NLPCC 2020 Shared Task 2 Guideline: Multi-Aspect-based Multi-Sentiment Analysis (MAMS)

1. Task overview

Aspect-based sentiment analysis (ABSA) aims at identifying the sentiment polarity towards the 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 decor is not special at all but their amazing food makes up for it", there are two aspect terms "decor" and "food", and they are associated with negative and positive sentiment respectively.

So far, several ABSA datasets have been constructed, including SemEval-2014 Restaurant Review dataset, Laptop Review dataset (Pontiki et al., 2014) and Twitter dataset (Dong et al., 2014). 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 ABSA degenerate to 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 many recent ABSA methods. On the other hand, even advanced ABSA methods 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. Considering merely the sentence-level sentiment of the samples may fail to achieve good performance on MAMS dataset. The MAMS task 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.

(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, given a sentence "The salmon is tasty while the waiter is very rude", as shown in the figure below, the sentence contains two aspect terms "salmon" and "waiter", the sentiment polarity towards the two aspect terms is *positive* and *negative*.

(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, given the same sentence "The salmon is tasty while the waiter is very rude", as shown in the figure below, the sentence contains two aspect categories "food" and "service", the sentiment polarity towards the two aspect categories is *positive* and *negative,* respectively.

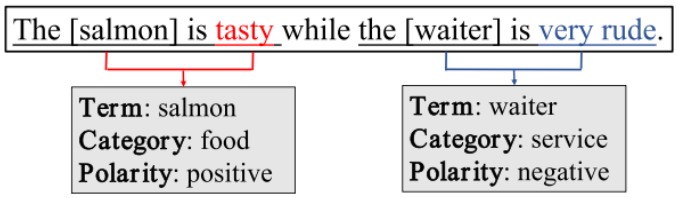

Figure 1. An example for the ATSA and ACSA tasks.

2. Dataset Construction

Similar to SemEval-2014 Restaurant Review dataset (Pontiki et al., 2014), we annotate sentences from the Citysearch New York dataset collected by (Ganu et al., 2009). We split each document in the corpus into a few sentences, and remove the sentences consisting more than 70 words.

For ATSA, we invited three experienced researchers who work on natural language processing (NLP) to extract aspect terms in the sentences and label 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 in a sentence for each aspect term.

For ACSA, we pre-defined eight coarse aspect categories: food, service, staff, price, ambience, menu, place and miscellaneous. Five aspect categories are adopted in

SemEval-2014 Restaurant Review Dataset. We add three more aspect categories to deal with some confusing situations. Three experienced NLP researchers were asked to identify the aspect categories described in given sentences and determine the sentiment polarities towards these aspect categories. We only keep the sentences which consist of at least two unique aspect categories with different sentiment polarities.

We will provide training and development sets to participating teams to build their models. The data is stored in XML format uniformly, as shown in the Figure 2. It contains sentences, aspect terms with their sentiment polarities, and aspect categories with their sentiment polarities. In total, the ATSA dataset consists of 11,186 training samples and 2,668 development samples. The ACSA dataset consists of 7,090 training samples and 1,789 development samples.

```
<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>
```

Figure 2. Dataset format of MAMS task.

3. Evaluation Metrics

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

$$\text{Precision (P)} = TP / (TP+FP)$$

$$\text{Recall (R)} = TP / (TP+FN)$$

$$F1 = 2*P*R / (P+R)$$

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 value. The final result is the average result of Macro-F1 values on the two sub-tasks (i.e., ATSA and ACSA).

4. Registration Requirement

The MAMS task is open to all sectors of society, regardless of age, nationality, colleges, universities, research institutes, and corporate practitioners can register for the competition. Students and employees of relevant units participating in the organization cannot win prizes.

Each participating team is limited to 1 to 5 people and a team leader must be designated. The registration must be completed by filling in the registration form and sending it to the designated email address. All members of the team must provide basic personal information and pass real-name authentication. In principle, all participating teams need to participate in both ATSA and ACSA subtasks, and the final ranking is determined by the averaged Macro-F1 scores of these two subtasks.

5. Submission Format

We will release the test set later, and all participating teams need to submit the results for the test data. We save the result file as **submit.csv** and each line in the result file is formatted as follows:

Sentence_ID; Sentence; Aspect_Term/Aspect_Category; Sentiment_Polarity

For the predict sentiment "Sentiment_Polarity", we use 1 to represent *positive*, 0 to represent *neutral,* and -1 to represent *negative*.

6. Dataset Statement

The competition dataset is owned by the data provider. The competition task and dataset are free and open source on the official competition platform. The data provider authorizes the participants to use the provided data for model training of the specified competition. The participants must not use the data for any commercial purpose. If it is used for scientific research, please indicate that the data comes from the relevant data providing unit and cite the following paper:

[1] Jiang, Qingnan, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. "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.

