Sentiment Expression Boundaries in Sentiment Polarity Classification

Anonymous EMNLP submission

Abstract

We investigate the effect of using sentiment expression boundaries in predicting sentiment polarity. We manually annotate a freely available sentiment polarity dataset with these boundaries and carry out a series of experiments which demonstrate that high quality sentiment expressions can drastically boost the performance of polarity prediction.

1 Introduction

Sentiment analysis has become a popular focus of natural language processing research, yet it is far from a solved problem, especially at finer-grained levels than the document and sentence. In this work, we aim to improve the performance of target-specific sentiment polarity prediction. We investigate whether identifying the parts of the sentence which carry the sentiment and focusing on them rather than the entire text sentence can be beneficial for this task.

To this end, we manually annotate a data set of about 8K sentences with sentiment expression boundaries. The data was originally released for SemEval 2014 Task 4 (Pontiki et al., 2014), and consists of two review data sets, one of laptops and one of restaurants. The sentiment targets in these data sets, known as *aspect terms*, are various features of laptops and restaurants. For example, in 1, *battery life*, a laptop feature, is an aspect term, a positive sentiment towards which is expressed via *is amazing*.

(1) *The **battery life** is amazing.*

Once the data is annotated with sentiment expressions, we use the annotation to augment aspect-based polarity prediction models and show that this information can greatly increase polarity prediction accuracy. We then build models which predict these annotations and use the outcome in polarity prediction to find how they can be helpful in a more realistic scenario in which gold-standard sentiment expressions are not available.

2 Related Work

To the best of our knowledge, the only data set containing manually annotations of opinion expressions is MPQA (Wiebe et al., 2005). Opinion expressions are categorized into two types: *direct subjective* and *expressive subjective*, the former mentioning the opinions explicitly and the latter implicitly. For example, in (2), *said* is a direct subjective and *full of absurdities* is an expressive subjective opinion expression.

(2) *"The report is full of absurdities," Xirao-Nima said.*

Our annotation scheme does not differentiate between these two types. For this example, our guidelines would annotate *is full of absurdities* as the sentiment expression towards *The report*, as it is clearly the source negative sentiment expressed by the speaker (*Xirao-Nima*).

A related work in terms of utilizing opinion expressions for other opinion mining tasks is (Johansson and Moschitti, 2013), who use features extracted from MPQA opinion expressions in product attribute identification (i.e. finding sentiment targets) and also document polarity classification. These features used in the second task, which is more relevant to this work, include the individual opinion expression words combined with the polarity or type of the expressions. Their results show that information extracted from opinion expressions can help improve polarity classification compared to when only bag-of-word features

are used and when sentiment polarity lexicons are used. In this work, we only use the *boundaries* of the expressions.

In this work, we aim at a more straightforward guideline for annotating opinion or sentiment expressions, where the main rule is to find part of sentence which independently carries the sentiment. Therefore, *said* in 2 would be ignored in our annotation as it does not help recognize the sentiment towards *The report*.

3 Annotating Sentiment Expressions

We define a sentiment expression to be the part(s) of the sentence which expresses the sentiment towards a certain target, implicit or explicit, in the sentence. Our goal is to create a hand-crafted data set which annotates the boundaries of these expressions for each target.

The following sections describe the annotation guidelines and the data used. In all the examples, targets are in bold font and sentiment expressions are underlined.

3.1 Annotation Guidelines

For each target, a span of tokens in the sentence which carries the sentiment expressed towards that target is annotated.[1]

We distinguish between neutral polarities which are associated with a lack of opinion towards the target and which usually do not have a sentiment expression to be annotated, e.g. (3), and neutral polarities which express a neutral opinion, e.g. (4).

We use the following three sentiment expression (SE) labels:

1. *SE* is used when there is only one continuous span of tokens carrying the sentiment.
2. *SE-pcomp* is used to annotate the preceding component of the sentiment expression when it is split into two discontinuous spans, e.g. *Knowledge* in (5).
3. *SE-scomp* is used to annotate the succeeding component of the sentiment expression when it is split into two discontinuous spans, e.g. (*are below average*) in (5).

During the annotation process, the following rules are adhered to:

[1]We use the *brat* annotation tool which is available at http://brat.nlplab.org.

