

Research Article

Multidimensional Self-Attention for Aspect Term Extraction and Biomedical Named Entity Recognition

Xinyu Song , Ao Feng , Weikuan Wang , and Zhengjie Gao 

Chengdu University of Information Technology, Chengdu, China

Correspondence should be addressed to Ao Feng; fengao@cuit.edu.cn

Received 25 August 2020; Revised 3 November 2020; Accepted 18 November 2020; Published 14 December 2020

Academic Editor: Feng Zhang

Copyright © 2020 Xinyu Song et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Wide attention has been paid to named entity recognition (NER) in specific fields. Among the representative tasks are the aspect term extraction (ATE) in user online comments and the biomedical named entity recognition (BioNER) in medical documents. Existing methods only perform well in a particular field, and it is difficult to maintain an advantage in other fields. In this article, we propose a supervised learning method that can be used for much special domain NER tasks. The model consists of two parts, a multidimensional self-attention (MDSA) network and a CNN-based model. The multidimensional self-attention mechanism can calculate the importance of the context to the current word, select the relevance according to the importance, and complete the update of the word vector. This update mechanism allows the subsequent CNN model to have variable-length memory of sentence context. We conduct experiments on benchmark datasets of ATE and BioNER tasks. The results show that our model surpasses most baseline methods.

1. Introduction

With the rapid growth of Internet data, people urgently need to obtain valuable information from massive unstructured text. In this context, the task of named entity recognition (NER) has attracted people's attention. However, most of the research studies are based on the extraction of place names or organization names from general datasets and pay little attention to specific fields. This paper explores aspect term extraction in user online comments and the biomedical named entity recognition in medical documents.

Aspect term extraction (ATE) is an important subtask in aspect-based sentiment analysis [1, 2]. It aimed to detect opinion targets explicitly appearing in the sentences. For example, "Screen is crystal clear and the system is very responsive." The words "Screen" and "system" should be extracted as aspect terms. The target words we extracted must be cooccurring with the emotional words. Therefore, the context is very important. Only emotional entities are the target words that need to be extracted. The purpose of biomedical named entity recognition (BioNER) is to extract specific entities in medical texts such as disease, gene, and

protein names. BioNER is an important subtask in the medical information extraction task. Comprehensive and accurate identification of entities in medical texts is helpful for the full utilization of medical information. At the same time, BioNER is also one of the basic tasks of constructing medical knowledge graphs [3].

In previous works, ATE and BioNER are considered as a token-classification task. Traditional machine learning models like the hidden Markov model (HMM) [4] and conditional random field (CRF) [5, 6] are used to solve the problems. Although these methods achieved reasonable performance, they use handcrafted features, which cannot be applied to other tasks. Recently, deep learning methods have been widely used in natural language processing (NLP). Neural models have become the de-facto standard for high-performance systems and widely used in natural language processing tasks. Many researchers have applied deep learning methods to success in the NER task. Similarly, the named entity recognition task in a special field can still be regarded as a token-level sequence labeling task.

However, it is still a challenging task because there are some problems that need to be solved for NER in special

fields. First, taking the ATE task as an example, the ATE task requires the extraction of terms with sentimental tendencies. For instance, in the sentence “*Sapphire is the only Indian restaurant I go to when I’m in NYC,*” we cannot extract the word “*Sapphire*” because it does not carry the user’s emotions. The model needs to be able to effectively distinguish whether the entity is emotionally inclined. Second, taking the BioNER task as an example, there are a lot of abbreviations in medical texts, and they are often ambiguous. With different contexts, their meanings often change. For instance, CAT can represent “*chloramphenicol acetyl transferase*” and “*computer automated tomography*” [7]. Third, the expansion of the medical exploration field and the rapid growth of the medical entity term have brought great difficulties for the biomedical named entity recognition task. For example, with the spread of COVID-19, “COVID-19” has become a medical entity with practical meaning. The key to solving the problem of NER in special fields is to better mining of the relationship between entities and context. In previous studies, many researchers used the convolutional neural network (CNN) and long-short term memory network (LSTM) to obtain good results in a particular specialty field. But, their mining of the relationship between entities and context is insufficient.

