Semi-Supervised Learning for Aspect-Based Sentiment Analysis

Hang Zheng*[†], Jianhui Zhang*
*School of Computer Science and Technology,
Hangzhou Dianzi University
HangZhou, China
{zh97, jh_zhang}@hdu.edu.cn

Yoshimi Suzuki[†], Fumiyo Fukumoto[†], Hiromitsu Nishizaki[†]
[†]Integrated Graduate School of Medicine, Engineering,
and Agricultural Sciences, Faculty of Engineering
University of Yamanashi
Kofu, Japan
{ysuzuki, fukumoto, hnishi}@yamanashi.ac.jp

***Abstract*—Aspect-based sentiment analysis is a rapidly growing domain in natural language processing which is a fine-grained study. Within this broad field, most existing studies use large amounts of labeled data by deep learning methods. However, obtaining massive quantities of labeled data to train a deep neural network model is frequently time-consuming and laborious. In this paper, we focus on semi-supervised learning based on ACSA with few labeled data in restaurant reviews and scholarly paper reviews. In order to leverage information from unlabeled data, the semi-supervised learning method-Ladder network is proposed to fix the problem. Furthermore, the pre-trained language models BERT, ALBERT and Longformer are used for text pre-processing and feature extraction. Extensive experiments on both datasets demonstrate the superiority of the Longformer based Ladder Network compared with supervised learning methods and other semi-supervised learning methods including Γ-Model and VAT.**

***Keywords*-Aspect-based sentiment analysis; semi-supervised learning; Ladder Networks; Longformer**

I. Introduction

As a popular area of Natural Language Processing (NLP), Sentiment Analysis (SA), also named opinion mining, aims to discover opinions, evaluations, attitudes and emotions within the context [1]. Hu et al. [2] argue that SA has three levels - document, sentence and aspect. Document-level SA refers to whether the whole document is positive, negative or neutral. Sentence-level SA focuses on the overall polarity of a sentence. They assume that a piece of text has only one sentiment. However, a document/sentence typically contains many different topics or aspects which may have opposite attitudes. Therefore, aspect-level SA concerns each particular aspect of a document/sentence. The Aspect-Based Sentiment Analysis (ABSA) helps understand fine-grained information of a specific aspect or entity in the text. Entities include person, products, events and papers, which have several aspects. For example, the character of a person, the keyboard of a laptop and the substance of a research paper. Based on whether aspects appear explicitly in the text, ABSA can be divided into two subtasks [3]: (i) Aspect term sentiment analysis (ATSA), (ii) Aspect category sentiment analysis (ACSA). TABLE I shows the main difference between them.

Table I
Difference between two subtasks of ABSA

Subtask	Sentence	Sentiment
ATSA	I liked the food and service, but the price was too high.	food: positive service: positive price: negative
ACSA	The menu was great, but the decor is not special.	{food, staff, ambience} food: positive ambience: negative

In this paper, we focus on semi-supervised learning based ACSA with few labeled data in restaurant reviews and research paper reviews. We modify the semi-supervised model Ladder Network (LN)[4] to leverage more information from unlabeled reviews to improve the performance of reviews aspect rating prediction. The key contributions of this paper can be summarized as follows:

(1) We collects and organizes 250 ACL2017 peer reviews with a 1-5 rank for each aspect of papers (e.g., substance, clarity and originality) in PeerRead [5] and 13,000 ICLR2017-2019 reviews without aspect score in OpenReview [6].

(2) We introduces a semi-supervised Ladder Network model to predict various aspect ranks of restaurant reviews and peer reviews with few labeled training data. We focus on fine-grained analysis rather than rating for the whole review.

(3) We discuss the feature extraction capabilities of BERT [7] and BERT-like methods including ALBERT [8] and Longformer [9].

(4) By comparing supervised learning methods with other semi-supervised methods, the experimental results show that the Ladder Network can effectively improve the model prediction in the absence of labeled data in the traditional ABSA dataset SemEval2014 [10] and the dataset we proposed.

