Coupling Global and Local Context for Unsupervised Aspect Extraction

Ming Liao[1,2]*, **Jing Li**[3,†], **Haisong Zhang**[4], **Lingzhi Wang**[1,2], **Xixin Wu**[1], **Kam-Fai Wong**[1,2]
[1]The Chinese University of Hong Kong, Hong Kong, China
[2]MOE Laboratory of High Confidence Software Technologies, China
[3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[4]Tencent AI Lab, Shenzhen, China
[1,2]{mliao,lzwang,wuxx,kfwong}@se.cuhk.edu.hk
[3]jing-amilia.li@polyu.edu.hk, [4]hansonzhang@tencent.com

Abstract

Aspect words, indicating opinion targets, are essential in expressing and understanding human opinions. To identify aspects, most previous efforts focus on using sequence tagging models trained on human-annotated data. This work studies *unsupervised* aspect extraction and explores how words appear in **global context** (on sentence level) and **local context** (conveyed by neighboring words). We propose a novel neural model, capable of coupling global and local representation to discover aspect words. Experimental results on two benchmarks, laptop and restaurant reviews, show that our model significantly outperforms the state-of-the-art models from previous studies evaluated with varying metrics. Analysis on model output show our ability to learn meaningful and coherent aspect representations. We further investigate how words distribute in global and local context, and find that aspect and non-aspect words do exhibit different context, interpreting our superiority in unsupervised aspect extraction.

1 Introduction

Opinion, one of the main factors shaping human behavior, is crucial to our daily activities (Liu, 2012). Every choice we make in our life, ranging from where to go for a Friday dinner to which job offer to pick up, is largely influenced by what other people think. To help individuals navigate decision-making processes, there exists growing attentions on opinion mining algorithms that distill massive opinion-rich texts — such as digital product reviews (Poddar et al., 2017) and social media discussions (Dusmanu et al., 2017) — into the opinionated information we need.

[R1]: It's truly a great **laptop** for the **price**.
[R2]: If you have the money, I suggest going for the **i7**

Table 1: Two sample laptop review sentences. **Aspect words** are in boldface and blue. Wavy underlines indicate local context words indicating aspect word "*i7*".

Towards human opinion understanding, it is essential to figure out what target the opinion centers around. After all, previous studies have long pointed out that human language mostly conveys opinion with aspect and sentiment words (Liu, 2012). In this work, we focus on aspect extraction, targeting at the recognition of words indicating opinion aspects (henceforth **aspect words**). We believe developing effective aspect extraction models will benefit a broad range of compelling applications, such as aspect-based sentiment classification (Tang et al., 2016), opinion summarization (Wu et al., 2016), trending event tracking (Feng et al., 2016), and so forth.

To date, most progress made in aspect extraction has focused on training sequence tagging models on human-annotated data (Li and Lam, 2017; Xu et al., 2018; Wang and Pan, 2018). However, acquiring manual labels will inevitably undergo an expensive data annotation process and is hence difficult to scale for datasets from new domain or language. In this work, we explore how aspect words can be discovered in a fully *unsupervised* manner. We are inspired by the linguistic phenomenon that *aspect words generally distinguish themselves from other words in their occurrence patterns within global and local context*. Here **global context** refers to how pairs of words co-occur with each other at sentence level (without considering word order and can be extended to capture document-level context), while **local context** means what neighbors a word has.

To illustrate why global and local context can work together to indicate aspect words, Table 1

*This work was partially done when Ming Liao was an intern at Tencent AI Lab, Shenzhen, China.

†Jing Li is the corresponding author and conducted most of the work at Tencent AI Lab, Shenzhen, China.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4578–4588, Hong Kong, China, November 3–7, 2019.

shows two sentences from laptop review benchmark (Pontiki et al., 2016). As can be seen from R1, aspect words "*price*" and "*laptop*" tend to appear together in R1-like sentences concerning "*laptop price*". As for R2, its aspect words "*i7*", though not co-occurring with other aspects, have similar neighbors "*for the*" in local context with "*price*", which reveals its high likelihood of being aspect words, the same as "*price*".

Inspired by the phenomenon above, we propose a novel *unsupervised* model capable of coupling global and local context to discover aspect word clusters. Our model is built on the success of topic models in aspect extraction (Lin and He, 2009; Brody and Elhadad, 2010; Zhao et al., 2010). It is attributed to their ability to form latent topics with words likely to co-occur in a subset of sentences instead of widely appearing in the entire corpus (Blei et al., 2003). These words happen to exhibit similar patterns of how aspect words occur on sentence level (Lin and He, 2009). However, the above methods, only exploiting global context, are arguably suboptimal for largely ignoring the rich information delivered by local context. Some recent work (He et al., 2017), on the other way around, focus on using local context, yet ignore its coupled effects with global context.

