Expert Systems With Applications 72 (2017) 221–230

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN

Tao Chen[a], Ruifeng Xu[a,b,*], Yulan He[c], Xuan Wang[a]

[a] Shenzhen Engineering Laboratory of Performance Robots at Digital Stage, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[b] Guangdong Provincial Engineering Technology Research Center for Data Science, Guangzhou, China
[c] School of Engineering and Applied Science, Aston University, Birmingham, UK

ARTICLE INFO

Article history:
Received 9 July 2016
Revised 1 October 2016
Accepted 21 October 2016
Available online 9 November 2016

Keywords:
Natural language processing
Sentiment analysis
Deep neural network

ABSTRACT

Different types of sentences express sentiment in very different ways. Traditional sentence-level sentiment classification research focuses on one-technique-fits-all solution or only centers on one special type of sentences. In this paper, we propose a divide-and-conquer approach which first classifies sentences into different types, then performs sentiment analysis separately on sentences from each type. Specifically, we find that sentences tend to be more complex if they contain more sentiment targets. Thus, we propose to first apply a neural network based sequence model to classify opinionated sentences into three types according to the number of targets appeared in a sentence. Each group of sentences is then fed into a one-dimensional convolutional neural network separately for sentiment classification. Our approach has been evaluated on four sentiment classification datasets and compared with a wide range of baselines. Experimental results show that: (1) sentence type classification can improve the performance of sentence-level sentiment analysis; (2) the proposed approach achieves state-of-the-art results on several benchmarking datasets.

1. Introduction

Sentiment analysis is the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text (Liu, 2015). With the rapid growth of social media on the web, such as reviews, forum discussions, blogs, news, and comments, more and more people share their views and opinions online. As such, this fascinating problem is increasingly important in business and society.

One of the main directions of sentiment analysis is sentence-level sentiment analysis. Much of the existing research on this topic focused on identifying the polarity of a sentence (e.g. positive, negative, neutral) based on the language clues extracted from the textual content of sentences (Liu, 2012; Pang & Lee, 2004; Turney, 2002). They solved this task as a general problem without considering different sentence types. However, different types of sentences express sentiment in very different ways. For example, for the sentence "*It is good.*", the sentiment polarity is definitely positive; for the interrogative sentence "*Is it good?*", the sentiment polarity is obscure, and it slightly inclined to the negative; for the comparative sentence "*A is better than B.*", we even cannot decide its sentiment polarity, because it is dependent on which opinion target we focus on (*A* or *B*).

Unlike factual text, sentiment text can often be expressed in a more subtle or arbitrary manner, making it difficult to be identified by simply looking at each constituent word in isolation. It is argued that there is unlikely to have a one-technique-fits-all solution (Narayanan, Liu, & Choudhary, 2009). A divide-and-conquer approach may be needed to deal with some special sentences with unique characteristics, that is, different types of sentences may need different treatments on sentence-level sentiment analysis (Liu, 2015).

There are many ways in classifying sentences in sentiment analysis. Sentences can be classified as subjective and objective which is to separate opinions from facts (Wiebe & Wilson, 2002; Wiebe, Bruce, & O'Hara, 1999; Yu & Hatzivassiloglou, 2003). Some researchers focused on target-dependent sentiment classification, which is to classify sentiment polarity for a given target on sentences consisting of explicit sentiment targets (Dong et al., 2014; Jiang, Yu, Zhou, Liu, & Zhao, 2011; Mitchell, Aguilar, Wilson, & Durme, 2013; Tang, Qin, Feng, & Liu, 2015a; Vo &

[*] Corresponding author. Fax.: 8618988759558.
E-mail addresses: chentao1999@gmail.com (T. Chen), xuruifeng@hitsz.edu.cn (R. Xu), hey9@aston.ac.uk (Y. He), wangxuan@hitsz.edu.cn (X. Wang).

http://dx.doi.org/10.1016/j.eswa.2016.10.065

Zhang, 2015). Others dealt with mining opinions in comparative sentences, which is to determinate the degree of positivity surround the analysis of comparative sentences (Ganapathibhotla & Liu, 2008; Jindal & Liu, 2006b; Yang & Ko, 2011). There has also been work focusing on sentiment analysis of conditional sentences (Narayanan et al., 2009), or sentences with modality, which have some special characteristics that make it hard for a system to determine sentiment orientations (Liu, Yu, Chen, & Liu, 2013).

In this paper, we propose a different way in dealing with different sentence types. In particular, we investigate the relationship between the number of opinion targets expressed in a sentence and the sentiment expressed in this sentence; propose a novel framework for improving sentiment analysis via sentence type classification. **Opinion target** (hereafter, target for short) can be any entity or aspect of the entity on which an opinion has been expressed (Liu, 2015). An opinionated sentence can express sentiments without a mention of any target, or towards one target, two or more targets. We define three types of sentences: **non-target sentences, one-target sentences** and **multi-target sentences**, respectively. Consider the following examples from the movie review sentence polarity dataset v1.0 (hereafter, MR dataset for short) (Pang & Lee, 2005)[1]:

Example 1. A masterpiece four years in the making.

Example 2. If you sometimes like to go to the movies to have fun, Wasabi is a good place to start.

Example 3. Director Kapur is a filmmaker with a real flair for epic landscapes and adventure, and this is a better film than his earlier English-language movie, the overpraised Elizabeth.

Example 1 is a non-target sentence. In order to infer its target, we need to know its context. Example 2 is a one-target sentence, in which the sentiment polarity of the target *Wasabi* is positive. Example 3 is a multi-target sentence, in which there are three targets: *Director Kapur, film* and *his earlier English-language movie, the overpraised Elizabeth*. We can observe that sentences tend to be more complex with more opinion targets, and sentiment detection is more difficult for sentences containing more targets.

Based on this observation, we apply a deep neural network sequence model, which is a bidirectional long short-term memory with conditional random fields (henceforth BiLSTM-CRF) (Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016), to extract target expressions in opinionated sentences. Based on the targets extracted, we classify sentences into three groups: non-target, one-target and multi-target. Then, one-dimensional convolutional neural networks (1d-CNNs) (Kim, 2014) are trained for sentiment classification on each group separately. Finally, the sentiment polarity of each input sentence is predicted by one of the three 1d-CNNs.

We evaluate the effectiveness of our approach empirically on various benchmarking datasets including the Stanford sentiment treebank (SST)[2] (Socher et al., 2013) and the customer reviews dataset (CR)[3] (Hu & Liu, 2004). We compare our results with a wide range of baselines including convolutional neural networks (CNN) with multi-channel (Kim, 2014), recursive auto-encoders (RAE) (Socher, Pennington, Huang, Ng, & Manning, 2011), recursive neural tensor network (RNTN) (Socher et al., 2013), dynamic convolutional neural network (DCNN) (Kalchbrenner, Grefenstette, & Blunsom, 2014), Naive Bayes support vector machines (NBSVM) (Wang & Manning, 2012), dependency tree with conditional random fields (tree-CRF) (Nakagawa, Inui, & Kurohashi, 2010) et al. Experimental results show that the proposed approach achieves state-of-the-art results on several benchmarking datasets. This shows that sentence type classification can improve the performance of sentence-level sentiment analysis.

The main contributions of our work are summarized below:

- We propose a novel two-step pipeline framework for sentence-level sentiment classification by first classifying sentences into different types based on the number of opinion targets they contain, and then training 1d-CNNs separately for sentences in each type for sentiment detection;
- While conventional sentiment analysis methods largely ignore different sentence types, we have validated in our experiments that learning a sentiment classifier tailored to each sentence type would result in performance gains in sentence-level sentiment classification.

