Overview of the NLPCC 2020 Shared Task: Multi-Aspect-Based Multi-Sentiment Analysis (MAMS)

Lei Chen[1], Ruifeng Xu[2], and Min Yang[1(✉)]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
{lei.chen,min.yang}@siat.ac.cn
[2] Harbin Institute of Technology (Shenzhen), Shenzhen, China
xuruifeng@hitsz.edu.cn

Abstract. In this paper, we present an overview of the NLPCC 2020 shared task on Multi-Aspect-based Multi-Sentiment Analysis (MAMS). The evaluation consists of two sub-tasks: (1) aspect term sentiment analysis (ATSA) and (2) aspect category sentiment analysis (ACSA). We manually annotated a large-scale restaurant reviews corpus for MAMS, in which each sentence contains at least two different aspects with different sentiment polarities. Thus, the provided MAMS dataset is more challenging than the existing aspect-based sentiment analysis (ABSA) datasets. MAMS attracted a total of 50 teams to participate in the evaluation task. We believe that MAMS will push forward the research in the field of aspect-based sentiment analysis.

Keywords: Multi-Aspect-based Multi-Sentiment Analysis · Aspect term sentiment analysis · Aspect category sentiment analysis

1 Introduction

Aspect-based sentiment analysis has attracted increasing attention recently due to its broad applications. It aims at identifying the sentiment polarity towards a specific aspect in a sentence. A target aspect refers to a word or a phrase describing an aspect of an entity. For example, in the sentence "The salmon is tasty while the waiter is very rude", there are two aspect terms "salmon" and "waiter", and they are associated with "positive" and "negative" sentiment, respectively.

Recently, neural network methods have dominated the study of ABSA since these methods can learn important features automatically from the input sequences and be trained in an end-to-end manner. [1] proposed to model the preceding and following contexts for the target via two separate long-short term memory (LSTM) networks. [2] proposed to learn an embedding vector for each aspect, and these aspect embeddings were used to calculate the attention weights

© Springer Nature Switzerland AG 2020
X. Zhu et al. (Eds.): NLPCC 2020, LNAI 12431, pp. 579–585, 2020.
https://doi.org/10.1007/978-3-030-60457-8_48

to capture important information for aspect-level sentiment analysis. [3] developed the deep memory network to compute the importance degree and text representation of each context word with multiple attention layers. [4] introduced the interactive attention networks (IAN) to interactively learn attention vectors for the context and target, and generated the representations for the target and context words separately. [5] extracted sentiment features with convolutional neural networks and selectively output aspect-related features for sentiment classification with gating mechanisms. Subsequently, Transformer [6] and BERT-based methods [7] have achieved noticeable success on ABSA task. [8] combined the capsule network with BERT to improve the performance of ABSA. There are also several studies attempting to simulate the process of human reading cognition to further improve the performance of ABSA [9,10].

So far, several ABSA datasets have been constructed, including SemEval-2014 Restaurant and Laptop review datasets [11], and Twitter dataset [12]. Although these three datasets have since become the benchmark datasets for the ABSA task, most sentences in these datasets consist of only one aspect or multiple aspects with the same sentiment polarity, which makes the ABSA task degenerate to the sentence-level sentiment analysis. Based on our empirical observation, the sentence-level sentiment classifiers (TextCNN and LSTM) without considering aspects can still achieve competitive results with more advanced ABSA methods (e.g., GCAE [5]). On the other hand, even advanced ABSA methods (e.g., AEN [13]) trained on these datasets can hardly distinguish the sentiment polarities towards different aspects in the sentences that contain multiple aspects and multiple sentiments.

In NLPCC 2020, we manually annotated a large-scale restaurant reviews corpus for MAMS, in which each sentence contains at least two different aspects with different sentiment polarities, making the provided MAMS dataset more challenging compared with existing ABSA datasets [8]. Considering merely the sentence-level sentiment of the samples would fail to achieve good performance on MAMS dataset.

