香港城市大學
City University of Hong Kong

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

Author: Ning Li, Chi-Yin Chow, Jia-Dong Zhang

Presenter: Ning Li

City University of Hong Kong

Aspect Mining (An Example)

香港城市大學
City University of Hong Kong

- Given a laptop's review:

  "I love the operating system and preloaded software."

- Our goal: find the aspects of the reviewed entity,

  e.g., *operating system* and *preloaded software*.

Introduction

香港城市大學
City University of Hong Kong

- Aspect mining as sequence labeling

I	love	the	operating	system	and	preloaded	software
O	O	O	B	I	O	B	I

  - **B**: the beginning of the aspect;
  - **I**: the continuation of the aspect;
  - **O**: the out of the aspect.

Introduction

香港城市大學
City University of Hong Kong

- Aspect Mining
  - Unsupervised: (e.g., LDA, word embeddings)
    - Do not need labeled reviews, however, it is hard to control a totally unsupervised model to only show the concerned aspects.
  - Supervised: (e.g., HMM, CRF, RNN, CNN)
    - Manual annotation of training data is usually very costly, especially for deep learning models on domain dependent aspects (i.e., different domains may have different aspect spaces).

Introduction

香港城市大學
City University of Hong Kong

- Semi-supervised Aspect Mining
  - One direction: to guide the unsupervised models by encoding prior domain knowledge, e.g., Seeded-BTM uses a set of seed words to help to extract related aspects.
  - The other direction: to enhance the supervised models with unlabeled reviews in corresponding domains, e.g., DE-CNN learns domain-specific word vectors from unlabeled reviews; and combines them with general embedding to train a CNN classifier. We call it two-phase (pre-training and supervised learning) model.