The rest of this article is organized as follows: we review related work in Section 2; and then present our approach in Section 3; experimental setup, evaluation results and discussions are reported in Section 4; finally, Section 5 concludes the paper and outlines future research directions.

2. Related work

2.1. Sentence type classification for sentiment analysis

Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing (NLP) (Liu, 2015; Ravi & Ravi, 2015). It is a multifaceted problem with many challenging and interrelated sub-problems, including sentence-level sentiment classification. Many researchers realized that different type of sentence need different treatment for sentiment analysis. Models of different sentence types, including subjective sentences, target-dependent sentences, comparative sentences, negation sentences, conditional sentences, sarcastic sentences, have been proposed for sentiment analysis.

Subjectivity classification distinguishes sentences that express opinions (called subjective sentences) from sentences that express factual information (called objective sentences) (Liu, 2015). Although some objective sentences can imply sentiments or opinions and some subjective sentences may not express any opinion or sentiment, many researchers regard subjectivity and sentiment as the same concept (Hatzivassiloglou & Wiebe, 2000; Wiebe et al., 1999), i.e., subjective sentences express opinions and objective sentences express fact. Riloff and Wiebe (2003) presented a bootstrapping process to learn linguistically rich extraction patterns for subjective expressions from a large unannotated data. Rill, Reinel, Scheidt, and Zicari (2014) presented a system to detect emerging political topics on twitter and the impact on concept-level sentiment analysis. Appel, Chiclana, Carter, and Fujita (2016) proposed a hybrid approach using SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010) and fuzzy sets to estimate the semantic orientation polarity and intensity of sentiment words, before computing the sentence level sentiments. Muhammad, Wiratunga, and Lothian (2016) introduced a lexicon-based sentiment classification system for social media genres, which captures contextual polarity from both local and global context. Fernández-Gavilanes, Álvarez-López, Juncal-Martínez, Costa-Montenegro, and González-Castaño (2016) proposed a novel approach to predict sentiment in online texts based on an unsupervised dependency parsing-based text classification method.

Most previous target related works assumed targets have been given before performing sentiment classification (Dong et al., 2014; Jiang et al., 2011; Mitchell et al., 2013; Vo & Zhang, 2015). Little research has been conducted on classifying sentence by the target number although there is a large body of work focusing on opinion target extraction from text.

[1] Available at: https://www.cs.cornell.edu/people/pabo/movie-review-data/.

[2] http://nlp.stanford.edu/sentiment/.

[3] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets.

A comparative opinion sentence expresses a relation of similarities or differences between two or more entities and/or a preference of the opinion holder based on some shared aspects of the entities. Jindal and Liu (2006a) showed that almost every comparative sentence had a keyword (a word or phrase) indicating comparison, and identified comparative sentences by using class sequential rules based on human compiled keywords as features for a naive Bayes classifier. Ganapathibhotla and Liu (2008) reported they were the first work for mining opinions in comparative sentences. They solved the problem by using linguistic rules and a large external corpus of Pros and Cons from product reviews to determine whether the aspect and sentiment context were more associated with each other in Pros or in Cons. Kessler and Kuhn (2014) presented a corpus of comparison sentences from English camera reviews. Park and Yuan (2015) proposed two linguistic knowledge-driven approaches for Chinese comparative elements extraction.

Negation sentences occur fairly frequently in sentiment analysis corpus. Many researchers considered the impact of negation words or phrases as part of their works (Hu & Liu, 2004; Pang, Lee, & Vaithyanathan, 2002); a few researchers investigated negation words identification and/or negative sentence processing as a single topic. Jia, Yu, and Meng (2009) studied the effect of negation on sentiment analysis, including negation term and its scope identification, by using a parse tree, typed dependencies and special linguistic rules. Zhang, Ferrari, and Enjalbert (2012) proposed a compositional model to detect valence shifters, such as negations, which contribute to the interpretation of the polarity and the intensity of opinion expressions. Carrillo-de Albornoz and Plaza (2013) studied the effect of modifiers on the emotions affected by negation, intensifiers and modality.

Conditional sentences are another commonly used language constructs in text. Such a sentence typically contains two clauses: the condition clause and the consequent clause. Their relationship has significant impact on the sentiment orientation of the sentence (Liu, 2015). Narayanan et al. (2009) first presented a linguistic analysis of conditional sentences, and built some supervised learning models to determine if sentiments expressed on different topics in a conditional sentence are positive, negative or neutral. Liu (2015) listed a set of interesting patterns in conditional sentences that often indicate sentiment, which was particularly useful for reviews, online discussions, and blogs about products.

Sarcasm is a sophisticated form of speech act widely used in online communities. In the context of sentiment analysis, it means that when one says something positive, one actually means negative, and vice versa. Tsur, Davidov, and Rappoport (2010) presented a novel semi-supervised algorithm for sarcasm identification that recognized sarcastic sentences in product reviews. González-Ibánez, Muresan, and Wacholder (2011) reported on a method for constructing a corpus of sarcastic Twitter messages, and used this corpus to investigate the impact of lexical and pragmatic factors on machine learning effectiveness for identifying sarcastic utterances. Riloff et al. (2013) presented a bootstrapping algorithm for sarcasm recognition that automatically learned lists of positive sentiment phrases and negative situation phrases from sarcastic tweets.

Adversative and concessive structures, as another kind of linguistical feature, are constructions express antithetical circumstances (Crystal, 2011). A adversative or a concessive clause is usually in clear opposition to the main clause about the fact or event commented. Fernández-Gavilanes et al. (2016) treated the constructions as an extension of intensification propagation, where the sentiment formulated could be diminished or intensified, depending on both adversative/concessive and main clauses.

2.2. Opinion target detection

Hu and Liu (2004) used frequent nouns and noun phrases as feature candidates for opinion target extraction. Qiu, Liu, Bu, and Chen (2011) proposed a bootstrapping method where a dependency parser was used to identify syntactic relations that linked opinion words and targets for opinion target extraction. Popescu and Etzioni (2005) considered product features to be concepts forming certain relationships with the product and sought to identify the features connected with the product name by computing the point wise mutual information (PMI) score between the phrase and class-specific discriminators through a web search. Stoyanov and Cardie (2008) treated target extraction as a topic coreference resolution problem and proposed to train a classifier to judge if two opinions were on the same target. Liu, Xu, and Zhao (2014) constructed a heterogeneous graph to model semantic relations and opinion relations, and proposed a co-ranking algorithm to estimate the confidence of each candidate. The candidates with higher confidence would be extracted as opinion targets. Poria, Cambria, and Gelbukh (2016) presented the first deep learning approach to aspect extraction in opinion mining using a 7-layer CNN and a set of linguistic patterns to tag each word in sentences.

Mitchell et al. (2013) modeled sentiment detection as a sequence tagging problem, extracted named entities and their sentiment classes jointly. They referred this kind of approach open domain targeted sentiment detection. Zhang, Zhang, and Vo (2015) followed Mitchell et al.'s work, studied the effect of word embeddings and automatic feature combinations on the task by extending a CRF baseline using neural networks.

2.3. Deep learning for sentiment classification

Deep learning approaches are able to automatically capture, to some extent, the syntactic and semantic features from text without feature engineering, which is labor intensive and time consuming. They attract much research interest in recent years, and achieve state-of-the-art performances in many fields of NLP, including sentiment classification.