In our paper, we propose a multidimensional self-attention CNN (MDSA-CNN) model for the named entity recognition task in a special field. We use the self-attention mechanism to calculate the dependencies between elements in the input sequence, and the calculation results reflect the importance of the elements. Through the training parameter matrix, it can capture the dependent elements that have a significant contribution to the task, regardless of the distance between the elements in the sequence. After that, the current word vector is updated through the weight, so that the word vector of the current element better expresses the true meaning in the context.

We use two tasks ATE and BioNER to verify the effectiveness of our model. For the ATE task, with extensive experiments on the SemEval-2014 dataset and the SemEval-2016 dataset, the results indicate that our MDSA-CNN model is superior to other baseline methods in aspect term extraction for aspect-level sentiment analysis tasks. For the BioNER task, we use the following datasets for experiments: NCBI-disease and BC2GM. The results show that our model can still obtain results that exceed the benchmark scores on the medical datasets. For the tasks ATE and BioNER on two completely different special field datasets, our models have achieved gratifying results.

2. Related Work

There are few pieces of research on NER in multiple special fields, and more researchers only conduct research in a single special field [8–10]. We take the ATE task and BioNER as examples to describe the development of NER in special fields.

ATE is the first step of aspect-based sentiment analysis. Early winners [5, 6, 11] of SemEval aspect-based sentiment analysis (ABSA) challenges employed traditional sequence

models, such as CRFs and maximum entropy (ME), to extract target words. They are heavily dependent on feature engineering. The aspect terms and opinion terms are extracted together based on the complex syntax pattern [12]. Recently, neural network-based models have become the mainstream method of aspect extraction. A recurrent neural network based on semantic synthesis tasks [13] uses a syntactic analysis tree to identify the emotions of phrases and sentences in order. Irsoy and Cardie [14] apply deep Elman-type RNN to extract opinion and aspect term expressions. LSTM-based [15] and CNN-based [16, 17] methods have achieved good results. Later on, Wang et al. [18] and He et al. [19] employed the attention mechanism to select and focus on the relevant parts of the input for ATE tasks.

BioNER is an important part of the medical text understanding system and plays a decisive role in accurately understanding medical texts. Early researchers used rule-based methods to deal with BioNER tasks [20, 21]. This method relies on a large number of manual features and is difficult to adapt to the rapid development of information in the medical field. After that, CRF-based methods became mainstream [22]. The complexity of medical terms has prompted researchers to use neural networks to handle BioNER tasks. Among them, deep learning methods, especially multitask learning methods, have received extensive attention from researchers [21, 23]. These methods only perform well in a particular field, and it is difficult to maintain an advantage in other fields.

3. Model

3.1. Problem Definition and Notations. The NER in a special field task can be formulated as a token-level sequence labeling problem. Assuming the input is a sequence of word indexes $X = \{x_1, \dots, x_T\}$, we should predict a label sequence $Y = \{y_1, \dots, y_T\}$, where each y_i comes from a finite label set $\mathbb{G} = \{B, I, O\}$. Note that an aspect term can be a phrase, B and I indicate the beginning word and the nonbeginning word of a term phrase, and O indicates other words.

3.2. Model Description. As shown in Figure 1, our model contains two key components, multidimensional self-attention and CNN-based model.

First, we introduce the structure of multidimensional attention. The traditional attention mechanism only calculates a score for each word vector, and multidimensional attention will change this result. It can get a vector with the same length as x_i . The vector is associated with

$$f(x_i, x_j) = W^T \sigma(W^{(1)}x_i + W^{(2)}x_j + b^{(1)}) + b. \quad (1)$$

We initialize $W, W^{(1)}, W^{(2)} \in \mathbb{R}^{d_e \times d_e}$ and use two bias terms before applying activation function. $f(x_i, x_j)$ represents the dependency between x_i and x_j from the same source \mathbf{x} .