This NLPCC 2020 shared task on MAMS has attracted a total of 50 teams to register, and 17 teams submitted the final results. We provide training and development sets to participating teams to build their models in the first stage and evaluate the final results on the test set in the second stage. The final ranking list is based on the average Macro-F1 scores of the two sub-tasks (i.e., ATSA and ACSA).

2 Task Description

Conventional sentiment classification aims to identify the sentiment polarity of a whole document or sentence. However, in practice, a sentence may contain multiple target aspects in a single sentence or document. For example, the sentence "the salmon is tasty while the waiter is very rude" expresses negative sentiment towards the "service" aspect, but contains positive sentiment concerning the "food" aspect. Considering merely the document- or sentence-level sentiment cannot learn the fine-grained aspect-specific sentiment.

Aspect-based sentiment analysis [11], which aims to automatically predict the sentiment polarity of the specific aspect in its context, has gained increasing popularity in recent years due to many useful applications, such as online customer review analysis and conversations monitoring. Similar to SemEval-2014 Task 4, NLPCC-2020 MAMS task also includes two subtasks: (1) aspect term sentiment analysis (ATSA) and (2) aspect category sentiment analysis (ACSA). Next, we will describe the two subtasks in detail.

2.1 Aspect Term Sentiment Analysis (ATSA)

The ATSA task aims to identify the sentiment polarity (i.e., positive, negative or neutral) towards the given aspect terms which are entities presented in the sentence. For example, as shown in the Fig. 1, the sentence "the salmon is tasty while the waiter is very rude" contains two aspect terms "salmon" and "waiter", the sentiment polarities towards the two aspect terms are positive and negative, respectively. Different from the ATSA task in SemEval-2014 Task 4, each sentence in MAMS contains at least two different aspect terms with different sentiment polarities, making the our ATSA task more challenging.

2.2 Aspect Category Sentiment Analysis (ACSA)

The ACSA task aims to identify the sentiment polarity (i.e., positive, negative or neutral) towards the given aspect categories that are pre-defined and may not presented in the sentence. We pre-defined eight aspect categories: food, service, staff, price, ambience, menu, and miscellaneous. For example, the sentence "the salmon is tasty while the waiter is very rude" contains two aspect categories "food" and "service", the sentiment polarities towards the two aspect categories are positive and negative, respectively. For our NLPCC-2020 ACSA task, each sentence contains at least two different aspect categories with different sentiment polarities.

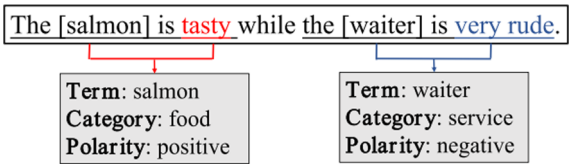

Fig. 1. An example for the ATSA and ACSA tasks.

3 Dataset Construction

Similar to SemEval-2014 Restaurant Review dataset [11], we annotate sentences from the Citysearch New York dataset collected by [14]. We split each document

in the corpus into a few sentences, and remove the sentences consisting more than 70 words. The original MAMS dataset was presented in [8]. In NLPCC-2020 shared task, we relabel the MAMS dataset by providing more high-quality validation and test data.

For the ATSA subtask, we invited three experienced researchers who work on natural language processing (NLP) to extract aspect terms in the sentences and assign the sentiment polarities with respect to the aspect terms. The sentences that consist of only one aspect term or multiple aspects with the same sentiment polarities are deleted. We also provide the start and end positions for each aspect term in the sentence.

For the ACSA subtask, we pre-defined eight coarse aspect categories: food, service, staff, price, ambience, menu, place and miscellaneous. Five aspect categories (i.e., food, service, price, ambience, anecdotes/miscellaneous) are adopted in SemEval-2014 Restaurant Review Dataset. We add three more aspect categories (i.e., staff, menu, place) to deal with some confusing situations. Three experienced NLP researchers were asked to identify the aspect categories described in the given sentences and determine the sentiment polarities towards these aspect categories. We only keep the sentences that consist of at least two unique aspect categories with different sentiment polarities.