Socher et al. (2011) introduced semi-supervised recursive autoencoders for predicting sentiment distributions without using any pre-defined sentiment lexica or polarity shifting rules. Socher et al. (2013) proposed a family of recursive neural network, including recursive neural tensor network (RNTN), to learn the compositional semantic of variable-length phrases and sentences over a human annotated sentiment treebank. Kalchbrenner et al. (2014) and Kim (2014) proposed different CNN models for sentiment classification, respectively. Both of them can handle the input sentences with varying length and capture short and long-range relations. Kim (2014)'s model has little hyper parameter tuning and can be trained on pre-trained word vectors. Irsoy and Cardie (2014a) presented a deep recursive neural network (DRNN) constructed by stacking multiple recursive layers for compositionality in Language and evaluated the proposed model on sentiment classification tasks. Tai, Socher, and Manning (2015) introduced a tree long short-term memory (LSTM) for improving semantic representations, which outperforms many existing systems and strong LSTM baselines on sentiment classification. Tang et al. (2015c) proposed a joint segmentation and classification framework for sentence-level sentiment classification. Liu, Qiu, and Huang (2016) used a recurrent neural network (RNN) based multitask learning framework to jointly learn across multiple related tasks. Chaturvedi, Ong, Tsang, Welsch, and Cambria (2016) proposed a deep recurrent belief network with distributed time delays for learning word dependencies in text which uses Gaussian networks with time-delays to initialize the weights of each hidden neuron. Tang, Qin, and Liu (2015b) gave a survey on this topic.

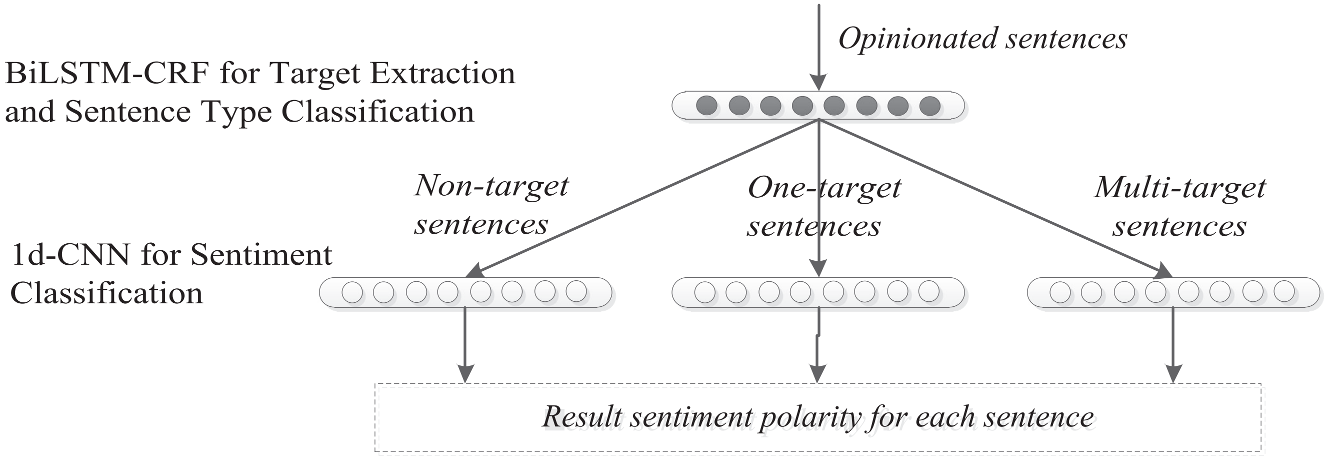
BiLSTM-CRF for Target Extraction and Sentence Type Classification
Opinionated sentences
Non-target sentences
One-target sentences
Multi-target sentences
1d-CNN for Sentiment Classification
Result sentiment polarity for each sentence

Fig. 1. Framework of sentence type classification based sentiment analysis using BiLSTM-CRF and 1d-CNN.

Table 1
An example sentence with labels in IOB format. The target is *the act*, the label *B* indicates the beginning of a target, *I* indicates that the word is inside a target, and *O* indicates a word belongs to no target.

Words:	*Yet*	*the*	*act*	*is*	*still*	*charming*	*here*	*.*
Labels:	*O*	*B*	*I*	*O*	*O*	*O*	*O*	*O*

3. Methodology

We present our approach for improving sentiment analysis via sentence type classification in this section. An overview of the approach is shown in Fig. 1. We first introduce the BiLSTM-CRF model which extracts target expressions from input opinionated sentences, and classifies each sentence according to the number of target explicitly expressed in it (Section 3.1). Then, we describe the 1d-CNN sentiment classification model which predicts sentiment polarity for non-target sentences, one-target sentences and multi-target sentences, separately (Section 3.2).

3.1. Sequence model for sentence type classification

We describe our approach for target extraction and sentence type classification with BiLSTM-CRF. Target extraction is similar to the classic problem of named entity recognition (NER), which views a sentence as a sequence of tokens usually labeled with IOB format (short for Inside, Outside, Beginning). Table 1 shows an example sentence with the appropriate labels in this format.

Deep neural sequence models have shown promising success in NER (Lample et al., 2016), sequence tagging (Huang, Xu, & Yu, 2015) and fine-grained opinion analysis (Irsoy & Cardie, 2014b). BiLSTM-CRF is one of deep neural sequence models, where a bidirectional long short-term memory (BiLSTM) layer (Graves, Mohamed, & Hinton, 2013) and a conditional random fields (CRF) layer (Lafferty, McCallum, & Pereira, 2001) are stacked together for sequence learning, as shown in Fig. 2. BiLSTM incorporates a forward long short-term memory (LSTM) layer and a backward LSTM layer in order to learn information from preceding as well as following tokens. LSTM (Hochreiter & Schmidhuber, 1997) is a kind of recurrent neural network (RNN) architecture with long short-term memory units as hidden units. Next we briefly describe RNN, LSTM, BiLSTM and BiLSTM-CRF.

RNN (Elman, 1990) is a class of artificial neural sequence model, where connections between units form a directed cycle. It takes arbitrary embedding sequences $x = (x_1, \ldots, x_T)$ as input, uses its internal memory network to exhibit dynamic temporal behavior. It consisting of a hidden unit h and an optional output y. T is the last time step. It is also the length of input sentence in this text sequence learning task. At each time step t, the hidden state h_t of the RNN is computed based on the previous hidden state h_{t-1} and the input at the current step x_t:

$$h_t = g(Ux_t + Wh_{t-1}) \tag{1}$$

where U and W are weight matrices of the network; $g(\cdot)$ is a non-linear activation function, such as an element-wise logistic sigmoid function. The output at time step t is computed as $y_t = \text{softmax}(Vh_t)$, where V is another weight parameter of the network, softmax is an activation function often implemented at the final layer of a network.

LSTM is a variant of RNN designed to deal with vanishing gradients problem (Hochreiter & Schmidhuber, 1997). The LSTM used in the BiLSTM-CRF (Lample et al., 2016) has two gates (an input gate i_t, an output gate o_t) and a cell activation vectors c_t.

BiLSTM uses two LSTMs to learn each token of the sequence based on both the past and the future context of the token. As shown in Fig. 2, one LSTM processes the sequence from left to right, the other one from right to left. At each time step t, a hidden forward layer with hidden unit function $\overrightarrow{h}$ is computed based on the previous hidden state $\overrightarrow{h}_{t-1}$ and the input at the current step x_t and a hidden backward layer with hidden unit function $\overleftarrow{h}$ is computed based on the future hidden state $\overleftarrow{h}_{t+1}$ and the input at the current step x_t. The forward and backward context representations, generated by $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ respectively, are concatenated into a long vector. The combined outputs are the predictions of teacher-given target signals.