Our work, to the best of our knowledge, is the first to explore *how global and local context jointly indicate aspect words*. Moreover, taken advantage of the recent advances in neural topic models (Miao et al., 2017; Srivastava and Sutton, 2017), we enable end-to-end learning of global and local representation, where the interaction between them contributed to aspect recognition can be automatically captured.

In experiments, we first compare our model with existing unsupervised models on aspect extraction. The results on restaurant and laptop reviews show that our model outperforms state-of-the-art approaches using global or local context only. For example, we achieve 36.1 F1 on laptop dataset, compared with 32.9 produced by He et al. (2017). Further discussions demonstrate our capability of capturing meaningful representations from global and local context, which interprets our superiority in aspect extraction. In addition, we empirically analyze global and local word context on our datasets. The results confirm that aspect words indeed vary in their global and local context compared with non-aspect ones, hence providing useful clues for aspect identification.

2 Related Work

Our work is mainly in the line with aspect extraction research. On this task, early studies mostly focus on the design of hand crafted rules (Hu and Liu, 2004; Zhuang et al., 2006; Qiu et al., 2011) or features (Jin et al., 2009; Li et al., 2010). Recently, the propose of neural models enables automatic representation learning without labor-intensive feature engineering (Wang et al., 2016, 2017; Li and Lam, 2017; Xu et al., 2018; Wang and Pan, 2018). These supervised models, rely on manually annotated data, thus restricted in their scaling ability for new domain or language. Instead, our work, focusing on *unsupervised* aspect extraction, can discover aspect words via exploiting how words occur in global and local context.

Our work is inspired by the unsupervised methods capturing latent aspect factors with LDA-style topic models (Lin and He, 2009; Brody and Elhadad, 2010; Zhao et al., 2010). We are also related with *non-neural* models incorporating word embeddings (encoding local context) to learn latent topics (discovered from global context) (Nguyen et al., 2015; Li et al., 2016; Shi et al., 2017). Compared with them — relying on expertise to customize inference algorithms, our model — in a neural architecture — does not require model-specific derivation, and enables interactions between global and local representations to be automatically learned. Though some neural models were recently proposed for our task (Wang et al., 2015; He et al., 2017), they focus on local context, unable to leverage global information. Distinguishing from them, we examine *how the coupled effects of global and local context can signal aspect words*, which have never been studied before in previous work.

3 Our Neural Model Coupling Global and Local Context

This section describes our neural model coupling the force of global and local context for aspect extraction. Figure 1 shows our overall architecture. There are two modules composed, one for local context modeling and the other for global.

In the following, we first describe the formulation of input and output in Section 3.1. Then the local and global context modeling process will be in turn given in Section 3.2 and 3.3.

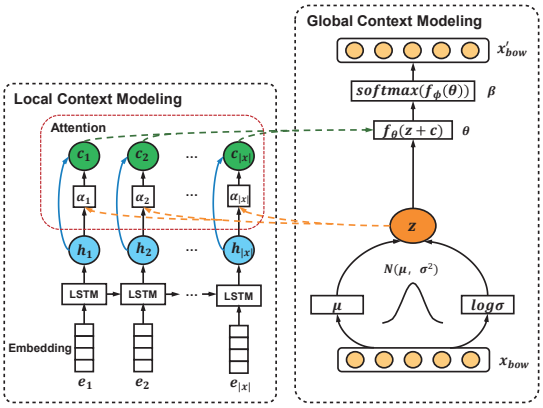


Figure 1: Our overall architecture with one module for local context modeling (on the left) and the other for global (on the right).

3.1 Input and Output

Before touching details to reveal how our model works, we first describe our input and output.

Formally, given a corpus $\mathcal{C}$ with $|C|$ sentences, $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{|C|}\}$, we process each sentence $\mathbf{x}$ into two forms: word sequence form $\mathbf{x}_{seq}$ and bag-of-words (BoW) form $\mathbf{x}_{bow}$. $\mathbf{x}_{seq} = \langle w_1, w_2, ..., w_{|\mathbf{x}|} \rangle$, where w_n indicates word index of the n-th word and $|\mathbf{x}|$ denotes the number of words. $\mathbf{x}_{bow}$ is the BoW term vector over the vocabulary V. Here $\mathbf{x}_{seq}$ (considering word order) is fed for modeling local context and learning how words co-occur with their neighbors, while $\mathbf{x}_{bow}$ (following the bag-of-words assumption in most topic models (Blei et al., 2003; Miao et al., 2017)) serves as the input for global context modeling and capture sentence-level word co-occurrence.

Our goal is to output distributional clusters of aspect words. Then following Qiu et al. (2011)'s practice, the top N nouns (N as a hyperparameter) from each cluster (ranked by likelihood) are selected as the extracted aspect words, considering most aspects are nouns.

