AlaBaSoAya: Aspect Based Sentiment Analysis in SemEval 2016 Task 5

Wei Fang, Juei-Yang Hsu

June 2, 2018

1 Introduction

With the rise of social networking and other online platforms, analyzing sentiment and opinions in user-generated content, such as reviews, are becoming increasingly important. In recent years, most sentiment research has focused on classifying the overall sentiment of a document into positive or negative. However, we would often like to understand the sentiments toward specific aspects of entities, for example, a review "The noodles are delicious but service is horrible" contains positive sentiment for the entity "food" and negative sentiment for the entity "service". Aspect Based Sentiment Analysis (ABSA) was proposed to address this problem. This paper describes our approach to the SemEval 2016 Task 5, Aspect Based Sentiment Analysis. In this task, an aspect category is defined as a combination of an entity type E and an attribute type A. E can be the reviewed entity itself (e.g., laptop), a part or component of it (e.g., battery or customer support), or another relevant entity (e.g., the manufacturer of it), while A is a particular attribute (e.g., durability, quality) of E. The task will be evaluated in two phases. In phase A, systems are tested for the aspect category identification slot and the opinion target expression (OTE) detection slots. We will only submit results for the aspect category identification slot. In phase B, the golden labels for the aspect category slot and the OTE slot are provided, and systems must assign a polarity for each opinion. The dataset includes review texts for two domains: restaurants and laptops.

2 Related Work

In last year's edition of this SemEval task (Pontiki et al., 2015), the aspect category identification slot attracted 6 teams for the laptops dataset and 9 teams for the restaurants dataset. The two systems with better F-1 scores in both domains in the aspect slot were from the NLANGP and the Sentinue teams. The NLANGP system modeled the aspect category extraction as a multiclass classification problem with features based on n-grams, parsing, and word clusters learnt from Amazon and Yelp data. The Sentinue system used a MaxEnt classifier with bag-of-word-like features for each entity and each attribute, and then applied heuristics to the output of the classifiers. The sentiment polarity slot attracted 10 teams for laptops and 12 teams for restaurants. The top two systems in both domains were from the Sentinue and ECNU teams. The top system, submitted by Sentinue, applied a MaxEnt classifier along with features based on n-grams, POS tagging, lemmatization, negation words, and publicly available sentiment lexica. The system of ECNU used features based on n-grams, PMI scores, POS tags, parse trees, negation words and scores

based on 7 sentiment lexica.
A type of recurrent neural network, Long Short-Term Memory (LSTM) networks, have superior ability to preserve sequence information over time. The linear chain structured LSTMs have performed well on various tasks, including sentiment classification. A generalization of the LSTM, the tree-structured LSTM (Tree-LSTM) network topology, was proposed, and it outperformed all existing systems on the Stanford Sentiment Treebank task (Tai et al., 2015). Our approach to the SemEval ABSA task combines this topology with other models and classifiers.
Recently, attention-based models that make multiple computation steps using explicit storage have performed well on question answering (QA) tasks. Most natural language processing tasks can be cast into QA problems. Our approach to this ABSA task uses the end-to-end memory network (Sukhbaatar et al., 2015) topology along with other models.

3 Task Description

3.1 Subtask 1: Sentence-level ABSA

Given an opinionated document about a target entity, the goal is to identify all the opinion tuples with the following types of information:
Slot 1: Aspect Category Detection. Identify every entity E and attribute A pair E#A towards which an opinion is expressed in the given text. E and A should be chosen from predefined inventories of entity types and attribute labels. The inventory for the restaurants domain contains 12 E#A pairs, and the inventory for the laptops domain contains 81 E#A pairs.
Slot 2: We will not submit results for this slot so not described here.
Slot 3: Each identified E#A, OTE tuple has to be assigned one of the following polarity labels: positive, negative, or neutral (mildly positive or mildly negative sentiment).

3.2 Subtask 2: Text-level ABSA

Given a set of customer reviews about a target entity, the goal is to identify a set of aspect, polarity tuples that summarize the opinions expressed in each review.

3.3 Subtask 3: Out-of-domain ABSA

Systems will be tested in a previously unseen domain for which no training data will be made available.

3.4 Evaluation

The models are evaluated separately for each slot and subtask on each domain. F1 measure will be used for the evaluation of aspect category detection, and Accuracy for sentiment polarity classification.

4 Models

4.1 Aspect Model

Given a sentence, the aspect-based models predict the possible categories mentioned in this sentence. This is the task of Slot 1 described in the previous section. We use Neural Network (NN), SVM, Maximum Entropy (MaxEnt) as our classifiers.

