Improvements for NLP Models by Considering Language Differences

Ruiqi Zhang
School of Computer Science, Trinity College Dublin, Dublin, Ireland

*Corresponding author's email: zhangr2@tcd.ie

Abstract. Aspect extraction has been playing a critical role in opinion mining of product reviews. Most of the existing works tackling various Natural Language Processing tasks focus on English corpus and barely talked about the effect of language differences when applying the algorithm to other languages. Based on a previous aspect extraction neural model and its application to text data in Chinese, this paper analyzed possible reasons and proposed plausible approaches to address the importance of handling language differences in Natural Language Processing tasks. By bringing the importance of adaptiveness of language models to attention, future NLP research work might produce more robust and comprehensive work regardless of the impact of language differences.

Keywords. aspect extraction, natural language processing, adaptive language model, language differences

1. Introduction

Aspect extraction is one of the key tasks in sentiment analysis, it aims to extract entity aspects on which opinions have been expressed [1]. For instance, in the sentence “The picture quality is great.”, the aspect term is “picture quality”. Aspect extraction is divided into two sub-tasks: (1) extracting all aspect terms (e.g., “picture quality”) from a review corpus, (2) grouping aspect terms into categories where each category is a single aspect consisted of words with similar meaning.

A previous work by [2] presented an attention-based aspect extraction (ABAE) model that applies neural word embeddings and attention mechanism to extract aspects. Their goal is to learn a set of aspect embeddings which will be interpreted as the aspect terms for each aspect. They used word embeddings to represent each word in the vocabulary. The vectors of each word correspond to the rows of a word embedding matrix. Then sentence embeddings will be constructed using these word embeddings and autoencoder. ABAE then learn an embedding matrix for a set of aspects, which share the same embedding space with the words. By reconstructing and transforming the sentence embeddings from the aspect embeddings with the least distortion, the model obtains the final aspect embeddings.

We intend to bring the language differences in building language models and NLP algorithms into consideration when approaching various NLP tasks. It will be more difficult to develop solutions to an NLP problem and solve that problem despite the impact of language. Similar to word2vec [3], we propose more language-specific word representations that takes language characters into consideration to tackle this problem.
2. Problem

In the original ABAE model, though the extracted aspect terms show a high level of coherence, its ability to identify aspects accurately could be improved. In addition to that, ABAE model uses neural word embeddings to capture the distribution of word co-occurrences, however, the extracted aspect terms still contain some meaningless words that do not help to identify the aspects. It fails when applied to a Chinese corpus in education domain and yields unsatisfactory aspect identification results in our experiment (Figure 1).

**Figure 1.** Aspect identification results on the restaurant domain taken from [2].

While sufficient researches of aspect extraction and identification have been conducted in recent years. However, most of them used online product review data and are not domain independent. Furthermore, most studies focus on English, and few in Chinese. Considering the semantic and structural differences between English and Chinese, experiments have shown that models that perform well on English corpus do not always produce the same good result.

We have applied the ABAE model on our datasets consisted of college students’ review for teachers starting from 2014 to 2016, which are filtered to contain only Chinese characters. By extracting 7 features from the text data and contrast the results obtained from minimum preprocessing and data with segmented words. The results are shown in Figure 2 and Figure 3, it demonstrates that a different sentence representation is effective in discovering meaningful aspects and significantly improves the aspect identification accuracy.