	Laptop		Restaurant	
	Train	Test	Train	Test
# sentences	3045	800	3041	800
# aspect terms	2358	654	3693	1134

Table 1: Number of sentences, aspect terms and their polarity distributions in the data sets

- The annotated span should be as short as possible but should respect the syntactic structure when possible (i.e. a span which is a syntactic constituent is preferred).
- The target term is excluded from the annotation span unless it is inside the span as in (6).
- When there are multiple, independent sentiment expressions targeting the same term, they are annotated using *SE-pcomp* and *SE-scomp* as in (7).
- Copulas connecting adjectives to the targets are included in the OE span in structures such as (8).

(3) *We had **lunch** in that restaurant last week.*

(4) *The **food** was OK.*

(5) *Knowledge of the **chef** and the waitress are below average.*

(6) *The ease of **set up** was terrific.*

(7) *It gives me the power and **speed** that I need to run all the programs I use to edit.*

(8) *The **food** was amazing.*

3.2 Data

The annotation[2] is carried out on the data released for subtask 2 of task 4 of SemEval 2014 (Pontiki et al., 2014) (henceforth known as SE14). The task was to predict the sentiment polarity (*positive, negative, neutral, conflict*) towards a target, or *aspect term*, in the sentence. There are two datasets, one from laptop and one from restaurant consumer reviews, both of which are used here. Examples of aspect terms are laptop screen or restaurant ambiance. Table 1 shows the number of sentences and aspect terms in each set. While some sentences contain multiple aspect terms, some have none. Most sentences, however, contain only one.

During the sentiment expression annotation, the annotator is given an aspect term, the sentence containing it and the sentiment polarity towards

[2]The annotation will be made publicly available.

the aspect term, all of which are already annotated for the shared task. Sentences containing multiple aspect terms are presented multiple times, once per aspect term, making the number of data items equal to the number of aspect terms (7839 in total).

At the first stage of the annotation, 100 aspect terms, 50 per domain, were randomly selected to calibrate the guidelines and train the annotator. Once full agreement was achieved on this subset, the rest of the data underwent annotation.

4 Experiments

The main purpose of sentiment expression annotation is to utilize it in predicting the sentiment polarity. We conduct a series of experiments to examine the degree to which the sentiment expressions can help boost polarity prediction performance. We first measure the upper bound of the improved performance using gold-standard sentiment expressions (*G-SE* henceforth). We also attempt to automatically identify the sentiment expressions and use them instead of gold-standard ones in polarity prediction.

4.1 Polarity Prediction with G-SE

To examine the effect of sentiment expressions, we first evaluate the performance of plain polarity prediction when they are not present - this is our baseline. We then add SE to the model and compare its performance with the baseline.

The polarity prediction (*PP*) systems are built using LSTM networks (Hochreiter and Schmidhuber, 1997). For the baseline, the input layer is the concatenation of an embedding layer, which uses pre-trained *GloVe* (Pennington et al., 2014) word embeddings (1.9M vocabulary *Common Crawl*), concatenated with a layer of binary-value vectors which identify if the token is inside an aspect term span. For the SE-augmented systems, these layers are also concatenated with a 3-dimensional vector which identify if the token is the beginning, inside or outside of the SE boundary.

To tune the hyper-parameters, a development set is randomly sub-sampled from each training set. Table 2 summarizes the hyper-parameters for the baseline $\texttt{PP}_{baseline}$ and gold-standard SE-augmented $\texttt{PP}_{GSE}$ systems. Both unidirectional and bidirectional LSTMs are tried during tuning. In addition to these parameters, we use the *Adam* (Kingma and Ba, 2014) algorithm for optimization, a softmax function at the output layer, and cross-entropy as the loss function. The models are built using *Keras*[3] with a *TensorFlow*[4] backend.

Table 4 displays the accuracy of the baseline polarity prediction system and the one augmented with gold-standard SE tags ($\texttt{PP}_{GSE}$). A substantial rise in the classification accuracy can be seen when SE boundaries are used, ranging from about 12 to 22 percent. This shows that knowing which words in the sentence are part of the sentiment expression is a valuable source of information for polarity classification. In the next section, we try to build a model which can automatically identify these boundaries.

4.2 Sentiment Expression Identification

We build sentiment expression identification (*SEI*) models using LSTM networks, similar to polarity prediction models. The input layer of the SEI models is the same as that of the polarity prediction models described in the previous section. The labels are the `BIO` tags assigned to each token in the sentence, identifying tokens at the beginning, inside and outside of the sentiment expression. We experiment with two models: one where all three tags have equal weights in computing training loss (`SEI`) and one where the `O` tag is down-weighted to account for its dominance in the data ($\texttt{SEI}_w$). The hyper-parameters used after tuning for both models are shown in Table 2 and the evaluation results are presented in Table 3. For weighting the `O` tag, two values were tried , 0.5 and 0.7, the former performing better on the laptop data set and the latter on the restaurant data set. The results in Table 3 are those of the best systems.