Introduction

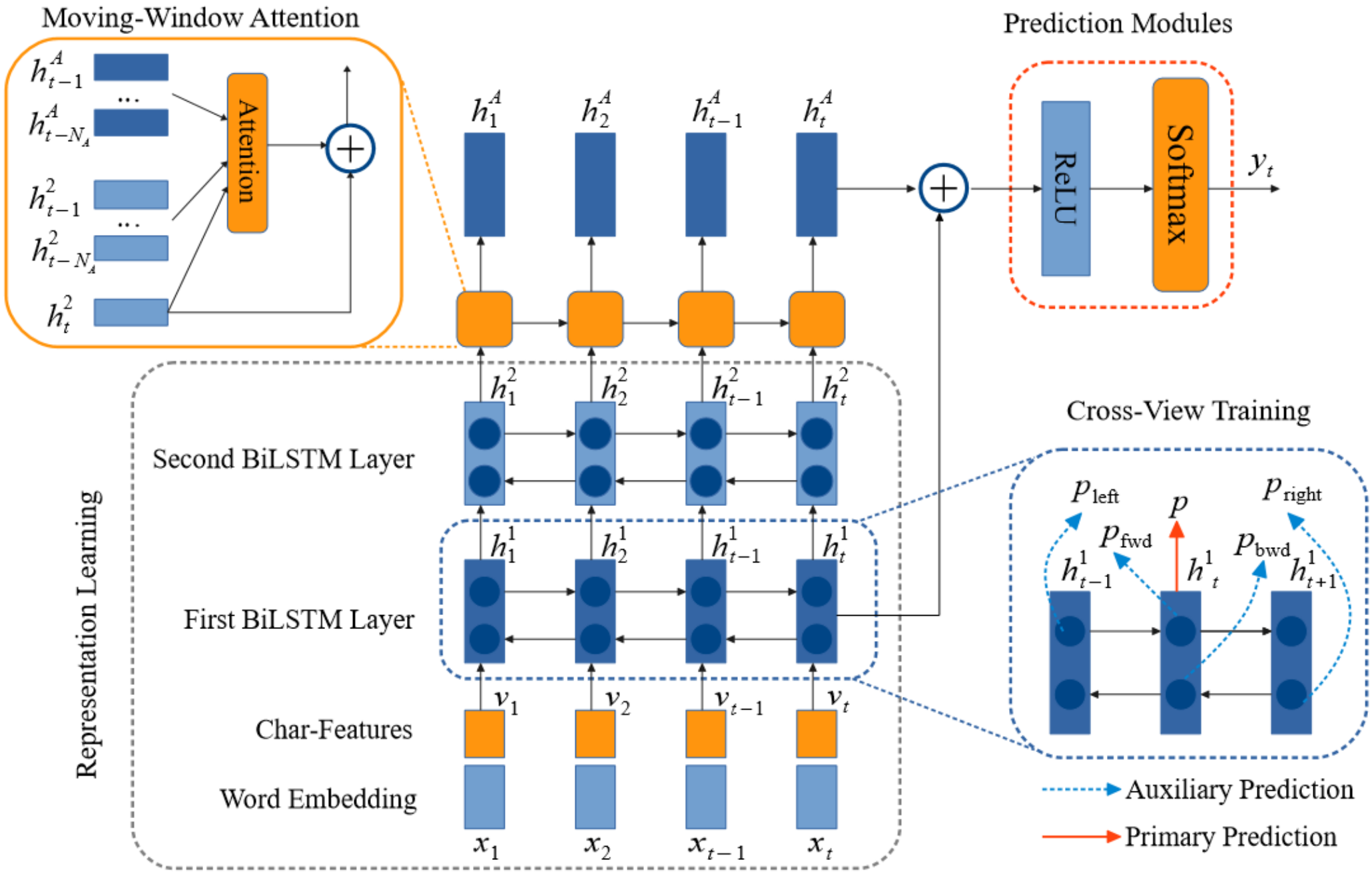
香港城市大學
City University of Hong Kong

- Semi-supervised Aspect Mining
  - Our consideration: 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.
  - To this end, we propose a new semi-supervised End-to-end MOVing-window Attentive framework (called EMOVA) to enhance aspect mining on customer reviews.

Our Framework: EMOVA

香港城市大學
City University of Hong Kong

- Four Main Components
  - Representation learning
  - Cross-view training
  - Moving-window attention
  - Primary and auxiliary prediction modules


Moving-Window Attention
Attention
Prediction Modules
ReLU
Softmax
Second BiLSTM Layer
First BiLSTM Layer
Char-Features
Word Embedding
Representation Learning
Cross-View Training
Auxiliary Prediction
Primary Prediction

Our Framework: EMOVA

CityU
香港城市大學
City University of Hong Kong

- Representation Learning
  - We combine general embeddings and char-features (from the output of a character-level CNN).
  - The new vectors are denoted by: $V = \{v_1, ..., v_T\}$.

Our Framework: EMOVA

香港城市大學
City University of Hong Kong

- Representation Learning
  - Further, the concatenation vector is fed into two BiLSTM layers.

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

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

Our Framework: EMOVA

香港城市大學
City University of Hong Kong

- Moving-Window Attention
  - In the aspect mining, 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.
  - We develop a moving-window attention component to capture such past nearby significance.
  - Moving-window attention only caches the most recent N_A hidden states.

Our Framework: EMOVA

香港城市大學
City University of Hong Kong

- Moving-Window Attention
  - The normalized significance score s_i^t of each cached state h_i^2 $(i \in [t - N_A, t - 1])$ as follows:

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

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

Our Framework: EMOVA

- Prediction Modules
  - The primary prediction module:

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

Our Framework: EMOVA

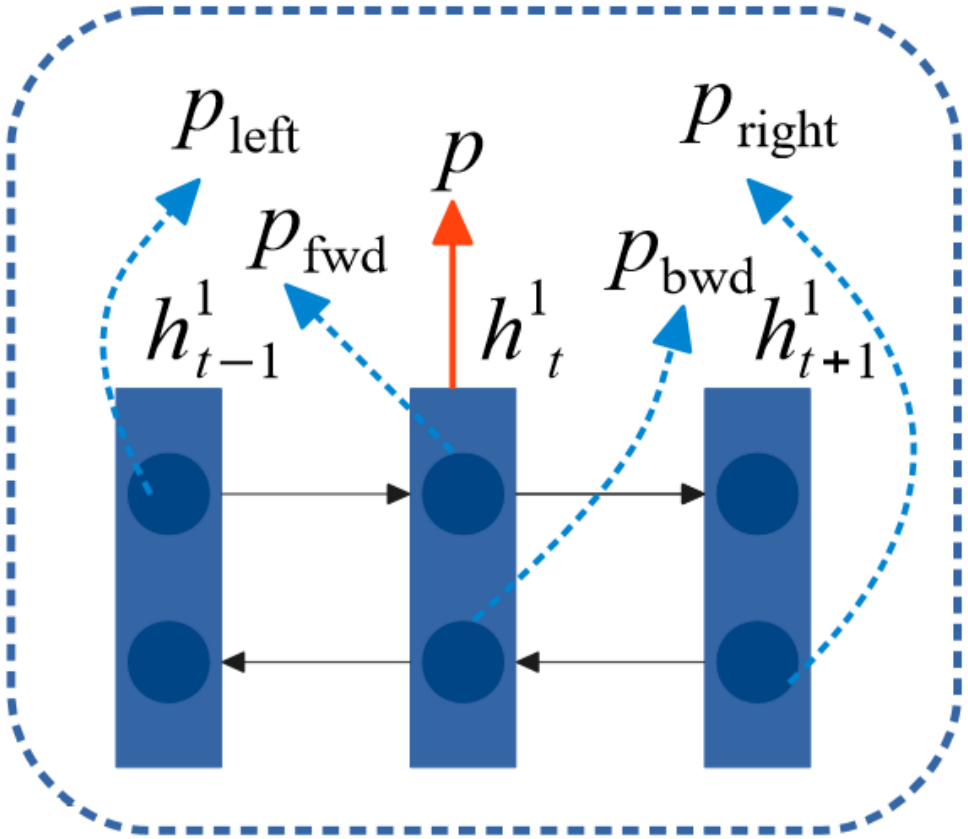
香港城市大學
City University of Hong Kong

- Prediction Modules
  - The auxiliary prediction modules:

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

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

Cross-View Training



Auxiliary Prediction

Primary Prediction

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

$$p_{\text{fwd}}(y_t|x_t) = nn_{\text{fwd}}(\overrightarrow{h}^1_t)$$

$$p_{\text{bwd}}(y_t|x_t) = nn_{\text{bwd}}(\overleftarrow{h}^1_t)$$

$$p_{\text{right}}(y_t|x_t) = nn_{\text{right}}(\overleftarrow{h}^1_{t+1})$$

Our Framework: EMOVA

香港城市大學
City University of Hong Kong

- Cross-View Training
  - For the labeled reviews, Cross-Entropy (CE) loss is utilized to train the primary prediction module:

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

Our Framework: EMOVA

香港城市大學
City University of Hong Kong

- Cross-View Training
  - For the unlabeled reviews, the framework first infers $p(y_i|x_i)$ 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:

$$L_{\mathrm{CVT}} = \frac{1}{D_u} \sum_{x_i \in D_u} \sum_j KL(p(y_i|x_i), p_j(y_i|x_i))$$

$$j \in \{\mathrm{left}, \mathrm{fwd}, \mathrm{bwd}, \mathrm{right}\}$$

Our Framework: EMOVA

香港城市大學
City University of Hong Kong

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

$$L = L_{\text{SUP}} + L_{\text{CVT}}$$

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

Experiments

香港城市大學
City University of Hong Kong

- Labeled Reviews

	$D_{\textbf{laptop1}}$		$D_{\textbf{laptop2}}$		$D_{\textbf{rest1}}$		$D_{\textbf{rest2}}$	
	Train	Test	Train	Test	Train	Test	Train	Test
Sentences	3,045	800	3,041	800	1,315	685	2,000	675
Labeled Aspects	2,358	654	1,743	1,134	1,192	542	1,743	622

Experiments

香港城市大學
City University of Hong Kong

- Unlabeled Reviews
  - Laptop reviews from Amazon Review Dataset (230,373 sentences);
  - Restaurant reviews from Yelp Review Dataset (2,677,025 sentences).