As another widely used sequence model, conditional random fields (CRF) is a type of discriminative undirected probabilistic graphical model, which represents a single log-linear distributions over structured outputs as a function of a particular observation input sequence.

Given observations variables X whose values are observed, random variables Y whose values the task requires the model to predict, and a undirected graph G where Y are connected by undirected edges indicating dependencies. CRF defines the conditional probability of a set of output values $y \in Y$ given a set of input values $x \in X$ to be proportional to the product of potential functions on cliques of the graph (McCallum, 2003),

$$p(y|x) = \frac{1}{Z_x} \prod_{s \in S(y,x)} \Phi_s(y_s, x_s) \tag{2}$$

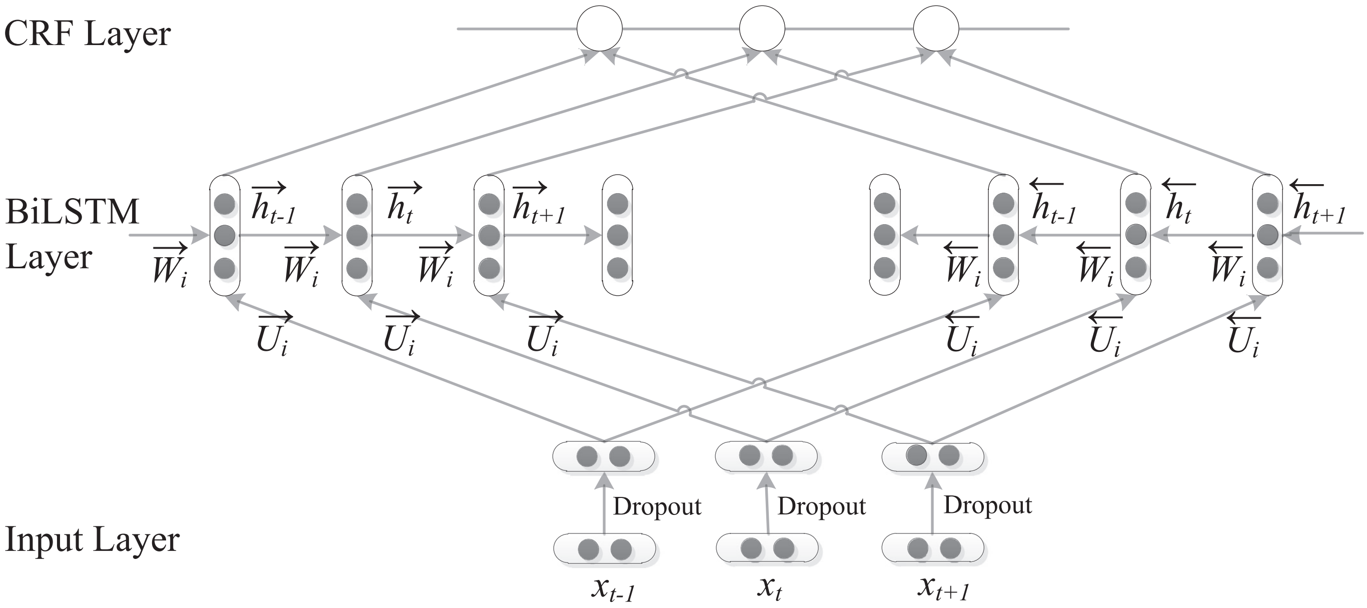

Fig. 2. An illustration of BiLSTM-CRF for target extraction and sentence type classification. BiLSTM layer incorporates a forward LSTM layer and a backward LSTM layer.

where Z_x is a normalization factor overall output values, $S(y, x)$ is the set of cliques of G, $\Phi_s(y_s, x_s)$ is the clique potential on clique s.

Afterwards, in the BiLSTM-CRF model, a softmax over all possible tag sequences yields a probability for the sequence y. The prediction of the output sequence is computed as follows:

$$y* = \mathrm{argmax}_{y \in Y} \sigma(X, y) \tag{3}$$

where $\sigma(X, y)$ is the score function defined as follows:

$$\sigma(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{4}$$

where A is a matrix of transition scores, $A_{y_i, y_{i+1}}$ represents the score of a transition from the tag y_i to y_{i+1}. n is the length of a sentence, P is the matrix of scores output by the BiLSTM network, P_{i, y_i} is the score of the y_i^{th} tag of the ith word in a sentence.

As shown in Fig. 2, dropout technique is used after the input layer of BiLSTM-CRF to reduce overfitting on the training data. This technique is firstly introduced by Hinton, Srivastava, Krizhevsky, Sutskever, and Salakhutdinov (2012) for preventing complex co-adaptations on the training data. It has given big improvements on many tasks.

After target extraction by BiLSTM-CRF, all opinionated sentences are classified into non-target sentences, one-target sentences and multi-target sentences, according to the number of targets extracted from them.

3.2. 1d-CNN for sentiment classification on each sentence type

1d-CNN, firstly proposed by Kim (2014), takes sentences of varying lengths as input and produces fixed-length vectors as output. Before training, word embeddings for each word in the glossary of all input sentences are generated. All the word embeddings are stacked in a matrix M. In the input layer, embeddings of words comprising current training sentence are taken from M. The maximum length of sentences that the network handles is set. Longer sentences are cut; shorter sentences are padded with zero vectors. Then, dropout regularization is used to control over-fitting.

In the convolution layer, multiple filters with different window size move on the word embeddings to perform one-dimensional convolution. As the filter moves on, many sequences, which capture the syntactic and semantic features in the filtered n-gram, are generated. Many feature sequences are combined into a feature map. In the pooling layer, a max-overtime pooling operation (Collobert et al., 2011) is applied to capture the most useful local features from feature maps. Activation functions are added to incorporate element-wise non-linearity. The outputs of multiple filters are concatenated in the merge layer. After another dropout process, a fully connected softmax layer output the probability distribution over labels from multiple classes.

CNN is one of most commonly used connectionism model for classification. Connectionism models focus on learning from environmental stimuli and storing this information in a form of connections between neurons. The weights in a neural network are adjusted according to the training data by some learning algorithm. It is the greater the difference in the training data, the more difficult for the learning algorithm to adapt the training data, and the worse classification results. Dividing opinionated sentences into different types according to the number of targets expressed in them can reduce the differences of training data in each group, therefore, improve overall classification accuracy.

4. Experiment

We conduct experiments to evaluate the performance of the proposed approach for sentence-level sentiment classification on various benchmarking datasets. In this section, we describe the experimental setup and baseline methods followed by the discussion of results.

4.1. Experimental setup

For training BiLSTM-CRF for target extraction and sentence type classification, we use the MPQA opinion corpus v2.0 (MPQA dataset for short) provided by Wiebe, Wilson, and Cardie (2005)[4] since it contains a diverse range of sentences with various numbers of opinion targets. It contains 14,492 sentences from a wide variety of news sources manually annotated with opinion target at the phrase level (7,026 targets). All the sentences are used to train BiLSTM-CRF.

[4] http://mpqa.cs.pitt.edu/.

