EMOVA: A Semi-supervised End-to-End Moving-Window Attentive Framework for Aspect Mining

Ning Li, Chi-Yin Chow(✉), and Jia-Dong Zhang

Department of Computer Science,
City University of Hong Kong, Kowloon, Hong Kong
ning.li@my.cityu.edu.hk, {chiychow,jzhang26}@cityu.edu.hk

Abstract. Aspect mining or extraction is one of the most challenging problems in aspect-level analysis on customer reviews; it aims to extract terms from a review describing aspects of a reviewed entity, e.g., a product or service. As aspect mining can be formulated as the sequence labeling problem, supervised deep sequence learning models have recently achieved the best performance. However, these supervised models require a large amount of labeled data which are usually very costly or unavailable. To this end, we propose a semi-supervised End-to-end MOVing-window Attentive framework (called **EMOVA**) that has three key features for aspect mining. (1) Two neural layers with Bidirectional Long Short-Term Memory (BiLSTM) are employed to learn representations of reviews. (2) Cross-View Training (CVT) is used to improve the representation learning over a small set of labeled reviews and a large set of unlabeled reviews from the same domain in a unified end-to-end architecture. (3) Since past nearby information in a text provides important semantic contexts for a prediction task in aspect mining, a moving-window attention component is proposed in EMOVA to enhance prediction accuracy. Experimental results over four review datasets from the SemEval workshops show that EMOVA outperforms the state-of-the-art models for aspect mining.

Keywords: Aspect mining · Semi-supervised learning · Cross-View training · Moving-window attention · End-to-end learning

1 Introduction

To achieve aspect-level analysis on product or service reviews, the first task is aspect mining (or aspect extraction), which aims to extract aspect terms from a review, e.g., "operating system" and "preloaded software" from a laptop's review *"I love the operating system and preloaded software"*. Existing aspect mining techniques can be divided into three categories, namely unsupervised, supervised, and semi-supervised.

Unsupervised learning models based on Latent Dirichlet Allocation (LDA) [13,36] and word embeddings [9] do not need labeled reviews. However, it is hard to

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12085, pp. 811–823, 2020.
https://doi.org/10.1007/978-3-030-47436-2_61

control a totally unsupervised model to only show the concerned aspects. Supervised sequential learning methods such as Hidden Markov Models (HMM) [12] and Conditional Random Fields (CRF) [11,29] are applied to extract aspects, as the task can be formulated as a sequence labeling problem. Currently, some supervised deep learning models [17,25,31,32] can achieve better performances than previous works by introducing additional supervision from lexicons and other hand-crafted features. However, we insist that the automated feature learning is always preferred. Moreover, because the manual annotation of training data is usually very costly, especially for domain dependent aspects (i.e., different domains may have different aspect spaces), researchers are motivated to develop more effective semi-supervised learning models for aspect mining.

Semi-supervised approaches include two directions, one is to guide the unsupervised models by encoding prior domain knowledge [2,3,15,20], and the other is to enhance the supervised models with unlabeled reviews in corresponding domains [33]. The latter approach outperforms the former as it benefits from both labeled and unlabeled reviews. However, the existing model [33] is trained in two separated phases: pre-train on unlabeled review in corresponding domains; and then perform supervised learning on labeled reviews. The representations (or embeddings) learned in the pre-training phase do not take advantages of labeled reviews, i.e., they only learn domain specific but task free representations. Our consideration is whether we can learn task and domain specific representations from both labeled and unlabeled reviews at the same time and perform aspect mining in an end-to-end architecture.

