Negative Focus Detection via Contextual Attention Mechanism

Longxiang Shen[1], **Bowei Zou**[12]*; **Yu Hong**[1],
Qiaoming Zhu[1], **Guodong Zhou**[1], **Ai Ti Aw**[2]
[1]School of Computer Scienceand Technology, Soochow University, China
[2]Aural & Language Intelligence Department, Institute for Infocomm Research, Singapore
lxshen.scu@gmail.com, zoubowei@suda.edu.cn,
{yhong, qmzhu, gdzhou}@suda.edu.cn, aaiti@i2r.a-star.edu.sg

Abstract

Negation is a universal but complicated linguistic phenomenon, which has received considerable attention from the NLP community over the last decade, since a negated statement often carries both an explicit negative focus and implicit positive meanings. For the sake of understanding a negated statement, it is critical to precisely detect the negative focus in context. However, how to capture contextual information for negative focus detection is still an open challenge. To well address this, we come up with an attention-based neural network to model contextual information. In particular, we introduce a framework which consists of a Bidirectional Long Short-Term Memory (BiLSTM) neural network and a Conditional Random Fields (CRF) layer to effectively encode the order information and the long-range context dependency in a sentence. Moreover, we design two types of attention mechanisms, word-level contextual attention and topic-level contextual attention, to take advantage of contextual information across sentences from both the word perspective and the topic perspective, respectively. Experimental results on the SEM'12 shared task corpus show that our approach achieves the best performance on negative focus detection, yielding an absolute improvement of 2.11% over the state-of-the-art. This demonstrates the great effectiveness of the two types of contextual attention mechanisms.

1 Introduction

As a linguistic phenomenon that reverses the polarity of a statement or its property into opposite, negation is ubiquitous in human languages. Consider following sentence S1. A negative affix *-n't* negates the statement that *mutual fund trades take effect until the market close*. Negation processing has been shown critical for a number of NLP applications, such as biomedical information extraction (Morante, 2010; Mehrabi et al., 2015) and sentiment analysis (Wiegand et al., 2010; Jimenez-Zafra et al., 2017).

(S1) *Mutual fund trades don't* ***take*** *effect {until the market close}.*[1]

In general, negation is marked by a word or an affix and interact with other parts in sentence. Blanco and Moldovan (2011) demonstrated that over 65% of negated statements in PropBank (Palmer et al., 2005) convey implicit positive meanings. For instance, underneath sentence S1, the statement that *mutual fund trades take effect* is true, when removing the propositional clause *until the market close* from the original sentence. As the definition of negation focus in SEM'12 (Morante and Blanco, 2012), a negative focus is the most prominently negated part of sentence (Huddleston and Pullum, 2002). In sentence S1, the negative focus is the propositional clause *until the market close*, yielding the interpretation that *mutual fund trades take effect, but not until the market close*. As discussed above, correctly understanding a negated statement requires precise detection of the negative focus. In this paper, we focus on detecting the negative focus in a sentence with a negative keyword and its corresponding dominated verb.

Not as simple as that in speech, negative focus detection poses considerable challenges on text understanding without knowing stress or intonation information. Previous studies (Blanco and Moldovan, 2011; Zou et al., 2014) pointed out that the contextual information plays a critical role in negative focus detection, since the real negative intention often relates to what the author repeatedly

* corresponding auther

[1] #13 of the test set in *SEM 2012 shared task corpus. Throughout this paper, the negated verbs are marked in bold, and the negative foci are in curly braces.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2251–2261,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

state in adjacent sentences. The same negative statement might be interpreted discriminately depending on different contexts. For example, in following three scenarios, consider what is the more likely negative focus of sentence S1 with different contexts:

I. ..., *but exchange fund trades do.*
negative focus: *mutual fund trades*
II. ..., *unless it reboots.*
negative focus: *take effect*
III. ..., *for the shareholders stayed put today.*
negative focus: *until the market close*

In scenario I, since *exchange fund trades* is emphasized in context, the negative focus should be the phrase *mutual fund trades*, yielding the interpretation that *the trades which don't take effect are mutual fund trades, but not exchange fund trades*. Unlike scenario I, the interpretation of negative statement from scenario II is that *Mutual fund trades don't take effect without reboot*, with the context *unless it reboots*, and the phrase *take effect* is the negative focus. Similarly, scenario III points out the reason of *not take effect* is that *the shareholders stayed put today*, which corresponds to the negative focus *until the market close*. In addition, it is worth noting that a negative focus is often not syntactically determined by sentence structure, instead pragmatically judged by authors' intentions conveyed in context.

This is consistent with the attention mechanism in neural networks which has been proven effective to improve the performances for many NLP tasks, such as sentiment analysis (Wang et al., 2016) and relation classification (Zhou et al., 2016). In this paper, we come up with two types of attention mechanisms to enforce the model to take full advantage of the contextual information for negative focus detection.