For sentiment classification with 1d-CNN, we test our approach on different datasets:

- **MR**: Movie review sentence polarity dataset v1.0. It contains 5331 positive snippets and 5331 negative snippets extracted from Rotten Tomatoes web site pages where reviews marked with “fresh” are labeled as positive, and reviews marked with “rotten” are labeled as negative. 10-fold cross validation was used for testing.
- **SST-1**: Stanford sentiment treebank contains 11,855 sentences also extracted from the original pool of Rotten Tomatoes page files. These sentences are split into 8544/1101/2210 for train/dev/test. Each of them is fine-grained labeled (very positive, positive, neutral, negative, very negative).
- **SST-2**: Binary labeled version of Stanford sentiment treebank, in which neutral reviews are removed, very positive and positive reviews are labeled as positive, negative and very negative reviews are labeled as negative (Kim, 2014). It contains 9613 sentences split into 6920/872/1821 for train/dev/test.
- **CR**: Customer reviews of 5 digital products contains 3771 sentences extracted from amazon.com, including 2405 positive sentences and 1366 negative sentences. 10-fold cross validation was used for testing.

Following Kim (2014)’s work, we use accuracy as the evaluation metric to measure the overall sentiment classification performance.

During training a BiLSTM-CRF for target extraction in a sentence, the input sequence x_t is set to the t-th word embedding (a distributed representation for a word (Bengio, Ducharme, Vincent, & Jauvin, 2003)) in a input sentence. Publicly available word vectors trained from Google News[5] are used as pre-trained word embeddings. The size of these embeddings is 300. U, W, V and h_0 are initialized to a random vector of small values, h_{t+1} are initialized to a copy of h_t recursively. A back-propagation algorithm with Adam stochastic optimization method is used to train the network through time with learning rate of 0.05. After each training epoch, the network is tested on validation data. The log-likelihood of validation data is computed for convergence detection.

For training CNN, we use: CNN-non-static model, ReLU as activation function, Adadelta decay parameter of 0.95, dropout rate of 0.5, the size of initial word vectors of 300. We use different filter windows and feature maps for different target classes. For non-target sentences, we use filter windows of 3, 4, 5 with 100 feature maps each; For one-target sentences, we use filter windows of 3, 4, 5, 6 with 100 feature maps each; For multi-target sentences, we use filter windows of 3, 4, 5, 6, 7 with 200 feature maps each.

4.2. Baseline methods

We benchmark the following baseline methods for sentence-level sentiment classification, some of them have been previously used in Kim (2014):

- **MNB**: Multinomial naive Bayes with uni-bigrams.
- **NBSVM**: SVM variant using naive Bayes log-count ratios as feature values proposed by Wang and Manning (2012).
- **Tree-CRF**: Dependency tree based method for sentiment classification using CRF with hidden variables proposed by Nakagawa et al. (2010).
- **RAE**: Semi-supervised recursive autoencoders with pre-trained word vectors from Wikipedia proposed by Socher et al. (2011).
- **MV-RNN**: Recursive neural network using a vector and a matrix on every node in a parse tree for semantic compositionality proposed by Socher, Huval, Manning, and Ng (2012).

Table 2
Example sentences in each target class of Stanford sentiment treebank. T0, T1 and T2+ refer to non-target sentences, one-target sentences and multi-target sentences recognized by BiLSTM-CRF, respectively. In each target class, we show 3 example sentences (one positive, one neutral, one negative sentence, respectively), $s1$ to $s9$ are the order numbers of the examples.

Class	Example sentences
T0	s1: *...very funny, very enjoyable ...*
	s2: *Dark and disturbing, yet compelling to watch.*
	s3: *Hey, who else needs a shower?*
T1	s4: *Yet the act is still charming here.*
	s5: *As a director, Mr. Ratliff wisely rejects the temptation to make fun of his subjects.*
	s6: *Notorious C.H.O. has oodles of vulgar highlights.*
T2+	s7: *Singer/composer Bryan Adams contributes a slew of songs – a few potential hits, a few more simply intrusive to the story – but the whole package certainly captures the intended, er, spirit of the piece.*
	s8: *You Should Pay Nine Bucks for This: Because you can hear about suffering Afghan refugees on the news and still be unaffected.*
	s9: *...while each moment of this broken character study is rich in emotional texture, the journey doesn’t really go anywhere.*

- **RNTN**: Recursive deep neural network for semantic compositionality over a sentiment treebank using tensor-based feature function proposed by Socher et al. (2013).
- **Paragraph-Vec**: An unsupervised algorithm learning distributed feature representations from sentences and documents proposed by Le and Mikolov (2014).
- **DCNN**: Dynamic convolutional neural network with dynamic k-max pooling operation proposed by Kalchbrenner et al. (2014).
- **CNN-non-static**: 1d-CNN with pre-trained word embeddings and fine-tuning optimizing strategy proposed by Kim (2014).
- **CNN-multichannel**: 1d-CNN with two sets of pre-trained word embeddings proposed by Kim (2014).
- **DRNN**: Deep recursive neural networks with stacked multiple recursive layers proposed by Irsoy and Cardie (2014a).
- **Multi-task LSTM**: A multi-task learning framework using LSTM to jointly learn across multiple related tasks proposed by Liu et al. (2016).
- **Tree LSTM**: A generalization of LSTM to tree structured network topologies proposed by Tai et al. (2015).
- **Sentic patterns**: A concept-level sentiment analysis approach using dependency-based rules proposed by Poria, Cambria, Winterstein, and Huang (2014).

4.3. Results

4.3.1. Qualitative evaluations

In Table 2, we show the example sentences in each target class of Stanford sentiment treebank. It is observed that many non-target sentences are small imperative sentences, have direct subjective expressions (DSEs) which consist of explicit mentions of private states or speech events expressing private states (Irsoy & Cardie, 2014b), e.g., *funny* and *enjoyable* in $s1$, *dark and disturbing* in $s2$. For some non-target sentences, it is difficult to detect its sentiment without context, e.g., it is unclear whether the word *shower* in $s3$ conveys positive or negative sentiment. Non-target sentences tend to be short comparing with two other types of sentences. Many one-target sentences are simple sentences, which contain basic constituent elements forming a sentence. The subject is mostly the opinionated target in a one-target sentence, e.g., *the act* in $s4$, *Mr. Ratliff* in $s5$ and *C.H.O.* in $s6$. Almost all the multi-target sentences are compound/complex/compound-complex sentences, which have two or more clauses, and are very complex in expressions. Many of them have coordinating or subordinating conjunctions, which make it difficult to identify the sentiment of a whole sentence, e.g., *but* in $s7$, *because* and *and* in $s8$, *while* in $s9$.

[5] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing.

Table 3
Experimental results of sentiment classification accuracy. % is omitted. The best results are highlighted in bold face. The results of the top 10 approaches have been previously reported by Kim (2014). The top 3 approaches are conventional machine learning approaches with hand-crafted features. *Sentic patterns* is rule based approach. Other 11 approaches, including our approach, are deep neural network (DNN) approaches, which can automatically extract features from input data for classifier training without feature engineering.

Model	MR	SST-1	SST-2	CR
MNB	79.0	–	–	80.0
NBSVM	79.4	–	–	81.8
Tree-CRF	77.3	–	–	81.4
Sentic patterns	–	–	86.2	–
RAE	77.7	43.2	82.4	–
MV-RNN	79.0	44.4	82.9	–
RNTN	–	45.7	85.4	–
Paragraph-Vec	–	48.7	87.8	–
DCNN	–	48.5	86.8	–
CNN-non-static	81.5	48.0	87.2	84.3
CNN-multichannel	81.1	47.4	88.1	85.0
DRNN	–	49.8	86.6	–
Multi-task LSTM	–	49.6	87.9	–
Tree LSTM	–	**50.6**	86.9	–
Our approach	**82.3**	48.5	**88.3**	**85.4**

Table 4
The class-by-class classification results using sentence type classification as well as without using sentence type classification on the four datasets. #train and #test are the word number of sentences in training and test dataset, respectively; $l_{\max}$ and l_{avg} are max and average word length of sentences, respectively; Acc_{CNN} is the experimental result that we do sentiment classification directly on the four datasets using 1d-CNN (non-static) without sentence type classification, and statistic the accuracy on each target class the same with the target class recognized by BiLSTM-CRF. Acc_{our} is the experimental result of our approach on each target class, which using both sentence type classification and 1d-CNN (non-static). Δ is the relative improvement ratio calculates. In the Acc_{CNN}, Acc_{our} and Δ columns, % is omitted for conciseness.

