Aspect Based Sentiment Analysis for Myanmar Language

Dr. Yadanar Oo
Natural Language Processing Lab
University of Computer Studies, Yangon, Myanmar
yadanaroo@ucsy.edu.mm

Dr. Khin Mar Soe
Natural Language Processing Lab
University of Computer Studies, Yangon, Myanmar
khinmarsoe@ucsy.edu.mm

***Abstract*—It is a fact of modern life that most customers continuously review the goods or services and rely on that review in order to make purchasing decision of that products. Sentiment analysis has been used by numerous researchers to address the issue in relation to particular features or aspects. This paper, insight into Aspect Based Sentiment Analysis (ABSA) on text level in restaurant review for Myanmar language. Proposed design isolates the undertaking in two subtasks: aspect term extraction and aspect specific sentiment extraction. Because it allows for independent attention to each subtask, this strategy is adaptable. As an initial step, this paper only focus on Aspect-term extraction. To perform Aspect-term extraction, it is considered as a sequence labeling task. The results from this research suggests that analysis of the restaurant reviews. The system utilizes experimentally improved pretrained word embedding features to improve performance.**

Keywords-Aspect-based Sentiment Analysis; Neural Network; Natural Language Processing; Restaurant Review; Aspect Term Extraction;

I. Introduction

Some people have questioned their decision to purchase goods or services in recent decades due to the proliferation of online information. Customers rely on those reviews to highlight the significance of the experience, and writing reviews for a variety of products or services is a rewarding experience. In addition, the majority of customers are able to justify the expenditure of time and money due to online reviews.

The process of classifying a sentence or text according to its polarity is known as sentiment analysis or opinion mining. Text is divided into negative, positive and neutral categories by polarity. Natural language processing(NLP) includes sentiment analysis as a key task, and it has attracted significant attention in recent years. Traditional sentiment analysis focus on extracting general opinion polarities on specific domain. Additionally, customers frequently highlight particular characteristics or aspects of various products or services. It is turning out to be progressively significant for analysis on specific features or aspect. Consequently, numerous analysts have been investigating and refining traditional methodologies.

Recent studies enhance traditional sentiment analysis approach by assuming that sentiments are expressed toward individual aspects or features. This type of analysis is Aspect Based Sentiment Analysis(ABSA). Aspects are mostly domain-dependent and can be defined as the target that the opinion is expressed toward. For example, “I loved this restaurant because of modern environment”. The positive sentiment (“loved, modern”) is expressed toward a certain aspect (“environment”).

In our language, Myanmar, although there are a lot of sentiment analysis processes appear under the complex structure of the language. Most of these early researches are classified on document level and sentence level polarity of sentiment analysis. It groups an input text or a sentence according to their expressing sentiment. There is no research relating to Aspect Based Sentiment Analysis(ABSA) in our language. In this study, we propose a neural network based approach that tackle ABSA for restaurant reviews in Myanmar Language. This is the initial step of ABSA. Moreover, we focus only on the Aspect Term Extraction of the Restaurant Dataset.

II. Background Theory

In the NLP community, neural network models recently attained state-of-the-art performance. Additionally, neural networks are built to mimic the capacities of human decision making.

A. Convolutional Neural Network

CNN is a form of deep, feedforward neural networks. CNN employs a technology similar to a multilayer perceptron that is optimized for low processing demands. Sentences or documents specified as a matrix are the input for CNN in NLP operations. Each rows of the matrix represented a single token, word or letter and might be a syllable as well. This implies that a vector us used to represent each word. These vectors are typically word embeddings. The input, output and hidden layers of a CNN are made up of various convolutional layers, pooling layers, fully connected layers and normalization layers.

Deep learning has dramatically enhanced many NLP processes, including word segmentation, part-of-speech tagging and named entity detection. Neural models have produced state-of-the-art results on a variety of sequence labeling problems. Convolutional Neural Network in Aspect Term Extraction is used in this method to do sequence labeling.

