Joint Aspect Extraction and Sentiment Analysis with Directional Graph Convolutional Networks

Guimin Chen[♡*], **Yuanhe Tian**[♥*], **Yan Song**[♠♡†]
[♡]Shenzhen Research Institute of Big Data [♥]University of Washington
[♠]The Chinese University of Hong Kong (Shenzhen)
[♡]chenguimin@sribd.cn [♥]yhtian@uw.edu [♠]songyan@cuhk.edu.cn

Abstract

End-to-end aspect-based sentiment analysis (EASA) consists of two sub-tasks: the first extracts the aspect terms in a sentence and the second predicts the sentiment polarities for such terms. For EASA, compared to pipeline and multi-task approaches, joint aspect extraction and sentiment analysis provides a one-step solution to predict both aspect terms and their sentiment polarities through a single decoding process, which avoids the mismatches in between the results of aspect terms and sentiment polarities, as well as error propagation. Previous studies, especially recent ones, for this task focus on using powerful encoders (e.g., Bi-LSTM and BERT) to model contextual information from the input, with limited efforts paid to using advanced neural architectures (such as attentions and graph convolutional networks) or leveraging extra knowledge (such as syntactic information). To extend such efforts, in this paper, we propose directional graph convolutional networks (D-GCN) to jointly perform aspect extraction and sentiment analysis with encoding syntactic information, where dependency among words are integrated into our model to enhance its ability to represent input sentences and help EASA accordingly. Experimental results on three benchmark datasets demonstrate the effectiveness of our approach, where D-GCN achieves state-of-the-art performance on all datasets.[1]

1 Introduction

End-to-end aspect-based sentiment analysis (EASA) aims to extract aspect terms in the text and predict their sentiment polarities so as to understand targeted sentiment towards particular objects. For example, in the sentence "*The ambiance is minimal but food is not phenomenal*", the aspect terms are "*ambiance*" and "*food*" and the sentiment polarities towards them are positive and negative, respectively. In general, there are mainly three types of approaches for this task, i.e., pipeline, multi-task, and joint-label approaches. Pipeline approaches (Mitchell et al., 2013; Zhang et al., 2015; Hu et al., 2019) perform aspect extraction and sentiment analysis in a sequence, which is not straightforward and suffers from error propagation among different steps; multi-task approaches (Mitchell et al., 2013; Zhang et al., 2015; Ma et al., 2018; Luo et al., 2019; He et al., 2019; Hu et al., 2019) apply an encoder to the input and use a separate decoding process to extract aspects and predict their sentiments, where there could be mismatches between the two decoding results. As a comparison, joint-label approaches (Mitchell et al., 2013; Zhang et al., 2015; Li and Lu, 2017; Li et al., 2019a; Hu et al., 2019) extract aspect terms and predict their sentiments simultaneously through a unified labeling scheme, which not only provides an one-step solution to EASA, but also avoids the aforementioned problems in the other two approaches.

In most cases, previous studies demonstrate that a good modeling of contextual information is effective in improving EASA performance. However, these studies mainly rely on powerful encoders (e.g., Bi-LSTM, CNN, BERT) (Zhang et al., 2015; Ma et al., 2018; Schmitt et al., 2018; Li et al., 2019a; Li et al., 2019b; Luo et al., 2019; He et al., 2019; Hu et al., 2019) and pre-trained embedings (e.g., GloVe, word2vec, FastText) (Schmitt et al., 2018; Li et al., 2019a) to learn contextual information, with limited effort paid to leveraging advanced architectures and extra knowledge for this task. To extend such effort, graph convolutional networks (GCN) was proposed and shows its effectiveness in conventional sentiment analysis (Zhang et al., 2019; Sun et al., 2019), as well as other tasks, e.g., text classification (Kipf and Welling, 2016), neural machine translation (Bastings et al., 2017), semantic role labeling (Marcheggiani and Titov, 2017), etc. Moreover, consider that discriminatively modeling the contextual features of a given word according to their positional relations to the word is helpful in text representation learning

[*]Equal contribution.
[†]Corresponding author.
[1]The code and models are released at https://github.com/cuhksz-nlp/DGSA.