Experiments

香港城市大學
City University of Hong Kong

- Main Results

	Models	D_{laptop1}	D_{laptop2}	D_{rest1}	D_{rest2}
1	**IHS_RD**	74.55	79.62	-	-
	DLIREC	73.78	84.01	-	-
	EliXa	-	-	70.04	-
	NLANGP	-	-	-	67.12
	EMOVA	**81.72**	**85.80**	**72.26**	**75.18**

Experiments

香港城市大學
City University of Hong Kong

- Main Results

	Models	D_{laptop1}	D_{laptop2}	D_{rest1}	D_{rest2}
	CRF	74.01	82.33	67.54	69.56
2	**WDEmb**	75.16	84.97	69.73	-
	LSTM	75.17	82.01	68.26	70.35
	EMOVA	**81.72**	**85.80**	**72.26**	**75.18**

Experiments

香港城市大學
City University of Hong Kong

- Main Results

	Models	D_{laptop1}	D_{laptop2}	D_{rest1}	D_{rest2}
3	**CMLA**	77.80	85.29	70.73	72.77
	MIN	77.58	-	-	73.44
	DE-CNN	81.59	-	-	74.37
	BERT	78.71	85.12	70.85	73.23
	EMOVA	**81.72**	**85.80**	**72.26**	**75.18**

Experiments

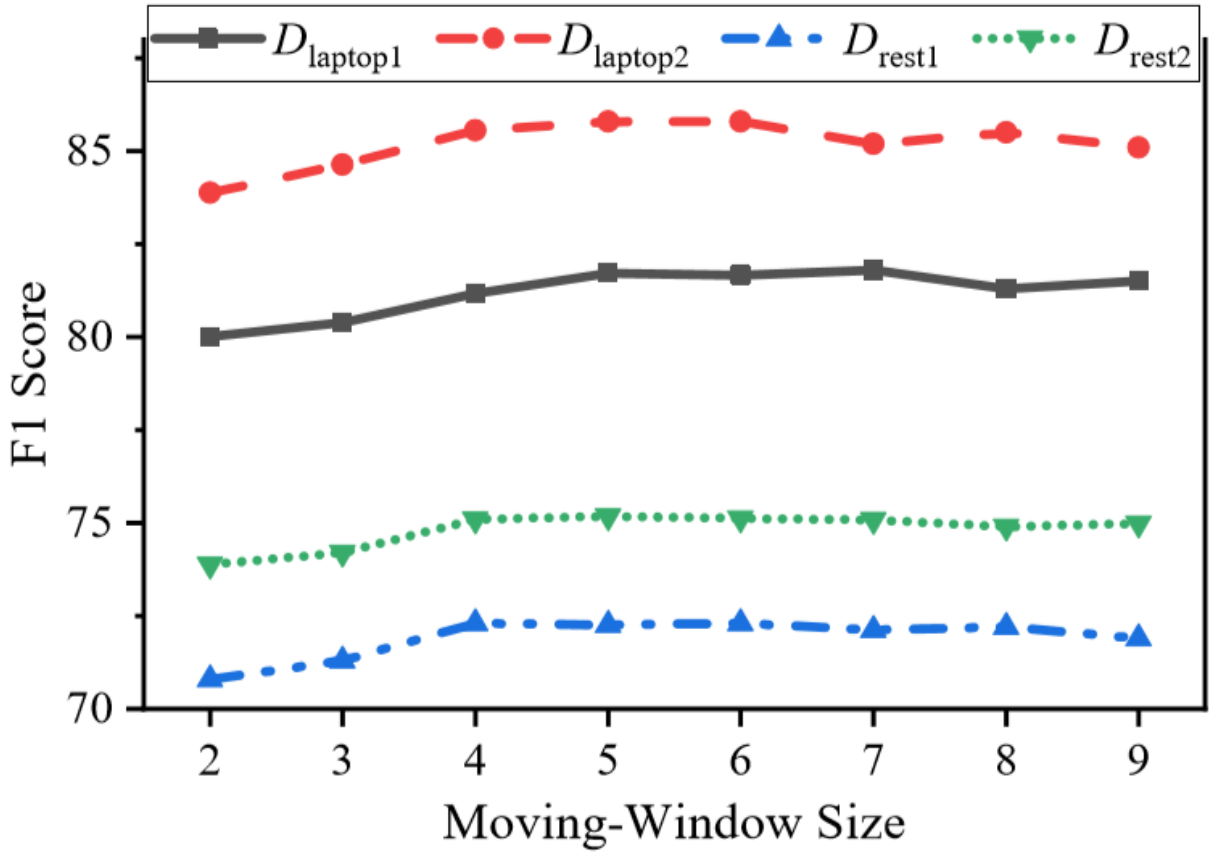
香港城市大學
City University of Hong Kong

- Main Results

	Models	D_{laptop1}	D_{laptop2}	D_{rest1}	D_{rest2}
4	**EMOVA-S**	77.32	83.48	70.10	72.35
	EMOVA-G	77.89	84.22	71.43	73.62
	EMOVA	**81.72**	**85.80**	**72.26**	**75.18**