II. Related work

A. Aspect Based Sentiment Analysis

Aspect based sentiment analysis is a fundamental task in the sentiment analysis research field. Most of the exiting works generally rely on supervised machine learning [11] and deep learning technology [12]. Although these

2021 International Conference on Cyberworlds (CW) | 978-1-6654-4065-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/CW52790.2021.00042

2642-3596/21/$31.00 ©2021 IEEE
DOI 10.1109/CW52790.2021.00042

technologies are effective, they require a large number of domain-specific datasets which are labeled manually such as SemEval2014 and Twitter. Different from the above works, peer review does not have enough aspect labeled data which is annotated by reviewers during the scientific publishing process. Therefore, we aim at predicting the aspect ratings by few raw data using semi-supervised learning.

B. *Ladder Network*

Valpola [13] firstly proposed the Ladder Network and demonstrated that it improved the performance over unsupervised learning. Rasmus et al.[4] combined Ladder Network with supervised learning and then reached state-of-the-art performance in semi-supervised MNIST and CIFAR-10 classification in image processing. Pan et al.[14] optimized the Ladder Network to sentiment analysis and demonstrated that the model could improve the performance of sentence-level SA tasks with few samples. Our work follows previous work to solve reviews ACSA problem by Ladder Network with word embedding layer before inputs.

C. *Review Rating Prediction*

Pang et al.[15] firstly studied the rating-inference problem(i.e., review rating prediction) who describes it as a multi-class classification/regression task. Most of the subsequent review ratings used IMDb and Yelp review datasets. Furthermore, with the release of PeerRead dataset, more and more studies focus on peer review sentiment analysis including paper acceptance decision and paper review score [16]. However, all tasks predict the overall sentiment to make a final decision or give a rank that is different from our aspect-based review rating.

III. Model

Figure 1 illustrates the structure of the Ladder Network. It consists of two encoders on each side of the figure and one decoder in the middle.

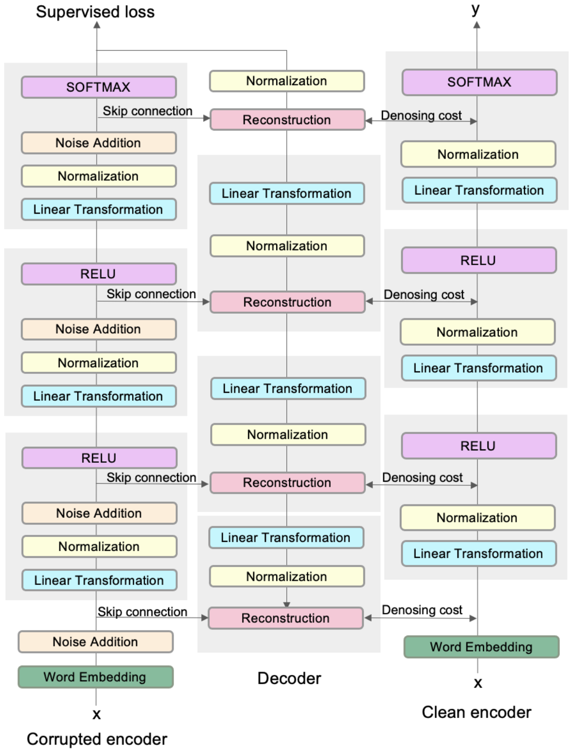

Figure 1. The structure of the Ladder Network

A. *Input Layer*

Before the Ladder Network, we have an input layer to extract features from peer review text named word embedding at the bottom of Figure 1. Recently, various language models pre-training with extensive unannotated corpus have demonstrated the promising capability of representation extraction such as BERT and its variant ALBERT. Considering that peer reviews are usually longer than other reviews(e.g., movie reviews in IMDb and restaurant reviews in Yelp), we also use the Longformer that can learn contextual features from longer reviews for comparison.

