Aspect Based Sentiment Analysis using syntactic dependencies

Team Members:
15CS30037 TYSS Santosh
16CE31005 R K S Jaswanth
15CS30035 Sumeet Shirgure
16CS10033 Vishal Reddy

Problem statement:

Aspect-level sentiment analysis is a fine-grained task in sentiment analysis, which aims to identify the sentiment polarity (i.e., negative, neutral, or positive) of a specific opinion target expressed in a comment/review by a reviewer. For example, given a sentence “The price is reasonable although the service is poor,'' the sentiment polarity for aspects ''price”and “service” are positive and negative respectively.

Introduction:

Recent years has seen rapid growth of research on sentiment analysis. Sentiment analysis has both business importance and academic interest. So far, most sentiment analysis research has focused on classifying the overall sentiment of a document into positive or negative. We would, however,often like to understand what are the specific sentiments towards different aspects of an entity, e.g. a restaurant review ”Food is decent but service is so bad.”contains positive sentiment towards aspect food but strong negative sentiment towards aspectservice. Classifying the overall sentiment as negative would neglect the fact that food was actually good. In 2010, a new framework named aspect-based sentiment analysis(ABSA) was proposed to address this problem. Here an aspect refers to an attribute or component of an entity, e.g., the screen of a cell phone, or the picture quality of a camera. An ABSA task typically involves several sub-tasks,including identifying relevant entities and aspects, determining the corresponding sentiment/polarity. Traditional methods for aspect-level sentiment analysis mainly focus on designing a set of features (such as bag-of-words, sentiment lexicons,and linguistic features) to train a classifier for sentiment classification. However, such kind of feature engineering work often relies on human ingenuity, which is a time-consuming process and lacks generalization.In recent years,more and more neural network based models have been proposed and obtained the state-of-the-art results. As previous research reveals that 40% of sentiment classification errors are caused by not considering targets in sentiment classification, recent works tend to focus on fusing the information of the targets and the contexts. Wang et al. (2016) and Tang et al. (2016a) both concatenated the aspect embeddings and embeddings of each word as inputs to a LSTM based model so as to introduce the information of the target into the model. Tay et al. (2017) adopted circular convolution and circular correlation to model the similarity between aspect and contextual words.Ma et al. (2017) and Zheng et al. (2018) both employed a bidirectional attention operation to achieve the representations of targets and contextual words determined by each other. Huang et al. (2018) introduced an attention-over-attention based network to model the aspects and contexts in a joint way and explicitly capture the interaction between aspects and the context.

Motivation:

Generally, we find that a dependency tree shortens the distance between the aspects and opinion words of a sentence, captures the syntactic relations between words, and offers discriminative syntactic paths on arbitrary sentences for information propagation across the tree. These properties allow neural network models to capture long-term syntactic dependencies effortlessly. Besides, dependency trees have graph-like structures bringing to play the recent class of neural networks, namely, graph convolutional networks (GCN). These observations motivate us to develop a neural model which can operate on the dependency tree of a sentence, with the aim to make accurate sentiment predictions with respect to specific aspects. Specifically, we propose a convolution over a dependency tree (CDT) model which exploits a GCN to model the structure of a sentence through its dependency tree, where node(word) embeddings of the tree are initialized by means of a Bi-directional Long Short Term Mem-ory (Bi-LSTM) network. We hypothesize that the architecture of CDT allows the Bi-LSTM account for contextual information between successive words, while the GCN enhances the embeddings by modeling the dependencies along the syntactic paths of the dependency tree. Such operations allow information to be transferred from opinion words to aspect words, implying that the encoding for aspectwords is sufficient for supervision in the classification task.

