基于双向 LSTM 与 CRF 融合模型的否定聚焦点识别

沈龙骧，邹博伟，叶静，周国栋，朱巧明

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘要：否定表达作为自然语言文本中常见的语言现象，对自然语言处理上层应用，如情感分析、信息抽取等，具有十分重要的意义。否定聚焦点识别任务是更细粒度的否定语义分析，其旨在识别出句子中被否定词修饰和强调的文本片段。本文将该任务作为序列标注问题，提出了一种基于双向长短期记忆网络结合条件随机场（BiLSTM-CRF）的否定聚焦点识别模型，其中，BiLSTM 网络能够充分利用上下文信息并抓取全局特征，CRF 层能够有效学习输出标签之间的前后依赖关系。在*SEM2012 评测任务数据集上的实验结果表明，基于 BiLSTM-CRF 的否定聚焦点识别方法的准确率（Accuracy）达到 69.58%，与目前最好的系统相比，性能提升了 2.44%。

关键词：否定聚焦点；BiLSTM-CRF 模型；序列标注

中图分类号：TP391　　　　**文献标识码：**A

Negation Focus Identification via Bi-directional LSTM-CRF Model

SHEN Longxiang , ZOU Bowei , YE Jing , ZHOU Guodong and ZHU Qiaoming

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Negative expressions are common phenomena in natural language text and play a critical role in various applications of natural language processing, such as sentiment analysis, information extraction. Negation focus identification task is a finer-grained negative semantic analysis, which aims at identifying the text fragment modified and emphasized by a negative keyword. We regard the negation focus identification as a sequence labeling task, and propose a bidirectional Long Short-Term Memory network with a Conditional Random Field layer (BiLSTM -CRF). Such model can not only learn the contextual information bi-directionally and capture the global features by the BiLSTM network, but also learn the dependency between the output tags by the CRF layer. Experimental results on the *SEM2012 dataset shows that the performance of our approach achieves the accuracy of 69.58% with the improvement of 2.44%, compared to the state-of-the-art system for negation focus identification.

Key words: Negation focus; BiLSTM-CRF model; Sequence labeling

1　引言

否定语义在自然语言中十分普遍，通常用于表示人们对某一观点的态度。否定表述通常包含一个否定运算符（例如“不”、“没有”），该运算符对表述本身或其某一方面的语义进行了反转。否定聚焦点是指在否定表述中，最显著被否定的文本片段，它是否定词特别强调的语义，从更细粒度上对文本中的否定含义进行界定。在相同句子中，根据描述者所强调的否定语义之间的差别，其对应的否定聚焦点也不相同。如下例所示：

基金项目：国家自然科学基金(61703293,61672367)；江苏省科技计划(BK20151222)

作者简介：沈龙骧（1995——），男，硕士研究生，主要研究方向：信息抽取；邹博伟（1984——），男，博士，主要研究方向：信息抽取、篇章分析；叶静（1994——），女，硕士研究生，主要研究方向：信息抽取；周国栋（1967——），男，教授，博士生导师，主要研究方向：自然语言处理、多语言跨文本信息抽取；朱巧明（1963——），男，教授，博士生导师，主要研究方向：中文信息处理、Web 信息处理。

例 1 牛**不**会用叉子吃草。

在例 1 中，根据不同解释，否定词“不”的聚焦点可能对应以下三种情形之一[1]：

- 否定聚焦点 1：牛
  解释：<u>牛</u>**不**用叉子吃草，但其它动物可以。
- 否定聚焦点 2：草
  解释：牛**不**用叉子吃<u>草</u>，但是吃其它事物。
- 否定聚焦点 3：用叉子
  解释：牛吃草，但是**不**<u>用叉子</u>。

前两种情形不符合常识及人们在使用语言时的习惯，因此例 1 中否定词“不”的聚焦点应该为“用叉子”。根据以上分析可以看出，否定聚焦点识别不仅要考虑否定结构的句法特点，更重要的是其语义表示，甚至是常识及语境。Blanco 和 Moldovan 在否定聚焦点语料的标注工作中指出，其人工标注一致性仅为 0.72[1]，这从另一个角度说明了否定聚焦点识别任务的难度。

现有的否定聚焦点识别方法主要集中于基于规则的方法[2]和基于特征工程的方法[1]，而这些传统方法大多依赖于领域专家进行模版或特征设计，需要耗费很多的人力和时间代价。与传统方法相比较，深度学习技术可以自动地学习特征，最小化特征工程的代价。基于深度学习的方法在自然语言处理的各个任务中已经被证明是有效的，如机器翻译[3-4]、情感分析[5-6]、关系抽取[7-8]等。近些年来，循环神经网络（Recurrent Neural Network，RNN）[9]及其变体长短期记忆网络（Long Short-Term Memory，LSTM）[10-11]和门控循环网络（Gated Recurrent Unit，GRU）[12]在序列化数据建模方面取得了较大成功。