In particular, we introduce a framework which consists of a bidirectional long short-term memory neural network and a conditional random fields layer (BiLSTM-CRF) to effectively encode the order information and the long-range context dependency in sentence. On this basis, we propose two attention mechanisms to capture contextual information across sentences. One is to additionally append word-level vectors of adjacent sentences into the input vectors. Such attention can automatically focus on the words in context related to the negative focus. The other is to concatenate the topic-level representations into the input vectors to better capture the contextual information. Experimentation on the SEM'12 shared task corpus (Morante and Blanco, 2012) shows that our approach achieves the best performance on negative focus detection task, yielding an absolute improvement of 2.11% over the state-of-the-art.

The rest of this paper is organized as follows. In Section 2, we give a brief review of related work. In Section 3, we introduce the details of the proposed approach. We show our experimental results and discussions in Section 4. Finally, Section 5 concludes with possible directions for future work.

2 Related Work

Earlier studies of negation processing in computational linguistics mainly lie in biomedical information extraction (Chapman et al., 2001; Goldin and Chapman, 2003). With the release of the BioScope corpus (Vincze et al., 2008), the negation processing techniques received a substantial boost (Morante et al., 2008; Apostolova et al., 2011; Zou et al., 2013). In addition, negation also was studied as a critical factor in polarity shifting in sentiment and opinion analysis with various lexical and syntactic features (Socher et al., 2013; Cruz et al., 2016).

In general, existing studies of negation processing mainly concentrate on three aspects: 1) cue detection, which finds negative triggers or expressions in text; 2) scope resolution, which determines the grammatical scope in a sentence affected by a negative cue; and 3) focus detection, which identifies the most prominently/explicitly negated part in a negative statement. A negative focus could be considered as the semantic part in scope that is intended to be interpreted as false to make other parts to be true. Among above these subtasks, negative focus detection is most extremely challenging in negation processing.

In the literature, research on negative focus detection was pioneered by Blanco and Moldovan (2011), who proposed a supervised learning model with a set of highly hand-crafted features as a benchmark for this task. On the basis, Blanco and Moldvan (2013) incorporates negation into other existing semantic representations. The dataset annotated by Blanco and Moldovan is used as a standard evaluation corpus for the SEM'12 shared task (Morante and Blanco, 2012). In this shared task, Rosenberg and Bergler (2012) employed three

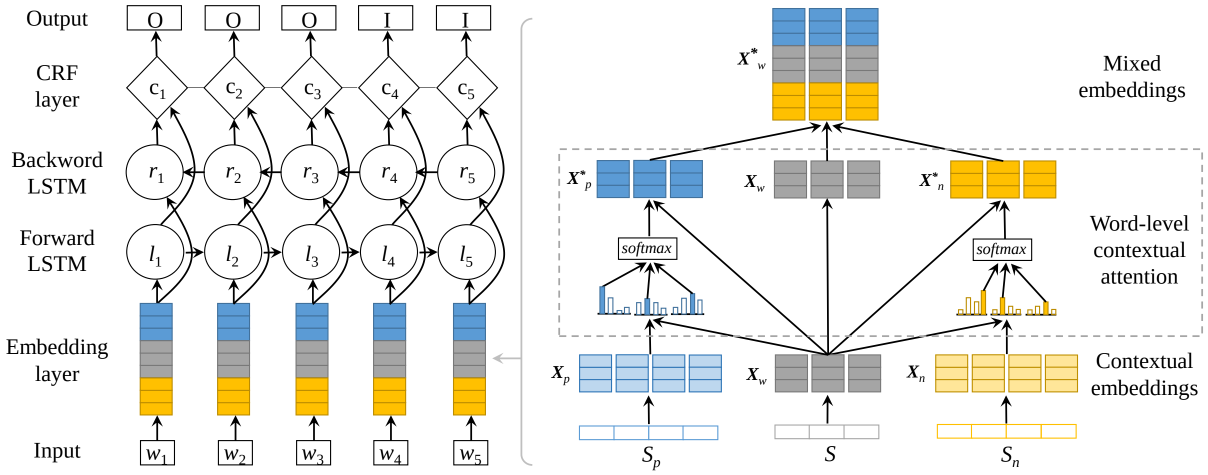

Figure 1: Architecture of BiLSTM-CRF network with word-level contextual attention mechanism for negative focus detection.

kinds of heuristic rules for negative focus detection. Moreover, Zou et al. (2014; 2015) proposed graph-based models to enrich intra-sentence features with inter-sentence features from both lexical and topic perspectives, respectively. Almost all of the above studies have demonstrated that the information contained in context plays a critical role in negative focus detection. In this paper, we explore the effectiveness of different representations and processing manners for modeling context in this task.

In addition to the SEM'12 shared task corpus, there are several annotated datasets for negative focus detection. Anand and Martell (2012) made a point about how to describe the contribution of negative focus to a sentence via the theory of question under discussion, and reannotated a part of the SEM'12 shared task corpus. Matsuyoshi et al. (2014) annotated 1,327 instances of negative foci in Japanese text, and proposed a heuristic rule-based approach to identify them. Banjade and Rus (2016) annotated the DT-Neg corpus on tutorial dialogue genre, which contains 1,088 of negative focus instances. Considering the scale and accessibility of the dataset, we utilize the SEM'12 shared task corpus for experimentation.

