11747 Assignment3 Report
Topic: Aspect Extraction

Fuyu Tang fuyut

Yifan Song yifanson

Jiameng Du jiamengd

Abstract

Aspect-Based Sentiment Analysis and its sub-task Aspect Extraction play an important role in real commerce and industry. In this report, we first introduce the task with concrete examples. We then go over the history and important related works of this topic, from traditional methods to deep learning. After that, we report the results of the re-implementations for the two selected State-of-the-Art models: DE-CNN and BERT-PT. At last, we discuss the remaining errors and limitations of the current models and propose potential solutions for improvements in the next final assignment.

1 Introduction

Knowing how people feel about the products is extremely important for industries. Thus, sentiment analysis on product reviews and customer opinions is increasingly viewed as crucial commercial success. Most approaches of sentiment analysis attempt to detect the overall polarity of a text, paragraph, or sentence. However, it is also valuable to learn the sentiment polarity on specific aspects (screen, battery, food quality, service) for target entities (laptop, restaurant). This task is called Aspect Based Sentiment Analysis (ABSA), whose goal is to identify aspects and the sentiment for each aspect (Pontiki et al., 2014, 2016). For instance, in the sentence "*The laptop has an incredible speed.*", the aspect is "*speed*" and a positive sentiment is mentioned towards it. The ABSA task can then be divided into two sub-tasks - Aspect Extraction (AE) and Aspect Sentiment Classification (ASC).

In AE task, the objective is to extract all aspects of the target entity, such as "*I liked the* ***food****, but the* ***service*** *was bad.*" The aspect can also be a multi-word term (but should be treated as a single aspect) such as "*The* ***hard disk*** *is too noisy.*" The AE task can be formalized as a sequence labeling task, where each word will be assigned one of the three labels - beginning word of aspect terms (*B*), in the aspect terms (*I*), or out of the aspect terms (*O*). For example, "*The* ***retina display*** *is great.*" can be labeled as "*O B I O O O*" where "*retina display*" is the aspect we would like to extract.

In ASC task, the objective is to determine the sentiment polarity of each aspect term which can be either *positive*, *negative*, or *neutral*. For example:

- "I loved their **fajitas**. (**fajitas**: *positive*)"
- "I hated their **fajitas**. (**fajitas**: *negative*)"
- "The **fajitas** are their first plate. (**fajitas**: *neutral*)".

There are many great works on this vital Natural Language Processing (NLP) task or one of its sub-tasks, especially AE which is a key challenge in ABSA. To narrow down the topic, we'll only focus on the AE task in this report, including the literature survey and reproduced models.

2 Related Work

Aspect Based Sentiment Analysis was first proposed by Hu and Liu (2004). As one of key subtask of ABSA, Aspect Extraction has been studied for more than a decade.

The AE task has been performed by both unsupervised and supervised approaches. One main unsupervised approach is frequent item mining and syntactic rule-based extraction (Zhuang et al., 2006; Qiu et al., 2011). These models depend on pre-defined rules so they only work well when aspects are restricted to a small group of nouns. The other unsupervised approach is to use topic modeling and Latent Dirichlet Allocation (LDA) based models (Titov and McDonald, 2008; Zhao et al., 2010). In these models, corpus is a mixture of topics (aspects), and topics are distributions over words. However, the corpus might be well described by the mixture of aspects but it's easy to have poor-quality individual aspects.

The traditional supervised approach typically uses Hidden Markov Models (HMM) and Conditional Random Fields (CRF) (Jakob and Gurevych, 2010). On top of this basis, part-of-speech and named entity features are incoporated in Chernyshevich (2014) while syntactic features and word embeddings are used in Toh and Su (2016). They won the aspect extraction task in 2014 and 2016 SemEval Challenge respectively.

In recent years, deep learning has become one of the most emerging techniques. Deep neural networks have also been applied to Aspect Extraction. Liu et al. (2015) may be the first work to use vanilla Long Short-Term Memory (LSTM), while Convolutional Neural Networks (CNN) may be the first to propose use on AE in Poria et al. (2016). Later, He et al. (2017) has incorporated the attention mechanism to improve the coherence of aspects by exploiting the distribution of word co-occurrences through neural word embeddings. A joint model has been proposed in Wang et al. (2017) to use multy-layer attention mechanism to jointly extract aspect and opnion terms. Li et al. (2018) has further strengthened the joint model using truncated history-attention and selective transformation network. In Li et al. (2019), the authors have designed a network to transfer aspect knowledge learned from a coarse-grained network. More recently, Graph Convolutional Networks (GCN) has been brought into this field (Zhao et al., 2020).