B. *Implementation of Ladder Network*

We adopt a fully connected MLP network with rectified linear units for all encoders in the forwarding path, but the topmost layer is softmax instead. We adopt batch normalization for both encoders and decoders to speed up convergence and prevent the decoder from encouraging the trivial solution. At each corrupted encoder, we add isotropic Gaussian noise n when input representation x and after batch normalization:

$$\tilde{x} = \tilde{h}^{(0)} = x + n^{(0)} \tag{1}$$

$$\tilde{z}^{(l)}_{pre} = W^{(l)}\tilde{h}^{(l-1)} \tag{2}$$

$$\tilde{z}^{(l)} = N_B(\tilde{z}^{(l)}_{pre}) + n^{(l)} \tag{3}$$

$$\tilde{h}^{(l)} = activation(\gamma^{(l)}(\tilde{z}^{(l)} + \beta^{(l)})) \tag{4}$$

$$C_s = -\frac{1}{N}\sum_{n=1}^{N} logP(\tilde{y} = y^*(n)|x(n)) \tag{5}$$

where N_B is the batch normalization mentioned, l is the number of layers and the parameters $\gamma^{(l)}$ and $\beta^{(l)}$ are responsible for shifting and scaling before activation function respectively. When removing the Gaussian noise and tilde symbols from the above equations, we will obtain the version of the clean encoder. The supervised cost C_s is equivalent to the negative average log probability of corrupted encoder output matching the true label $y^*(n)$ when the input is $x(n)$.

In decoder path, the $\tilde{z}^{(l)}$ in each layer is rebuilt by following equations:

$$u_{pre}^{(l+1)} = V^{(l)}\hat{z}^{(l+1)} \tag{6}$$

$$u^{(l+1)} = N_B(u_{pre}^{(l+1)}) \tag{7}$$

$$\hat{z}^{(l)} = g(\tilde{z}^{(l)}, u^{(l+1)}) \tag{8}$$

Especially, $u_{pre}^{(L)} = \tilde{y}$ at the top. And the parameter matrix $V^{(l)}$ is the transpose of $W^{(l)}$ in the encoder. The denosing function input $\tilde{z}^{(l)}$ by skip connection and $u^{(l+1)}$ after batch normalization and output the reconstruction value for denosing cost C_d :

$$C_d = \sum_{l=0}^{L} \frac{\lambda_l}{m_l} \frac{1}{N} \sum_{n=1}^{N} \|z^{(l)}(n) - \hat{z}_{BN}^{(l)}\|^2 \tag{9}$$

where λ_l is a hyperparameter to indicate the importance of each decoder, m_l is the width of layer and N is the number of training data. So the final loss function is $C_{total} = C_s + C_d$.

C. Γ-*Model and VAT*

Γ-Model is a variant of the Ladder Network. By removing the most of the decoders except for the top one (i.e., $\lambda_l = 0$, if $l \neq L$) while the encoder in the Γ-Model still consists of clean and noisy paths. Virtual Adversarial Training(VAT) [17] is another semi-supervised learning method that exploits information from unlabeled data by adding perturbations to the word embedding layer. These two methods serve as a control for our experiments.

IV. Experiment

A. *Dataset*

To validate the semi-supervised Ladder Network for ABSA task, we conduct extensive experiments on two datasets. One is a typical dataset called SemEval2014. There are two domain-specific datasets: Laptop14 and Restaurants14. But we only use Restaurants14 four aspects with positive, negative and neutral polarities. The other is a dataset combine 250 ACL2017 labeled reviews in PeerRead with 13,000 ICLR2017-2019 unlabeled reviews in OpenReview. The ACL2017 review data provides several aspects with a 1-5 rank and we choose five aspects: (i) Substance, (ii) Meaningful comparison, (iii) Clarity, (iv) Originality, (v) Soundness/Correctness. Both of them are imbalanced datasets. For example, the proportion of positive, negative, neutral in terms of food aspect in SemEval2014 is about 0.7, 0.2, 0.1. And the data distribution of Clarity aspect in ACL2017 is 0.008, 0.080, 0.180, 0.440, 0.292.