The detailed statistics of the datasets for ATSA and ACSA subtasks are reported in Table 1. The released datasets are stored in XML format, as shown in the Fig. 2. Each sample contains the given sentence, aspect terms with their sentiment polarities, and aspect categories with their sentiment polarities. In total, the ATSA dataset consists of 11,186 training samples, 2,668 development samples, and 2,676 test samples. The ACSA dataset consists of 7,090 training samples, 1,789 development samples, and 1,522 test samples.

Table 1. Statistics of MAMS dataset.

Dataset		Positive	Negative	Neutral	Total
ATSA	Training	3,380	2,764	5,042	11,186
	Development	803	654	1,211	2,668
	Test	1,046	545	1,085	2,676
ACSA	Training	1,929	2,084	3,077	7,090
	Development	486	522	781	1,789
	Test	562	348	612	1,522

4 Evaluation Metrics

Both ATSA and ACSA tasks are evaluated using Macro-F1 value that is calculated as follows:

$$Precision(P) = \frac{TP}{TP + FP} \tag{1}$$

```
<sentence id="2846">
  <text>
    Not only was the food outstanding, but the little 'perks' were great.
  <\text>
  <aspectTerms>
    <aspectTerm term="food" polarity="positive" from="17" to="21" />
    <aspectTerm term="perks" polarity="positive" from="51" to="56" />
  </aspectTerms>
  <aspectCategories>
    <aspectCategory category="food" polarity="positive" />
    <aspectCategory category="service" polarity="positive" />
  </aspectCategories>
</sentence>
```

Fig. 2. Dataset format of MAMS task.

$$Recall(R) = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 * \frac{P * R}{P + R} \tag{3}$$

where TP represents true positives, FP represents false positives, TN represents true negatives, and FN represents false negatives. We average the F1 value of each category to get Macro-F1 score. The final result for the MAMS task is the averaged Macro-F1 scores on the two sub-tasks (i.e., ATSA and ACSA).

5 Evaluation Results

In total, there are 50 teams registered for the NLPCC-2020 MAMS task, and 17 teams submitted their final results for evaluation. Table 2 shows the Macro-F1 scores and ranks of these 17 teams. The Macro-F1 results confirmed our expectations. It is noteworthy that we have checked the technique reports of the top three teams and reproduced their codes. Next, we briefly introduce the implementation strategies of the top-3 teams.

The best average Macro-F1 score (82.4230%) was achieved by the *Baiding* team. They tackle the MAMS task as a sentence pair classification problem and employed pre-trained language models as the feature extractor. In addition, the bidirectional gated recurrent unit (Bi-GRU) is connected to the last hidden layer of pre-trained language models, which can further enhance the representation of aspects and contexts. More importantly, a weighted voting strategy is applied to produce an ensemble model that combines the results of several models with different network architectures, pre-trained language models, and training steps.

The *Just a test* team won the 2nd place in the MAMS shared task. They achieved a Macro-F1 score of 85.2435% on the ATSA task and 79.4187% on the ACSA task. The averaged Macro-F1 score was 82.33%, which was slightly worse than that of the *Baiding* team. The RoBERTa-large is used as the pre-trained language model. The *Just a test* team added a word sentiment polarity prediction

task as an auxiliary task and simultaneously predicted the sentiment polarity of all aspects in a sentence to enhance the model performance. In addition, a data augmentation via EDA (Easy data augmentation) [15] is adopted to further improve the performance, which doubled the training data.

The *CUSAPA* team won the third place, which achieved a Macro-F1 score of 84.1585% on the ATSA task and 79.7468% on the ACSA task. The averaged Macro-F1 score was 81.9526%. The *CUSAPA* team employs a joint learning framework to train these two sub-tasks in a unified framework, which improves the performance of both tasks simultaneously. Furthermore, three BERT-based models are adopted to capture different aspects of semantic information of the context. The best performance is achieved by combing these models with a stacking strategy.

