a sense for a given word. The decoder predicts context words based on the chosen sense and the given word. Zhang et al. (2018) aimed to solve knowledge graph powered question answering. Three neural networks are used to parameterize probabilities of a topic entity given a query and an answer, an answer based on a query and a predicted topic, and the topic given the query. Chen et al. (2018) aimed to infer missing links in a knowledge graph. Three neural networks are used to parameterize probabilities of a latent path given two entities and a relation, a relation based on two entities and the chosen latent path, and the relation given the latent path. Our method also uses neural networks to parameterize two discrete distributions but aims to solve the DMSC task.

6 Conclusion

In this paper, we propose a variational approach to weakly supervised DMSC. We extract many target-opinion word pairs from dependency parsers using simple rules. These pairs can be "supervision" signals to predict sentiment polarity. Our objective function is to predict an opinion word given a target word. After introducing the sentiment polarity as a latent variable, we can learn a sentiment polarity classifier by optimizing the variational lower bound. We show that we can outperform weakly supervised baselines by a large margin and achieve comparable results to the supervised method with hundreds of labels per aspect, which can reduce a lot of labor work in practice. In the future, we plan to explore better target-opinion word extraction approaches to find better "supervision" signals.

Acknowledgments

This paper was supported by the Early Career Scheme (ECS, No. 26206717) from Research Grants Council in Hong Kong. Ziqian Zeng has been supported by the Hong Kong Ph.D. Fellowship. We thank Intel Corporation for supporting our deep learning related research. We also thank the anonymous reviewers for their valuable comments and suggestions that help improve the quality of this manuscript.

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *Proceedings of OSDI*, pages 265–283.

Kenneth Bloom, Navendu Garg, Shlomo Argamon, et al. 2007. Extracting appraisal expressions. In *Proceedings of NAACL-HLT*, pages 308–315.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP*, pages 740–750.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of EMNLP*, pages 452–461.

Wenhu Chen, Wenhan Xiong, Xifeng Yan, and William Wang. 2018. Variational knowledge graph reasoning. In *Proceedings of NAACL-HLT*, pages 1823–1832.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, pages 1606–1615.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of KDD*, pages 168–177.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *Proceedings of EMNLP*, pages 1035–1045.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *ICLR*.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics*, 2(1-2):83–97.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289.

Himabindu Lakkaraju, Richard Socher, and Chris Manning. 2014. Aspect specific sentiment analysis using hierarchical deep learning. In *Proceedings of NIPS workshop on Deep Learning and Representation Learning*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of EMNLP*, pages 107–117.

Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of ACL*, pages 410–419.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of ACL*, pages 1106–1115.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of EMNLP*, pages 2886–2892.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of EMNLP*, pages 1433–1443.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of IJCAI*, pages 4068–4074.

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Transactions of the Association for Computational Linguistics*, 4:231–244.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *Proceedings of ICDM*, pages 1020–1025.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 983–992. ACM.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Aitor García Pablos, Montse Cuadros, and German Rigau. 2015. V3: Unsupervised aspect based sentiment analysis for semeval2015 task 12. In *Proceedings of SemEval*, pages 714–718.

Ana-Maria Popescu and Orena Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of EMNLP-HLT*, pages 339–346.

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

Arti Ramesh, Shachi H Kumar, James Foulds, and Lise Getoor. 2015. Weakly supervised models of aspect-sentiment for online course discussion forums. In *Proceedings of ACL*, pages 74–83.

Jonathon Read and John Carroll. 2009. Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of CIKM workshop on Topic-sentiment Analysis for Mass Opinion*, pages 45–52. ACM.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of ICML*, pages 1278–1286.

Simon Šuster, Ivan Titov, and Gertjan van Noord. 2016. Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Proceedings of NAACL-HLT*, pages 1346–1356.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING*, pages 3298–3307.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of EMNLP*, pages 214–224.

Ivan Titov and Ehsan Khoddam. 2015. Unsupervised induction of semantic roles within a reconstruction-error minimization framework. In *Proceedings of NAACL-HLT*, pages 1–10.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of KDD*, pages 783–792.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of KDD*, pages 618–626.

Jingjing Wang, Jie Li, Shoushan Li, Yangyang Kang, Min Zhang, Luo Si, and Guodong Zhou. 2018. Aspect sentiment classification with both word-level and clause-level attention networks. In *IJCAI*, pages 4439–4445.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016a. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of EMNLP*, pages 616–626.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of AAAI*, pages 3316–3322.

Yequan Wang, Minlie Huang, Li Zhao, et al. 2016b. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of EMNLP*, pages 606–615.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of EMNLP-HLT*, pages 347–354.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *Proceedings of EMNLP-CoNLL*, pages 1335–1345.

Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of EMNLP*, pages 2034–2044.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *Proceedings of AAAI*, pages 6069–6076.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of CIKM*, pages 43–50.