本文将否定聚焦点识别作为序列标注任务，采用双向长短期记忆网络（Bidirectional LSTM, BiLSTM）学习否定词上下文中前向和后向的远距离特征，同时，在该网络输出层后增加条件随机场（Conditional Random Field，CRF）结构，学习输出标签之间的前后依赖关系。本文首先将句子中的词进行向量化表示，同时将每个词对应的相关特征（词性、位置、句法信息、语义角色等）向量化，并进行组合，将组合后的向量送入 BiLSTM 网络中进行训练，最后通过 CRF 层解码出全局最优标注序列。

本文提出的基于 BiLSTM-CRF 模型的否定聚焦点识别方法在*SEM2012 数据集上进行测试，准确率达到 69.58%，取得了目前最好的性能。此外，相关实验验证了语义角色信息对否定聚焦点识别的有效性。

本文组织结构如下：第二节介绍否定聚焦点识别的相关研究及 BiLSTM-CRF 模型的相关工作；第三节详细描述本文提出的基于 BiLSTM-CRF 模型的否定聚焦点识别方法；第四节介绍实验设置，并对实验结果进行分析；第五节给出本文结论。

2 相关研究

本节分别介绍否定聚焦点识别研究的进展，以及 BiLSTM-CRF 模型在自然语言处理研究中的相关工作。

2.1 否定聚焦点识别

否定聚焦点识别任务由德克萨斯大学的 Blanco 和 Moldovan 于 2011 年首次提出[1]，他们从语义角度对否定聚焦点的概念进行了定义和描述，并基于 PropBank 语料库[13]标注了否定聚焦点数据集。同时，提出决策树模型对否定聚焦点进行识别。然而，该方法采用了 22 类复杂繁琐的词法和句法特征，对特征工程依赖严重，需要大量人工参与和领域知识。

目前，针对否定聚焦点识别任务的研究相对匮乏，一方面原因是该任务本身难度较大（人工标注一致性仅为 0.72[1]），另一方面，还未有充足的否定聚焦点识别语料供现有模型进行训练（Blanco 标注的数据集规模为 3993 句）。*SEM2012 评测任务将否定聚焦点识别作为其子任务之一[14]。Rosenberg 和 Bergler 采用基于启发式规则的方法来识别否定聚焦点[2]，该方法不仅需要语言专家参与制定模版，在领域适应性方面也存在一定限制。Zou 等人利用上下文特征，提出基于“词–主题”结构的双层图模型[15]。该方法首先需要借助海量文本建立主题模型，同

[1] 本文用粗体表示否定运算符，用下划线表示否定聚焦点。

时在训练过程学习大量参数，而调参方法大多基于个人经验，导致该方法扩展性较差。

不同于以往的传统模型，本文基于双向长短期记忆网络与条件随机场模型来识别否定词聚焦点。该模型能够充分利用上下文信息，并有效捕获相邻词的潜在依赖关系；此外，也摆脱了对特征工程以及基于经验的大量参数学习的依赖，而由神经网络自动学习参数及特征表示。据我们所知，本文首次尝试采用神经网络模型解决否定聚焦点识别问题。

2.2 BiLSTM-CRF 模型

近年来，深度学习在自然语言处理的各个任务中均取得突破性进展。其中，循环神经网络（RNN）作为一类典型的序列标注网络，最早由 Goller 和 Kuchler 在 1996 年提出[9]；而由于 RNN 受限于梯度消失和梯度爆炸问题[16-17]，Hochreiter 和 Schmidhuber 提出了 RNN 的变体长短期记忆网络（LSTM）[10]；之后，由于 LSTM 只能获取单方向的上下文信息，Graves 等人提出了双向 LSTM（BiLSTM）并将其应用于语音识别[18-19]，该模型可以在特定时间范围内有效利用过去和未来的特征。另一方面，条件随机场（CRF）是由 Lafferty 等人于 2001 年提出[20]，近些年其在自然语言处理领域中得到了广泛应用。在序列标注任务中，CRF 可以对输出的相邻标签之间的前后依赖关系加以考虑。

基于以上原因，一些工作尝试将 BiLSTM 与 CRF 连接起来对序列化数据进行建模。Huang 等人首次将 BiLSTM 与 CRF 的混合模型用于NLP的序列标注任务上[21]；Ma 等人将 BiLSTM、CRF、CNN 三种模型进行融合并应用于端到端的序列标注任务中[22]；Lample 等人将 BiLSTM-CRF 模型用于命名实体识别任务中[23]。该模型在序列标注任务上的有效性逐渐得到证实。