In this paper, we propose a new semi-supervised End-to-end MOVing-window Attentive framework (called **EMOVA**) to enhance aspect mining on customer reviews. Instead of separately pre-training and supervised learning, EMOVA alternately learns a model on a mini-batch of labeled reviews and unlabeled reviews from the same domain based on Cross-View Training (CVT) [5]. Specifically, EMOVA derives the representations of reviews based on two neural layers with Bidirectional Long Short-Term Memory (BiLSTM) [8] by considering two important observations in reviews: (1) Customer reviews often contain misspelling words; (2) Multiple aspects may coordinately appear in one sentence. To this end, EMOVA derives char-features from words as extra embeddings, because general pre-trained word embeddings (e.g., GloVe [21]) may not cover all misspelling words. Moreover, the past nearby words provide useful semantic clues for finding new aspects. For instance, under the coordinate structure, the previous aspect (e.g., "operating system") should be more significant than other words to guide the extraction of subsequent aspects (e.g., "preloaded software"). To capture these context significances, EMOVA employs an attention mechanism to encode the information within a moving-window.

In general, the contributions of this paper can be summarized as below.

– We are the first to propose a semi-supervised deep learning framework for aspect mining, which introduces CVT to use unlabeled reviews to improve the representation learning within a unified end-to-end architecture.

– We first attempt to develop a moving-window attention mechanism after two BiLSTM layers to capture significant past nearby information for the aspect prediction.
– We conduct extensive experiments to evaluate the performance of EMOVA based on four real-world review datasets. Experimental results show that EMOVA performs better than the state-of-the-art techniques.

The reminder of this paper is organized as follows. Section 2 discusses related works. Then, we present our EMOVA framework in Sect. 3. Section 4 shows the experimental results. Finally, Sect. 5 concludes this paper.

2 Related Works

2.1 Aspect Mining as Sequence Labeling

Sequence labeling is a very common problem in natural language processing (e.g., part-of-speech tagging and named-entity recognition) and aims to assign a label to each element in a sequential input. The aspect mining task can be formulated as a sequence labeling problem, in which a label (whether an aspect or not) is given to each word in the review. Formally, the problem can be described as predicting a label sequence $\{y_1...y_n\}$ for a given word sequence $\{x_1...x_n\}$, where $y_i \in \{ASPECT,\ NONASPECT\}$. For instance, the reference [12] defines a set of labels to distinguish feature aspects, component aspects and function aspects, and train HMM to label each word in the review. However, the researchers [11] simplify these labels and apply $\{B, I, O\}$ scheme, where B identifies the beginning of an aspect, I for the continuation of the aspect, and O for other words. The $\{B, I, O\}$ scheme can well handle aspects expressing in phrases and has been applied for aspect mining [17,33] and aspect-opinion term co-extraction [31,32]. Our EMOVA also uses the same $\{B, I, O\}$ labeling scheme.

2.2 Semi-supervised Approaches

Our EMOVA framework relates to the semi-supervised models for aspect mining. Most existing methods use prior knowledge to guide an unsupervised topic model. For instance, some methods manually choose domain specified seed words [15,20] or *must* and *cannot* sets [3] for topic modeling. By introducing lifelong topic modeling [2], researchers propose a continually modeling system that can automatically mine knowledge from previous results to supervise the following tasks. However, this kind of methods often need manually defined domain knowledge and do not fully use existing labeled reviews. Another direction of semi-supervised learning is to take the advantage of unlabeled reviews in the same domain to improve the supervised model. The idea of pre-training has been applied in the aspect mining model [33] to learn domain specific word embeddings from unlabeled reviews in advance; these word embeddings have better representations than the general word embeddings and are fed into normal supervised models. However, these pre-trained domain specific representations are still not

specific enough for the aspect mining task. Nevertheless, our EMOVA framework can learn both task and domain specific representations of reviews in an unified framework, which then enhance the aspect mining.

2.3 Cross-View Training

Normally, a deep learning model works best when trained on a large amount of data with reliable labels. However, for domain (or even entity) dependent aspects, manual annotation could be a huge investment. One solution is to apply effective semi-supervised learning to leverage unlabeled reviews. Current semi-supervised deep learning models separate the training process into two phases: pre-training and supervised learning. A key disadvantage of such models is that the first phase on representation learning does not benefit from labeled reviews.