III. PROPOSED SYSTEM

In this study, sub process of the text-level aspect-based sentiment analysis (ABSA) was proposed. Normally, our ABSA architecture divides the task into two subtasks.

Aspect Term Extraction and Aspect Specific Sentiment Extraction. In the first step, the system extracts precisely represented aspects from a given text. A sequence tagging task is carried out by the system in order to extract aspects from the text. In the second step, each extracted aspect term is processed separately and the predicted sentiment label is also assigned to the context of the aspect term. To perform ABSA, both steps are considered sequence labeling tasks. This approach is flexible and it is allowed to work on each subtask individually.

A. Aspect Term Extraction

The extraction of stated aspect phrases is the first step in our process of extracting aspect-based sentiment analysis. In this step, we consider aspect term extraction as a typical syllable-level tagging problem. In this, we use the BIO tagging approach to encode expressed aspect terms. After that, we divide aspect into four groups. There is **food category**, **price category**, **location category** and **service category**. According to these aspects in restaurant reviews, we label the different aspect into BIO format such as at the beginning (B-aspect), inside (I-aspect) or outside of the aspect. Food category is tagged as (B-F, I-F), price category is tagged as (B-P, I-P), location category is labelled as (B-L, I-L) and service category is tagged as (B-S, I-S).

TABLE I. DATASET STATISTICS FOR EACH CATEGORY OF ASPECT TERMS.

Dataset	Train	Dev	Test	Total
Food Category	2,950	680	753	4,383
Price Category	2,080	542	475	3,097
Service Category	1,837	428	287	2,552
Location Category	945	163	138	1,246

Table 1 describes the data statistic for each category of aspect terms. It can be seen that the most concerning category for restaurant dataset is food category, followed by price and service category. There were relatively few comments on location category.

ဒီဆိုင်က အချဉ်ရည် က အရသာ ကောင်းတယ် ။

အချဉ်ရည်/Food Category

အရသာ/Food Category

ကောင်းတယ်/Positive

Fig. 1. Example of aspect-based sentiment analysis for food category.

Figure 1 shows the example of Aspect Term Extraction for food category.

အရသာတော့ အဆင်ပြေပါတယ် serviceလည်းမဆိုးဘူး

အရသာ/Food Category

အဆင်ပြေပါတယ်/Positive

service/ Service Category

မဆိုးဘူး/Positive

Fig. 2. Example of aspect-based sentiment analysis for two different categories.

ဒီနေ့ Time City ကဆိုင်မှာသွားစားလာတာပါ ဈေးနှုန်းနဲ့ အရသာ ဟင်းပွဲ menu အရလက်ခံလို့ရပေမယ့် သန့်ရှင်းမှုပိုင်း တော်တော်အားနည်းပါတယ်

Time City/Location Category

ဈေးနှုန်း/Price Category

အရသာ/Food Category

ဟင်းပွဲ menu/Food Category

လက်ခံလို့ရ/Positive

သန့်ရှင်းမှုပိုင်း/Service Category

တော်တော်အားနည်းပါတယ်/Negative

Fig. 3. Example of aspect-based sentiment analysis for four categories.

There are two different aspect terms food category and service category in figure 2 and four different aspect terms in figure 3. Our system considers aspect term extraction as a typical sequence tagging problem over segmented word and it can also identify the aspect boundary and extract multiple aspect categories.

B. Aspect Specific Sentiment Extraction

The second phase in a two-step architecture for aspect-based sentiment analysis is the extraction of the predicted polarity label given a previously identified aspect term. We use a convolution network to address this aspect specific sentiment extraction and this architecture is very similar to the initial stage of aspect term extraction. Therefore, we use a similar tagging technique as we tag each word in the input sentence with the BIO format. Three labels- positive, negative and neutral – are used in this step.

This paper mainly focuses on Aspect Term Extraction process. As we proposed earlier, we expressed aspect terms as the BIO tagging scheme and group aspect into four different categories.