B. *Experimental Settings*

Unlike most reviews, peer review is usually longer than others. So we intercept 300 words for BERT-large and ALBERT-large to avoid exceeding 512 tokens and introduce Longformer-large to do feature extraction for the full text. For the restaurant dataset, we use 100 labeled data with 200 unlabeled for training and 100 for testing in all aspects. All experiments use ALBERT for feature extraction. In the peer review experiments, 100 labeled data with 10,000 unlabeled data for training and 150 data for testing. We construct a 4-layer Ladder Network in which hyperparameters are decided by Optuna.

C. *Results and Discussion*

We used two evaluation metrics, i.e., accuracy and Macro-F1. Accuracy is the most common metric but Macro-F1 is more suitable for measuring the performance of multi-class classification tasks, especially in cases of imbalanced datasets(such as the peer review dataset we used).

Table II
Comparison between Ladder network and other methods

Aspect		**NB**	**SVM**	**Γ-Model**	**VAT**	**Ladder**
Food	ACC	54.00%	56.00%	60.00%	60.00%	**62.00**%
	F1	50.10%	49.69%	54.66%	54.77%	**57.76**%
Ambience	ACC	53.00%	57.00%	69.00%	63.00%	**73.00**%
	F1	46.27%	43.48%	47.51%	46.73%	**48.79**%
Service	ACC	57.00%	63.00%	69.00%	67.00%	**72.00**%
	F1	49.97%	49.30%	53.59%	55.21%	**55.48**%
Anecdote	ACC	46.00%	53.00%	55.00%	55.00%	**57.00**%
	F1	45.92%	49.86%	51.42%	52.23%	**53.01**%

Table III
Performance of Ladder Network with different sizes of data

Aspect		**Labeled/Unlabeled/Test**		
		100/200/100	**100/400/100**	**100/800/100**
Food	ACC	62.00%	65.00%	**68.00**%
	F1	57.76%	60.63%	**62.09**%
Anecdote	ACC	57.00%	59.00%	**61.00**%
	F1	53.01%	56.67%	**58.40**%

TABLE II and III show the main results on SemEval2014. In TABLE II, we compare Ladder Network with two supervised methods: Naive Bayes(NB), Support Vector Machine(SVM) and two semi-supervised methods: Γ-Model and VAT. Each supervised method uses 100 labeled data and the semi-supervised approach has an additional 200 unlabeled data. Generally speaking, semi-supervised methods are better than supervised. Compared with Γ-Model and VAT, the Ladder Network achieves the best performance in terms of accuracy and F1 which means a stronger ability to exploit unlabeled data. With the number of unlabeled data increases to 800 in TABLE III, the accuracy and F1 improved by 4%, 6%; 5.19%, 4.33% for food and anecdote aspects respectively. This also demonstrates the effectiveness of Ladder Network with small amounts of labeled data.

In TABLE IV, we experiment with BERT, ALBERT and Longformer as the input layer to model in peer review dataset. A 300 words window is used to ensure that the requirements of BERT and ALBERT are not exceeded. Longformer uses this way as a comparison. For all experiments here, we collect 100 labeled data and 10,000 unlabeled data for training and 150 for testing. The results

Table IV
COMPARISON OF WORD EMBEDDING METHODS IN PEER REVIEW

Aspect		**BERT** (300)	**ALBERT** (300)	**Longformer** (300)	**Longformer** (ALL)
Substance	ACC	49.33%	53.33%	53.33%	**59.33%**
	F1	20.30%	22.67%	21.89%	**24.93%**
Comparison	ACC	52.67%	54.67%	55.33%	**60.00%**
	F1	23.38%	30.42%	27.54%	**34.40%**
Clarity	ACC	36.67%	42.00%	40.67%	**43.33%**
	F1	24.24%	31.27%	27.05%	**35.70%**
Originality	ACC	38.67%	40.00%	44.67%	**44.67%**
	F1	26.32%	34.36%	30.54%	**38.22%**
Soundness	ACC	43.33%	46.67%	49.33%	**49.33%**
	F1	37.78%	38.34%	35.92%	**44.50%**

show that using the full text for training improves the performance of all aspects in peer review and the Substance is a difficult aspect for the model to correctly identify.