BERT is one of the key innovations in recent NLP research (Devlin et al., 2019). With the power of BERT, BERT-based models have demonstrated very competitive performance in Aspect Extraction. For instance: Sun et al. (2019) constructs a sequence of auxiliary sentences using the sequence of aspects then fine-tunes BERT with both sequences; Rietzler et al. (2020) first uses self-supervised fine-tuning on domain specific data, then follows the second task-specific fine-tuning stage. More BERT-based works are coming out, and it will not be surprising that they dominate the State of the Art (SOTA) list of Aspect Extraction.

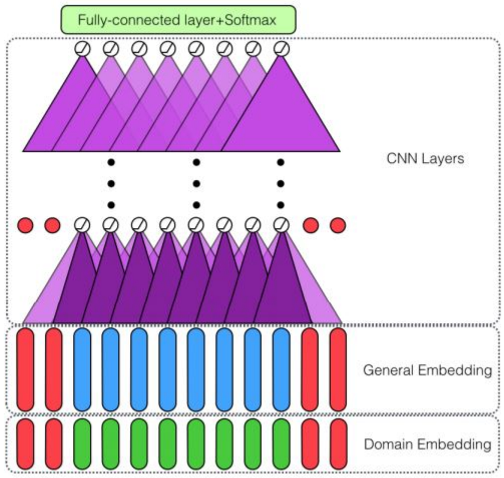

Figure 1: Overview of DE-CNN, figure from original paper

3 Model Reproduction

After reviewing the history and relevant works of AE, we'll pick two typical recent AE models that achieve competitive performance to reproduce, one non-BERT based model and one BERT based model. More impressively, these two highly-cited works are from the same authors.

3.1 DE-CNN

Before BERT came out, Dual Embeddings CNN (DE-CNN), which integrates GloVe and domain-specific embeddings, was the SOTA model of Aspect Extraction (Xu et al., 2018). This work has shown that employing both (pre-trained) general-purpose embeddings and embeddings trained from the domain-specific data would produce a competitive performance even with a relatively simple model to avoid over sophisticated models like many other previous works that cause problems in real deployment. The performance of this model is still very close to the current SOTA and higher than the performance of vanilla BERT and multiple BERT-based models, while requiring much less time and computation resource for training.

3.2 BERT-PT

The second model is called BERT Post-Training (BERT-PT), which is currently one of the SOTA models for Aspect Extraction, and yet another simple but brilliant approach (Xu et al., 2019). In BERT-PT, the authors propose a joint post-training technique that takes BERT's pre-trained

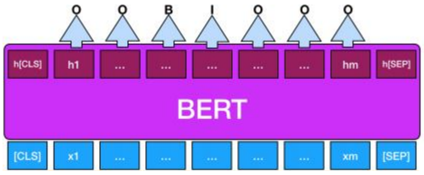

Figure 2: Overview of BERT for AE, figure from original paper

Algorithm 1: Post-training Algorithm

Input: $\mathcal{D}_{\text{DK}}$: one batch of DK data;
$\mathcal{D}_{\text{MRC}}$ one batch of MRC data;
u: number of sub-batches.

$\nabla_{\Theta}\mathcal{L} \leftarrow 0$
$\{\mathcal{D}_{\text{DK},1}, \dots, \mathcal{D}_{\text{DK},u}\} \leftarrow \text{Split}(\mathcal{D}_{\text{DK}}, u)$
$\{\mathcal{D}_{\text{MRC},1}, \dots, \mathcal{D}_{\text{MRC},u}\} \leftarrow \text{Split}(\mathcal{D}_{\text{MRC}}, u)$
for $i \in \{1, \dots, u\}$ **do**
  $\mathcal{L}_{\text{partial}} \leftarrow \frac{\mathcal{L}_{\text{DK}}(\mathcal{D}_{\text{DK},i}) + \mathcal{L}_{\text{MRC}}(\mathcal{D}_{\text{MRC},i})}{u}$
  $\nabla_{\Theta}\mathcal{L} \leftarrow \nabla_{\Theta}\mathcal{L} + \text{BackProp}(\mathcal{L}_{\text{partial}})$
end
$\Theta \leftarrow \text{ParameterUpdates}(\nabla_{\Theta}\mathcal{L})$

Figure 3: Post-training algorithm, *DK* represents domain knowledge data and *MRC* represents task knowledge data, figure from original paper

weights as the initialization for basic language understanding and post-train BERT with both domain knowledge and task knowledge data before fine-tuning on the end-tasks.