3 基于 BiLSTM-CRF 的否定聚焦点识别

本节首先介绍 BiLSTM-CRF 模型，然后给出基于该模型的否定聚焦点识别方法的细节。

3.1 BiLSTM-CRF 模型

LSTM 单元 循环神经网络（RNN）适合为序列化数据建模，该模型利用前一时刻的隐藏状态和当前输入决定最终的输出结果。然而，在实际应用中，RNN 受限于梯度消失和梯度爆炸问题[16-17]，为解决该问题，Hochreiter 和 Schmidhuber 提出了一个 RNN 的变体，LSTM 网络[10]。

图 1 给出了 LSTM 记忆单元的结构，其由输入门（input gate）、输出门（output gate）、遗忘门（forget gate）和一个细胞状态（cell）组成，它们控制着当前信息以一定的比例传递到下一时刻，或者舍弃。因此，LSTM 能够有效利用长距离依赖关系，并消除冗余的上下文信息。

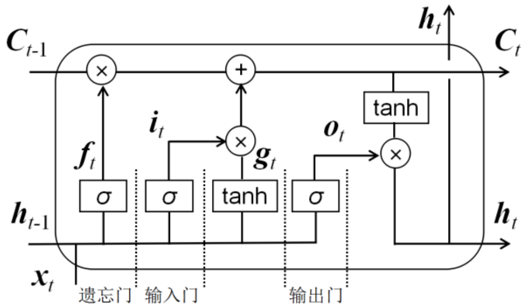

图 1 LSTM 记忆单元结构

从图 1 中可以看出，输入门控制着输入新信息按比例保存到细胞状态中，遗忘门控制着细胞状态所保留的历史信息，输出门决定了最终的输出信息，t 时刻的一个 LSTM 单元的更新公式如下：

$$
\begin{aligned}
&\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{ii}\boldsymbol{x}_t+\boldsymbol{b}_{ii}+\boldsymbol{W}_{hi}\boldsymbol{h}_{(t-1)}+\boldsymbol{b}_{hi})\\
&\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{if}\boldsymbol{x}_t+\boldsymbol{b}_{if}+\boldsymbol{W}_{hf}\boldsymbol{h}_{(t-1)}+\boldsymbol{b}_{hf})\\
&\boldsymbol{g}_t=\tanh(\boldsymbol{W}_{ig}\boldsymbol{x}_t+\boldsymbol{b}_{ig}+\boldsymbol{W}_{hg}\boldsymbol{h}_{(t-1)}+\boldsymbol{b}_{hg})\\
&\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{io}\boldsymbol{x}_t+\boldsymbol{b}_{io}+\boldsymbol{W}_{ho}\boldsymbol{h}_{(t-1)}+\boldsymbol{b}_{ho})\\
&\boldsymbol{C}_t=\boldsymbol{f}_t\otimes\boldsymbol{C}_{(t-1)}+\boldsymbol{i}_t\otimes\boldsymbol{g}_t\\
&\boldsymbol{h}_t=\boldsymbol{o}_t\otimes\tanh(\boldsymbol{C}_t)
\end{aligned}
\tag{1}
$$

其中，$\otimes$ 表示元素级乘法运算，$\boldsymbol{h}_t$ 表示 t 时刻的隐藏状态，$\boldsymbol{C}_t$ 表示 t 时刻的细胞状态，$\boldsymbol{h}_{(t\text{-}1)}$表示 t-1 时刻隐藏状态，$\boldsymbol{i}_t$、$\boldsymbol{f}_t$、$\boldsymbol{o}_t$ 分别表示 t 时刻的输入门、遗忘门和输出门，σ 表示 sigmoid 激活函数，$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 表示相对应的权重矩阵和偏置向量。

双向 LSTM 在 LSTM 中仅考虑了单一方向

的上下文信息，却忽略了另一个方向。一个有效的解决方案是双向 LSTM（BiLSTM），该模型采用两个相反方向的并行层——前向层和后向层，分别从序列的始端和末端开始运行，因此，可以捕获正向与反向的上下文信息。本文将两个 LSTM 层输出的隐藏状态进行拼接作为 BiLSTM 网络的输出。