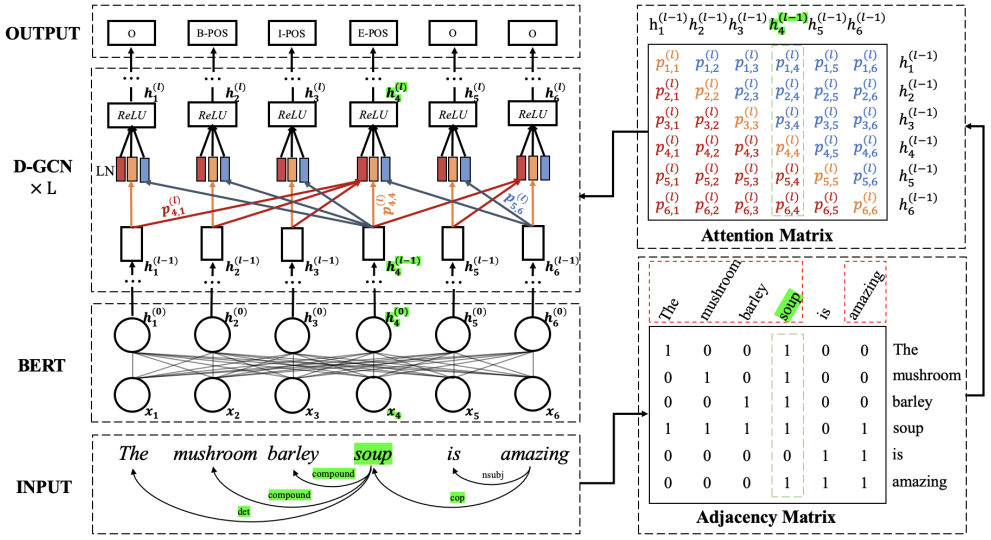


Figure 1: The overall architecture of our approach, where the graph is built upon the dependency tree of an input sentence, with all edges in the graph illustrated in the adjacency matrix. The red, blue, and orange colors illustrate our modeling for contextual features in *left*, *right*, and *self* positional relations with a specific word, respectively.

(Zhang et al., 2017; Song et al., 2018; Shaw et al., 2018; Tian et al., 2020b), any encoder for EASA could also be beneficial from adding such treatment to model the input text. Therefore, it is expected to enhance conventional GCN with directional information for different parts of input, so that one can distinguish them and appropriately model the contextual information for EASA.

In this paper, we propose directional graph convolutional networks (D-GCN) for EASA, which performs the task following the sequence labeling paradigm and models dependency relations among words in the input with an appropriate architecture. Specifically, for an input sentence, we firstly build the word relation graph upon its auto-processed dependency trees; then, we apply a direction mechanism in GCN, where for each word, we separately encode its associated contextual features (which are suggested by the graph) with respect to different positional relations (i.e., on the left, right or self). To further distinguish the importance of different contextual features, we also propose an attention mechanism, in which we assign different weights to such features that are computed according to the comparisons among them, so as to emphasize important ones for EASA. To illustrate the effectiveness of our approach, experiments are performed on three benchmark datasets, where the results confirm that D-GCN is an appropriate model in leveraging dependency-based word relations for EASA, with state-of-the-art performance observed on all datasets.

2 The Approach

The overall architecture of our approach is illustrated in Figure 1, which follows the sequence labeling paradigm for EASA, where an input sentence $\mathcal{X} = x_1 \cdots x_i \cdots x_n$ is tagged by a corresponding joint label[2] sequence $\widehat{\mathcal{Y}} = \widehat{y}_1 \cdots \widehat{y}_i \cdots \widehat{y}_n$. For the D-GCN, there are L layers placed in between the context encoder (i.e., BERT) and the output layer, where to feed them, an adjacency matrix (shown at the lower right side of Figure 1) representing the graph is built on the dependency tree of the input sentence and an attention matrix (shown at the upper right side of Figure 1) is applied to the edges in the graph to weight the contextual features associated with a specific word, i.e., "*soup*". In the following text, we firstly introduce normal GCN, then elaborate our proposed D-GCN, and finally illustrate EASA labeling with D-GCN.