After that, we use the softmax function to convert $f(x_i, x_j)$ into a probability distribution:

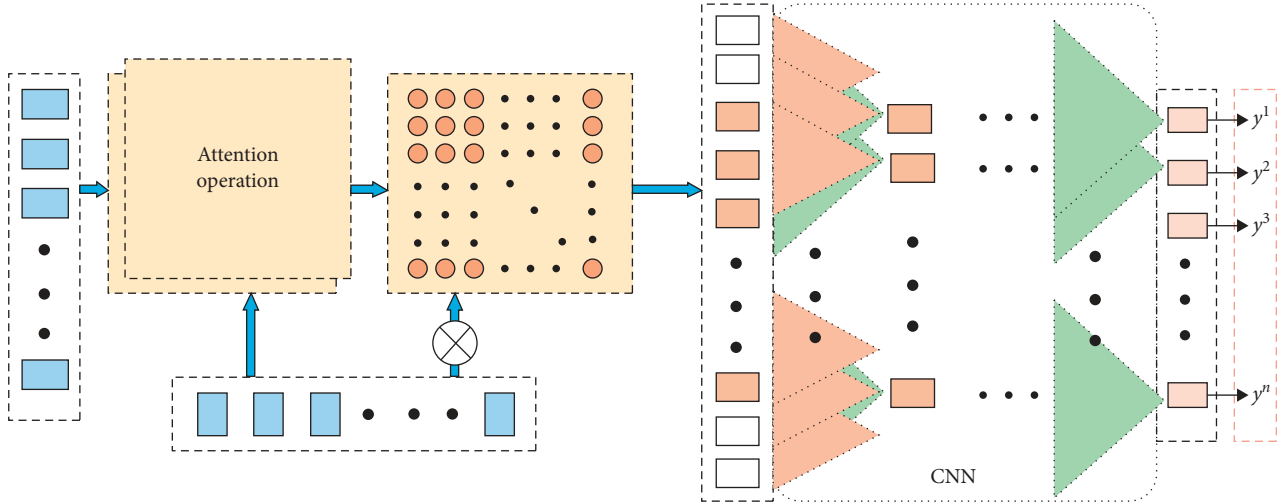


FIGURE 1: The overall architecture of multidimensional self-attention CNN (MDSA-CNN).

$$p(z_k = i | \mathbf{x}, x_j) = \frac{\exp(f(x_i, x_j))}{\sum_{i=1}^n \exp(f(x_i, x_j))}. \quad (2)$$

The value of $p(z_k = i | \mathbf{x}, x_j)$ represents the correlation of the k^{th} dimension of token i to x_j . The larger $p(z_k = i | \mathbf{x}, x_j)$ represents the higher correlation of the two-word vectors.

We use P_{ki}^j to refer to $p(z_k = i | \mathbf{x}, x_j)$. The output s^j can be written as

$$s_j = \left[\sum_{i=1}^n P_{ki}^j \mathbf{x}_{ki} \right]_{k=1}^{d_c} = \left[\mathbb{B}_{i \sim p(z_k | \mathbf{x}, x_j)} \mathbf{x}_{ki} \right]_{k=1}^{d_c}. \quad (3)$$

The output of multidimensional attention for all elements from \mathbf{x} is $\mathbf{s} = [s_1, s_2, \dots, s_n] \in \mathbb{R}^{d_c \times n}$. We use s_i to represent the word vector of each word, and \mathbf{s} is the sentence encoding output. Each word vector is updated by the attention mechanism. At this point, the word vector containing the valid information of the context will be able to more truly express the meaning of the word itself.

We use the word vector processed by the attention mechanism as the network input of the four-layer CNN. We employ one-dimensional convolution with kernel size $k = 2c + 1$ at each layer.

Each filter extracts valid information about a position i in a sentence. Based on the kernel size, we determine that a filter can extract information about the i^{th} word and the context of the window length $2c$. After the CNN network layer, we can get the output vector corresponding to the input vector one to one. For the first CNN ($l = 1$) layer, we use two different filter sizes. For the remaining 3 CNN layers, we only use one filter size. We will explain the experimental parameter setting in detail in the hyperparameter section.

The results of the CNN layer will be processed by the output layer. The calculated results represent the label distribution for each location. The output layer contains the fully connected layer, the softmax function, and the dropout method.

4. Experiments

4.1. Datasets. In our ATE experiment, we used two datasets provided by SemEval ABSA challenge [2, 24] to evaluate our model. Table 1 shows the details of the two benchmark datasets, including the number of sentences and aspects. These two datasets are from subtask 1 of SemEval-2014 Task 4 and subtask 1 of SemEval-2016 Task 5, respectively.