3.2 Local Context Modeling

As mentioned above, local context modeling module takes word sequence form, $\mathbf{x}_{seq}$, as its input. In this module, each word $w_n \in \mathbf{x}_{seq}$ is first processed with an embedding layer and converted into an embedding vector $\mathbf{e}_n$. Then we employ long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) network to explore local context. Word embeddings $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{|\mathbf{x}|}$ are processed into hidden states via recurrently exploring word co-occurrence with left neighbors. Specifically, for word w_n, its hidden states $\mathbf{h}_n$ is:

$$\mathbf{h}_n = f_{LSTM}(\mathbf{e}_n, \mathbf{h}_{n-1}) \quad (1)$$

where $f_{LSTM}(\cdot)$ refers to an LSTM unit. The hidden states $\langle \mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{|\mathbf{x}|} \rangle$ are considered as the local representation, further leveraged in global context modeling and described later.

3.3 Global Context Modeling

Our global context modeling module is inspired by previous practice that discovers aspect words with LDA-fashion Bayesian graphical models (Lin and He, 2009). We assume there are K latent aspect factors embedded in the the given corpus $\mathcal{C}$. Each factor ϕ_k ($k = 1, 2, ..., K$) is represented with a distributional word cluster over the vocabulary V.

Also Inspired by neural topic models (Miao et al., 2017), we adopt a variational auto-encoder (VAE) (Kingma and Welling, 2013), with an encoder and a decoder, to resemble the topic model-style data generation process. In doing so, we enable latent aspects, capturing word co-occurrence in both global and local context, to be learned in neural architecture. There are two main steps involved: First, the input sentence $\mathbf{x}$ (in BoW form $\mathbf{x}_{bow}$) is encoded to global representation $\mathbf{z}$. Conditioned on $\mathbf{z}$, together with local representation $\mathbf{h}_n$ (defined in Section 3.2 and $n = 1, 2, ..., |\mathbf{x}|$), decoder further generates $\mathbf{x}'_{bow}$, the BoW-form reconstruction of $\mathbf{x}_{bow}$. In the rest of this section, we first introduce how global representation $\mathbf{z}$ is learned by encoder from global context. Then we introduce how global and local representations are coupled to work together for data generation.

Global Representation Encoding. The encoder is employed to learn global representation, $\mathbf{z}$, from $\mathbf{x}_{bow}$. Following Miao et al. (2017), words in global context are assumed to satisfy Gaussian distribution, prior on mean μ and standard deviation σ. Their estimation formula are defined as:

$$\mu = f_\mu(f_e(\mathbf{x}_{bow})), \log \sigma = f_\sigma(f_e(\mathbf{x}_{bow})) \quad (2)$$

where $f_*(\cdot)$ is a neural perceptron performing a linear transformation operation followed by a non-linear ReLU activation (Nair and Hinton, 2010).

Coupling Global and Local Context. Recall we obtain the hidden states $\mathbf{h}_n$ ($n = 1, 2, ..., |\mathbf{x}|$) from local context modeling, and here we describe how we couple them with global representation $\mathbf{z}$.

Concretely, we employ attention mechanism (Bahdanau et al., 2015) over the hidden states in local representation, which, in aware

of global information, aims to identify words in $\mathbf{x}$ that can usefully indicate its aspect factors. We design attention weight α_n to measure the similarity between the semantic meaning of word w_n and $\mathbf{x}$'s global representation $\mathbf{z}$:

$$\alpha_n = \mathbf{z}^T f_h(\mathbf{h}_n) \tag{3}$$

where $f_h(\cdot)$ is a ReLU activation function. The context vector of this attention, namely **globally-scoped local representation**, is defined as:

$$\mathbf{c} = \sum_n^{|\mathbf{x}|} \alpha_n \mathbf{h}_n \tag{4}$$

In the next step, the decoder will use $\mathbf{c}$ for learning corpus-level aspect factors and reproducing $\mathbf{x}_{bow}$. Below comes more details.

Decoding Process. Given the global representation $\mathbf{z}$ and the globally-scoped local representation $\mathbf{c}$, the decoder carries out the data generation process conditioned on both of them. For each input sentence $\mathbf{x}$, we assume each word $w_n \in \mathbf{x}_{bow}$ is sampled conditioned on its aspect mixture, θ, a K-dim distribution reflecting $\mathbf{x}$'s composition of aspect factors. θ is then estimated with both $\mathbf{z}$ and $\mathbf{c}$, conveying global and local context of $\mathbf{x}$'s words. The story describing $\mathbf{x}$'s generation process is:

- Draw global representation $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$.
- Obtain globally-scoped local representation $\mathbf{c}$ with Eq. 4.
- Aspect mixture $\theta = softmax(f_\theta(\mathbf{z} + \mathbf{c}))$.
- For the n-th word in $\mathbf{x}$:
  - $\beta_n = softmax(f_\phi(\theta))$.
  - Draw the word $w_n \sim Multi(\beta_n)$.