2.1 Graph Convolutional Networks

The representation of an input sentence always plays an important role in achieving good model performance when it is fed to different natural language processing (NLP) tasks (Song et al., 2017; Babanejad et al., 2020). Contextual features, such as n-grams and syntactic information, have been demonstrated to be highly useful to enhance text representation and thus improve NLP model performance (Huang et al., 2007; Jiang et al., 2009; Wang et al., 2011; Song and Xia, 2012; Song et al., 2012; Song and Xia, 2013; Dong et al., 2014; Miller et al., 2016; Bastings et al., 2017; Marcheggiani and Titov, 2017; Yoon et al., 2018; Seyler et al., 2018; Kumar et al., 2018; Diao et al., 2019; Sun et al., 2019; Zhang et al., 2019; Huang and Carley, 2019; Margatina et al., 2019; De Cao et al., 2019; Tian et al., 2020a; Tian et al., 2020c; Tian et al., 2020d). In addition, it is also proved that GCN could be a powerful model to capture context features suggested by the graph-alike signals, e.g., dependency tree, of an input sentence.

[2] A joint label contains two parts: aspect boundary identifier (i.e., B, I, E, O) and the sentiment mark (i.e., POS, NEG, NEU).

Normal GCN models usually have L layers, and its input graph can be built upon the dependency tree of the input sentence, where an edge is added to every two words, i.e., x_i, and x_j, if there exists a dependency relation between them. In general, a 0-1 adjacency matrix $\mathbf{A} = \{a_{i,j}\}_{n \times n}$ is used to represent the graph where $a_{i,j} = 1$ if there is an edge between x_i and x_j and $a_{i,j} = 0$ otherwise. Based on $\mathbf{A}$, for any x_i in $\mathcal{X}$, the l-th GCN layer takes the output $\mathbf{h}_i^{(l-1)}$ from the $(l-1)$-th GCN layer[3] and computes its output by

$$\mathbf{h}_i^{(l)} = \textit{ReLU}\left(\sum_{j=1}^{n} a_{i,j}\left(\mathbf{W}^{(l)} \cdot \mathbf{h}_j^{(l-1)} + \mathbf{b}^{(l)}\right)\right) \tag{1}$$

where $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are trainable matrix and bias for the l-th layer. Therefore, all contextual features associated with x_i (i.e., all x_j satisfying $a_{i,j} = 1$) are treated equally in normal GCN models.

2.2 Directional Graph Convolutional Networks

The motivation of D-GCN is to separately model contextual features that have different positional relationships with their associated word, and further weight such features according to the comparison among them. Therefore, following the same notations in normal GCN, in the l-th D-GCN layer, our approach to compute the output $\mathbf{h}_i^{(l)}$ for x_i is formalized by

$$\mathbf{h}_i^{(l)} = \textit{ReLU}\left(\sum_{j=1}^{n} p_{i,j}\left(\mathbf{W}_{dir}^{(l)} \cdot \mathbf{h}_j^{(l-1)} + \mathbf{b}^{(l)}\right)\right) \tag{2}$$

where $\mathbf{W}_{dir}^{(l)}$ and $p_{i,j}$ (which correspond to $\mathbf{W}^{(l)}$ and $a_{i,j}$ in Eq.(1), respecitively) show our improvement to normal GCN through the direction modeling and attention mechanism. For the direction information, $\mathbf{W}_{dir}^{(l)}$ encodes the positional relationship of all x_j with respect to x_i and have three choices, i.e., $\mathbf{W}_{left}^{(l)}$, $\mathbf{W}_{right}^{(l)}$, and $\mathbf{W}_{self}^{(l)}$ for different i and j. For example, $\mathbf{W}_{dir}^{(l)} = \mathbf{W}_{left}^{(l)}$ if $j < i$. Then, instead of treading all contextual features equally as that in Eq.(1), attention (through $p_{i,j}$) is applied to the edge between x_i and x_j to weight different contextual features. Specifically, $p_{i,j}$ is computed via

$$p_{i,j}^{(l)} = \frac{a_{i,j} \cdot exp(\mathbf{h}_i^{(l-1)} \cdot \mathbf{h}_j^{(l-1)})}{\sum_{j=1}^{n} a_{i,j} \cdot exp(\mathbf{h}_i^{(l-1)} \cdot \mathbf{h}_j^{(l-1)})} \tag{3}$$

where $\mathbf{h}_i^{(l-1)} \cdot \mathbf{h}_j^{(l-1)}$ computes the interaction between x_i and x_j through inner product. Note that we also apply $a_{i,j}$ from $\mathbf{A}$ to computing $p_{i,j}$ so that the attention for any two words can be easily ignored if there is not an edge between them ($a_{i,j} = 0$).