3 Contextual Attention-based Negative Focus Detection

In this section, we first introduce the BiLSTM-CRF framework. On the basis, we then come up with two kinds of attention mechanisms to model the contextual information, i.e. word-level attention and topic-level attention.

3.1 BiLSTM-CRF Framework

In this paper, we recast negative focus detection problem as an I/O tagging task, and apply the IO label scheme, where **I** denotes the token is **I**nside the scope of negative focus in a sentence, and **O** indicates the token is **O**utside of this scope. For example, the sentence S1 are tagged as below.

mutual/O *fund*/O *trades*/O *don't*/O ***take***/O *effect*/O *until*/I *the*/I *market*/I *close*/I

The left part of Figure 1 illustrates the BiLSTM-CRF framework for negative focus detection. First, the sequence of embeddings ($\boldsymbol{x}_i$) is given as input to BiLSTM networks, which generates a representation of the left context ($\boldsymbol{l}_i$) and the right context ($\boldsymbol{r}_i$) for each token in a sentence. Then, these representations are concatenated ($\boldsymbol{c}_i$) and linearly projected onto a CRF layer to take into account neighboring tags, yielding the final prediction for every token (y_i).

Embedding Layer We build an embedding layer to encode words and semantic role labels. Given an input sentence $S = (w_1, w_2, ..., w_n)$, we first transform each word into a real-valued vector $\boldsymbol{x}_w \in \mathbb{R}^{d_w}$. In addition, according to the annotated guideline (Blanco and Moldovan, 2011), semantic roles in sentence often have a close relationship with negative focus and negated verb. To capture such informative features, we then map the semantic role tag of each word to a real-valued vector $\boldsymbol{x}_{sr}$ of dimension d_{sr} using a embedding matrix $\boldsymbol{SR} \in \mathbb{R}^{d_{sr} \times |V_{sr}|}$, where V_{sr} is the set of semantic role label vocabulary. Finally, we represent an input sentence as a vector sequence $\boldsymbol{x} = [\boldsymbol{x}_w; \boldsymbol{x}_{sr}]$

with the embedding dimension $d = (d_w + d_{sr})$.

BiLSTM Layer We employ LSTM networks (Hochreiter and Schmidhuber, 1997) which incorporate memory-cells to capture long contextual dependencies in sequence. Formally, each cell in LSTM can be computed as follows:

$$\begin{aligned}
\boldsymbol{X}_t &= [\boldsymbol{h}_{t-1} \;\; \boldsymbol{x}_t]^T \\
\boldsymbol{i}_t &= \sigma(\boldsymbol{W}_i \boldsymbol{X}_t + \boldsymbol{b}_i) \\
\boldsymbol{f}_t &= \sigma(\boldsymbol{W}_f \boldsymbol{X}_t + \boldsymbol{b}_f) \\
\boldsymbol{o}_t &= \sigma(\boldsymbol{W}_o \boldsymbol{X}_t + \boldsymbol{b}_o) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tanh(\boldsymbol{W}_c \boldsymbol{X}_t + \boldsymbol{b}_c) \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t)
\end{aligned} \tag{1}$$

where σ is the sigmoid function, and $\odot$ is the element-wise multiplication. $\boldsymbol{W}_i$, $\boldsymbol{W}_f$, and $\boldsymbol{W}_o$ are the weighted matrices, and $\boldsymbol{b}_i$, $\boldsymbol{b}_f$, and $\boldsymbol{b}_o$ are the biases, which parametrize the transformations of input, forget, and output gates, respectively. $\boldsymbol{h}_t$ is the vector of hidden state.

Moreover, to benefit from both previous and future information, a forward LSTM and a backward LSTM are employed to generate a representation $\overrightarrow{\boldsymbol{h}}_t$ of the left context and $\overleftarrow{\boldsymbol{h}}_t$ of the right, respectively. The word representation $\boldsymbol{h}_t$ is obtained by concatenating the left and right outputs $\left[\overrightarrow{\boldsymbol{h}}_t; \overleftarrow{\boldsymbol{h}}_t\right]$.

CRF Layer To learn the dependencies across output labels of the BiLSTM layer, we model them jointly using a conditional random field (CRF) layer (Lafferty et al., 2001). For an input sentence $\boldsymbol{x}$, we denote $\boldsymbol{C}$ as the matrix of output by BiLSTM. $\boldsymbol{C}$ is of size $n \times k$, where k is the number of distinct tags, and $c_{i,j}$ corresponds to the score of the j^{th} tag of the i^{th} token in a sentence. For a sequence of predictions $\boldsymbol{y}$, we define its score to be

$$s(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} c_{i, y_i}, \tag{2}$$

where $A_{i,j}$ denotes the score of a transition from the tag i to the tag j. y_0 and y_{n+1} are the additional tags of *START* and *END*, respectively. A softmax layer over all possible tag sequences yields a probability for the sequence $\boldsymbol{y}$:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp(s(\boldsymbol{x}, \boldsymbol{y})), \tag{3}$$

where $Z(\boldsymbol{x}) = \sum_Y \exp(s(\boldsymbol{x}, Y))$, and Y denotes all possible tag sequences.