Here $f_*(\cdot)$ is a ReLU-activated neural perceptron described above. Particularly, the weight matrix of $f_\phi(\cdot)$ (with the softmax normalization) are employed as the aspect-word distributions (distributional word clusters), ϕ, used to represent the latent aspect factors and serves as our main output.

Learning Objective. We design the learning objective of our entire framework as:

$$\mathcal{L} = D_{KL}(p(\mathbf{z})||q(\mathbf{z}\,|\,\mathbf{x})) - \mathbb{E}_{p(\mathbf{z})}[p(\mathbf{x}\,|\,\mathbf{z})] \tag{5}$$

where $p(\mathbf{z})$ is a standard Gaussian prior. The first term reflects encoding loss while the second estimation likelihood (for decoding). We refer the readers to Miao et al. (2017) for more details.

4 Experimental Setup

Datasets. We conduct experiments on two benchmark datasets constructed for the SemEval aspect-based sentiment analysis (ABSA) challenge (Pontiki et al., 2014, 2015, 2016) with human annotated aspects — one gathers restaurant reviews (henceforth **restaurant**) and the other consists of laptop reviews (henceforth **laptop**). For model training and evaluation, we combine the training and test datasets for 2014-2016 ABSA (except for 2015 without laptop data released). Following common practice (Wagner et al., 2014), a review sentence is considered as a data sentence for input. The statistics of our datasets are displayed in Table 2. We can see that aspect words take around 25.0% of the vocabulary, yet less than 19.1% of the words per sentence. It is indicated the sparsity and diversity of aspect words, further suggesting the challenging of our task.

	# of sen	|Voc|	Avg len per sen	|Apt|	# of apt per sen
Laptop					
Train	6,355	3,374	14.51	837	2.35
Test	800	1,866	13.17	430	2.52
Rest					
Train	6,359	5,166	13.20	1,404	2.36
Test	2,161	3,800	13.51	948	2.41

Table 2: Statistics of laptop and restaurant (rest) datasets. |Voc|: the vocabulary size (including stop words). Avg len: average number of tokens. |Apt|: the number of distinct aspects. # of apt per sen: average number of aspects in a sentence.

Preprocessing. Here are our preprocessing steps. First, we adopted NLTK toolkit for text tokenization.[1] Then, we normalized all letters into their lower cases. Next, we removed words appearing less than five times. Finally, for BoW-form input, we removed all stop words and punctuation following common practice in topic models (Blei et al., 2003).

Parameter Setting. We applied pre-trained GloVe embedding (Pennington et al., 2014) for initialization in local context modeling.[2] The embedding dimension is set to 300, and batch size to 128. For the number of latent aspect factors, K, we tuned it on training data with five-fold validation and set it to 40. As for N, the number of nouns to be selected from each aspect word cluster, we set it to 30 following Qiu et al. (2011). In model training, we employ Adam optimizer (Kingma and Ba, 2015), with learning rate set to

[1] https://www.nltk.org

[2] https://nlp.stanford.edu/projects/glove/

$1e-3$, and run 20 epochs with early stop strategy adopted. Dropout strategy (Srivastava et al., 2014) is also adopted to avoid overfitting.

Evaluation Metrics. To ensure comparable performance, for clustering-based approaches, we select the top 30 nouns from 40 aspect clusters, same as our set up. For the rest, the top $1,200$ nouns are extracted. Here we adopt two sets of evaluation metrics. First, we follow Qiu et al. (2011) to test sentence-level aspect extraction, where the intersection of our selection and the words appear in a review sentence are considered as the extracted aspects. In this evaluation, we report precision, recall, and F1 scores. Second, we evaluate our ability to build aspect lexicon (a.k.a. corpus-level extraction) following Hamilton et al. (2016). We consider all the annotated aspects as gold standard lexicon and adopt accuracy for evaluation.

Comparisons. We first consider a simple baseline that randomly selects nouns as aspect words (henceforth RANDOM). We also compare with extracting- and clustering-based baselines — TF-IDF (Bahdanau et al., 2015), K-MEANS (Lloyd, 1982) (implemented with sklearn toolkit[3] and taking Glove embedding for similarity measure), and BTM[4] (Yan et al., 2013), state-of-the-art in short text topic modeling and well-performed in aspect extraction (He et al., 2017).

In addition, we consider the following recently proposed *unsupervised* models in comparison: LF-LDA (Nguyen et al., 2015), LDA topic model incorporating word embeddings (GloVe is applied here), and ABAE (He et al., 2017), the state-of-the-art attention-based model for unsupervised aspect extraction. Besides the existing models, we also compare with our variant that only models global context with neural topic model (henceforth GBC ONLY). The full model coupling global and local context is hence referred to as LCC+GBC.