Methodology:

Now we describe the CDT model which takes as input a dependency tree of a sentence. Node embeddings of the dependency tree are initially modeled by means of a BiLSTM, and the embeddings are further enhanced via a GCN. Finally, an aggregator is applied over the enhanced aspect embeddings to distill a dense vector embedding for the classification task. In particular, we aim to extract embeddings which encode both contextual and dependency information between a specific aspect expression and opinion words, providing supervisory signals for the aspect-based classification task. We briefly describe the BiLSTM model, which takes as input the sentences with n ordered word embeddings. The BiLSTM integrates context information in the word embeddings by keeping track of dependencies along the chain of words.Given an aspect-sentence pair(a,s), where a={a1,a2,...,al} is a sub-sequence of the sentence s={w1,w2,...,wn}. The sentence s has corresponding word embeddings x={x1,x2,...xn}.The LSTM learns hidden state representation in the forward direction on the word embeddings in x. This allows contextual information to be captured in a forward direction.In a similar fashion, a backward LSTM will learn representations on x. Finally, we can concatenate the corresponding parallel representations modeled by both forward and backward LSTMs into higher dimensional representations {h1,h2,...,hn}, which contains the sub-sequence {ha1,ha2,...,hal} corresponding to the aspect expression a. In doing so, we capture contextual information between opinion words and aspects. Besides, we integrate dependency information in the contextualized embeddings using aGCN which operates directly on the dependency tree of the sentence. The dependency tree can be interpreted as a graph G with n nodes, where nodes represent words in the sentence and edges represent syntactic dependency paths between words in the graph. The nodes of the dependency tree are given by real valued vectors modeled by BiLSTM as described above. This structure allows a GCN to operate directly on the graph to model dependencies that exist between words. To allow the GCN to model node embeddings efficiently, we allow G to have self-loops. The GCN approach ensures that the sentence structure represented by the dependency tree is encoded efficiently, whereby the representations for nodes encode the local position of opinion words and the target words in the dependency tree.The dependency tree G for any arbitrary sentence can be represented as an n×n adjacency matrix A, with entries Aij signaling if node i isconnected to node j by a single dependency path in G. Together with node embeddings modeled by BiLSTMs, wecan exploit a GCN capable of operating directly on graphs. The GCN makes efficient use of dependency paths to transform and propagate information across the paths, and update node embeddings by aggregating the propagated information. In such an operation, the GCN only considers the first order neighborhood of a node when modeling its embeddings. However,k successive GCN operations result in the propagation of information across the k-th order neighborhood. A single node embedding update takes the

$$h_i^{(k+1)} = \phi\left(\sum_{j=1}^{n} c^i A_{ij}\left(W^{(k)} h_j^{(k)} + b^{(k)}\right)\right)$$

In extracting a final embedding for the classification task, we exploit a simple aggregator. For our framework, we choose an average pool which aggregates information over the aspect vectors. We choose to aggregate only the aspect vectors because we believe that these vectors encode contextual and dependency information owing to theBiLSTM and the GCN respectively. The BiLSTM and the GCN can be interpreted as message passing networks. Specifically, the BiLSTM allow aspect words of an arbitrary sentence to be contextualized, while the GCN finds the local position of aspect words in the syntactic dependency tree. The local position within the dependency tree encodes dependency information of a word with respect to its neighbors. As a result, the BiLSTM and the GCN allow embeddings for aspect words to have discriminative features, providing supervisory signals for the classification task. Moreover, we perform an average pool to retain most of the information in the aspect vectors.

The aspect-based representation is passed to a fully connected softmax layer whose output is a probability distribution over the different sentiment polarities. The model is trained end-to-end through a backpropagation, where the objective function to be minimized is the cross entropy error.

Experiments:

In this section, we conduct experiments to validate our model which we denote as CDT on bench-mark datasets. We also present restricted versions of our model denoted as ASP-BiLSTM and ASP-GCN. Unlike our main model, ASP-BiLSTM only exploits BiLSTM to model contextual information with respect to a specific aspect expression, while ASP-GCN exploits a GCN to model dependencies between words. Both models extract a final embedding on the aspect vectors. We propose these two models to observe the performance of GCN and BiLSTM, as well as the performance when we stack a GCN on a BiLSTM which forms the CDTmodel. To distinguish CDT as the new state-of-the-art in aspect-based sentiment classification, we compare CDT with several well established models, showing that CDT outperforms the very recent models in the classification task. In particular, we perform case studies with visualizations to verify our approach of aggregating only aspect vectors for the final embedding. We further present visualizations on case examples showing how GCN improves on a simple BiLSTM model.