2.3 Tagging with Directional Graph Convolutional Networks

In our approach, we use BERT (Devlin et al., 2019) to encode the input $\mathcal{X}$ and obtain the hidden vector $\mathbf{h}_i^{(0)}$ for each x_i. Then we feed $\mathbf{h}_i^{(0)}$ to L layers of D-GCN and obtain the corresponding output $\mathbf{h}_i^{(L)}$. Afterwards, we use a trainable matrix $\mathbf{W}$ to align $\mathbf{h}_i^{(L)}$ to the output space by $\mathbf{o}_i = \mathbf{W} \cdot \mathbf{h}_i^{(L)}$. Finally, we apply a *softmax* decoder to $\mathbf{o}_i$ to predict the joint label $\hat{y}_i$ for aspect extraction and sentiment analysis via

$$\hat{y}_i = \arg\max \frac{exp(\mathbf{o}_i^t)}{\sum_{t=1}^{|\mathcal{T}|} exp(\mathbf{o}_i^t)} \tag{4}$$

where $\mathcal{T}$ denotes the label set and o_i^t refers to the value at dimension t in $\mathbf{o}_i$.

3 Experiment

3.1 Settings

In our experiments, we use three benchmark datasets, including restaurant (REST) dataset from SemEval ABSA challenges (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016), laptop (LPTP) dataset from Pontiki et al. (2014), and Twitter (TWTR) dataset from Mitchell et al. (2013). All these datasets contain the ground truth labels of target aspect and their sentiment polarities. Following (Li et al., 2019a; Li et al., 2019b; He et al., 2019; Hu et al., 2019), we only consider three sentiment polarities, i.e., *positive*, *negative*, and *neutral*, where all cases with *conflict* label in REST and LPTP dataset are ignored. We report the statistics (the number of sentences, aspects with respect to *positive*, *neutral*, and *negative* sentiment polarities) of the three datasets in Table 1. For TWTR dataset, since there is no standard train-test split, we only report its total statistics and follow (Mitchell et al., 2013;

[3]The input to the first GCN layer is the hidden vector $\mathbf{h}_i^{(0)}$ obtained from the context encoder, i.e., BERT.

Dataset	LPTP			REST			TWTR
	Train	Test	Total	Train	Test	Total	Total
# Sentence	3,045	800	3,845	3,877	2,158	6,035	2,350
# Aspect	2,300	634	2,934	4,310	2,288	6,598	3,223
# POS	987	339	1,326	2,609	1,524	4,133	698
# NEG	861	130	991	1,035	501	1,536	271
# NEU	452	165	617	666	263	929	2,254

Table 1: The statistics of three benchmark datasets, where number of total sentences and aspect terms, as well as the number of them with positive (POS), negative (NEG), and neutral (NEU) sentiment polarities are reported.

Models	REST	LPTP	TWTR	AVG.
BERT-Base	73.46	60.75	55.95	63.39
+ GAT	75.71	61.23	58.00	64.45
+ 1 D-GCN layer	76.04	65.75	60.21	67.33
+ 2 D-GCN layers	76.36	65.38	59.64	67.12
+ 3 D-GCN layers	**76.75**	**66.61**	**60.66**	**68.00**
+ 4 D-GCN layers	76.69	64.50	59.87	67.02

(a)

Models	REST	LPTP	TWTR	AVG.
BERT-Large	76.37	65.53	58.76	66.89
+ GAT	76.53	66.18	59.28	66.75
+ 1 D-GCN layer	**77.81**	**68.53**	**62.26**	**69.53**
+ 2 D-GCN layers	77.41	68.20	62.03	69.21
+ 3 D-GCN layers	77.78	68.32	62.12	69.40
+ 4 D-GCN layers	77.31	67.49	61.92	68.90

(b)

Table 2: Experimental results (F1 scores) of models with and without D-GCN, as well as a baseline using GAT, on the test sets of three benchmark datasets. For all models, we try BERT-Base (a) and BERT-Large (b) encoders. For our D-GCN models, we try different numbers (1 to 4) of D-GCN layers. An average score column (AVG.) is added to demonstrate the overall performance of different models on the three datasets.