During training, we maximize the log-probability of the correct tag sequence:

$$\mathcal{L}_c = \max \log(p(\boldsymbol{y}|\boldsymbol{x})). \tag{4}$$

While during decoding, we predict the output sequence that obtains the maximum score given by

$$\boldsymbol{y}^* = \arg\max_{Y} s(\boldsymbol{x}, Y). \tag{5}$$

3.2 Word-level Contextual Attention

As discussed in Section 1, contextual information is vital for negative focus detection. However, the BiLSTM-CRF model might fail to identify the important information related to a negative focus in contexts. To address this issue, we propose two kinds of attention mechanisms, i.e. the word-level contextual attention and the topic-level contextual attention.

The right part of Figure 1 illustrates the architecture of the word-level contextual attention mechanism. Given previous sentence S_p and next sentence S_n, we respectively encode each word $\boldsymbol{w}_{p_i}$ and $\boldsymbol{w}_{n_i}$ into real-valued vectors $\boldsymbol{x}_{p_i}$ and $\boldsymbol{x}_{n_i} \in \mathbb{R}^{d_w}$ by the word embeddings described in Subsection 3.1. Suppose that N_p, N, and N_n denote the lengths of previous sentence, current sentence, and next sentence, $\boldsymbol{X}_p \in \mathbb{R}^{d_w \times N_p}$, $\boldsymbol{X}_w \in \mathbb{R}^{d_w \times N}$, and $\boldsymbol{X}_n \in \mathbb{R}^{d_w \times N_n}$ are made up of all word embeddings in S_p, S, and S_n, respectively.

This means, the word-level contextual attention mechanism produces attention weight vectors $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_n$, and weighted hidden representations $\boldsymbol{X^*}_p$ and $\boldsymbol{X^*}_n$ as follows:

$$\boldsymbol{\alpha}_p = softmax(\max_{col}(\frac{\boldsymbol{X}_p^T \cdot \boldsymbol{X}_w}{\sqrt{d_w}})) \tag{6}$$

$$\boldsymbol{\alpha}_n = softmax(\max_{col}(\frac{\boldsymbol{X}_n^T \cdot \boldsymbol{X}_w}{\sqrt{d_w}})) \tag{7}$$

$$\boldsymbol{X^*}_p = \boldsymbol{X}_w \odot (\boldsymbol{\alpha}_p \otimes \boldsymbol{e}_{d_w}) \tag{8}$$

$$\boldsymbol{X^*}_n = \boldsymbol{X}_w \odot (\boldsymbol{\alpha}_n \otimes \boldsymbol{e}_{d_w}) \tag{9}$$

where the contextual attention weights vectors $\boldsymbol{\alpha}_p$, $\boldsymbol{\alpha}_n \in \mathbb{R}^N$, and the contextual representation matrices $\boldsymbol{X^*}_p$, $\boldsymbol{X^*}_n \in \mathbb{R}^{d_w \times N}$. The operator $\max_{col}(\cdot)$ in Eq.(6) and Eq.(7) means extracting the maximum value for each column in matrix to generate a vector. Besides, we apply the scaling factor of $1/\sqrt{d_w}$ to counteract the effect that the dot products grow large in magnitude (Vaswani et al., 2017). The operation $\boldsymbol{\alpha} \otimes \boldsymbol{e}_{d_w}$ in Eq.(8) and