And the bib format is as follows:

```
@inproceedings{jiang2019challenge,
  title={A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis},
  author={Jiang, Qingnan and Chen, Lei and Xu, Ruifeng and Ao, Xiang and Yang, Min},
  booktitle={ EMNLP-IJCNLP},
  pages={6281--6286},
  year={2019}}
```

7. Competition Rules

(1) Participants are forbidden to improve their performance rankings by using bad loopholes such as rule loopholes or technical loopholes outside the scope of the designated assessment of technical capabilities. It is forbidden to copy other people's works, exchange answers, and use multiple trumpet in the competition. The results will be cancelled and the participants will be seriously dealt with.

(2) Participants are forbidden to use external data (e.g., Twitter, SemEval-2014, Yelp, Movie reviews, Amazon reviews) to improve the results. However, the pretrained language models such as BERT can be used in this competition.

(3) Participants are required to provide their source code, technical documentation and other materials for review by the organizing committee.

(4) The organizing committee reserves the right to modify the rules of the competition, the right to judge and dispose of cheating in the competition, and the right to withdraw or refuse awards to the participating teams that violate the competition rules and fairness.

8. Organizer

Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences and Harbin Institute of Technology (Shenzhen).

If you have any questions or concerns, please do not hesitate to contact Min Yang (min.yang@siat.ac.cn) and Ruifeng Xu (xuruifeng@hit.edu.cn) by email.

Reference:

[1] 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.

[2] 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), volume 2, pages 49–54.

[3] Gayatree Ganu, Nomie Elhadad, and Amlie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In WebDB.

中文版：

1. 任务概述

属性级别情感分析（ABSA）目标在于判别句子中特定属性所表达的情感极性。一个目标属性代指用于描述特定属性实体的单词或者词组，例如，在句子“The decor is not special at all but their amazing food makes up for it”中，包含“decor”和“food”这两个属性，并且分别表达了“负面”和“正面”情感极性。

到目前为止，主要有三个公开的 ABSA 任务数据集，包括 SemEval-2014 Restaurant Review、Laptop Review (Pontiki et al., 2014）以及 Twitter（Dong et al., 2014）数据集。即使这三个数据集已经成为 ABSA 任务的基准数据集，但这些数据集中的大部分句子都只包含一个属性或者多个具有相同情感极性的属性，这使得 ABSA 任务退化成句子级别情感分析。实验结果证明，一些没有考虑特定属性信息的句子级别情感分析模型（例如 TextCNN 和 LSTM）也能在上述三个 ABSA 数据集上表现出不错的效果，甚至与一些最近提出的 ABSA 模型具有相近的结果。另外，大多现有的 ABSA 模型难以处理一个句子中包含多个具有不同情感极性的属性的情况。

在 NLPCC-2020 中，我们定义了多属性多情感分析任务（MAMS），并且人工标注了一个大规模的餐厅评论数据集，其中每个句子都包含至少两个属性，并且这些属性具有不同的情感极性。因此，MAMS 任务相较于以往的数据集更具有

挑战性，如果仅仅考虑句子级别的情感极性将难以在 MAMS 数据集上取得良好的表现。本次 MAMS 评测任务包含两个子任务：（1）基于属性词的情感分析（ATSA）和（2）基于属性类别的情感分析（ACSA）。接下来，我们将详细描述这两个子任务。

（1）基于属性词的情感分析（ATSA）

ATSA 任务的目标是判别给定的属性词所表达的情感极性（正面、负面或者中性），其中属性词是指句子中出现的描述某个属性的实体词或短语。例如，如图 1 所示，给定一个句子“The salmon is tasty while the waiter is very rude”，句子包含两个属性词（分别为“salmon”和“waiter”），所表达的情感极性分别是“正面”和“负面”。

（2）基于属性类别的情感分析（ACSA）

ACSA 任务的目标是判别给定的属性类别所表达的情感极性（正面、负面或者中性），其中属性类别是我们预先定义好的，并且属性类别可能不会出现在给定句子中。我们设定了八个属性类别，分别是“*food*”、“*service*”、“*staff*”、“*price*”、“*ambience*”、“*menu*”、“*place*”、“*miscellaneous*”。例如，给定句子“The salmon is tasty while the waiter is very rude”，如图 1 所示，句子中包含两个属性类别（分别为“food”和“service”），所表达的情感极性分别是“正面”和“负面”。

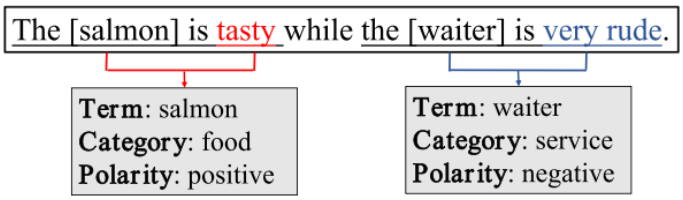

图 1. MAMS 评测任务示例

2. 数据集

与 SemEval-2014 Restaurant Review 数据集（Pontiki et al., 2014）类似，我们从 Citysearch New York 数据集（Ganu et al., 2009）中选取餐厅评论进行数据标注的。我们将语料库中的每个文档分成多个句子，并且删除包含 70 个以上单词的句子。