Table 2. Macro-F1 scores (%) on the MAMS dataset.

Team	ATSA	ACSA	Average	Rank
Baiding	84.3770	**80.4689**	**82.4230**	**1**
Just a test	**85.2435**	79.4187	82.3311	**2**
CUSAPA	84.1585	79.7468	81.9526	**3**
PingAnPai	84.5463	79.1408	81.8436	4
DUTSurfer	84.1994	78.5792	81.3893	5
wesure01	83.3898	78.3331	80.8615	6
Xiao Niu Dui	83.9645	76.5508	80.2576	7
To be number one	82.4616	76.8539	79.6577	8
AG4MAMS	82.1669	77.0149	79.5909	9
rain2017	80.1005	78.6458	79.3732	10
NLPWUST	81.2856	75.7212	78.5034	11
CABSA	81.6573	72.4605	77.0589	12
MXH42	80.9779	72.1240	76.5510	13
FuXi-NLP	77.9562	73.5253	75.7407	14
YQMAMS	84.0473	47.1836	65.6154	15
W and Triple L	61.3888	63.4616	62.4252	16
HONER	55.9910	49.3538	52.6724	17

6 Conclusion

In this paper, we briefly introduced the overview of the NLPCC-2020 shared task on Multi-Aspect-based Multi-Sentiment Analysis (MAMS). We manually annotated a large-scale restaurant reviews corpus for MAMS, in which each sentence contained at least two different aspects with different sentiment polarities, making the provided MAMS dataset more challenging compared with existing

ABSA datasets. The MAMS task has attracted 50 teams to participate in the competition and 17 teams to submit the final results for evaluation. Different approaches were proposed by the 17 teams, which achieved promising results. In the future, we would like to create a new MAMS dataset with samples from different domains, and add a new cross-domain aspect-based sentiment analysis task.

References

1. Tang, D., Qin, B., Feng, X., et al.: Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING, 3298–3307 (2016)
2. Wang, Y., Huang, M., Zhu, X., et al.: Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 606–615 (2016)
3. Tang, D., Qin, B., Liu, T.: Aspect level sentiment classification with deep memory network. arXiv preprint arXiv:1605.08900 (2016)
4. Ma, D., Li, S., Zhang, X., et al.: Interactive attention networks for aspect-level sentiment classification. arXiv preprint arXiv:1709.00893 (2017)
5. Xue, W., Li, T.: Aspect based sentiment analysis with gated convolutional networks. In: Meeting of the Association for Computational Linguistics, vol. 1, pp. 2514–2523 (2018)
6. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
7. Devlin, J., Chang, M.W., Lee, K., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Jiang, Q., Chen, L., Xu, R., et al.: A challenge dataset and effective models for aspect-based sentiment analysis. 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), pp. 6281–6286 (2019)
9. Lei, Z., Yang, Y., Yang, M., et al.: A human-like semantic cognition network for aspect-level sentiment classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6650–6657 (2019)
10. Yang, M., Jiang, Q., Shen, Y., et al.: Hierarchical human-like strategy for aspect-level sentiment classification with sentiment linguistic knowledge and reinforcement learning. Neural Netw. **117**, 240–248 (2019)
11. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: SemEval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval), pp. 27–35 (2014)
12. Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., Xu, K.: 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), vol. 2, pp. 49–54 (2014)
13. Song, Y., Wang, J., Jiang, T., Liu, Z., Rao, Y.: Attentional encoder network for targeted sentiment classification. arXiv preprint arXiv:1902.09314 (2019)
14. Ganu, G., Elhadad, N., Marian, A.: Beyond the stars: improving rating predictions using review text content. In: WebDB (2009)
15. Wei, J., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196 (2019)