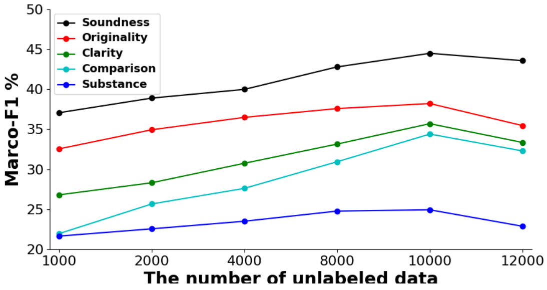

Figure 2. F1 against the number of unlabeled data

Figure 2 shows the F1 on different numbers of unlabeled data in peer review dataset. Until 10,000 unlabeled data, F1 grows steadily as the number of data increases. The semi-supervised noise from unlabeled data, which has a different distribution than the labeled data, affects the model's performance as the unlabeled data continues to increase.

V. CONCLUSION

In this paper, we introduced a semi-supervised model to address the problem of insufficient labeled data of ABSA. The model leveraged 200 additional unlabeled restaurant reviews to improve performance(7.66%, 7.09%, 5.51%, 2.52% in terms of F1 on SemEval2014) over the corresponding supervised learning baseline respectively. In peer review dataset, Longformer using the full review text as input can obtain good performance gain in both accuracy and Marco-F1. We also found that the model suffers when more unlabeled data is used. Future work will include: (i) investigating multi-task learning in peer review, (ii) using other long text feature extraction methods instead of Longformer.

ACKNOWLEDGMENT

This work was supported by JKA through its promotion founds from KEIRIN RACE.

REFERENCES

[1] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5(1)*, 1-167.

[2] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

[3] Xue, W., & Li, T. (2018). Aspect based sentiment analysis with gated convolutional networks. *arXiv preprint arXiv:1805.07043*.

[4] Rasmus, A., Valpola, H., Honkala, M., Berglund, M., & Raiko, T. (2015). Semi-supervised learning with ladder networks. *arXiv preprint arXiv:1507.02672*.

[5] Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*.

[6] Tran, D., Valtchanov, A., Ganapathy, K., Feng, R., Slud, E., Goldblum, M., & Goldstein, T. (2020). An Open Review of OpenReview: A Critical Analysis of the Machine Learning Conference Review Process. *arXiv preprint arXiv:2010.05137*.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[8] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

[9] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

[10] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 27-35).

[11] Vo, D. T., & Zhang, Y. (2015, June). Target-dependent twitter sentiment classification with rich automatic features. In *Twenty-fourth international joint conference on artificial intelligence*.

[12] Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems, 108*, 42-49.

[13] Valpola, H. (2015). From neural PCA to deep unsupervised learning. In *Advances in independent component analysis and learning machines* (pp. 143-171). Academic Press.

[14] Pan, Y., Chen, Z., Suzuki, Y., Fukumoto, F., & Nishizaki, H. (2020, September). Sentiment analysis using semi-supervised learning with few labeled data. In *2020 International Conference on Cyberworlds (CW)* (pp. 231-234). IEEE.

[15] Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

[16] Leng, Y., Yu, L., & Xiong, J. (2019, October). Deepreviewer: Collaborative grammar and innovation neural network for automatic paper review. In *2019 International Conference on Multimodal Interaction* (pp. 395-403).

[17] Miyato, T., Dai, A. M., & Goodfellow, I. (2016). Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.