5 Experimental Result

In this section, we first discuss comparison results with unsupervised aspect extraction models in Section 5.1. Section 5.2 shows what our model learns and interprets why it can discover aspect words. Next, in Section 5.3, we carry out an empirical study over how global and local context indicate aspect words. Last, we further discuss our parameter effects and main error in Section 5.4.

[3] https://scikit-learn.org/stable/

[4] https://github.com/xiaohuiyan/BTM

Models	Restaurant				Laptop			
	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc
Comparisons								
RANDOM	24.9	20.4	22.4	17.3	24.1	39.0	29.8	20.8
TF-IDF	28.6	24.8	26.6	24.3	22.5	18.3	20.2	21.9
K-MEANS	28.1	40.0	33.0	19.0	23.0	35.6	27.9	23.5
BTM	30.6	56.4	39.7	21.2	25.8	48.8	33.7	31.3
LF-LDA	30.2	60.3	40.2	24.8	26.3	50.1	34.4	28.4
ABAE	30.9	57.8	40.2	23.6	25.4	46.5	32.9	32.0
Our models								
GBC ONLY	30.5	57.9	39.9	24.2	25.6	49.8	33.8	29.8
LCC+GBC	**31.2**	**60.5**	**41.2**	**26.0**	**28.0**	**50.2**	**36.1**	**33.7**

Table 3: Precision (pre), Recall (rec), F1, and Accuracy (Acc) produced by various unsupervised models. Acc measures corpus-level extraction while others sentence-level. Our model significantly outperform others in F1 and Acc measure (paired t-test, $p < 0.05$).

5.1 Aspect Extraction Results

Main Comparison Results. In Table 3, we report the aspect extraction results on two datasets. Several interesting observations can be drawn:

• *All models tend to yield better F1 on restaurant yet better Acc on laptop.* We find that restaurant exhibits generally worse model performance on corpus-level extraction than that on sentence level. The opposite findings is whereas drawn on laptop. It might be because restaurant contains more distinct aspects (shown with the larger |Apt| in Table 2). It is possibly due to the prominence of rare aspects, which is challenging to be discovered.

• *Simple baselines do not work well.* Both RANDOM and TF-IDF perform poorly, indicating the challenge of unsupervised aspect extraction.

• *Global context can well indicate aspects.* We observe that approaches based on topic models (BTM, LF-LDA, and our models) perform better than others. The results indicate that aspect words do vary from other words in global context distributions. Topic model-based approaches, via exploiting sentence-level word co-occurrence, can thus effectively identify aspect words.

• *Coupling global and local context is effective.* By combining topic models (global context) with word embeddings (local context), LF-LDA produces the second best F1. Also, our full model LCC+GBC outperforms its variant GBC ONLY in F1. These observations indicate the benefit of joint modeling of global and local context to discover aspect words.

Besides, by comparing model performance over the two datasets, we observe that all models perform worse on laptop. An intuitive explanation is that laptop reviews generally concern wide

GBC ONLY	LCC+GBC
good, **menu**, little, kind, **noise**, play, daniel, **fare**, **jelly**, much, details, father, **neighborhood**, **wine**, door, possible, murray, keep, vagan, heaviness, *cool*, wins, angel, upper, *romantic*, takes, **avenue**, **fruit**, pink, strips	**value**, **wine**, *date*, evening, **food**, **block**, **avenue**, **line**, *fun*, years, *love*, yes, hang, knows, must, **cheese**, *favorite*, **course**, *romantic*, **tip**, jeans, servers, cold, **pastrami**, **atmosphere**, *fine*, **counter**, word, **sauce**, phone
crash, **keyboard**, encounter, 39, **battery**, **wires**, **memory**, photographs, *sooner*, fits, **feel**, things, steve, overheat, would, seem, **pro**, **touch**, **cpu**, **mouse**, figure, user, *better*, users, ran, days, question, apple, worth, sit	needed, **skype**, advise, **keyboard**, remote, daily, matches, **high**, *love*, originally, wife, loud, *excellent*, **macbook**, freezes, *well*, anytime, got, friend, *weird*, even, **gui**, weeks, recently, **internet**, **8.1**, **laptops**, tapping, **speakers**, noon

Table 4: Top 30 words of sample latent aspects learned by our variant GBC ONLY (on the left) and full model LCC+GBC (on the right). The top displays the outputs on restaurant dataset and the bottom laptop. **Aspect words** are in boldface and blue (annotated in least one sentence), while *sentiment words* are in italic and red.

range of aspects (e.g., screens, battery, etc.), while restaurant reviews tend to be centered around general aspects (e.g., food and service). Aspect words thus exhibit sparse occurrence patterns in laptop reviews, rendering generally worse model performance. Section 5.3, we will discuss more. For the same reason, local context helps LCC+GBC obtain larger margin on laptop, compared with models relying on global word co-occurrence.