Implementation and parameter settings

For fairness in model comparison, we use similar parameters in compared models. Specifically, we exploit 300-dimensional Glove vectors for the word embeddings, as well as a30-dimensional part-of-speech (POS) em-beddings, 30-dimensional position embeddings,which is used to identify the relative position of each word with respect to the aspect in the sentence. We concatenate both word, POS and position embeddings, and learn a 50-dimensional BiL-STM embeddings for each word. The GCN operates on the dependency tree of the sentence to enhance the BiLSTM embeddings. All sentences are parsed by the Stanford parser.To encourage the GCN to model dependencies between words,we randomly dropout 10% of neurons per layer and about 0.7at the input layer. The GCN model is trained for 100epochs with batch size 32. We use the adam optimizer with learning rate 0.01 for all datasets.

- Baselines:
- Majority assigns the sentiment polarity with most frequent occurrences in the training set to each sample in test set.
- TD-LSTM adopts two LSTMs to model the left context with target and the right context with target respectively (Tang et al., 2016a); It takes the hidden states of LSTM at last time-step to represent the sentence for prediction.

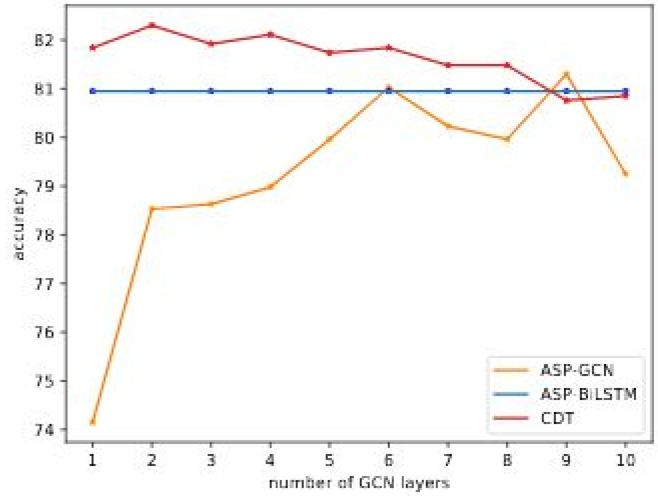
- MemNet (Tang et al., 2016b) applies attention multiple times on the word embeddings, and the output of last attention is fed to softmax for prediction.
- IAN (Ma et al., 2017) interactively learns attentions in the contexts and targets, and generates the representations for targets and contexts separately.
- RAM (Chen et al., 2017) is a multilayer architecture where each layer consists of attention-based aggregation of word features and a GRU cell to learn the sentence representation.
- LCR-Rot(Zheng et al., 2018) employs three Bi-LSTMs to model the left context, the target and the right context. Then they propose a rotatory attention mechanism which models the relation between target and left/right contexts.
- AOA-LSTM(Huang et al., 2018) introduces an attention-over-attention (AOA) based network to model aspects and sentences in a joint way and explicitly capture the interaction between aspects and context sentences
- TNet (Li et al., 2018a): In this work,BiLSTM embeddings are transformed into target specific embeddings, and a CNNmodel is used to extract a final embedding.
- PRET+MULT (He et al., 2018b): A multi-task framework based on LSTMs is proposed to transfer knowledge from a document-level model task to an aspect-level model task.
- SA-LSTM-P (Wang and Lu, 2018): This work first learn embeddings using BiLSTM and model structural dependencies between words by means of a segmentation attention mechanism.
- LSTM+SynATT+TarRep (He et al., 2018a):This method models target representation asa weighted sum of aspect embeddings, andmodels the syntactic structure of the sentence using an attention mechanism.
- MGANA (Fan et al., 2018b): A BiLSTM is exploited to capture contextual information in the sentence, while a multi-grained attention mechanism is proposed to extract an embedding which effectively captures the interaction between the aspect and the context.
- MGANB (Li et al., 2018b): This work integrates an alignment mechanism in a multi-task model comprising of an aspect-term taskand an aspect-category task to effectively extract aspect-specific representations.
- HSCN (Lei et al., 2019): A model is proposed to capture interactions between the context and target, select target words and extract target-specific contextual representation, while measuring the deviation between target specific contextual representation and target representations.