4.1.1 Feature Selection and Model Description

Because the categories are not dependent on time, it's natural to use non-causal features such as bag-of-words (BOW) or bag-of-word-vectors. With regard to BOW, we tested bi-gram and tri-gram models to build the document-term matrix for training corpus. Also, we used Singular Value Decomposition (SVD) to reduce the dimensionality of the document-term matrix, speeding up the training of different Neural Network models for the same data. SVD was not applied in the cases of the MaxEnt and SVM classifiers. For the bag-of-word-vectors representation, the sentences are first parsed with the Stanford Parser (Socher et al., 2013), and are converted into word vectors by GloVe (Pennington et al., 2014). Each classifier will output the probability distribution of categories, and a threshold determines whether a category should be predicted or not for the sentence.

4.1.2 SVM

We experimented with the linear kernel and the non-linear kernel (RBF) for SVM classifiers, and then chose fine-tuned thresholds for different aspects.

4.1.3 NN

The network was setup with 2-4 hidden layers and 200-500 neurons for each layer during training.

4.2 Sentiment Models

Given a sentence and its golden aspect and OTE labels, the aspect-based sentiment models predicts a polarity for each E#A pair. This is the task of Slot 3 described in the previous section.

4.2.1 The end-to-end LSTM Model (LSTM-NN)

The end-to-end LSTM model consists of two stages. In the first stage of the model, the word vectors of a sentence are fed into an LSTM network. In the second stage, the output of the LSTM network is concatenated with the aspect vector that is encoded in 1-of-N representation, and is then sent through a fully connected network. Finally, a softmax layer is applied to obtain the probabilities of the 3 polarity choices. The model is trained end-to-end. The overall model is shown in Figure 1.

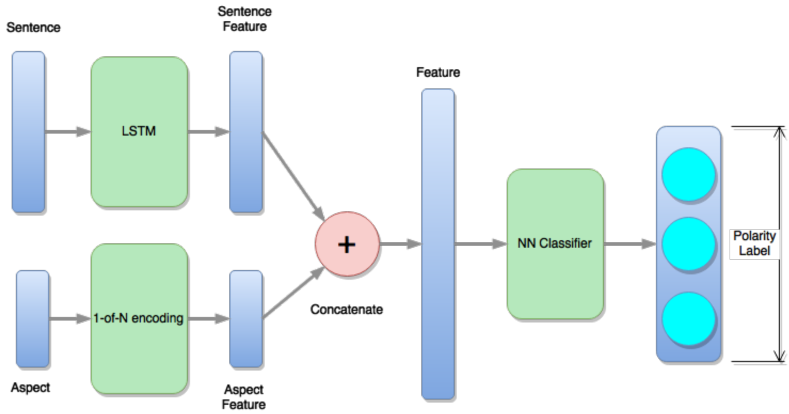

Figure 1: end-to-end LSTM model for polarity detection

4.2.2 The end-to-end Memory Network Model (MemN2N)

The end-to-end Memory Network model follows the topology proposed by Sukhbaatar et al. We modeled this subtask as a QA problem, where the question is the aspect and the answer would be one of the 3 polarities. When the aspect query q is input, it is embedded to obtain an internal state u. At the same time, each word in the sentence is converted into a GloVe vector, and is also embedded into memory vectors m_i. The match, or attention, is computed between u and each memory m_i by taking the inner product followed by a softmax. The attention is then used as weights to extract evidence from another memory embedding, and is summed to get the memory output o. The final prediction is generated by passing the sum of o and u through a fully connected network. Multiple computational steps, or hops, can be performed by passing the sum of o and u through the memory embedding and making another inference before feeding it through the neural network. The model is shown in Figure 2.

4.2.3 The end-to-end LSTM Memory Network Model (LSTM-MemN2N)

A modified end-to-end memory network model is experimented, where the words in a sentence are fed through a LSTM network before being embedded as memory.

4.2.4 The Pre-trained Tree-LSTM Models

The Pre-trained Tree-LSTM models are similar to the LSTM model, except that a Tree-LSTM network, as proposed by Tai et al., replaces the LSTM network in the first stage. The Tree-LSTM is pre-trained by sentences with non-conflicting polarities, that is, excluding those containing conflicting polarities on different aspect categories. The target polarities during pre-training is set to 5, such that for sentences that have the same polarities on different aspects, the polarity label is enhanced. The task dataset does not have fine-grained annotation at each node of the parse tree, so the labels are only applied to the root node. During the second stage of training, the sentences are passed through the pre-trained Tree-LSTM networks, and the features are concatenated with the 1-of-N aspect vector and then fed into a classifier, just like the LSTM model. We experimented with both NN classifiers and SVM classifiers.