Zhang et al., 2015; Li et al., 2019a; Luo et al., 2019; Hu et al., 2019) to use ten-fold cross-validation on it in our experiments. We use an off-the-shelf system, i.e., Standard CoreNLP Toolkits (SCT)[4] to obtain the dependency tree for each sentence to construct its D-GCN graph, since SCT is a well-known NLP toolkit that has been used in many previous studies (Huang and Carley, 2019; Sun et al., 2019; Tian et al., 2020a). We use uncased version of BERT-Base and BERT-Large[5] (Devlin et al., 2019) under their default settings. All trainable parameters in our D-GCN model are randomly initialized. Following previous studies (Li et al., 2019a; Li et al., 2019b; Luo et al., 2019; He et al., 2019; Hu et al., 2019), we evaluate all models by F1 score.

3.2 Results

In the main experiments, we run our models and baselines with and without D-GCN, and try different numbers (i.e., from 1 to 4) of D-GCN layers. We also run a baseline that uses graph attention networks (GAT) (Veličković et al., 2017) for references. Table 2 shows the results (F1 scores) on all datasets. There are several observations. First, D-GCN works well with both base and large BERT, where consistent improvement is observed over the baselines (including the GAT baseline) across datasets. Second, for models using BERT-Base, three layers of D-GCN achieve the best result, where more or fewer layers cause inferior performance. One possible explanation could be that although we only model the contextual features directly linked to a specific word in each D-GCN layer, contextual information in the larger range can be leveraged indirectly across layers when the number of D-GCN layers increases, so that EASA performance is improved accordingly. However, further adding layer could lead to over-fitting and introduce more noises and thus harm the EASA results. Different from BERT-Base, models using BERT-Large require fewer D-GCN layers to achieve the best performance because BERT-Large is more powerful in encoding contextual information so that they rely less on the long-range contextual information encoded by higher layers of D-GCN. Moreover, we also compare our best model using BERT-base and BERT-large with previous studies, where the results (F1 scores) are presented in Table 3. It is found that our models (especially with BERT-Large) outperform all previous EASA studies. Particularly, although the pipeline approach shows a surprising good performance over other previous studies, we prove that an appropriate model design could effectively take full advantage of the joint approach.

3.3 Ablation Study

To explore the effect of the direction feature (DIR) and the attention mechanism (ATT) applied in our D-GCN, we conduct an ablation study on our best model (BERT-Large) with 1 layer D-GCN, where either DIR or ATT is ablated. The results (F1 scores) on different datasets are reported in Table 4, where the scores from baseline with (ID: 4) and without (ID: 5) normal GCN are also presented. It is clearly indicated that the ablation of either DIR and ATT (ID: 2, 3) hurts model performance, which suggests both parts contribute to improving the EASA task. In

[4]We use the version of 3.9.2 downloaded from `https://stanfordnlp.github.io/CoreNLP/`.

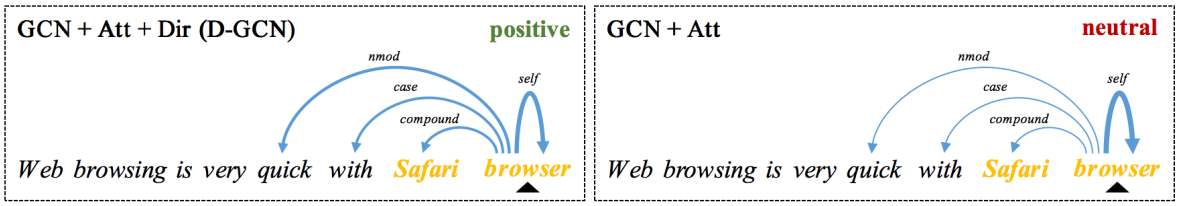
[5]We download different BERT models from `https://github.com/huggingface/transformers`.