Cross-View Training (CVT) [5] semi-supervises the learning by alternately switching the training process on labeled data and unlabeled data. It restricts the views on input data while training on unlabeled examples. Through auxiliary prediction modules, CVT can improve the representation learning of the supervised model. The idea of CVT is as follows: (1) A primary prediction module is trained with the standard supervised learning on labeled examples; (2) On unlabeled examples, a number of auxiliary prediction modules with different views on the input data are trained to agree with the primary prediction module; (3) By alternatively training on labeled data and unlabeled data, both representation learning and prediction modules get improved. Our EMOVA framework is based on the idea of CVT but has one more task specific architecture (e.g., moving-window attentions on two BiLSTM layers) for aspect mining.

3 The Framework EMOVA

In this section, we present our semi-supervised deep learning framework for aspect mining. First, we formulate the aspect mining task into a sequence labeling problem. Then, we present the technical details of the four key components in EMOVA. The architecture of our EMOVA framework is shown in Fig. 1.

3.1 Problem Statement

Suppose we have a set of labeled (D_l) and unlabeled (D_u) reviews for an entity. The aspect mining task is to learn a classifier from both D_l and D_u to extract a set of aspects for the entity. This task can be formulated as a sequence labeling problem by using $\{B, I, O\}$ scheme, where B, I, and O indicate the beginning of, the continuation of, and the out of the aspect, respectively (refer to Sect. 2.1). Each word x_i in the sentence $X = \{x_1, ..., x_T\}$ must be assigned as one of $\{B, I, O\}$. For instance, the input sentence *"I love the operating system and preloaded software"* may have the label sequence of $\{O, O, O, B, I, O, B, I\}$.

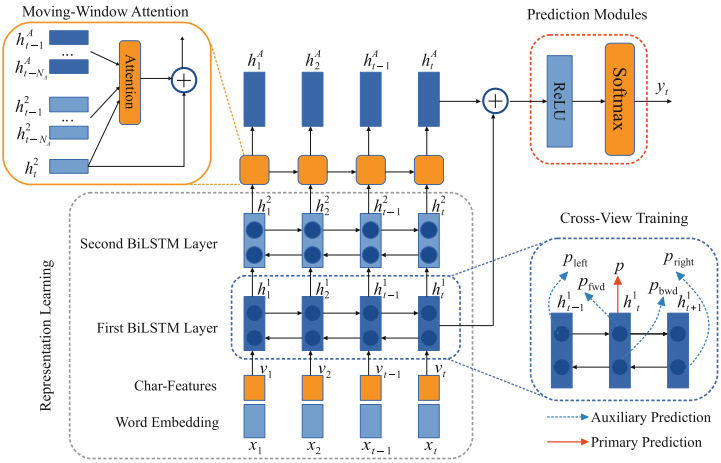

Fig. 1. The architecture of our EMOVA framework.

3.2 Representation Learning

As recurrent neural networks can naturally represent the sequential information, our framework employs BiLSTM [8] to build the memory of contextualized representations for sequence labeling in the aspect mining task. Because combining general embeddings and char-features can help to handle misspelling words [19], we represent each word in the input sequence as the concatenation of an embedding vector and the char-feature from the output of a character-level Convolutional Neural Network (CNN). Further, the concatenation vector is fed into two BiLSTM layers which often achieve the best performance on building the memory in many sequential tasks [26]. Let $V = \{v_1, ..., v_T\}$ be the concatenation vectors of words. Their hidden representations are derived by concatenating the outputs of both forward $\overrightarrow{LSTM}$ and backward $\overleftarrow{LSTM}$ as follows:

$$h_t^1 = [\overrightarrow{LSTM}(v_t) \oplus \overleftarrow{LSTM}(v_t)], t \in [1, T], \text{ and} \tag{1}$$

$$h_t^2 = [\overrightarrow{LSTM}(h_t^1) \oplus \overleftarrow{LSTM}(h_t^1)], t \in [1, T], \tag{2}$$

in which $\oplus$ denotes the concatenation operation, h_t^1 is the hidden representations from the first BiLSTM layer at time step t, and h_t^2 is from the second layer.