		#train	#test	$l_{\max}$	l_{avg}	Acc_{CNN}	Acc_{our}	Δ
MR	T0	6,426	698	52	18.8	83.3	84.5	1.44
	T1	2,552	290	56	22.7	77.9	78.8	1.16
	T2+	618	78	51	27.1	73.9	75.1	1.62
SST-1	T0	5,436	1,367	51	17.5	49.8	50.9	2.21
	T1	2,495	655	56	21.0	45.1	46.1	2.22
	T2+	613	189	52	25.5	38.0	39.8	4.74
SST-2	T0	4,373	1,134	50	16.9	87.6	89.6	2.28
	T1	2,047	534	53	20.3	83.9	86.7	3.34
	T2+	500	153	51	24.9	82.0	84.1	2.56
CR	T0	1,982	191	75	17.4	85.5	88.4	3.39
	T1	1,140	152	105	19.1	80.9	83.2	2.84
	T2+	273	33	95	27.9	74.1	78.0	5.26

Table 5
Experimental results of different sequence models. % is omitted for conciseness. The best results are highlighted in bold face.

Model	MR	SST-1	SST-2	CR
CRF	81.7	47.6	87.4	84.4
LSTM	81.3	47.5	87.6	84.1
BiRNN	81.7	48.1	87.9	84.8
BiRNN-CRF	81.8	48.3	87.9	84.9
BiLSTM	82.0	48.3	88.0	85.3
BiLSTM-CRF	82.3	48.5	88.3	85.4

Overall, as the result of the qualitative evaluations, the difficulty degree of sentiment classification on each sentence type is $T2+ > T1 > T0$, i.e., multi-target sentences are most difficult, while non-target sentences are much easier for sentiment classification. The experimental results listed in the next subsection validate this observation.

4.3.2. Overall comparison

Table 3 shows the results achieved on the MR, SST-1, SST-2 and CR datasets. It is observed that comparing with three hand-crafted features based methods, although RAE and MV-RNN perform worse on MR dataset, two CNN based methods gives better results on both MR and CR datasets. This indicates the effectiveness of DNN approaches. Among 11 DNN approaches, our approach outperforms other baselines on all the datasets except SST-1, i.e., our approach gives relative improvements of 0.98% compared to CNN-non-static on MR dataset, 0.23% and 0.47% relative improvements compared to CNN-multichannel on SST-2 and CR dataset, respectively. Comparing with two CNN based methods, our sentence type classification based approach gives superior performance on all the four datasets (including SST-1 dataset). These validate the influences of sentence type classification in terms of sentence-level sentiment analysis.

4.3.3. Comparison on each target class

Table 4 shows the statistics and comparison of each target class on the MR, SST-1, SST-2 and CR datasets. The relative improvement ratio Δ calculates as follows:

$$\Delta = (Acc_{our} - Acc_{CNN}) \div Acc_{CNN} \times 100 \tag{5}$$

It is obvious that the performance for every target class is improved using sentence type classification. Yet, the improvement for the multi-target sentences (T2+) is more significant than other two target classes on three of the four dataset, e.g. the relative improvement ratio of T2+ class on the SST-1 and CR datasets are 4.75% and 5.26%, respectively, which are about twice higher than the relative improvement ratio of T1 class. Table 4 is a clear indication that the proposed sentence type classification based sentiment classification approach is very effective for complex sentences. Both the Acc_{CNN} and Acc_{our} use 1d-CNN (non-static) and pre-trained Google News word embedding, our approach achieves better performance because the divide-and-conquer approach, which first classifies sentences into different types, then optimize the sentiment classifier separately on sentences from each type.

4.3.4. Comparison with different sequence models

We have also experimented with different sequence models, including CRF, LSTM, BiRNN (Schuster & Paliwal, 1997), BiRNN-CRF, BiLSTM and BiLSTM-CRF, for sentence type classification. For CRF, we use CRFSuite (Okazaki, 2007) with word, Part-Of-Speech tag, prefix, suffix and a sentiment dictionary as features. For LSTM, BiRNN, BiRNN-CRF and BiLSTM, we also use Google News word embeddings as pre-trained word embeddings. For other parameters, we use default parameter settings.

Table 5 shows the experimental results on the MR, SST-1, SST-2 and CR datasets. It can be observed that BiLSTM-CRF outperforms all the other approaches on all the four datasets. It is because BiLSTM-CRF has more complicated hidden units, and offers better composition capability than other DNN approaches. CRF with hand-crafted features gives comparable performance to LSTM, but lower performance than more complex DNN models. BiRNN and BiLSTM gives better performance compared to LSTM because they can learn each token of the sequence based on both the past and the future context of the token, while LSTM only use the past context of the token. Comparing BiRNN and BiLSTM with BiRNN-CRF and BiLSTM-CRF, respectively, it is observed that combining CRF and DNN models can improve the performance of DNN approaches.

4.3.5. Evaluation on opinion target extraction with BiLSTM-CRF

One unavoidable problem for every multi-step approach is the propagation of errors. In our approach, we use a BiLSTM-CRF/1d-CNN pipeline for sentiment analysis. It is interesting to see how

Table 6
Experimental results of target extraction with BiLSTM-CRF on SemEval16 task 5 aspect based sentiment analysis dataset subtask 1 slot 2. *Best System* refers to the participation system with best performance submitted to SemEval16 task 5. *Baseline* refers to baseline model provided by the organizers; *C* refers to the model only uses the provided training data; *U* refers to the model uses other resources (e.g., publicly lexica) and additional data for training; "-" refers to no submissions were made. % is omitted for conciseness. The best results are highlighted in bold face.

Models		English	Spanish	French	Russian	Dutch	Turkish
Best System	U	72.34	68.39	66.67	33.47	56.99	–
Best System	C	66.91	68.52	65.32	30.62	51.78	–
Baseline	C	44.07	51.91	45.46	49.31	50.64	41.86
BiLSTM-CRF	C	**72.44**	**71.70**	**73.50**	**67.08**	**64.29**	**63.76**

the first stage of opinion target extraction impacts the final sentiment classification. Evaluation on target extraction with BiLSTM-CRF is a fundamental step for this work.

Lample et al. (2016) reported that BiLSTM-CRF model obtained state-of-the-art performance in NER tasks in four languages without resorting to any language-specific knowledge or resources. Specially, in CoNLL-2002 dataset, it achieved 85.75 and 81.74 F1 score in Spanish and Dutch NER tasks, respectively; In CoNLL-2003 dataset, it achieved 90.94 and 78.76 F1 score in English and German NER tasks, respectively.

We have also conducted experiments with BiLSTM-CRF using the SemEval-2016 task 5 aspect based sentiment analysis dataset (Pontiki et al., 2016). There are 3 subtasks in this task, each subtask contains several slots. We have conducted experiments on subtask 1 slot 2: sentence-level opinion target expression extraction, on the restaurants domain. F1 score is used as metric. The experimental results are shown in Table 6.