Model	Restaurant	Laptop
Majority	65.00	53.45

TD-LSTM	75.63	68.13
MemNet	79.98	70.33
IAN	78.60	72.10
RAM	80.23	74.49
LCR-Rot	81.34	75.24
AOA-LSTM	81.20	74.50
TNet	80.79	76.54
PRET+MULT	79.11	71.15
SA-LSTM-P	81.60	75.1
LSTM+SynATT+TarRep	80.63	71.94
MGANA	81.25	75.39
MGANB	81.49	76.21
HSCN	77.8	76.1
ASP-BiLSTM	80.95	74.22
ASP-GCN	81.30	74.53
CDT	82.30	77.19

From the table, we find that CDT generally outperforms all models for the different datasets. ASP-BiLSTM, ASP-GCN and CDT extract final representations from only the aspect vectors. Based on the performance, it seems as a sufficient technique for the classification task. We believe that the aspect vector is encoded with con-text and dependency information from the context and structure of the sentence by means of the BiL-STM and the GCN. The BiLSTM and GCN can be regarded as message passing networks, propa-gating information along a chain of sequence of words(BiLSTM) or along syntactic dependency path(GCN). Due to the fact that relevant informa-tion is passed to the aspect words, a simple average pool is all we need to retain information relevant to the classification task. Note that the informa-tion propagated in the network is learned there-fore only weighed information is encoded within the aspect words.

We conduct an experiment to demonstrate that the performance of our proposed models, namely CDT and ASP-GCN, depend on the number of layers of the GCN. We perform this experiment on the Restaurant dataset and present the result in Figure.

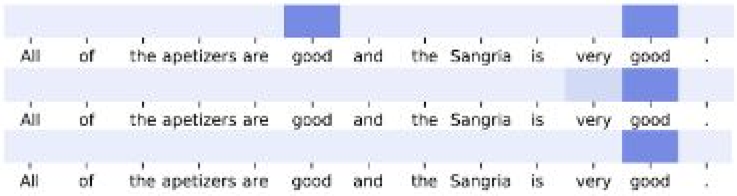


In our experimentation, we find that as we increase the number of layers the accuracy performance increase to an extent. In particular, ASP-GCN increase in model performance over 6 layers of the GCN. The performance becomes unstable after the 6-th layer. Since GCN passes information in the local neighborhood of any node, successive operations on the dependency tree allows ASP-GCN to pass information to the furthest node. The problem of overfitting takes effect when the layers rises beyond a threshold, explaining the accuracy curve after the 6-th layer in the figure. Another important observation is the convergence of accuracy performance of the ASP-BiLSTM and ASP-GCN at the 6-th layer. Note that ASP-BiLSTM only captures contextual information while ASP-GCN captures dependency information. However, both models converge in performance at the 6-th layer.Taking advantage of the GCN and the BiLSTMwe expect to improve performance, capturing both context and dependencies with respect to the as-pect expression. As seen in the accuracy curve of CDT, the GCN integrates dependency information in the contextualized embeddings to improve accuracy performance over just 2 layers, reducing the number of GCN layers needed.

