Aspect Based Sentiment analysis

Natural Language Process Project

Libin Cheng 19300740005

The Department of Environmental Science and Engineering

1. Introduction

Sentiment analysis is increasingly viewed as a vital task both from an academic and a commercial standpoint. The majority of current approaches, however, attempt to detect the overall polarity of a sentence, paragraph, or text span, regardless of the entities mentioned (e.g., laptops, restaurants) and their aspects (e.g., battery, screen; food, service). By contrast, this task is concerned with aspect based sentiment analysis (ABSA), where the goal is to identify the aspects of given target entities and the sentiment expressed towards each aspect. Datasets consisting of customer reviews with human-authored annotations identifying the mentioned aspects of the target entities and the sentiment polarity of each aspect will be provided.

This project is based on subtask 2 of SemEval-2014 Task 4: Aspect Based Sentiment Analysis. We are required to design some methods for sentiment classification specific to an aspect. For a given set of aspect terms within a sentence, determine whether the polarity of each aspect term is positive, negative, neutral or conflict (i.e., both positive and negative). We have two ABSA tasks, one is in the field of restaurant reviews and the other is in the field of laptop reviews. For example: “I loved their fajitas” → {fajitas: positive}; “I hated their fajitas, but their salads were great” → {fajitas: negative, salads: positive}; “The fajitas are their first plate” → {fajitas: neutral}; “The fajitas were great to taste, but not to see” → {fajitas: conflict}

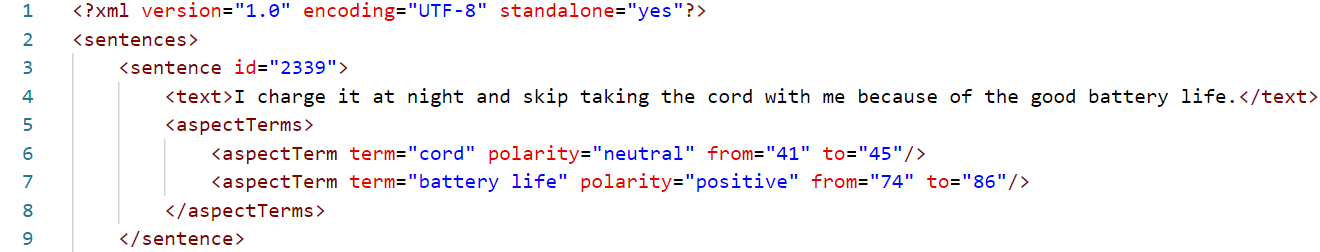
1. Dataset and Format

Two domain-specific datasets for laptops and restaurants, consisting of over 6K sentences with fine-grained aspect-level human annotations have been provided for training.

The dataset of restaurant reviews consists of over 3K English sentences from the restaurant reviews of Ganu et al. (2009). The original dataset of Ganu et al. included annotations for coarse aspect categories and overall sentence polarities; the data provider modified the dataset to include annotations for aspect terms occurring in the sentences, aspect term polarities, and aspect category-specific polarities. Experienced human annotators identified the aspect terms of the sentences and their polarities. Additional restaurant reviews, not in the original dataset of Ganu et al. (2009), are being annotated in the same manner, and they will be used as test data.

The dataset of laptop reviews consists of over 3K English sentences extracted from customer reviews of laptops. Experienced human annotators tagged the aspect terms of the sentences and their polarities. Part of this dataset will be reserved as test data.

The sentences in the datasets are annotated using XML tags. The following example illustrates the format of the annotated sentences of the restaurant’s dataset. The possible values of the polarity field are: “positive”, “negative”, “conflict”, “neutral”. The possible values of the category field are: “food”, “service”, “price”, “ambience”, “anecdotes/miscellaneous”. The format of the laptop’s dataset is the same as in the restaurant datasets, with the only exception that there are no annotations for aspect categories.

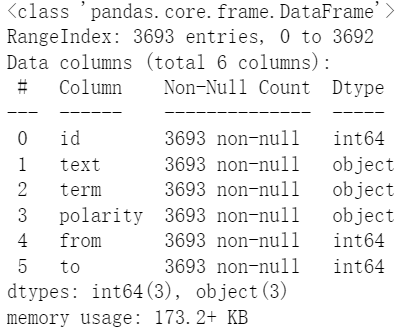
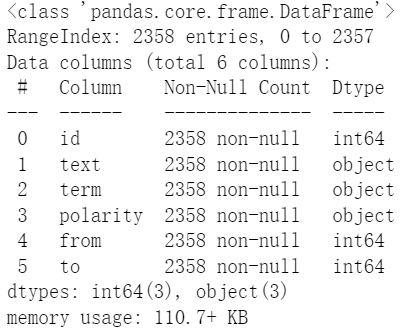


1. Data preprocessing

Firstly, we need to extract useful data from the XML file, which includes the ID of the sentence, the *<text>*, the content of each *<aspectTerm>* in *<aspectTerms>*, polarity, from, and to. Then we need to convert the information of these contents into a CSV file, with each column containing: *id*, *text*, *term*, *polarity*, *from*, and *to*. Then, we can further analyze the data and use *etree.ElementTree* in the *XML* library to parse the XML data, obtaining a list of *text\_ List* and a dictionary operation with term as the key and polarity as the value, using *nltk.FreqDist (operation).most\_common(t)* method obtains the most common *t* aspects, which can analyze which aspects the comment is most concerned about. For example, the following figure shows the top 20 areas that comments are most concerned about in the restaurant dataset.



While processing the file, we can see that there are some abnormal space characters in the file, with which we replace the symbols with spaces. So after preprocessing the file, we get the information of CSV file as follows (left is the dataset of restaurants and right is the dataset of laptop).

Since the test set given by the website does not have polarity, we set 90% of the data set as a training set, and of the remaining 10%, 40% are used as a verification set, and 60% are used as a final test set. The obtained training set, validation set and test set information are shown in the figure below.

1. Methods
2. Evaluations
3. Discussion
4. References

<http://alt.qcri.org/semeval2014/task4/>

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*.