5.2 Model Interpretation

Here we probe into our output and study why LCC+GBC model works.

Topic Coherence. We first analyze the coherence of our latent aspects, where C_V metrics, a widely-applied automatic topic coherence measure (Röder et al., 2015) is adopted. LCC+GBC's latent aspects achieves C_V coherence scores of 0.401 and 0.393 on restaurant and laptop dataset, respectively, compared to 0.382 and 0.377 produced by GBC ONLY. It hence suggests the joint effects of global and local context also helps produce coherent aspects.

Sample Latent Aspects. We further conduct a qualitative analysis on the produced latent aspects. Table 4 shows the top 30 words (ranked by likelihood) of the sample latent aspects. LCC+GBC's output aspects look more coherent, with words

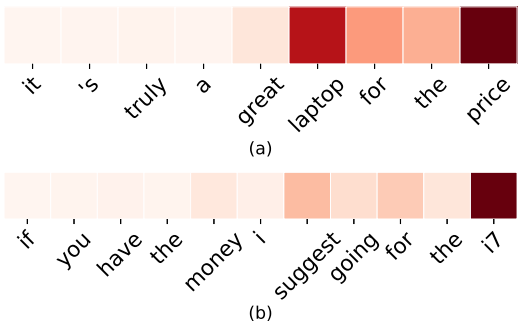


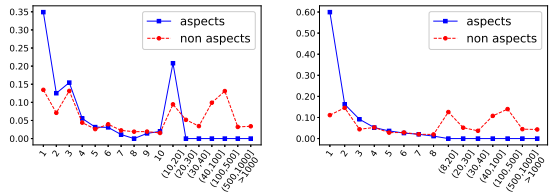
Figure 2: The globally-scoped local attention weights learned by our LCC+GBC model for the sample sentences in Table 1. Darker colors indicate higher values.

having similar semantics clustered together, such as "*course*", "*romantic*", and "*date*" learned from restaurant dataset, and "*skype*", "*remote*", and "*internet*" from laptop. The possible reason is that words conveying similar semantic meanings tend to appear in similar local context. LCC+GBC, via coupling local context with global one, is thus able to capture such semantic representations.

We also notice that LCC+GBC discovers more rare aspects, such as "*pastrami*" (from restaurant) and "*gui*" (from laptop). These words, though may exhibit sparse occurrence and unable to be discovered purely with global context, might be effectively indicated by their local context. This reveals the benefit of combining the effects of global and local context for aspect extraction.

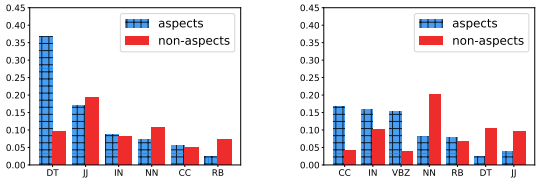
In addition, we notice that our output aspect clusters include some sentiment words. It is possibly because aspect words tend to co-occur with sentiment ones in both global and local context. Thus without supervision, it is likely our models discover them together. These findings suggest our potential benefit on extracting sentiment words — which can be easily separated from aspects by their POS tags (Qiu et al., 2011)). Such extension is beyond the scope of this paper but worth exploring in future work.

Case Study. To understand what LCC+GBC learns resulting its superiority in aspect extraction. We take the two samples in Table 1 as input. Figure 2 visualizes their globally-scoped attention weights (defined in Eq. 3). We observe that LCC+GBC assigns high attention weights for aspect words "*laptop*", "*price*", and "*i7*". Also highlighted are the neighboring words "*for the*" in both attentions, usefully signaling their next word to be aspect. The results indicate LCC+GBC learns meaningful information via exploring inter-



(a) Restaurant (b) Laptop

Figure 3: Sentence-level co-occurrence of aspect (in blue line) and non-aspect (in red line) word pairs distribution from two datasets. X-axis: word pair frequency. Y-axis: proportion of pairs. Aspect and non-aspect words exhibit diverse distributions.



(a) Left Context (b) Right Context

Figure 4: The distribution of neighboring POS tags in left and right context (laptop dataset). Aspect neighbors are in red and non-aspects' in blue. X-axis: neighboring POS tags. Y-axis: POS proportions. Aspects' and non-aspects' neighbors have different POS tags.

actions between global and local representation.

5.3 Analysis of Global and Local Context

To extensively understand the effects of global and local context on aspect identification, we carry out an empirical study on our datasets.