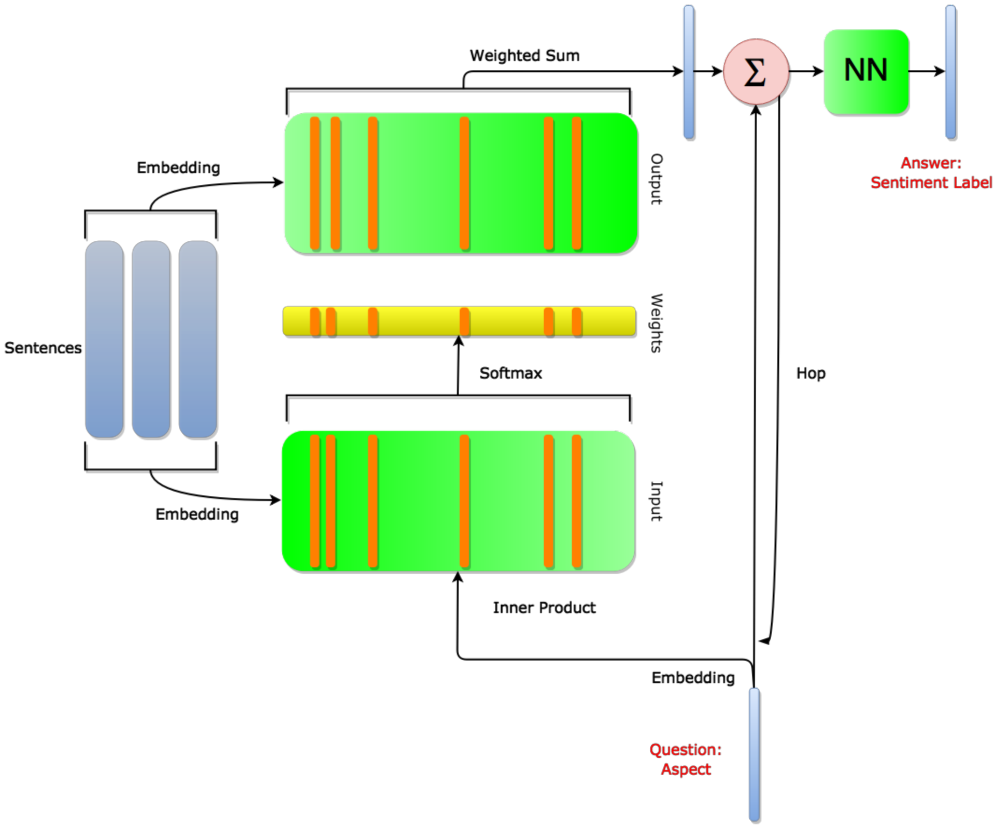

Figure 2: End-to-end Memory Network topology

5 Results

5.1 Aspect Category Detection

Our results for the aspect category detection are summarized in Table 1. Our best model is linear SVM with concatenation of BOW and bag of GloVe vectors, catching both critical entity and semantic relation simultaneously. Other models such as MaxEnt have similar cross-validation accuracies compared with SVM, but are slightly lower.

Model	Laptop	Restaurant
Linear SVM (BOW)	52.16	62.58
Linear SVM (BOW + GloVe)	51.05	64.17
Max Entropy (BOW + GloVe)	50.42	63.72
NN (GloVe)	40.53	62.15

Table 1: 10-fold cross-validation F1 scores for the aspect category detection slot

5.2 Polarity Detection

Our results for the polarity detection slot are summarized in Table 2. The LSTM-NN model outperforms the baseline BOW - Linear SVM, and replacing the NN with MemN2N results in a slightly better accuracy. Although the TreeLSTM models are not trained end-to-end, they outperform all models, producing a cross-validation accuracy about 6-8% higher than the baseline BOW - Linear SVM.

Model	Laptop	Restaurant
BOW - Linear SVM (Baseline)	75.9	76.8
LSTM - NN	77.2	80.1
MemN2N	76.4	79.3
LSTM - MemN2N	79.2	80.0
Pre-trained TreeLSTM - Linear SVM	84.1	82.5
Pre-trained TreeLSTM - NN	84.6	83.7

Table 2: 10-fold Cross-validation accuracies for the polarity detection slot

6 Conclusion & Future Work

In this paper, we developed neural network-based models to improve the detection of sentiments on different aspects for a given review. For the aspect category detection slot, linear SVM with BOW or GloVe features produce the best results. For the polarity detection slot, our best model, a tree-structured Long Short-Term Memory (TreeLSTM) network, outperforms the baseline by quite a margin. The model was pre-trained and is combined with a fully connected network. Because the results were obtained by training the networks separately, we suggest that training the TreeLSTM and the fully connected network end-to-end may result in an even better accuracy. An end-to-end TreeLSTM - MemN2N model could also be worth a test.

References

1. Sukhbaatar, Szlam, Weston, and Fergus. 2015. End-To-End Memory Networks. *arXiv:1503.08895v5* .

2. Tai, Socher, and Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *arXiv:1503.00075v3* .

3. Pontiki, Galanis, Papageorgiou, Manandhar, and Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).*

4. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

5. Richard Socher, John Bauer, Christopher D. Manning and Andrew Y. Ng. 2013. Parsing With Compositional Vector Grammars. *Proceedings of ACL 2013* .