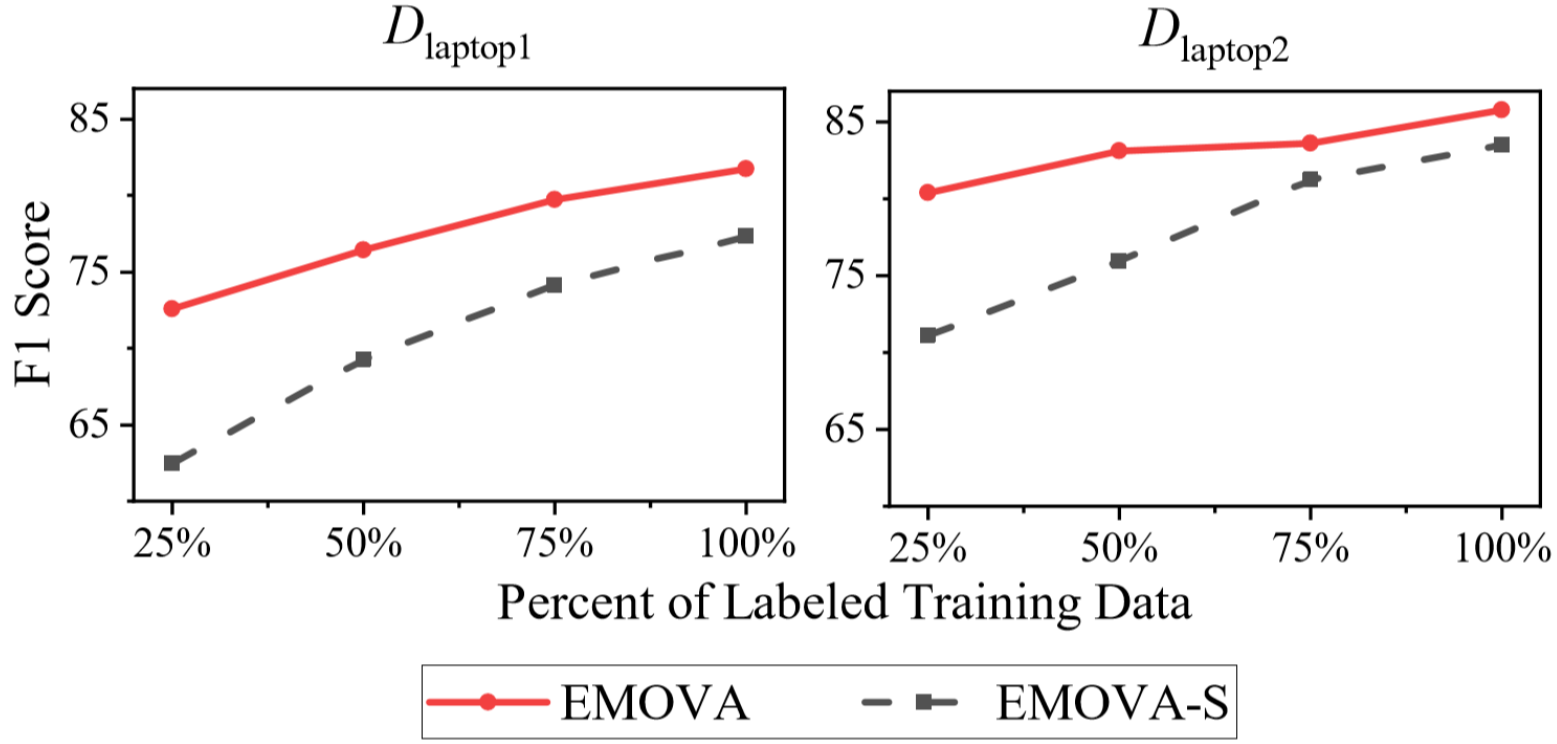
Experiments

- Effects of the moving-window size.



Experiments

- Performance vs. percent of the labeled training set.



Conclusion

- We 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 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.

香港城市大學
City University of Hong Kong

Thank You!