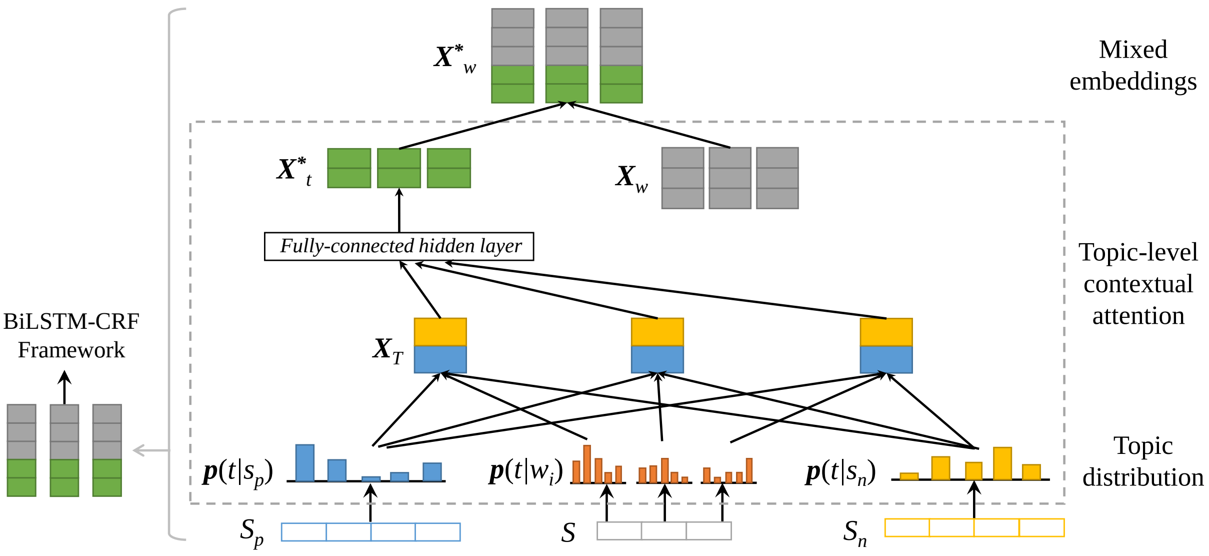

Figure 2: Architecture of topic-level contextual attention mechanism for negative focus detection.

Eq.(9) means that it repeatedly concatenates $\boldsymbol{\alpha}$ for d_w times, where $\boldsymbol{e}_{d_w}$ is a row vector with d_w of 1s.

Finally, the final sentence representation is obtained by

$$\boldsymbol{X^*}_w = [\boldsymbol{X^*}_p; \boldsymbol{X}_w; \boldsymbol{X^*}_n], \qquad (10)$$

and we replace $\boldsymbol{X}_w$ with $\boldsymbol{X^*}_w$ in the embedding layer of the BiLSTM-CRF framework.

3.3 Topic-level Contextual Attention

To acquire the latent topical distribution of words, we apply the GenSim topic modeling package[2] (Rehurek and Sojka, 2010) to generate an LSI semantic model (Bradford, 2008) and obtain two probabilities with the topic distribution derived from the Wikipedia corpus[3]: 1) $\boldsymbol{p}(t|w)$, the topic distribution given word w, and 2) $\boldsymbol{p}(t|s)$, the topic distribution given sentence s. The correlation between the word w and the sentence s can be calculated by the similarity between their corresponding topic distribution:

$$\boldsymbol{Sim}_{ws} = \boldsymbol{p}(t|w) \odot \boldsymbol{p}(t|s). \qquad (11)$$

Figure 2 illustrates the architecture of topic-level contextual attention mechanism. In this paper, we obtain the topic distribution of each word in current sentence $\boldsymbol{p}(t|w_i)$ and its adjacent sentences $\boldsymbol{p}(t|s_p)$ and $\boldsymbol{p}(t|s_n)$, respectively. The topic-level contextual attention mechanism produces the topical hidden representations

$$\boldsymbol{x}_{t_i} = [\boldsymbol{Sim}_{w_i s_p}; \boldsymbol{Sim}_{w_i s_n}], \qquad (12)$$

where $\boldsymbol{x}_{t_i} \in \mathbb{R}^{2d_T}$, and d_T is the dimension of topic distribution. Suppose that N denotes the length of current sentence, $\boldsymbol{X}_T \in \mathbb{R}^{2d_T \times N}$ is made up of all words' topical hidden representations. To better learn the latent features and expand the dimension of $\boldsymbol{X}_T$ from $2d_T$ to d_t, we utilize a fully-connected hidden layer:

$$\boldsymbol{X^*}_t = \tanh(\boldsymbol{W}\boldsymbol{X}_T + \boldsymbol{b}), \qquad (13)$$

where $\boldsymbol{W} \in \mathbb{R}^{d_t \times 2d_T}$ is the transpose of parameter matrix and $\boldsymbol{b} \in \mathbb{R}^{d_t}$. Note that d_t is a hyper-parameter are fine-tuned on the development set. As a result, we can obtain the sentence representation as follows

$$\boldsymbol{X^*}_w = [\boldsymbol{X}_w; \boldsymbol{X^*}_t], \qquad (14)$$

and replace $\boldsymbol{X}_w$ with $\boldsymbol{X^*}_w$ in the embedding layer of the BiLSTM-CRF framework.

4 Experimentation

In this section, we first introduce the details of dataset, hyper-parameters settings, and baselines. Then we compare our model with existing systems and baselines to demonstrate the effectiveness of the proposed contextual attention mechanisms. Finally, we perform in-depth analysis of the impact factors of our approach and conduct careful error analysis to better understand the limitation of our model.

[2] https://radimrehurek.com/gensim/index.html
[3] https://radimrehurek.com/gensim/wiki.html

Hyper-parameter	Value
word embed dim d_w	1,024
semantic role embed dim d_{sr}	200
topic embed dim d_t	300
topic number T	80
LSTM hidden layer size h	250
dropout probability p	0.3
learning rate η	0.015

Table 1: Hyper-parameter settings.

4.1 Dataset and Settings

All of experiments conduct on the SEM'12 shared task corpus[4] (Morante and Blanco, 2012) which is annotated on top of PropBank. According to the guideline (Blanco and Moldovan, 2011), negative focus is restricted to verbal negation (marked with MNEG labels), thus all of the words belonging to a semantic role most likely to correspond to the verbal negation are selected as negative focus. In total, this corpus contains 3,544 instances of negative focus. For fair comparison, we adopt the same partition as SEM'12 shared task in all experiments, i.e., with 2,304 for training, 531 for development, and 712 for testing. For each instance, the corpus provides the previous and next sentences as contexts. We evaluate our results in terms of accuracy (acc for short), i.e., for each negation, the predicted focus is considered correct if it is a perfect match with its gold annotation.