We use two datasets for the BioNER experiment. The two datasets are NCBI-disease [25] and BC2GM [26]. Table 2 shows the details of these two datasets, including the number of sentences and entities.

We used 300D GloVe 840B [27] as the initial word embedding of the model. Glove word vectors were pre-trained on a large amount of text data by the unsupervised methods. At the same time, we used domain embedding [28] as a supplement. The out-of-vocabulary (OOV) words in the training set are randomly initialized by uniform distribution between $(-0.05, 0.05)$. The word embeddings are fine-tuned during the training phase.

4.2. Baseline Methods. We perform a comparison of MDSA-CNN with some baselines using the standard evaluation of the SemVal-2014 Task 4 and SemVal-2016 Task 5:

CRF uses the conditional random field to process the glove word vector [27]. No other neural network structure is involved in training.

IHS RD [5] and NLANGP [29] represent the winners of the original competition [2, 24].

WDEmb [30] sends word embeddings, linear context embeddings, and dependency path embeddings to CRF.

LSTM [15, 31] represents the most basic Bi-LSTM model.

CNN [28] uses the basic CNN model to process glove word vectors for aspect term extraction. This model is to show that the attention mechanism is important.

TABLE 1: Dataset description with the number of sentences and number of aspect terms.

Datasets	Training		Testing	
	Sentences	Aspects	Sentences	Aspects
SemEval-14 laptop	3,045	2,358	800	654
SemEval-16 restaurant	2,000	1,743	676	622

TABLE 2: Dataset description with the number of sentences and number of entity mentions.

Datasets	NCBI-disease	BC2GM
Sentences	7,295	20,000
Entity mentions	6892	24583

Bi-LSTM-CNN-CRF [32] is the mainstream method in named entity recognition (NER). We use this method to prove that there is still room for optimization when the baseline method is applied to a special field.

RNCRF [33] used a recurrent neural network built by a dependency tree to solve the problem of joint extraction of aspect and opinion terms. In addition, CRF and manually extracted features are also used in the model.

CMLA [18] uses a multilayer coupled-attention network to complete the coextraction of aspect and opinion terms.

MIN [31] uses the multitask learning method. This framework is divided into two parts. The first part completes the extraction of aspect and opinion terms, and the second part determines whether the sentence is inductive.

DECNN [28] is a CNN model employing two types of pretrained embeddings for aspect extraction: general-purpose embeddings and domain-specific embeddings.

Benchmark [21]: source of benchmark performance scores of datasets is NCBI-disease.

Multitask learning approach [34] uses multitask methods to train multiple medical datasets simultaneously.

4.3. Hyperparameter. We set the dimension of attention to 2. The first layer of CNN consists of 128 filters with kernel size $k = 3$ and 128 filters with kernel size $k = 5$. The kernel size k of the next three layers of CNN is 5. Each CNN is composed of 256 filters. For window size c , when the kernel size $k = 3$, then $c = 1$, and when $k = 5$, then $c = 2$. We set dropout rate to 0.55. According to the characteristics of CNN training, we set the learning rate to 0.0001. We use the Adam optimizer [35] to optimize the model.

5. Results and Analysis

In Table 3, we find that the performance of MDSA-CNN achieved state-of-the-art results in laptop datasets. By

TABLE 3: Comparison of score results for ATE tasks.

Model	Laptop			Restaurant		
	P	R	F1	P	R	F1
CRF	75.32	72.74	74.01	—	—	—
IHS RD	76.09	73.07	74.55	—	—	—
NLANGP	—	—	—	72.56	72.12	72.34
WDEmb	76.78	73.61	75.16	—	—	—
LSTM	76.37	74.16	75.25	72.09	70.45	71.26
Bi-LSTM-CNN-CRF	78.79	76.83	77.80	73.37	71.65	72.50
RNCRF	79.22	77.64	78.42	—	—	—
CMLA	78.46	77.15	77.80	—	—	—
MIN	75.29	73.88	74.58	—	—	—
DECNN	82.47	80.73	81.59	75.21	73.55	74.37
Our model	82.62	81.19	81.90	73.29	73.91	73.60