In this table, for English, the best systems are NLANG (Toh & Su, 2016) (U) and UWB (Hercig, Brychcín, Svoboda, & Konkol, 2016) (C), respectively; For Spanish, GTI (Álvarez López, Juncal-Martínez, Fernández-Gavilanes, Costa-Montenegro, & González-Castaño, 2016) achieves both the best systems U and C; For French, they are IIT-T (Kumar, Kohail, Kumar, Ekbal, & Biemann, 2016) (U) and XRCE (Brun, Perez, & Roux, 2016) (C); For Russian, Danii achieves both the best systems U and C; For Dutch, they are IIT-T (Kumar et al., 2016) (U) and TGB (Çetin, Yıldırım, Özbey, & Eryiğit, 2016) (C).

It is observed that BiLSTM-CRF achieves the best performance on all the dataset using different languages, and outperforms the others by a good margin in 5 out of 6 languages. It indicates that BiLSTM-CRF is effective in opinion target expression extraction.

We have also evaluated the performance of BiLSTM-CRF on the MPQA dataset described in Section 4.1. We randomly select 90% sentences in MPQA dataset for training and the remaining 10% sentences for testing. BiLSTM-CRF achieves 20.73 F1 score on opinion target extraction. This is due to the complex nature of the data that many opinion targets are not simple named entities such as person, organization and location in typical NER tasks. Rather, the opinion targets could be events, abstract nouns or multi-word phrases. For example, "*overview of Johnson's eccentric career*" in sentence "*An engaging overview of Johnson 's eccentric career.*". Target number classification is much easier. It achieves 65.83% accuracy, when we classify the test sentences into 3 groups by the target numbers extracted from them. These results show that even though the performance of the first step of our approach is not very high, our pipeline approach still achieves the state-of-the-art results on most benchmarking datasets. If we can improve the performance of the sequence model for opinion target extraction, the final sentiment classification performance of our approach may be further improved.

We have also considered using other existing opinion target detection systems, which are specifically trained for this task. Unfortunately, it is not very easy to find an applicable one. Some opinion target detection systems, such as Liu et al. (2014), can also be regard as NER models.

4.3.6. Error analysis for sentence type classification

We have also done error analysis for sentence type classification. In this section, we list some result examples from the Stanford sentiment treebank. The *__O*, *__B* and *__I* concatenated after each word are the label predicted by BiLSTM-CRF.

Easy example 1: *Yet__O the__B act__I is__O still__O charming__O here__O .__O.*
Easy example 2: *The__B-MPQA movie__I-MPQA is__O pretty__O funny__O now__O and__O then__O without__O in__O any__O way__O demeaning__O its__O subjects__O .__O*
Easy example 3: *Chomp__O chomp__O !__O.*
Difficult example 1: *You__B 'll__O probably__O love__O it__B .__O*
Difficult example 2: *This__B is__O n't__O a__B new__I idea__I .__O.*
Difficult example 3: *An__O engaging__O overview__O of__O Johnson__O 's__O eccentric__O career__O .__O*

It is observed that sentences with basic constituent elements (*Easy example 1*), even if a litter long in length (*Easy example 2*), are relatively easier for target extraction with BiLSTM-CRF. One reason is that in these two sentences, the targets (*the art* and *the movie*) are commonly used nouns; Another reason is that the MPQA dataset, used for training BiLSTM-CRF model, is obtained from news sources. News text is usually more structured than the text from other sources, such as web reviews. Small imperative sentence (*Easy example 3*) is also relatively easier for target extraction, because many of them are non-target sentences.

Sentences containing pronouns, such as *you* and *it* in *Difficult example 1* and *this* in *Difficult example 2*, are relatively more difficult for target extraction with BiLSTM-CRF. Moreover, complex target, such as *overview of Johnson's eccentric career* in *Difficult example 3*, is also very difficult.

Example sentence: *Their computer-animated faces are very expressive.*
Result of CRF: *Their__O computer-animated__O faces__B are__O very__O expressive__O .__O*
Result of BiLSTM-CRF: *Their__B computer-animated__I faces__I are__O very__O expressive__O .__O*

We have also analyzed examples in which BiLSTM-CRF detects opinion targets better than CRF. As shown above, CRF can only identify a partial opinion target (*faces*), while BiLSTM-CRF can identify the whole opinion target more accurately (*their computer-animated faces*).

5. Conclusion

This paper has presented a novel approach to improve sentence-level sentiment analysis via sentence type classification. The approach employs BiLSTM-CRF to extract target expression in opinionated sentences, and classifies these sentences into three types according to the number of targets extracted from them. These three types of sentences are then used to train separate 1d-CNNs for sentiment classification. We have conducted extensive experiments on four sentence-level sentiment analysis datasets in comparison with 11 other approaches. Empirical results show that our approach achieves state-of-the-art performance on three of the four datasets. We have found that separating sentences containing different opinion targets boosts the performance of sentence-level sentiment analysis.

In future work, we plan to explore other sequence learning models for target expression detection and further evaluate our approach on other languages and other domains.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61370165, 61632011), National 863 Program of China 2015AA015405, Shenzhen Fundamental Research Funding JCYJ20150625142543470, Shenzhen Peacock Plan Research Grant KQCX20140521144507925 and Guangdong Provincial Engineering Technology Research Center for Data Science 2016KF09.

References

Álvarez López, T., Juncal-Martínez, J., Fernández-Gavilanes, M., Costa-Montenegro, E., & González-Castaño, F. J. (2016). GTI at SemEval-2016 task 5: Svm and crf for aspect detection and unsupervised aspect-based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 306–311). San Diego, California: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/S16-1049.

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems, 108*, 110–124.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the international conference on language resources and evaluation (LREC): vol. 10* (pp. 2200–2204).

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137–1155.

Brun, C., Perez, J., & Roux, C. (2016). XRCE at SemEval-2016 task 5: Feedbacked ensemble modeling on syntactico-semantic knowledge for aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 277–281). San Diego, California: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/S16-1044.

Carrillo-de Albornoz, J., & Plaza, L. (2013). An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. *Journal of the American Society for Information Science and Technology, 64*(8), 1618–1633.

Çetin, F. S., Yıldırım, E., Özbey, C., & Eryiğit, G. (2016). TGB at SemEval-2016 task 5: Multi-lingual constraint system for aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 337–341). San Diego, California: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/S16-1054.

Chaturvedi, I., Ong, Y.-S., Tsang, I. W., Welsch, R. E., & Cambria, E. (2016). Learning word dependencies in text by means of a deep recurrent belief network. *Knowledge-Based Systems, 108*, 144–154.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research, 12*, 2493–2537.

Crystal, D. (2011). *Dictionary of linguistics and phonetics*: vol. 30. John Wiley & Sons.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., & Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (ACL)* (pp. 49–54). Baltimore, Maryland: Association for Computational Linguistics.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211.

Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. J. (2016). Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications, 58*, 57–75.

Ganapathibhotla, M., & Liu, B. (2008). Mining opinions in comparative sentences. In *Proceedings of the 22nd international conference on computational linguistics (COLING): vol. 1* (pp. 241–248). Association for Computational Linguistics.

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (ACL): vol. 2* (pp. 581–586). Association for Computational Linguistics.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 6645–6649). IEEE.

Hatzivassiloglou, V., & Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on computational linguistics (COLING): vol. 1* (pp. 299–305). Association for Computational Linguistics.