Table 1 lists the hyper-parameters in our experiments. We utilize the pre-trained word embeddings by ELMo[5] (Peters et al., 2018). The semantic role embeddings are initialized randomly by a continuous uniform distribution. All the models are optimized using the stochastic gradient descent (SGD). We pick the parameters showing the best performance on the development set when no further improvement occurs within 50 epochs, and report the performances on the test set.

To verify the effectiveness of our approach, we compare with several baselines on negative focus detection, which are briefly introduced as follows.

LSTM: The LSTM model which contains only a forward LSTM layer with a fully-connected layer followed by a softmax classifier.

BiLSTM: The BiLSTM model employs the same architecture as LSTM, except a additional backward LSTM layer.

	System	Acc
Existing Methods	CLaC (Rosenberg (2012))	60.00
	FOC-DET (Blanco (2013))	65.50
	WTGM (Zou (2015))	68.40
Our Methods	LSTM	58.71
	BiLSTM	60.81
	BiLSTM-CRF	67.28
	W-Att BiLSTM-CRF	70.22
	T-Att BiLSTM-CRF	**70.51**
	WT-Att BiLSTM-CRF	69.80

Table 2: Performances of negative focus detection systems on the SEM'12 corpus.

BiLSTM-CRF: The BiLSTM-CRF model described in Subsection 3.1 without any contextual attention mechanism.

W-Att BiLSTM-CRF: The word-level contextual attention-based BiLSTM-CRF model described in Subsection 3.2.

T-Att BiLSTM-CRF: The topic-level contextual attention-based BiLSTM-CRF model described in Subsection 3.3.

WT-Att BiLSTM-CRF: The BiLSTM-CRF model with an input representation $[\boldsymbol{X}^*{}_p; \boldsymbol{X}_w; \boldsymbol{X}^*{}_n; \boldsymbol{X}^*{}_t]$. This system combines the above two types of contextual attention mechanisms.

CLaC: A heuristic-based system with three linguistic rules related to adverb constituent, passive, and negative trigger of focus, respectively (Rosenberg and Bergler, 2012).

FOC-DET: A decision tree based system with 22 kinds of manually designed features (Blanco and Moldovan, 2013).

WTGM: A graph model consists of a word layer and a topic layer to capture lexical contextual information (Zou et al., 2015).

4.2 Results

Table 2 shows the performance comparison of various negative focus detection models. We can see that all of contextual attention based models (row 7-9 in Table 2) achieve better perfomances than existing methods (row 1-3 in Table 2) and models without contextual attention (row 4-6 in Table 2). In addition, both word-level and topic-level contextual attention based models outperform the state-of-the-art system (WTGM in Table 2) with about 2% accuracy gain at least. The

[4] http://www.clips.ua.ac.be/sem2012-st-neg

[5] https://allennlp.org/elmo

Method	Context	Acc
No Attention	N/A	67.28
Word-level Attention	only previous sentence	69.24
	only next sentence	69.94
	previous+next sentences	**70.22**
Topic-level Attention	only previous sentence	69.38
	only next sentence	70.08
	previous+next sentences	**70.51**

Table 3: Performance comparison by using different contextual information into the BiLSTM-CRF framework.

results demonstrate the effectiveness of these two types of contextual attention mechanisms.

Moreover, to better quantify the contribution of the different attention mechanisms of our approach, we also conduct several attention variants. Comparing the three types of attention mechanisms (row 7-9 in Table 2), the topic-level attention based model achieves the best performance. We also observe that the word-level attention based model also achieves comparative performance with the topic-level one. However, when combining the two types of attention mechanisms, the performance declines. It indicates that both the word-level attention mechanism and the topic-level one can capture the contextual information effectively, but applying such information repeatedly might lead to feature redundancy.

In addition to the methods that take advantage of the contextual features in adjacent sentences, we also compare the performances of different frameworks for negative focus detection (row 4-6 in Table 2), which only apply the features in current sentence. The results indicate that the BiLSTM-CRF framework is a better fit for encoding order information and long-range context dependency for such sequence labeling task.

4.3 Analysis and Discussion

Impact of Contextual Information. We examine the impact of the context position on the performance for our approach. Table 3 shows the performance comparison when using different contextual information. We can see that, both the word-level and topic-level attention based models achieve the best performances by utilizing the contextual features from the previous and next sentences, which verifies the effectiveness of the two types of contexts. Besides, the results also indicate that contextual information is key for negative focus detection, especially the next sentence.

Method	Embeddings	Acc
BiLSTM-CRF	only word	67.28
	word + SR	68.82
W-Att BiLSTM-CRF	only word	68.26
	word + SR	70.22
T-Att BiLSTM-CRF	only word	68.40
	word + SR	70.51

Table 4: Performance comparison for systems only using word embeddings and those using word and SR embeddings (SR: semantic role).