3.3 Moving-Window Attention

In the aspect labeling task, the information from past nearby steps provide useful clues for a prediction, e.g., the label "I" cannot follow "O", and the previous aspects can guide the extraction of subsequent aspects. To capture such important past nearby information, our framework develops a moving-window attention component [16] after the two-layer BiLSTM network, while the attention mechanisms have become an essential component for various tasks

to model significances and dependencies of sequential terms [30]. Specifically, the moving-window attention only caches the most recent N_A hidden states. At step t, we calculate the normalized significance score s_i^t of each cached state h_i^2 ($i \in [t - N_A, t - 1]$) as follows:

$$s_i^t = Softmax(U^A \cdot tanh(W_1^A h_i^2 + W_2^A h_t^2 + W_3^A h_i^A)), \tag{3}$$

where $tanh$ is the activation function, h_i^2 and h_t^2 denote the cached past state and current state from the second BiLSTM layer, and h_i^A denotes the previous attentive representations in the moving-window. U^A, W_1^A, W_2^A, and W_3^A are the model parameters.

To calculate current moving-window attentive aspect representation h_t^A at step t, our framework computes the weighted sum of the cached previous moving-window attentive aspect representations h_i^A with the score weights s_i^t, applies the ReLU activation function, and stacks the result on current state h_t^2, given by

$$h_t^A = h_t^2 + ReLU(\sum_{i=t-N_A}^{t-1} s_i^t \times h_i^A). \tag{4}$$

3.4 Prediction Modules

In our framework, CVT trains labeled data with a primary prediction module. Suppose y_t is the label for the word $x_t \in X$. The primary prediction module determines the probability distribution $p(y_t|x_t)$ over labels from the results of the first BiLSTM layer (h_t^1) and moving-window attention layer (h_t^A) with a simple one-hidden-layer neural network (denoted by nn), given by

$$p(y_t|x_t) = nn(h_t^1 \oplus h_t^A) = Softmax(U^P \cdot ReLU(W^P(h_t^1 \oplus h_t^A)) + b), \tag{5}$$

where U^P and W^P are the model parameters.

Further, the proposed framework shares the first BiLSTM layer with the auxiliary prediction modules that have restricted views of unlabeled reviews. There are four different auxiliary prediction modules (p_{left}, p_{fwd}, p_{bwd}, and p_{right}) in the framework, where p_{left} means, for the prediction of current word, this module only has a view of all past words on the left of current word in the sentence; p_{fwd} has a view of left and current words; p_{bwd} sees current and words on the right; and p_{right} only sees all future words on the right, as shown in Fig. 1. BiLSTM can easily provide these restricted views without additional computation as follows:

$$p_{\text{left}}(y_t|x_t) = nn_{\text{left}}(\overrightarrow{h}^1_{t-1}),\ p_{\text{fwd}}(y_t|x_t) = nn_{\text{fwd}}(\overrightarrow{h}^1_t),$$

$$p_{\text{bwd}}(y_t|x_t) = nn_{\text{bwd}}(\overleftarrow{h}^1_t),\ \text{and}\ p_{\text{right}}(y_t|x_t) = nn_{\text{right}}(\overleftarrow{h}^1_{t+1}), \tag{6}$$

where nn_{left}, nn_{fwd}, nn_{bwd}, and nn_{right} denote neural networks with the same structure given in Eq. 5. Since the second BiLSTM layer has already seen all words, we can only feed the hidden representations $\overrightarrow{h}^1$ and $\overleftarrow{h}^1$ from the first BiLSTM layer to the auxiliary prediction modules in order to restrict their view on an input sequence.

3.5 Cross-View Training

The key idea of CVT is to use unlabeled reviews from the same domain of labeled reviews to enhance the representation learning. During CVT, the model alternately learns on a mini-batch of labeled reviews or unlabeled reviews.