Aspects vs. Non-aspects. We first compare the global and local word occurrence statistics in context of aspect and non-aspect words. This is to empirically analyze why LCC+GBC can effectively distinguish aspect and non-aspect words.

To examine global context, Figure 3 presents distributions of word pair co-occurrence in sentences, with two lines corresponding to aspect and non-aspect pairs. It is seen different distribution are exhibited by aspect and non-aspect pairs, with non-aspect ones flatly distributed over varying frequency while the aspect pairs are more sparse. This indicates global context, capturing how words co-appear in sentences, can indeed help distinguish aspect and non-aspect words.

We also notice that aspect pair distributions are slightly different on two datasets. On restaurant dataset, we observe a pulse on pairs occurring $10-20$ times, while the distribution on laptop is a long tail. This demonstrates the sparse aspect occurrence patterns in laptop dataset (probably owing to the broad range of aspects discussed there), also explains the general worse performance on it (compared to restaurant and shown in Table 3).

We then analyze local context and show the distribution of POS tags (predicted with NLTK toolkit) in left and right neighbors. We take laptop dataset as an example to discuss, and similar observations are drawn from restaurant. For better displays, from 34 POS tags in total, we pick up the top 5 tags in aspects' and non-aspects' neighbors respectively. Distributions are shown over their union set in Figure 4. In both left and right context, we observe aspects and non-aspects exhibit different distributions for their neighboring POS tags. For example, aspect words are likely to have "*DT*" (e.g., *an*, *this*) appearing as left while "*RB*" (e.g., *highly*, *barely*) frequently acting as non-aspects' left neighbors while rarely in aspects'.

In addition, we display in Table 5 the top 10 neighboring bigrams in left context. Although it is merely a qualitative human judgement at this point, we can draw some interesting observations from the results. For example, some opinioned bigrams, such as "a great" and "love the", are likely to appear on the left local context of aspects (opinion targets). Such patterns may usefully indicate aspect words and help distinguish them from non-aspect ones. As for right bigram neighbors, they exhibit sparse occurrence patterns, hence might provide less useful clues. We will analyze the effects of left and right local context next.

Aspects	Non-aspects
easy to, and the, of the, for the, with the, a great, that the, it 's, all the, love the	of the, it is, and the, is a, battery life, to use, it 's, with the, i have, for the

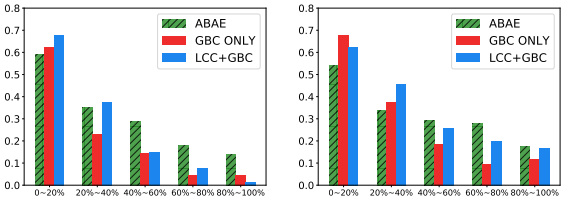
Table 5: Top 10 neighboring bigrams in left context of aspects and non-aspects (laptop dataset).

Local Context Modeling. We then compare and discuss the effectiveness of varying modules to capture local representation. The performance of our variants combined with varying local encoders are shown in Table 6. It is observed that all variants with attention, in aware of global context and put over local representation, yield better performance. This shows that attention mechanism is able to capture interactions between global and local representations, which are useful in discovering aspect words. We also notice that LSTM encoder performs better than CNN and BI-LSTM.

	Restaurant	Laptop
W/O LCC	39.9	33.6
AVG EMB	40.0	33.9
LSTM *(w/o att)*	40.1	34.4
CNN *(w/ att)*	40.4	34.0
BI-LSTM *(w/ att)*	40.6	34.8
LSTM *(w/ att)*	**41.2**	**36.1**

Table 6: F1 score of our variants with varying encoders for local context modeling. Here *att* refers to the attention to capture globally-scoped local representation (shown in Eq. 3 and 4). In the first column, W/O LCC refers to GBC ONLY variant. AVG EMB means average embedding. LSTM (*w/ att*) is our LCC+GBC model.

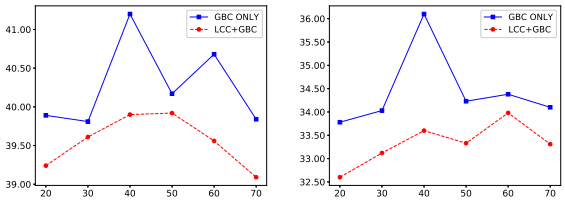
It is possibly because, in local context, left neighbors convey more useful clues for indicating aspect words, compared with right ones. As a result, CNN and BI-LSTM, equally considering left and right context, might be somehow affected by the noise in right context. They are thus outperformed by LSTM, which only models local context in left-to-right direction.



(a) Restaurant (b) Laptop

Figure 5: Recall scores for discovering varying quintiles (20%) of aspects (ranked by frequency). X-axis: quintiles of aspect frequency. Y-axis: recall scores.