Semantic Role	Train	Dev	Test
A1	980	222	309
AM-NEG	592	138	172
AM-TMP	161	35	46
AM-MNR	127	27	38
A2	112	28	36
A0	94	23	31
AM-ADV	78	23	26
Others	165	30	61
NONE	88	19	35
Total	2,304	531	712

Table 5: Statistics of the top seven semantic role labels of negative focus. NONE class: the negative focus has at least one word that does not belong to any role.

Impact of Semantic Role Information. Since a completed and exclusive semantic role that is most likely negated is annotated as a negative focus in sentence, according to the annotation guideline (Blanco and Moldovan, 2011), semantic role information plays a critical role in predicting the focus. Therefore, we compare the performances between the models adding semantic role information in embedding layer and those only using the word embeddings. As shown in Table 4, we can see that the performances of all three models are improved obviously, which demonstrates the effectiveness of semantic role information for this task.

In addition, Table 5 lists the top seven semantic role labels of negative focus in the SEM'12 shared task corpus. We can observe that the distribution of semantic role labels is relatively concentrated into A1, AM-NEG, AM-TMP, and AM-MNR. Unlike other negation processing task, such as scope resolution, which heavily relies on the

Method	Composition Manner	Acc
W-Att BiLSTM-CRF	within input layer	**70.22**
	within hidden layer	69.38
T-Att BiLSTM-CRF	within input layer	**70.51**
	within hidden layer	69.80

Table 6: Performance comparison for different composition manners of attention matrices.

Word Embedding	Dimension	Acc
Senna	50	70.08
Glove	100	69.66
Word2vec	300	69.10
BERT	768	70.22
ELMo	1,024	**70.51**

Table 7: Performance comparison for different pre-trained word embeddings on the T-Att BiLSTM-CRF model.

lexical or syntactic features (Zou et al., 2013; Qian et al., 2016), the negative focus detection task tends to consider more semantic relation between the focus and other words and phrases.

Impact of Attention Manner. In this paper, our contextual attention matrices are concatenated into the input embeddings. In addition, we also explore another manner of adding attention mechanisms, which directly concatenates the attention matrix into the hidden layer of BiLSTM-CRF framework. As shown in Table 6, it indicates that the composition manner within the input layer is more effective than that within the hidden layer for both word-level and topic-level attention mechanisms.

Impact of Pre-trained Word Embedding. To compare the impacts of different pre-trained word embeddings for negative focus detection task, we attempt to employ other pre-trained word embeddings. Table 7 show the performances of the T-Att BiLSTM-CRF model with different pre-trained word embeddings, including Senna[6], Glove[7], Word2vec[8], and BERT[9]. We can see that ELMo achieves the best performance, but the performance gaps between different pre-trained word embeddings and different dimensions are not significant.

Embeddings	Acc
word	67.28
word+chunking label	67.84
word+dependency label	68.12
word+PoS tag	68.26
word+semantic role	**68.82**
all features	68.54

Table 8: Performance comparison for different features with the BiLSTM-CRF model.

Impact of Different Features. Table 8 shows the comparison of the BiLSTM-CRF models with different feature embeddings. The reason we employ the basic model is to avoid that the impacts of different features are covered by contextual attention mechanisms. The results indicate that the semantic role feature is the most effective for negative focus detection, which confirms our analysis above. In addition, the performance of the system with all features is lower than that only using the semantic role feature. It might be that other features cannot capture more effective information than the semantic role features.

Error Analysis. We manually analyzed a total of 210 of the incorrect instances labeled by the T-Att BiLSTM-CRF system. The main error patterns include:

1) Focus is the negative trigger (50/210, 23.8%). According to the annotation guideline of negative focus (Blanco and Moldovan, 2011), when a negated statement does not carry any positive meaning, or if it cannot be inferred that a part of sentence is negated, the gold standard annotation regards the negative trigger (e.g., *not*) itself as the negative focus. In this case, neither the local information nor the contextual information could identify the negative focus. It is difficult to accurately identify such negative focus by our approach.

2) Annotation issue (84/210, 40%). As mentioned in the above, a negative focus always corresponds a single semantic role. However, we find that some of labeled instances do not meet this constraint. Such non-conformity contains three aspects: a) the focus has at least one word that does not belong to any role (35/84), b) the focus is labeled as a fragment of semantic role (36/84), and c) the focus is labeled across more than one roles (13/84). These situations may heavily limit the effects of the semantic features in embedding layer and the CRF layer which learns the depen-

[6] http://ronan.collobert.com/senna

[7] http://nlp.stanford.edu/projects/glove

[8] https://code.google.com/archive/p/word2vec

[9] https://github.com/google-research/bert

dencies across output labels. To verify the impact of this issue, we conduct the T-Att BiLSTM-CRF system on the processed test set that remove the 84 of instances. The performance achieves 79.94 in accuracy, yielding an improvement of 9.43% (Table 2). Therefore, the future work should fix the issue of corpus, or reconsider whether the guidelines could generalize the linguistic phenomenon of negative focus.