comparing with DECNN, we discover that the introduction of multidimensional attention improves performance. We believe that the self-attention mechanism makes up for CNN’s inability to obtain long-distance entity word features and helps the CNN model to capture more contextual body information. In addition, the multidimensional attention mechanism ensures that the model can fuse effective features from different angles. We noticed that after the multidimensional attention mechanism, the word embedding of the target has changed significantly. The change of the word embedding comes from the emotional word vector in the current sentence. For example, “Super light, super sexy and everything just works.” We calculated the MSE distance of the entity “works” word embedding before and after the multidimensional attention mechanism and found that there was a big difference. At the same time, we noticed that on the restaurant dataset, we are slightly behind the DECNN model results. We think the reason is that the restaurant dataset has a small sentence volume and our model is more complex than DECNN. This increases the risk of overfitting. The results in Table 4 show that our multidimensional self-attention mechanism still obtains promising results on medical datasets. The NCBI-disease and BC2GM datasets have exceeded the benchmark results.

TABLE 4: Comparison of score results for the BioNER task.

Model	NCBI-disease			BC2GM		
	P	R	F1	P	R	F1
Benchmark	85.10	80.80	82.90	—	—	—
Multitask learning approach	80.63	80.11	80.37	75.98	70.56	73.17
Our model	85.57	81.28	83.37	77.60	76.98	77.29

6. Conclusions

In order to solve the NER task in a specific field, we designed the MDSA-CNN network. This method introduces a multidimensional self-attention mechanism, which can obtain more comprehensive features of each word in a sentence. Because self-attention can update the characteristics of the current word from any word in the sentence, this method does not rely on the distance between words. For NER in a specific field, we select the ATE task in the online review field and the BioNER task in the medical field. These are two areas where there are huge differences. We conducted experiments on these two tasks. The results show that our model has achieved gratifying results on both tasks. In some datasets, our model achieved state-of-the-art results. This shows that our model is suitable for processing NER tasks in a specific field.

Data Availability

All data supporting this systematic review are from previously reported studies and datasets, which have been cited in the paper. The processed data are available from the first author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the Research Innovation Team Fund (award no. 18TD0026) from the Department of Education, Sichuan Province, and in part by the Sichuan Science and Technology Program (project nos. 2020YFG0168 and 2019YFG0189).

References

- [1] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, August 2004.
- [2] M. Pontiki, D. Galanis, H. Papageorgiou et al., "SemEval-2016 task 5: aspect based sentiment analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 19–30, San Diego, CA, USA, June 2016.
- [3] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, "Learning a health knowledge graph from electronic medical records," *Scientific Reports*, vol. 7, 2017.
- [4] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a high-performance learn-ing name-finder," in *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 194–201, Washington, NY, USA, Mar 1997.
- [5] M. Chernyshevich, "IHS r&d Belarus: cross-domain extraction of product features usingCREF," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval2014)*, pp. 309–313, Dublin, Ireland, August 2014.
- [6] Z. Toh and W. Wang, "DLIREC: aspect term extraction and term polarity classification sys-tem," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval2014)*, pp. 235–240, Dublin, Ireland, August 2014.
- [7] M. Stevenson and Y. Guo, "Disambiguation in the biomedical domain: the role of ambiguity type," *Journal of Biomedical Informatics*, vol. 43, no. 6, pp. 972–981, 2010.
- [8] M. Yin, C. Mou, K. Xiong, and J. Ren, "Chinese clinical named entity recognition with radical-level feature and self-attention mechanism," *Journal of Biomedical Informatics*, vol. 98, Article ID 103289, 2019.
- [9] J. Deng, L. Cheng, and Z. Wang, "Self-attention-based BiGRU and capsule network for named entity recognition," 2020, <https://arxiv.org/abs/2002.00735>.
- [10] S. Peshterliev, C. Dupuy, and I. Kiss, "Self-attention gazetteer embeddings for named-entity recognition," 2020, <https://arxiv.org/abs/2004.04060>.
- [11] I. S. Vicente, X. Saralegi, and R. Agerri, "EliXa: a modular and flexible absa platform," 2017, <https://arxiv.org/abs/1702.01944>.
- [12] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational Linguistics*, vol. 37, no. 1, pp. 9–27, 2011.
- [13] R. Socher, A. Perelygin, J. Wu et al., "Recursivedeep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, WA, USA, October 2013.
- [14] O. Irsoy and C. Cardie, "Opinion mining with deep recurrent neural networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, October 2014.
- [15] P. Liu, S. Joty, and H. Meng, "Fine-grained opinion mining with recurrent neural networks and word embeddings," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1433–1443, Lisbon, Portugal, September 2015.
- [16] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems*, vol. 108, pp. 42–49, 2016.
- [17] M. S. Akhtar, T. Garg, and A. Ekbal, "Multi-task learning for aspect term extraction and aspect sentiment classification," *Neurocomputing*, vol. 398, pp. 247–256, 2020.
- [18] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Coupled multi-layer attentions for co-extraction of aspect and opinion terms," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intel-ligence*, San Francisco, CA, USA, February 2017.
- [19] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "An unsupervised neural attention model foraspect extraction," in *Proceedings of the 55th Annual Meeting of the Association for Com-putational Linguistics (Volume 1: Long Papers)*, Florence, Italy, July 2019.
- [20] X. Wang, Y. Zhang, Q. Li, C. H. Wu, and H. Jiawei, "PENNER: Pattern-enhanced nested named entity recognition in