针对 ATSA 子任务，我们邀请了三位经验丰富的 NLP 研究人员来提取句子中的属性词，并为每个属性词标注情感极性。我们删除了仅包含一个属性词或多个

具有相同情感极性属性词的句子。此外，我们还提供了每个属性词在句子中的开始和结束位置。

针对 ACSA 子任务，我们预先设定了八个属性级别，分别为“*food*”、“*service*”、“*staff*”、“*price*”、“*ambience*”、“*menu*”、“*place*”以及“*miscellaneous*”。除了在 SemEval-2014 Restaurant Review 数据集中定义的五个属性类别外，我们新增加了三个属性类别来处理一些模棱两可的情况。我们要求标注人员确定句子中描述的属性类别，并给出这些属性类别对应的情感极性。与 ATSA 一样，我们只保留包含多个属性类别并且表达不同情感极性的句子。

第一阶段，我们将为参赛选手提供训练集和开发集用以训练模型。数据统一以 XML 格式进行存储，如图 2 所示，包含原始句子、属性词及其表达的情感极性、属性类别及其表达的情感极性。其中，ATSA 数据集包含 11186 条训练集样本和 2668 条开发集样本，ACSA 数据集共包含 7090 条训练集样本和 1789 条开发集样本。

```
<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>
```

图 2. MAMS 评测任务数据格式样例

3. 评测指标

针对 MAMS 评测的两个子任务 ATSA 和 ACSA，均采用 Macro-F1 值进行评测，计算方式如下：

$$\text{精准率 Precision (P)} = TP / (TP+FP)$$

$$\text{召回率 Recall (R)} = TP / (TP+FN)$$

$$\text{F1 值} = 2*P*R / (P+R)$$

其中 TP 是真阳例，FP 是假阳例，TN 是真阴例，FN 是假阴例。我们将首先计算测试集中每一个属性的 F1 值，然后对其求平均值得到模型在整个测试集上

的 Macro-F1 值。注意，每个参赛队伍需要同时完成 ATSA 和 ACSA 两个任务，最终结果为 ATSA 和 ACSA 这两个子任务上的平均 Macro-F1 值。

4. 注册要求

本次测评任务面向社会各界开放，不限年龄、国籍，高校、科研院所、企业从业人员均可报名参赛。另外，参赛者也可以以个人名义参赛，不要求必须提供单位信息。参与比赛组织工作的学生和员工可参与比赛但不可参与评奖。

每支参赛团队限 1-5 人，需指定一名队长，通过填写报名表格发送至指定邮箱完成报名。报名时需提供所有成员的个人基本信息，并通过实名认证。

5. 提交说明

在第二阶段，我们会发布测试集，参赛选手需要根据测试集给出自己的预测结果，并保存为 submit.csv。结果文件中每一行的格式为：

Sentence_ID; Sentence; Aspect_Term/Aspect_Category; Sentiment_Polarity

其中，情感极性预测项“Sentiment_Polarity”需用数字代替，1 代表“正面”，0 代表“中性”，-1 代表“负面”情绪。

6. 数据声明

竞赛数据归数据提供单位所有，赛题及数据在官方竞赛平台进行免费开源，数据提供方授权参赛人员使用提供的数据进行指定比赛的模型训练工作，参赛人员不得将数据用于任何商业用途。若做科研使用，请注明数据来源及引用相关论文。

7. 竞赛规则

（1） 参赛者禁止在指定考核技术能力的范围外，利用规则漏洞或技术漏洞等不良途径提高成绩排名，禁止在比赛中抄袭他人作品、使用多个小号等，经发现将取消比赛成绩并严肃处理。

（2） 本次竞赛不允许参赛选手使用外部数据（例如 Twitter，SemEval-2014，Yelp，Movie reviews，Amazon reviews 等）来提升比赛结果，但允许使用预训练语言模型 BERT 等。

（3） 决赛入围团队需提供代码、技术文档等材料供组委会审查。

（4） 组委会保留对比赛规则进行调整修改的权利、比赛作弊行为的判定权利和

处置权利、收回或拒绝授予影响组织及公平性的参赛团队奖项的权利。

8. 组织方

中国科学院深圳先进技术研究院和哈尔滨工业大学（深圳）。

如果您有任何疑问，请联系杨敏（min.yang@siat.ac.cn）和徐睿峰（xuruifeng@hit.edu.cn）。

参考文献：

[1] 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.

[2] 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), volume 2, pages 49–54.

[3] Gayatree Ganu, Nomie Elhadad, and Amlie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In WebDB.