3) Subjectivity of the annotation process (11/210, 5.2%). There are several golden annotated instances that we disagree with. For example,

(S2) Previous sentence: ..., *one equipment failure forces a complete plant shutdown*.

(S3) {*On some days*}, *the Nucor plant doesnt* ***produce*** *anything*.[10]

Consider the previous sentence S2, the author emphasize that *the Nucor plant will shut down*, thus the most likely negative focus should be *anything* in S3, whereas the prepositional phrase *on some days* only provides a temporal statement in our view. However, different annotators might have different understandings and interpretations for this case. Note that the inter-annotator agreement of the corpus is only 0.72 after all, which indicates the inherent challenge in negative focus detection task.

5 Conclusion

In this work, we discussed the unique necessity of modeling contextual information for negative focus detection. We designed an attention based neural architecture which can better capture contextual features by utilizing the word-level and topic-level attention mechanisms. Experiments on the SEM'12 shared task corpus demonstrate the effectiveness of our approach. The datasets and source code of this paper are publicly available at `http://github.com/longxshen/NegFocus`.

In the future, we intend to review the annotated guidelines for negative focus with the problematic annotation cases mentioned in this paper, and explore the ways to combine additional patterns and constraints from negative focus definition with neural network techniques to further improve negative focus detection.

[10] #287 of test set in SEM'12 shared task corpus.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 61703293, No. 61751206, and No. 61672368). This work was also supported by the joint research project of Alibaba and Soochow University. Finally, we would like to thank the anonymous reviewers for their insightful comments and suggestions.

References

Pranav Anand and Craig Martell. 2012. Annotating the focus of negation in terms of questions under discussion. In *Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, pages 65–69. Association for Computational Linguistics.

Emilia Apostolova, Noriko Tomuro, and Dina Demnerfushman. 2011. Automatic extraction of lexico-syntactic patterns for detection of negation and speculation scopes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 283–287. Association for Computational Linguistics.

Rajendra Banjade and Vasile Rus. 2016. Dt-neg: Tutorial dialogues annotated for negation scope and focus in context. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 3768–3771.

Eduardo Blanco and Dan Moldovan. 2011. Semantic representation of negation using focus detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 581–589. Association for Computational Linguistics.

Eduardo Blanco and Dan Moldovan. 2013. Retrieving implicit positive meaning from negated statements. *Natural Language Engineering (NLE)*, 20(4):501–535.

Roger Bradford. 2008. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *Proceedings of the 17th Conference on Information and Knowledge Management (CIKM)*, pages 153–162.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Noa P Cruz, Maite Taboada, and Ruslan Mitkov. 2016. A machine-learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*, 67(9):2118–2136.

Ilya M. Goldin and Wendy W. Chapman. 2003. Learning to detect negation with not in medical texts. In *Proceedings of the Workshop on Text Analysis and Search for Bioinformatics of SIGIR*.

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

Rodney D. Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Salud Maria Jimenez-Zafra, M Teresa Martin Valdivia, Eugenio Martinez Camara, and Luis Alfonso Urenalopez. 2017. Studying the scope of negation for spanish sentiment analysis on twitter. *IEEE Transactions on Affective Computing*.

John D. Lafferty, Andrew Mccallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Eighteenth International Conference on Machine Learning*, pages 282–289.

Suguru Matsuyoshi, Ryo Otsuki, and Fumiyo Fukumoto. 2014. Annotating the focus of negation in japanese text. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 1743–1750.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, et al. 2015. Deepen: A negation detection system for clinical text incorporating dependency relation into negex. *Journal of Biomedical Informatics*, 54:213–219.

Roser Morante. 2010. Descriptive analysis of negation cues in biomedical texts. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC)*, pages 1429–1436.

Roser Morante and Eduardo Blanco. 2012. *sem 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 256–274. Association for Computational Linguistics.

Roser Morante, Anthony Liekens, and Walter Daelemans. 2008. Learning the scope of negation in biomedical texts. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 715–724. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. pages 2227–2237.

Zhong Qian, Peifeng Li, Qiaoming Zhu, Guodong Zhou, Zhunchen Luo, and WeiLuo. 2016. Speculation and negation scope detection via convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 815–825.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Sabine Rosenberg and Sabine Bergler. 2012. Uconcordia: Clac negation focus detection at *sem 2012. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 294–300. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP)*, pages 1631–1642. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Veronika Vincze, Gyorgy Szarvas, Richard Farkas, Gyorgy Mora, and Janos Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(11):1–9.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 606–615. Association for Computational Linguistics.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andres Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 207–212. Association for Computational Linguistics.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2013. Tree kernel-based negation and speculation scope detection with structured syntactic parse features. In *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP)*, pages 968–976. Association for Computational Linguistics.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2014. Negation focus identification with contextual discourse information. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 522–530. Association for Computational Linguistics.

Bowei Zou, Qiaoming Zhu, and Guodong Zhou. 2015. Unsupervised negation focus identification with word-topic graph model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1632–1636. Association for Computational Linguistics.