Varying Aspect Frequency. Recall that the sample aspects in Table 4 suggest more rare aspects discovered by LCC+GBC compared with GBC ONLY. We conduct an analysis on model performance to discover aspects with varying frequency. Figure 5 compares the recall scores produced by ABAE, GBC ONLY, and LCC+GBC when retrieving varying aspect quintiles (5-quantile) (ranked by frequency). It shows ABAE performs better in discovering low-frequency aspects while GBC ONLY better at recognizing frequent ones. It suggests global context is more useful to indicate common aspects while local context better at signaling rare ones. LCC+GBC, capturing the coupled effects of global and local context, can identify both common and rare aspects, and thus yield superior performance.



(a) Restaurant (b) Laptop

Figure 6: F1 scores of our models (in y-axis) given varying aspect number K (in x-axis). LCC+GBC performs consistently better than GBC ONLY .

5.4 Further Discussion

Parameter Analysis In main results, we fix the number of latent aspects to $K = 40$. In Figure 6, we further examine how our models (GBC ONLY and LCC+GBC) perform given varying number of K. Here to ensure comparable performance, we set $N = 1200/K$. It is seen that LCC+GBC yields consistently better F1 than GBC ONLY. We also observe that both models do not exhibit monotomic curves, where LCC+GBC obtains the best performance given $K = 40$, consistent with the validation results.

Error Analysis. Here we analyze our main errors types. One major type is caused by wrongly identifying aspect phrases, such as "*windows 7*", where "*7*" is missed possibly ascribed to its sparse occurrence. We have such errors owning to modeling context in word level and sometimes fail to capture semantics in coarser grain. One possible solution is to extend our global context modeling module to learn phrase-level semantics (He, 2016). Another main errors occur when processing context-sensitive aspect words. For example, "*hard*" indicates aspect in "*The hard drive is fast.*", rather than "*It is hard to use that laptop.*". Our model, failing to distinguish "*hard*" in varying context, considers it as aspect for both sentences. To deal with such error, we can adopt context-aware decoders, such as Hu et al. (2017), to distinguish word semantics in different context.

6 Conclusion

We have presented a study of unsupervised aspect extraction via exploring the coupled effects of global and local context. A neural model has been proposed to learn the interactions between global and local representations indicative of aspect words. Experiment results on two benchmark datasets show our model outperform comparison approaches modeling local or global context only.

We find out three interesting points in empirical analysis over global and local context: First, aspects and non-aspects exhibit distinguishing distributions in either global and local context; Second, in local context, left neighbors can better indicate aspect words compared with the right; Third, local context can better indicate rare aspects while global signals common aspects better.

Acknowledgments

This work is partially supported by the following HK grants: RGC-GRF (14232816, 14209416, 14204118, 3133237), NSFC (61877020) & ITF (ITS/335/18). We thank the three anonymous reviewers for the insightful suggestions on various aspects of this work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812, Los Angeles, California. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 66–71, Berlin, Germany. Association for Computational Linguistics.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada. Association for Computational Linguistics.

Yulan He. 2016. Extracting topical phrases from clinical documents. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2957–2963. AAAI Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. A novel lexicalized hmm-based learning framework for web opinion mining. In *Proceedings of the 26th annual international conference on machine learning*, pages 465–472. Citeseer.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 165–174.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Yingju Xia, Shu Zhang, and Hao Yu. 2010. Structure-aware review mining and summarization. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 653–661.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2892, Copenhagen, Denmark. Association for Computational Linguistics.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 375–384, New York, NY, USA. ACM.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Stuart P. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2410–2419, International Convention Centre, Sydney, Australia. PMLR.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA. Omnipress.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Lahari Poddar, Wynne Hsu, and Mong Li Lee. 2017. Author-aware aspect topic sentiment model to retrieve supporting opinions from reviews. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 472–481, Copenhagen, Denmark. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 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.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA. ACM.

Bei Shi, Wai Lam, Shoaib Jameel, Steven Schockaert, and Kwun Ping Lai. 2017. Jointly learning word embeddings and latent topics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 375–384.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations (ICLR)*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307. The COLING 2016 Organizing Committee.

Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. 2014. Dcu: Aspect-based polarity classification for semeval task 4. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 223–229, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

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

Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, pages 616–625.

Wenya Wang and Sinno Jialin Pan. 2018. Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2171–2181. Association for Computational Linguistics.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626, Austin, Texas. Association for Computational Linguistics.

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 the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3316–3322.

H. Wu, Y. Gu, S. Sun, and X. Gu. 2016. Aspect-based opinion summarization with convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3157–3163.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598. Association for Computational Linguistics.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1445–1456, New York, NY, USA. ACM.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 56–65.

Li Zhuang, Feng Jing, and Xiaoyan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006*, pages 43–50.