For the labeled reviews D_l, the Cross-Entropy (CE) loss is utilized to train the primary prediction module $p(y_t|x_t)$:

$$L_{\text{SUP}} = \frac{1}{D_l} \sum\nolimits_{x_t, y_t \in D_l} CE(y_t, p(y_t|x_t)). \tag{7}$$

For the unlabeled reviews D_u, the framework first infers $p(y_i|x_i)$ $(x_i \in D_u)$ based on the primary prediction module and then trains the auxiliary prediction modules to match the primary prediction module by using the Kullback-Leibler (KL) divergence function as the loss:

$$L_{\text{CVT}} = \frac{1}{D_u} \sum\nolimits_{x_i \in D_u} \sum\nolimits_{j} KL(p(y_i|x_i), p_j(y_i|x_i)), \tag{8}$$

where $j \in \{\text{left}, \text{fwd}, \text{bwd}, \text{right}\}$ and the parameters of the primary prediction module are fixed during training. The auxiliary prediction modules can enhance the shared representations, because the new terms that are not in labeled reviews may have been encoded into the model and be useful for making predictions on some new aspects.

Further, we combine the supervised and CVT losses and minimize the total loss L with stochastic gradient descent:

$$L = L_{\text{SUP}} + L_{\text{CVT}}. \tag{9}$$

In particular, we alternately minimize L_{sup} over a mini-batch of labeled reviews and L_{CVT} over a mini-batch of unlabeled reviews.

4 Experiments

In this section, we evaluate the performance of our proposed EMOVA framework and compare it with the state-of-the-art supervised and semi-supervised approaches.

4.1 Experimental Settings

Datasets: We conduct experiments over four benchmark datasets from the SemEval workshops [22–24]. Table 1 shows their statistics. D_{laptop1} and D_{laptop2} contain reviews of the laptop domain, while D_{rest1} and D_{rest2} are for the restaurant domain. In these datasets, aspect words have been labeled by the task organizer.

The framework EMOVA needs unlabeled reviews for CVT. We collect unlabeled reviews corresponding to four labeled training datasets to train the

Table 1. Statistics of datasets.

	D_{laptop1}		D_{laptop2}		D_{rest1}		D_{rest2}	
	Train	Test	Train	Test	Train	Test	Train	Test
Number of sentences	3,045	800	3,041	800	1,315	685	2,000	675
Number of labeled aspects	2,358	654	1,743	1,134	1,192	542	1,743	622

model, which include laptop reviews from Amazon Review Dataset (230,373 sentences) [10] and restaurant reviews from Yelp Review Dataset (2,677,025 sentences) [34]. For comparison, we also train the model on a general unlabeled dataset (One Billion Word Language Model Benchmark) [1] to see whether performing CVT on general texts can improve the supervised model for aspect mining. As some sentences in the testing dataset may also appear in unlabeled reviews, we remove these sentences in unlabeled reviews to make the comparison fair.

Baselines: We compare our EMOVA with four groups of baselines. The first group is the winner of each dataset in the SemEval workshops, including **IHS_RD** [4] (D_{laptop1} winner), **DLIREC** [29] (D_{laptop2} winner), **EliXa** [27] (D_{rest1} winner), and **NLANGP** [28] (D_{rest2} winner). The second group is traditional supervised models including:

- **CRF** [14] is the most commonly used method for sequence labeling.
- **WDEmb** [35] is an enhanced CRF model with word embeddings, context embeddings, and dependency embeddings.
- **LSTM** [18] is a vanilla BiLSTM with domain embeddings.

The third group takes the advantages of gold-standard opinion terms, sentiment lexicons, and other additional resources for training.

- **CMLA** [32] applies a multi-layer architecture with coupled-attentions to model aspects and opinion words.
- **MIN** [17] consists of three LSTM layers for multi-task learning, in which a sentiment lexicon and dependency rules are used to find opinion words.
- **DE-CNN** [33] is the state-of-the-art model based on CNN and utilizes both general word embeddings and domain-specific embeddings for aspect mining.
- **BERT** [6] is one of the key innovations in the recent progress of language modeling and achieves the state-of-the-art performance on many natural language processing tasks, we fine-tune $\text{BERT}_{\text{BASE}}$ on the datasets as a baseline.