| Aspect  | Method | Precision | Recall | $F_1$ |
|---------|--------|-----------|--------|-------|
| Food    | LocLDA | 0.898     | 0.648  | 0.753 |
|         | ME-LDA | 0.874     | 0.787  | 0.828 |
|         | SAS    | 0.867     | 0.772  | 0.817 |
|         | BTM    | 0.933     | 0.745  | 0.816 |
|         | SERBM  | 0.891     | 0.854  | 0.872 |
|         | k-means| 0.931     | 0.647  | 0.755 |
|         | ABAE   | 0.953     | 0.741  | 0.828 |
| Staff   | LocLDA | 0.804     | 0.585  | 0.677 |
|         | ME-LDA | 0.779     | 0.540  | 0.638 |
|         | SAS    | 0.774     | 0.556  | 0.647 |
|         | BTM    | **0.828** | 0.579  | 0.677 |
|         | SERBM  | 0.819     | 0.582  | 0.680 |
|         | k-means| 0.789     | 0.685  | 0.659 |
|         | ABAE   | 0.802     | **0.728** | **0.757** |
| Ambience| LocLDA | 0.603     | 0.677  | 0.638 |
|         | ME-LDA | 0.773     | 0.558  | 0.648 |
|         | SAS    | 0.780     | 0.542  | 0.640 |
|         | BTM    | 0.813     | 0.599  | 0.685 |
|         | SERBM  | 0.805     | 0.592  | 0.682 |
|         | k-means| 0.730     | 0.637  | 0.677 |
|         | ABAE   | **0.815** | **0.698** | **0.740** |

**Figure 2.** The aspect extraction precision obtained from ABAE model when applied to Chinese corpus without sentence segmentation.
Our proposal: language-specific presentation

We propose that, instead of using a general word2vec [3] sentence representation, aspect extraction and possibly other NLP text related researchers should use language-specific representations before applying them in next stage step, e.g. use more Chinese-tailored language models to represent the text. More language-specific representation indicates better understanding of the meaning of the text, thus allow better performance of the algorithm that uses the sentence representations as input for further processing.

Despite the fact that domain difference partly accounts for the model performance decrease in Figure 1 and Figure 2, language difference can be the more important reason that reduced the model performance. For written language, Chinese does not have an alphabet but uses logographic systems, in which symbols represent the words themselves instead of various letters. Different from English, Chinese is an uninflected language and express meaning by word order, adverbials or shared understanding of the context [4]. Considering the differences of the two languages in many aspects, we propose that different text representations can be explored to apply the original aspect extraction model to Chinese corpuses.

Related work

Language differences have been widely considered in domain-dependent problems and NLP tasks. The fact that various toolkit (e.g. [5,6]) developed for other languages, specifically for Chinese language, proves that language differences indeed need to be taken into consideration in NLP tasks and its applications.

Recently, CNN have achieved excellent results in various natural language applications as the sentence encoder [7]. This motivates us to use a CNN encoder for learning sentence representations to better capture the important words in a sentence and consequently yield better results for aspect extraction. The CNN encoder performs convolution and pooling operations on an input sentence, then uses a fully-connected layer to produce a fixed-length encoding of the sentence as the sentence representations for further operations afterwards.

Next step

We used a real data set to present the importance of language-specific representations in aspect extraction as an example of NLP tasks. The related work shows a few researches in which language variations indeed affects the performance of an algorithm. We first need to identify if and to what extent language-specific representations for NLP processing really matters. We suggest to apply existing algorithms such as ABAE on similar datasets but in Chinese language. It should be interesting to how the models perform simply when changing the data language. To further explore the problem, experiments using different text representation on the same datasets can be carried out in order to find the best language-specific text representation model.

References
[1] M. Hu and B. Liu 2004 Mining opinion features in customer reviews *Proceedings of the 19th national conference on Artificial intelligence* pp 755-760
[2] R. He et al. 2017 An Unsupervised Neural Attention Model for Aspect Extraction (Vancouver, Canada: Association for Computational Linguistics)
[3] T. Mikolov et al. 2013 Distributed Representations of Words and Phrases and their Compositionality in *NIPS*
[4] K. Wong, W. Li, R. Xu, and Z. Zhang 2010 Book Review: Introduction to Chinese Natural Language Processing *Computational Linguistics* 36(4)
[5] X. Qiu, Q. Zhang, and X. Huang 2013 FudanNLP: A Toolkit for Chinese Natural Language Processing (Sofia, Bulgaria: Association for Computational Linguistics)
[6] J. Tao, F. Zhang, A. Li, and Y. Li 2009 Advances in Chinese Natural Language Processing and Language resources
[7] N. Kalchbrenner, E. Grefenstette, and P. Blunsom 2014 A Convolutional Neural Network for Modelling Sentences (Baltimore, Maryland: Association for Computational Linguistics)