Models		REST	LPTP	TWTR
Pipeline	Hu et al. (2019)	74.92	68.06	57.69
Multi-task	He et al. (2019)	-	58.37	-
	Luo et al. (2019)	72.78	60.35	51.37
Joint-label	Li et al. (2019a)	69.80	57.90	48.01
	Li et al. (2019b)	73.24	61.12	-
	Hu et al. (2019)	57.85	48.66	48.11
D-GCN (BERT-Base)		76.75	66.61	60.66
D-GCN (BERT-Large)		**77.81**	**68.53**	**62.26**

Table 3: Comparison of F1 scores between our best models (i.e., 3 D-GCN layers for BERT-Base and 1 D-GCN layer for BERT-Large) with previous studies on all three benchmark datasets.

ID	SETTING ATT	SETTING DIR	REST	LPTP	TWTR
1	√	√	**77.81**	**68.53**	**62.26**
2	×	√	77.16	66.51	61.19
3	√	×	77.23	66.56	61.39
4	×	×	72.91	55.87	52.42
5	Baseline		76.37	65.53	58.76

Table 4: Ablation results from our best model (i.e., BERT-Large encoder with 1 D-GCN layer). "DIR" and "ATT" denote direction modeling and attention mechanism, respectively. √ and × refer to whether a component is used.



(a) (b)

Figure 2: An example sentence with its sentiment outputs from our D-GCN model (with BERT-Large encoder) (a) and a reference model without the direction modeling (i.e., ID 3 in Table 4) (b). The predicted aspect term "*Safari browser*" is highlighted in yellow. The correct and incorrect predicted sentiment polarities are presented in green and red color. We visualize the weights assigned to the contextual features associated with "*browser*" on the arcs (including the arc linking "*browser*" itself) between them, where thicker arc refers to higher weights.

addition, directly using normal GCN (ID: 4) leads to further inferior results compared to baseline (ID: 5) without using it, which emphasizes the necessity of our design to weight different contextual features through D-GCN.

3.4 Case Study

To explore how the D-GCN model captures position information to improve model performance, we explore the effect of direction modeling by comparing the output of our D-GCN models with BERT-Large encoder and a reference model without the direction modeling (i.e., ID 3 in Table 4). In Figure 2, we show an example sentence with the outputs from two models, where both models correctly recognize the aspect term "*Safari browser*" (highlighted in yellow color). In addition, our D-GCN model also correctly predicts the sentiment polarity "*positive*" (in green), while the reference model fails to do so (its output "*neural*" is highlighted in red). In the figure, for "*browser*", we visualize the weight assigned to each of its associated word (i.e., the word that connected to "*browser*" by a dependency arc or "*browser*" itself) on the arc between them, where thicker arcs refer to higher weights. From Figure 2, it is found that the reference model (i.e., GCN + Att.) assigns the highest weight to "*browser*" itself, which makes its associated contextual features fail to contribute to the process of predicting the joint label for "*browser*". On the contrary, our D-GCN approach that considers the directional information allows the attention mechanism to assign higher weights to its contextual features, especially the context word "*quick*" that may provide useful cues to predict a "*positive*" sentiment polarity compared to the reference model. To summarize, this example shows a typical case that, by allowing the attention mechanism to assign appropriate weights to the contextual features, our D-GCN model can leverage the positional relationship between a word and its contextual features to improve the EASA task.

4 Conclusion

In this paper, we propose a joint approach for EASA with D-GCN, whose graph is built upon the dependency tree of the input sentence obtained from off-the-shelf toolkits. The novelty of this work lies in the direction modeling and attention applied in GCN, where in each D-GCN layer, for each word, we separately model its different contextual features by considering their direction to the word, and weight these features according to the comparisons among them. Experimental results on three widely used benchmark datasets illustrate the effectiveness of our approach, with the state-of-the-art performance achieved on all datasets. Further analysis confirms that both direction modeling and attention mechanism are helpful for the task.