IV. DATA PREPARATION

In our language, there is no standard corpus for sentiment analysis. In comparison to creating such a corpus in similar language, our language is more difficult and time-consuming. Because there are no clearly defined word boundaries in Myanmar language and no agreement on word division, therefore words are difficult to define. As a result, training corpus are useless because different researchers have different results. This approach performs a sequence labeling scheme at the syllable level, thus overcoming word segmentation difficulties and ambiguities.

In this system, we create corpus of ABSA for restaurant dataset. We labelled 8K sentences by hand. The first 80% for training set, the latter 10% for testing and development set.

V. EXPERIMENTAL RESULTS

Since many machine learning algorithms and the majority of deep learning architectures are unable to analyze plain texts in their raw form, word embedding is necessary in today's world where deep learning approaches are becoming more and more popular for NLP applications. The majority of current word embedding results are typically trained on Wikipedia articles or news items. The performance increases when we use pre-trained embedding with our own training data. Using pre-trained embedding in advance is one of the main goals of our design. We experiment on different word embedding to see how they affect in our aspect-term extraction task's accuracy.

TABLE II. RESULTS WITH BASED LINE , WITH NO EMBEDDING AND WORD2VEC EMBEDDING.

Name	Precision	Recall	F-Measure
Baseline	73.24	69.63	71.35
No Embedding	68.05	71.21	69.59
word2vec 300-d	76.06	70.71	73.25

Table 2 present our result compare to baseline embedding, with no embedding and with our own embedding. Among all the models, the CNN model without embedding had the worst performance in F1 value. The experiment found that the F1 value of the CNN model with own embedding was improved by 1.9% and 3.66% compared with the other two models.

As a different experiment, we assess the performance using various dimension ranging from 100-700.

TABLE III. RESULTS WITH DIFFERENT EMBEDDING DIMENSION.

Name	Precision	Recall	F-Measure
d-100	68.64	71.01	71.80
d-200	71.42	75.25	73.22
d-300	76.06	70.71	73.25
d-400	70.61	76.20	73.24
d-500	71.88	76.48	74.06
d-600	72.92	76.72	74.72
d-700	72.39	76.64	74.41

It is also observed that higher dimension lead to much better performance. Three main observations can be obtained from Table 3. First, CNN based model improvements the performance when we expend dimension of embedding vector. Second, the performance improves as the embedding table's dimension increases. Third, among the models, 600-d shows the best performance.

VI. CONCLUSION

In this study, we proposed aspect-based sentiment analysis tasks for restaurant dataset in Myanmar Language, which can be formulated as two steps sequence labeling problems and we investigate the performance result on our initial step of aspect boundary labels. We proposed our model as syllable level a sequence labeling scheme which can address the segmentation problem. We evaluate performance on CNN architecture. We also execute embedding using the same architecture in a CNN-based model that learns character- and syllable-level representations of syllables for Myanmar words. The comparison of outcomes with various dimensions reveals that performance improves with increasing dimension. In future work, we will increase the training data and we investigate the second step, aspect specific sentiment extraction.

REFERENCES

1. G. Eason, B. Noble, and I. N. Sneddon, "Aspect-Based Sentiment Analysis Using a Two-Step Neural Network Architecture," Soufian Jebbara and Philipp Cimiano. Germany, Bielefeld University, Semantic Computing Group Cognitive Interaction Technology Center of Excellence (CITEC)
2. Xuezhe Ma and Eduard H. Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF," CoRR, abs/1603.01354
3. "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam, 6 Nov 2022
4. Jie Yang and Yue Zhang. NCRF++: An Open-source Neural Sequence Labeling Toolkit. arXiv:1806.05626v2[cs.CL] 17 Jun 2018.
5. Jie Yang, Shuailong Liang, and Yue Zhang. "Design challenges and misconceptions in neural sequence labeling". In COLING, 2018.
6. https://code.google.com/p/word2vec/
7. Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. "All-butthe-top: Simple and effective postprocessing for word representations." arXiv preprint arXiv:1702.01417.