- biomedical literature,” in *Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine*, Madrid, Spain, December 2018.
- [21] R. Leaman and Z. Lu, “TaggerOne: joint named entity recognition and normalization with semi-Markov Models,” *Bioinformatics*, vol. 32, no. 18, pp. 2839–2846, 2016.
- [22] R. Leaman, C.-H. Wei, C. Zou, and Z. Lu, “Mining chemical patents with an ensemble of open systems,” *Database*, vol. 2016, p. baw065, 2016.
- [23] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, “Deep learning with word embeddings improves biomedical named entity recognition,” *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.
- [24] M. E. Peters, M. Neumann, M. Iyyer et al., “Deep contextualized word representations,” 2018, <https://arxiv.org/abs/1802.05365>.
- [25] R. I. Dogan and Z. Lu, “An improved corpus of disease mentions in PubMed citations,” in *proceedings of the 2012 workshop on biomedical Natural Language processing*, Montréal, Canada, June 2012.
- [26] R. K. Ando, “Biocreative II gene mention tagging system at IBM watson,” In *Second Biocreative Challenge Evaluation Workshop*, Centro nacional de Investigaciones oncologicas (CNIO), Madrid, Spain, 2007.
- [27] J. Pennington, R. Socher, and C. Manning, “GloVe: global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014.
- [28] H. Xu, B. Liu, L. Shu, and P. S. Yu, “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)*, pp. 592–598, Melbourne, Australia, July 2018.
- [29] Z. Toh and J. Su, “NLANGP at SemEval-2016 task 5: improving aspect based sentimentanalysis using neural network features,” in *Proceedings of the 10th International Work-Shop on Semantic Evaluation (SemEval-2016)*, pp. 282–288, San Diego, CA, USA, June 2016.
- [30] Y. Yin, F. Wei, L. Dong, K. Xu, M. Zhang, and M. Zhou, “Unsupervised word and dependency path embeddings for aspect term extraction,” in *Proceedings of the twenty-fifth internaional joint conference on artificial intelligence*, pp. 2979–2985, Alto, CA, USA, July 2016.
- [31] X. Li and W. Lam, “Deep multi-task learning for aspect term extraction with memory interac-tion,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2886–2892, Copenhagen, Den-mark, Sepember 2017.
- [32] N. Reimers and I. Gurevych, “Reporting score distributions makes a difference: performancestudy of LSTM-networks for sequence tagging,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348, Copenhagen, Denmark, September 2017.
- [33] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, “Recursive neural conditional random fieldsfor aspect-based sentiment analysis,” in *Proceedings of the 2016 Conference on Empiri-Cal Methods in Natural Language Processing*, pp. 616–626, Austin, TX, USA, November 2016.
- [34] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, “A neural network multi-task learning approach to biomedical named entity recognition,” *BMC Bioinformatics*, vol. 18, p. 368, 2017.
- [35] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” 2014, <https://arxiv.org/abs/1412.6980>.