Hercig, T., Brychcín, T., Svoboda, L., & Konkol, M. (2016). UWB at SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 342–349). San Diego, California: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/S16-1055.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *Computing Research Repository (CoRR), abs/1207.0580*. URL: http://arxiv.org/abs/1207.0580.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining (KDD)* (pp. 168–177). ACM.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991.

Irsoy, O., & Cardie, C. (2014a). Deep recursive neural networks for compositionality in language. In *Advances in neural information processing systems 27: annual conference on neural information processing systems (NIPS)* (pp. 2096–2104).

Irsoy, O., & Cardie, C. (2014b). Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 720–728).

Jia, L., Yu, C., & Meng, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM conference on information and knowledge management (CIKM)* (pp. 1827–1830). ACM.

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (ACL): vol. 1* (pp. 151–160). Association for Computational Linguistics.

Jindal, N., & Liu, B. (2006a). Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 244–251). ACM.

Jindal, N., & Liu, B. (2006b). Mining comparative sentences and relations. In *Proceedings, the 21st national conference on artificial intelligence and the 18th innovative applications of artificial intelligence conference (AAAI): vol. 22* (pp. 1331–1336).

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (ACL)* (pp. 655–665). Baltimore, Maryland: Association for Computational Linguistics.

Kessler, W., & Kuhn, J. (2014). A corpus of comparisons in product reviews. In *In proceedings of the 9th language resources and evaluation conference (LREC), Reykjavik, Iceland* (pp. 2242–2248). URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1001_Paper.pdf.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1746–1751).

Kumar, A., Kohail, S., Kumar, A., Ekbal, A., & Biemann, C. (2016). IIT-TUDA at SemEval-2016 task 5: Beyond sentiment lexicon: Combining domain dependency and distributional semantics features for aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 1129–1135). San Diego, California: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/S16-1174.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning (ICML), Williams College, Williamstown, MA, USA, June 28, - July 1, 2001* (pp. 282–289).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31th international conference on machine learning (ICML)* (pp. 1188–1196).

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers. doi:10.2200/S00416ED1V01Y201204HLT016.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions* . Cambridge University Press.

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.

Liu, K., Xu, L., & Zhao, J. (2014). Extracting opinion targets and opinion words from online reviews with graph co-ranking. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (ACL)* (pp. 314–324). Baltimore, Maryland: Association for Computational Linguistics.

Liu, Y., Yu, X., Chen, Z., & Liu, B. (2013). Sentiment analysis of sentences with modalities. In *Proceedings of the 2013 international workshop on mining unstructured big data using natural language processing (UnstructureNLP@CIKM)* (pp. 39–44). ACM.

McCallum, A. (2003). Efficiently inducing features of conditional random fields. In *Proceedings of the 19th conference on uncertainty in artificial intelligence (UAI)* (pp. 403–410). Morgan Kaufmann.

Mitchell, M., Aguilar, J., Wilson, T., & Durme, B. V. (2013). Open domain targeted sentiment. In *Proceedings of the 2013 conference on empirical methods in natural language processing, EMNLP 2013, 18–21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL* (pp. 1643–1654).

Muhammad, A., Wiratunga, N., & Lothian, R. (2016). Contextual sentiment analysis for social media genres. *Knowledge-Based Systems, 108*, 92–101.

Nakagawa, T., Inui, K., & Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics (NAACL)* (pp. 786–794). Association for Computational Linguistics.

Narayanan, R., Liu, B., & Choudhary, A. (2009). Sentiment analysis of conditional sentences. In *Proceedings of the 2009 conference on empirical methods in natural language processing (EMNLP): vol. 1* (pp. 180–189). Association for Computational Linguistics.

Okazaki, N. (2007). Crfsuite: a fast implementation of conditional random fields (CRFs).

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on association for computational linguistics (ACL)* (pp. 271–278). Association for Computational Linguistics.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics (ACL)* (pp. 115–124). Association for Computational Linguistics.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP)* (pp. 79–86).

Park, M., & Yuan, Y. (2015). Linguistic knowledge-driven approach to chinese comparative elements extraction. In *Proceedings of the 8th SIGHAN workshop on chinese language processing (SIGHAN-8)* (pp. 79–85). Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., Clercq, O. D., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jimnez-Zafra, S. M., & Eryiit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval)*. In *SemEval '16* (pp. 19–30). San Diego, California: Association for Computational Linguistics.

Popescu, A., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the 2005, human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP)* (pp. 9–28). Springer.

Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems, 108*, 42–49.

Poria, S., Cambria, E., Winterstein, G., & Huang, G.-B. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems, 69*, 45–63.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics, 37*(1), 9–27. doi:10.1162/coli_a_00034.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications . *Knowledge-Based Systems, 89*, 14–46.

Rill, S., Reinel, D., Scheidt, J., & Zicari, R. V. (2014). Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems, 69*, 24–33.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation.. In *Proceedings of the 2013 conference on empirical methods on natural language processing (EMNLP)* (pp. 704–714). Association for Computational Linguistics.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on empirical methods in natural language processing (emnlp)* (pp. 105–112). Association for Computational Linguistics.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45*(11), 2673–2681.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 1201–1211). ACL.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 151–161). Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP)* (pp. 1631–1642). Citeseer.

Stoyanov, V., & Cardie, C. (2008). Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd international conference on computational linguistics (COLING): vol. 1* (pp. 817–824). Association for Computational Linguistics.

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.

Tang, D., Qin, B., Feng, X., & Liu, T. (2015a). Target-dependent sentiment classification with long short term memory. arXiv preprint arXiv:1512.01100.

Tang, D., Qin, B., & Liu, T. (2015b). Deep learning for sentiment analysis: Successful approaches and future challenges . *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5*(6), 292–303. doi:10.1002/widm.1171. URL: http://dx.doi.org/10.1002/widm.1171.

Tang, D., Qin, B., Wei, F., Dong, L., Liu, T., & Zhou, M. (2015c). A joint segmentation and classification framework for sentence level sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(11), 1750–1761. doi:10.1109/TASLP.2015.2449071.

Toh, Z., & Su, J. (2016). NLANGP at SemEval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 282–288). San Diego, California: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/S16-1045.

Tsur, O., Davidov, D., & Rappoport, A. (2010). Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the 4th international conference on weblogs and social media (ICWSM)* (pp. 162–169).

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics (acl)* (pp. 417–424). Association for Computational Linguistics.

Vo, D.-T., & Zhang, Y. (2015). Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of the twenty-fourth international joint conference on artificial intelligence (IJCAI)* (pp. 1347–1353).

Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics (ACL): vol. 2* (pp. 90–94). Association for Computational Linguistics.

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the association for computational linguistics (ACL)* (pp. 246–253). Association for Computational Linguistics.

Wiebe, J., & Wilson, T. (2002). Learning to disambiguate potentially subjective expressions. In *Proceedings of the 6th conference on natural language learning (CoNLL)* (pp. 1–7). Association for Computational Linguistics.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, 39*(2–3), 165–210.

Yang, S., & Ko, Y. (2011). Extracting comparative entities and predicates from texts using comparative type classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (ACL): vol. 1* (pp. 1636–1644). Association for Computational Linguistics.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on empirical methods in natural language processing (EMNLP)* (pp. 129–136). Association for Computational Linguistics.

Zhang, L., Ferrari, S., & Enjalbert, P. (2012). Opinion analysis: the effect of negation on polarity and intensity. In *KONVENS workhop PATHOS-1st workshop on practice and theory of opinion mining and sentiment analysis* (pp. 282–290).

Zhang, M., Zhang, Y., & Vo, D.-T. (2015). Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP)* (pp. 612–621).