Similar to polarity classification, the results indicate that sentiment expression identification is easier on the restaurant data set, The performance on both data sets, however, is not very high, which will likely affect the downstream task of SE-augmented polarity prediction. This will be investigated in the next section. As can also be seen, down-weighting the `O` tag does not improve the performance.

4.3 Polarity Prediction with P-SE

At this stage, we apply the sentiment expression identification model (`SEI`) described in the previous section to the polarity prediction data sets, and use the predicted SE boundaries in building

[3] https://keras.io/
[4] https://www.tensorflow.org/

	Laptop								Restaurant							
	direction	epochs	batch size	#hidden layer	hidden layer size	learning rate	activation	dropout rate	direction	epochs	batch size	#hidden layer	hidden layer size	learning rate	activation	dropout rate
PP$_{baseline}$	uni	50	20	1	50	0.001	tanh	0.3	uni	50	40	3	150	0.001	tanh	0.3
PP$_{GSE}$	uni	50	20	2	50	0.01	tanh	0.3	uni	50	30	2	100	0.001	tanh	0.3
PP$_{GPSE}$	uni	100	40	1	50	0.01	tanh	0	uni	100	40	1	100	0.001	tanh	0
PP$_{PSE}$	uni	40	20	1	50	0.001	tanh	0.5	uni	40	20	1	50	0.001	tanh	0.5
SEI	bi	100	30	2	50	0.001	tanh	0.5	bi	100	30	2	100	0.001	tanh	0
SEI$_w$	bi	100	30	2	100	0.001	tanh	0.5	bi	100	30	2	100	0.001	tanh	0.5

Table 2: Hyper-parameters of the polarity prediction and sentiment expression identification models tuned on development sets

	Laptop						Restaurant					
	Dev			Test			Dev			Test		
	Prec.	*Rec.*	*F*1	*Prec.*	*Rec.*	*F*1	*Prec.*	*Rec.*	*F*1	*Prec.*	*Rec.*	*F*1
SEI	0.70	0.73	0.71	0.64	0.72	0.68	0.72	0.72	0.72	0.70	0.80	0.74
SEI$_w$	0.71	0.71	0.71	0.62	0.75	0.68	0.67	0.76	0.71	0.64	0.79	0.71

Table 3: Precision. recall and F1 of sentiment expression identification models

	Laptop		Restaurant	
	Dev	Test	Dev	Test
PP$_{baseline}$	67.10	64.37	71.96	74.43
PP$_{GSE}$	84.32	86.09	84.60	90.30
PP$_{GPSE}$	83.03	64.98	84.44	75.13
PP$_{PSE}$	81.75	65.29	81.85	75.40

Table 4: Accuracy of polarity prediction models

polarity prediction models, similar to PP$_{GSE}$ in Section 4.1. It should however be noted that the training and development sets have been seen by the SEI model. Therefore, inflated results are expected on the development set. We experiment with two models: one trained, tuned and tested using predicted SE boundaries (PP$_{GPSE}$) and one trained and tuned on gold-standard but tested using predicted SE boundaries PP$_{PSE}$. Table 2 contains the hyper-parameter sets for both models and Table 4 shows their performances.

As expected, the development set does not see much change between fully and partially gold-standard and fully predicted SE boundaries. The test set, on the other hand, observes a large degradation compared to when fully gold-standard boundaries are used. However, the accuracy scores are higher on these sets than the baseline, albeit mostly marginally. These results show that better SEI models are required to get the full benefit of the SE information in polarity prediction.

5 Conclusion

We build on an existing freely available sentiment-polarity-annotated dataset by explicitly marking those words in a sentence which are contributing towards the expression of opinion or sentiment towards a particular target. In experiments with this dataset, we demonstrate that knowledge of the boundaries of sentiment expressions can greatly simplify the task of polarity classification. We assume that this knowledge has the effect of reducing noise for the learner by de-emphasizing words in the input that are not contributing towards the sentiment. This suggests that it may be worth paying attention to the task of sentiment expression identification. Our preliminary experiments with this task show that a standard LSTM can achieve an F-score of approximately 70%, providing only a minor increase to polarity classification performance. Future work will involve further experiments with sentiment expression identification and also alternative ways of utilizing these boundaries in polarity prediction. Joint modelling of sentiment expression and polarity is another interesting direction for further research.

References

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics* 39(3):473–509.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. pages 27–35.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pages 1631–1642.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 1(2):0.