The fourth group is the variants of EMOVA.

- **EMOVA-S** is our supervised model but without CVT on unlabeled data, so it is a purely supervised learning model.
- **EMOVA-G** only performs CVT on the general unlabeled text (One Billion Word Language Model Benchmark) [1] which is not specific to the laptop or restaurant domain.

We report the results of these baselines in their original works, since we use exactly the same datasets.

Training Settings: We use pre-trained GloVe 840B 300-dimension vectors [21] to initialize the word embeddings, and the char-feature size is 50. All of the weight matrices except those in LSTMs are initialized from the uniform distribution $U(-0.2, 0.2)$. For the initialization of the matrices in LSTMs, we adopt the Glorot Uniform strategy [7]. We apply dropout while the rates are set as 0.5 for labeled reviews and 0.8 for unlabeled reviews. The hidden state size is set to 300, and the learning rate is 0.05. We set the mini-batch size as 50 sentences, and the moving-window size (i.e., the number of cached past nearby aspect representations) N_A is 5.

4.2 Experimental Results

Main Results: We report F1 score (%) in the Table 2. The result shows that EMOVA performs the best. Compared to those challenge winners (**IHS_RD** on D_{laptop1}, **DLIREC** on D_{laptop2}, **EliXa** on D_{rest1}, and **NLANGP** on D_{rest2}), EMOVA achieves absolute gains of 7.17%, 1.79%, 2.22%, and 2.84%, respectively. Even **EMOVA-S** (without CVT) can perform better than those supervised baselines in the first and second groups on three of the four datasets (except the second laptop dataset). The main reason should be the effectiveness of our moving-window attention layer which can help to discover some aspects under the guidance of frequent aspects in coordinate structures. The result also shows that **EMOVA-G** with general unlabeled texts can improve the pure supervised model **EMOVA-S**.

Table 2. Comparison results in F1 score.

	Models	D_{laptop1}	D_{laptop2}	D_{rest1}	D_{rest2}		Models	D_{laptop1}	D_{laptop2}	D_{rest1}	D_{rest2}
1	**IHS_RD**	74.55	79.62	–	–	3	**CMLA**	77.80	85.29	70.73	72.77
	DLIREC	73.78	84.01	–	–		**MIN**	77.58	–	–	73.44
	EliXa	–	–	70.04	–		**DE-CNN**	81.59	–	–	74.37
	NLANGP	–	–	67.12	–		**BERT**	78.71	85.12	70.85	73.23
2	**CRF**	74.01	82.33	67.54	69.56	4	**EMOVA-S**	77.32	83.48	70.10	72.35
	WDEmb	75.16	84.97	69.73	–		**EMOVA-G**	77.89	84.22	71.43	73.62
	LSTM	75.17	82.01	68.26	70.35		**EMOVA**	**81.72**	**85.80**	**72.26**	**75.18**

The third group of baselines is considered as some special cases of semi-supervised learning, as they all rely on additional resources (e.g., hand-craft features, lexicons, pre-trained domain embeddings, and pre-trained language models) to improve the performance. In the pre-training step of these two-phase models (e.g., **DE-CNN** and **BERT**), they do not take advantage of labeled reviews. More specifically, **BERT** learns better representations by training a

Table 3. Ablation study on the key components of EMOVA.

Models	D_{laptop1}	D_{laptop2}	D_{rest1}	D_{rest2}
EMOVA	**81.72**	**85.80**	**72.26**	**75.18**
W/o char-features	−0.06	−0.04	−0.07	−0.06
W/o attentions	−1.59	−1.73	−1.30	−1.16
W/o fwd & bwd	−0.32	−0.14	−0.21	−0.27
W/o left & right	−0.43	−0.55	−0.51	−0.60

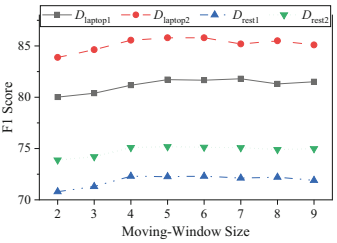

Fig. 2. Effects of the moving-window size N_A.

deep language model on large amounts of texts, and **DE-CNN** attempts to learn domain-specific but general-purpose representations rather than both domain and task specific representations in our EMOVA. As a result, EMOVA works better than the two-phase (i.e., pre-training and supervised learning) models.