Acknowledgements

This work is supported by The Chinese University of Hong Kong (Shenzhen) under University Development Fund UDF01001809.

References

Nastaran Babanejad, Ameeta Agrawal, Aijun An, and Manos Papagelis. 2020. A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5799–5810, Online, July.

Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph Convolutional Encoders for Syntax-aware Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question Answering by Reasoning Across Documents with Graph Convolutional Networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2306–2317.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2019. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. *ArXiv*, abs/1911.00720.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An Interactive Multi-Task Learning Network for End-to-End Aspect-Based Sentiment Analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546.

Binxuan Huang and Kathleen M Carley. 2019. Syntax-Aware Aspect Level Sentiment Classification with Graph Attention Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5472–5480.

Zhongqiang Huang, Mary Harper, and Wen Wang. 2007. Mandarin Part-of-speech Tagging and Discriminative Reranking. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1093–1102.

Wenbin Jiang, Liang Huang, and Qun Liu. 2009. Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and Pos Tagging – A Case Study. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 522–530.

Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR*, abs/1609.02907.

Abhishek Kumar, Daisuke Kawahara, and Sadao Kurohashi. 2018. Knowledge-Enriched Two-Layered Attention Network for Sentiment Analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 253–258.

Hao Li and Wei Lu. 2017. Learning Latent Sentiment Scopes for Entity-Level Sentiment Analysis. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A Unified Model for Opinion Target Extraction and Target Sentiment Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019b. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41.

Huaishao Luo, Tianrui Li, Bing Liu, and Junbo Zhang. 2019. DOER: Dual Cross-Shared RNN for Aspect Term-Polarity Co-Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 591–601.

Dehong Ma, Sujian Li, and Houfeng Wang. 2018. Joint Learning for Targeted Sentiment Analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4737–4742.

Diego Marcheggiani and Ivan Titov. 2017. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515.

Katerina Margatina, Christos Baziotis, and Alexandros Potamianos. 2019. Attention-based Conditioning Methods for External Knowledge Integration. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3944–3951.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-Value Memory Networks for Directly Reading Documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open Domain Targeted Sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654.

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.

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.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria 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.

Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth. 2018. Joint Aspect and Polarity Classification for Aspect-based Sentiment Analysis with End-to-End Neural Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1114.

Dominic Seyler, Tatiana Dembelova, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. A Study of the Importance of External Knowledge in the Named Entity Recognition Task. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 241–246.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468.

Yan Song and Fei Xia. 2012. Using a Goodness Measurement for Domain Adaptation: A Case Study on Chinese Word Segmentation. In *LREC*, pages 3853–3860.

Yan Song and Fei Xia. 2013. A Common Case of Jekyll and Hyde: The Synergistic Effect of Using Divided Source Training Data for Feature Augmentation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 623–631, Nagoya, Japan, October.

Yan Song, Prescott Klassen, Fei Xia, and Chunyu Kit. 2012. Entropy-based Training Data Selection for Domain Adaptation. In *Proceedings of COLING 2012: Posters*, pages 1191–1200, Mumbai, India, December.

Yan Song, Chia-Jung Lee, and Fei Xia. 2017. Learning Word Representations with Regularization from Prior Knowledge. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 143–152.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 175–180.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-Level Sentiment Analysis Via Convolution over Dependency Tree. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5683–5692.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020a. Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296, Online, July.

Yuanhe Tian, Yan Song, and Fei Xia. 2020b. Supertagging Combinatory Categorial Grammar with Attentive Graph Convolutional Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, November.

Yuanhe Tian, Yan Song, Fei Xia, and Tong Zhang. 2020c. Improving Constituency Parsing with Span Attention. In *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing*, November.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020d. Improving Chinese Word Segmentation with Wordhood Memory Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285, Online, July.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph Attention Networks. *arXiv preprint arXiv:1710.10903*.

Yiou Wang, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, Kentaro Torisawa, et al. 2011. Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-Analyzed Data. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 309–317.

Seunghyun Yoon, Joongbo Shin, and Kyomin Jung. 2018. Learning to Rank Question-Answer Pairs Using hierarchical Recurrent Encoder with Latent Topic Clustering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, June.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural Networks for Open Domain Targeted Sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4560–4570.