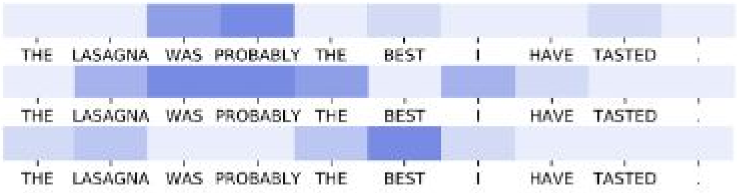
Case Study

We study the behaviour of ASP-BiLSTM, TNet and CDT on case examples. To this end we present visualizations showing the attention these models place on words. For a good model, we expect the model to attend to words which influence the sentiment inferred on a specific aspect. It is clear that GCN complements the BiLSTM to improve model performance. This means that the BiLSTM can identify opinion words within the context with respect to a specific aspect. However, in some complicated contexts, it might perform poorly. But the GCN can build upon BiLSTM to attend to the correct opinion words by leveraging the dependencies among words. Consider the case example shown

Attention visualization for ASP-BiLSTM(1st row), TNet(2nd row) and CDT (3rd row) for the aspect word ‘Sangria’

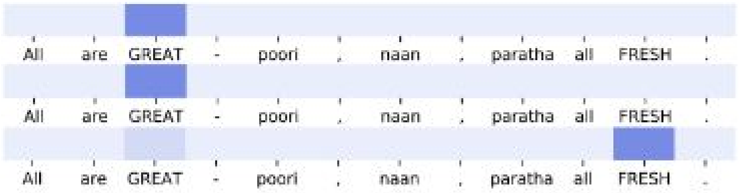


ASP-BiLSTM was clever to know that the word ‘good’ is an opinion word with respect to the aspect ‘Sangria’. But ASP-BiLSTM failed to identify whether ‘good’ on the far left is asso-ciated to the ‘Sangria’ or ‘good’ on the far right is associated to ‘Sangria’. Interestingly, we find that the GCN could analyze this further through the dependencies between words to identify that it is the ‘good’ on the far right. TNet on the other hand measures the association between ‘Sangria’ and ‘good’ in both directions to identify the correct ‘good’.



Attention visualization for ASP-BiLSTM(1st row), TNet (2nd row) and CDT (3rd row) for the aspect word ‘paratha’

even though the BiLSTM is able to identify the opinion word ‘GREAT’ which expresses an opinion on the aspect ‘parathra’, CDT is able to capture the opinion word ‘FRESH’ which directly expresses the sentiment towards the aspect. However, from the visualization is eas-ily observed that CDT still attends to ‘GREAT’.This suggests that the GCN is able to model the importance of the words with respect to the as-pect, placing larger weights to words directly expressing an opinion on the aspect. At the sametime, TNet misses the opinion word ‘FRESH’ and places attention to the word ‘GREAT’ just likeASP-BiLSTM.



Attention visualization for ASP-BiLSTM(1st row), TNet (2nd row) and CDT (3rd row) for the aspect word 'LASAGNA'

we find that ASP-BiLSTM places small attention on the opinion word 'BEST' which expresses the sentiment on the aspect word 'LASAGNA', while focusing its attention on 'WAS PROBABLY' which is not meaningful alone. Interestingly, CDT builds upon this little information and rely on the dependencies between the words through the dependency tree to learn that 'BEST' is the correct word to attend to. Similar to ASP-BiLSTM, TNetmisses the important word 'BEST' and places at-tention to 'WAS PROBABLY'. This result suggests that TNet heavily depends on the representations modeled by its BiLSTM layer, while CDT con-siders other information such as the dependencies among words to accurately identify words which expresses opinions on specific aspects.

Conclusion

Modeling representations for aspect-based senti-ment classification generally require capturing informative words which express the sentiment in-ferred on the target aspect. Leveraging neural networks are highly desirable for representation learning. BiLSTM-based models have been successful to capture contextual information in priorworks.In this paper, we integrate a GCN with a simple BiLSTM model, with the aim to capture structural and contextual information of sentences. We have shown that the GCN successfully performs convo-lutions on the dependency tree to refine BiLSTM embeddings.