Ablation Study: The key components of EMOVA include char-features, BiLSTM layers, moving-window attentions, primary and auxiliary prediction modules, as shown in Fig. 1. To show the significance of each component, we remove each of them and evaluate the F1 score, as depicted in Table 3. Firstly, we disable the char-features and the result shows only slight effect in the row for **w/o char-features**. Then, we remove the moving-window attention layer and the result drops significantly on all datasets in the row for **w/o attentions**, which shows the essentiality of moving-window attentions. To explore which auxiliary prediction modules are more important, we only enable two of them (p_{fwd} and p_{bwd}, or p_{left} and p_{right}) at each time. We find that EMOVA **w/o fwd & bwd** that do not see the current word is better than EMOVA **w/o left & right**, which may be caused by the more restricted view on the unlabeled input.

Effects of the Moving-Window Size: We also evaluated the effects of the size of moving-window in the attention layer of our EMOVA framework, the results are shown in Fig. 2. It is hard to improve the overall performance by simply increasing the moving-window size, i.e., EMOVA can achieve better aspect mining accuracy by focusing attention on a certain number of past nearby words. To reduce the computation cost, the moving-window size N_A is set to 5 in our experiments.

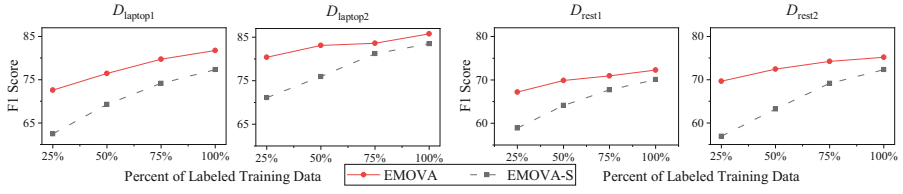

Fig. 3. Performance vs. percent of the labeled training set.

Less Labeled Training Data: A very common situation in aspect mining is some domains (or products) may not have large volumes of labeled data. To this end, we explore how EMOVA scales with less data by only feeding a subset (25%, 50%, 75%) of the labeled training data, as presented in Fig. 3. EMOVA with half of the training data can perform as well as **EMOVA-S** without CVT that sees all the training data. Thus, EMOVA is particularly useful when only a small set of labeled reviews is available, which greatly reduces the cost on manual annotations.

5 Conclusion

In this paper, we have proposed the first semi-supervised End-to-end MOVing-window Attentive framework (EMOVA) for aspect mining on customer reviews. The framework derives the representations of reviews based on two neural layers with Bidirectional Long Short-Term Memory (BiLSTM). The Cross-View Training (CVT) is employed to train auxiliary prediction modules on unlabeled reviews to improve the representation learning in a unified end-to-end architecture. Further, EMOVA exploits the moving-window attention mechanism to capture significant past nearby semantic contexts. Experimental results over four datasets from SemEval workshops show that EMOVA outperforms the state-of-the-art models, even on small labeled training datasets.