3.3 Datasets

Most AE works are performed on two benchmark datasets which consist of review sentences with aspect terms labeled: the *laptop* dataset from subtask 1 of SemEval-2014 Task 4 (Pontiki et al., 2014) and the *restaurant* dataset from subtask 1 of SemEval-2016 Task 5 (Pontiki et al., 2016). Our two models have also been experimented on these two datasets and the performance is measured by *F1 score*. The statistics of these two datasets are shown in Table 1.

3.4 Results

The reproduced results can be seen in Table 2. We have utilized the same hyper-parameters with the original papers to reproduce the two models. The

	Laptop	*Restaurant*
Training	3045 S./2358 A.	2000 S./1743 A.
Testing	800 S./654 A.	676 S./622 A.

Table 1: Summary statistics of datasets, S: number of sentences; A: number of aspects

Model	*Laptop (F1)*	*Restaurant (F1)*
BERT$_{our}$	77.02	72.96
DE-CNN	81.59	74.37
DE-CNN$_{our}$	81.35	74.45
BERT-PT	84.26	77.97
BERT-PT$_{our}$	82.98	77.50

Table 2: F1 results for the original models and our reproduced models

numbers are not exactly same with the original papers but within a very small range of difference. We believe this difference mainly comes from the different devices and different versions of libraries or embeddings we are using. We have also run a vanilla BERT model on the two datasets by fine-tuning with the AE task to serve as a baseline model - both models have outperformed this baseline significantly.

4 Discussion

Though achieving competitive performance, the two models both still have space for improvement. One remaining error that both models are suffering from is what we called "boundary" error. The "boundary" error can be classified into two types:

- the extracted aspect is not complete or only a part of it, for instance, in the sentence "*I tried several monitors with HDMI and this was the case each time.*", the model predicts only "*HDMI*" (or “*monitors*”) as the aspect but the true aspect should be "*monitors with HDMI*""
- the extracted aspect includes leading or tailing terms, typically adjectives, for example, in the sentence "*Buy the separate RAM memory and you will have a rock*", the model predict "*separate RAM memory*" as the aspect but the true aspect should be "*RAM memory*"

The boundary error seems to only cause a minor difference between the extracted aspects and the "gold" labels and doesn’t always appear. However, it still harm the actual performance significantly. One possible approach we plan to adopt is to add another fine-tuning stage after the original model that utilizes a new "boundary/position" neural network which can correct the boundary error by modifying the starting and ending position of the aspect extracted.

Another problem is that AE task doesn't have a large dataset as some other NLP tasks with natural data. The AE dataset requires actual humans to label manually which would be labor-demanding and time-consuming. In addition, both models, as well as most other AE models, perform on the two datasets separately based on category/domain (except the unsupervised ones). It would be interesting to explore if a model can be trained and generalized well across categories/domains. One way to collect more data is to leverage the power of crowdsourcing. As an alternative on the modeling side, to improve the performance, we can also include an Adversarial Training process that adds adversarial examples in the training stage which can help regularize the model and increase robustness.

References

Maryna Chernyshevich. 2014. Crossdomain extraction of product features using crf. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014*.

Jacob Devlin, Ming-Wei Chang, and Kenton Lee. 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)*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Mingqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Xin Li, Lidong Bing, Piji Li, Wai Lam, , and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.

Zheng Li, Ying Wei, Yu Zhang, Xiang Zhang, Xin Li, and Qiang Yang. 2019. Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Pengfei Liu, Shafiq R. Joty, and Helen M. Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Maria Pontiki, Dimitris Galanis, and Haris Papageorgiou. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval-2014)*.

Maria Pontiki, Dimitris Galanis, and Haris Papageorgiou. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. In *Knowledge-Based Systems*.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. In *Computational Linguistics*.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspecttarget sentiment classification. In *Proceedings of The 12th Language Resources and Evaluation Conference*.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. 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)*.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International World Wide Web Conference*.

Zhiqiang Toh and Jian Su. 2016. Nlangp at semeval2016 task 5: Improving aspect based sentiment analysis using neural network features. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016*.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Bert posttraining for review reading comprehension and aspect-based sentiment analysis. 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)*.

Pinlong Zhao, Linlin Hou, and Ou Wu. 2020. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. In *Knowledge-Based Systems*.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*.