References

1. Chelba, C., et al.: One billion word benchmark for measuring progress in statistical language modeling. In: Proceedings of INTERSPEECH (2013)
2. Chen, Z., Liu, B.: Topic modeling using topics from many domains, lifelong learning and big data. In: Proceedings of ICML (2014)
3. Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting domain knowledge in aspect extraction. In: Proceedings of EMNLP (2013)
4. Chernyshevich, M.: IHS R&D Belarus: cross-domain extraction of product features using CRF. In: Proceedings of SemEval (2014)
5. Clark, K., Luong, M.T., Manning, C.D., Le, Q.V.: Semi-supervised sequence modeling with cross-view training. In: Proceedings of EMNLP (2018)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT (2018)
7. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of AISTATS (2010)
8. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
9. He, R., Lee, W.S., Ng, H.T., Dahlmeier, D.: An unsupervised neural attention model for aspect extraction. In: Proceedings of ACL (2017)
10. He, R., McAuley, J.: Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proceedings of WWW (2016)

11. Jakob, N., Gurevych, I.: Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In: Proceedings of EMNLP (2010)
12. Jin, W., Ho, H.H., Srihari, R.K.: A novel lexicalized HMM-based learning framework for web opinion mining. In: Proceedings of ICML (2009)
13. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proceedings of WSDM (2011)
14. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML (2001)
15. Li, N., Chow, C.Y., Zhang, J.: Seeded-BTM: enabling biterm topic model with seeds for product aspect mining. In: Proceedings of IEEE HPCC/SmartCity/DSS (2019)
16. Li, X., Bing, L., Li, P., Lam, W., Yang, Z.: Aspect term extraction with history attention and selective transformation. In: Proceedings of IJCAI (2018)
17. Li, X., Lam, W.: Deep multi-task learning for aspect term extraction with memory interaction. In: Proceedings of EMNLP (2017)
18. Liu, P., Joty, S., Meng, H.: Fine-grained opinion mining with recurrent neural networks and word embeddings. In: Proceedings of EMNLP (2015)
19. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of ACL (2016)
20. Mukherjee, A., Liu, B.: Aspect extraction through semi-supervised modeling. In: Proceedings of ACL (2012)
21. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: Proceedings of EMNLP (2014)
22. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: Semeval-2015 task 12: aspect based sentiment analysis. In: Proceedings of SemEval (2015)
23. Pontiki, M., et al.: Semeval-2016 task 5: aspect based sentiment analysis. In: Proceedings of SemEval (2016)
24. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: Semeval-2014 task 4: aspect based sentiment analysis. In: Proceedings of SemEval (2014)
25. Poria, S., Cambria, E., Gelbukh, A.: Aspect extraction for opinion mining with a deep convolutional neural network. Knowl.-Based Syst. **108**, 42–49 (2016)
26. Reimers, N., Gurevych, I.: Optimal hyperparameters for deep LSTM-networks for sequence labeling tasks. arXiv preprint arXiv:1707.06799 (2017)
27. San Vicente, I., Saralegi, X., Agerri, R., Sebastián, D.S.: EliXa: a modular and flexible ABSA platform. In: Proceedings of SemEval (2015)
28. Toh, Z., Su, J.: NLANGP at semeval-2016 task 5: improving aspect based sentiment analysis using neural network features. In: Proceedings of SemEval (2016)
29. Toh, Z., Wang, W.: Dlirec: aspect term extraction and term polarity classification system. In: Proceedings of SemEval (2014)
30. Vaswani, A., et al.: Attention is all you need. In: Proceedings of Advances in NIPS (2017)
31. Wang, W., Pan, S.J., Dahlmeier, D., Xiao, X.: Recursive neural conditional random fields for aspect-based sentiment analysis. In: Proceedings of EMNLP (2016)
32. Wang, W., Pan, S.J., Dahlmeier, D., Xiao, X.: Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In: Proceedings of AAAI (2017)
33. Xu, H., Liu, B., Shu, L., Yu, P.S.: Double embeddings and CNN-based sequence labeling for aspect extraction. In: Proceedings of ACL (2018)
34. Yelp Dataset: Yelp Dataset Challenge. https://www.yelp.com/dataset/challenge. Accessed 05 Mar 2019

35. Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M., Zhou, M.: Unsupervised word and dependency path embeddings for aspect term extraction. In: Proceedings of IJCAI (2016)
36. Zhao, W.X., Jiang, J., Yan, H., Li, X.: Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of EMNLP (2010)