CRF 层 在序列标注中，一个词的标签通常与其周围词的标签存在关联。因此，在序列标注任务中，对给定句子，一种有效的方法是将句子中当前词与相邻词的标签的关系考虑在内，然后解码出全局最优的标签序列。基于此，本文在 BiLSTM 网络输出层后增加了一层条件随机场（CRF）结构。形式地，给定句子：

$$S=(x_1,x_2,x_3,...,x_n)$$

其预测标签序列为：

$$y=(y_1,y_2,y_3,...,y_n)$$

定义其得分为：

$$G(S,y)=\sum_{i=0}^{n}\boldsymbol{T}_{y_i,y_{i+1}}+\sum_{i=1}^{n}\boldsymbol{E}_{i,y_i} \tag{2}$$

其中，$\boldsymbol{T}$ 表示转移得分矩阵，$\boldsymbol{T}_{i,j}$ 表示从标签 i 到标签 j 的转移得分，y_0 与 y_{n+1} 是新增的句子起始标签和终止标签，$\boldsymbol{T}$ 的维度为 $(k+2)\times(k+2)$；$\boldsymbol{E}$ 是 BiLSTM 的输出得分矩阵，其维度为 $n\times k$，其中 k 为不同标签的数量，$\boldsymbol{E}_{i,j}$ 表示句子中第 i 个词的第 j 个标签的得分。在预测句子所有可能的标签序列时，采用柔性最大值（softmax）对结果进行归一化：

$$p(y|S)=\frac{e^{G(S,y)}}{\sum_{\bar{y}\in Y_S}e^{G(S,\bar{y})}} \tag{3}$$

在训练过程中，本文最大化正确标签序列的对数概率：

$$\begin{aligned}\log(p(y|S))&=G(S,y)-\log(\sum_{\bar{y}\in Ys}e^{G(S,\bar{y})})\\&=G(S,y)-\log\underset{\bar{y}\in Ys}{add}\,G(S,\bar{y})\end{aligned} \tag{4}$$

其中，Y_S 表示句子 S 所有可能的标签序列。从公式（4）可以看出，该模型生成概率最大的标签序列。解码时，获取最高得分的标签序列作为最终预测的输出序列：

$$\hat{y}=\arg\max_{\bar{y}\in Y_S}G(S,\bar{y}) \tag{5}$$

BiLSTM-CRF 模型 图 2 给出了 BiLSTM-CRF 模型框架。首先，将句子中的词与其特征进行向量化；其次，将特征向量送入 BiLSTM 模型从前向和后向两个方向学习上下文特征；然后，将 BiLSTM 的输出结果作为 CRF 层的输入；最终，由 CRF 层预测全局最优的标签序列。此外，为减小过拟合的影响，我们在 BiLSTM 模型两端各添加了一个 dropout 层。

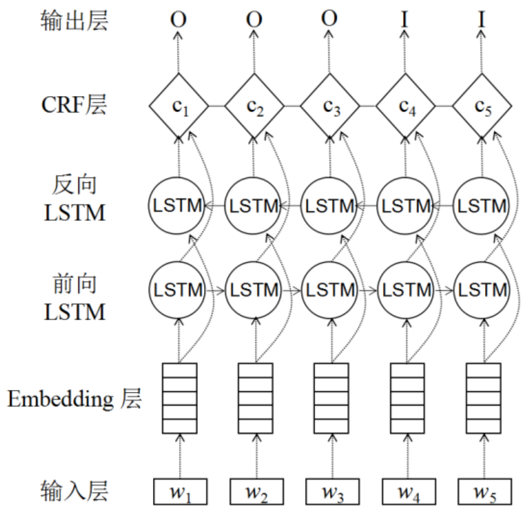

图 2 BiLSTM-CRF 模型框架

3.2 基于 BiLSTM-CRF 的否定聚焦点识别

标记方案 本文标注集合定义如下：

- 标记 **I**：句子中的词位于否定词对应的聚焦点内；
- 标记 **O**：句子中的词不属于否定聚焦点。

以*SEM2012 数据集中的一个标注实例为例，图 3 给出了每个词对应的标记。例如，该句子的否定聚焦点为 *most Americans*，因此这两个词对应的标签为 **I**，而句子中其余词对应的标签为 **O**。

Embedding 层 作为模型的输入，本文构建 embedding 层对词及对应的特征向量进行编码。对给定句子 $S=(w_1,w_2,w_3,...w_n)$，首先使用向量矩阵 $\boldsymbol{W_E}$ 将每个词转换成一个维度为 d_w 的实值向量，其中 $\boldsymbol{W_E}\in \mathrm{P}^{d_w\times|V|}$，$V$ 表示词表。

在自然语言处理领域的各任务中，相关研究探索了各种词法、句法、语义等特征

词	...	*believe*	*most*	*Americans*	*wo*	***n't***	*make*	*the*	*convenience*	*trade-off*	...
词性		VBP	RBS	NNPS	MD	RB	VB	DT	NN	NN	
相对位置		-5	-4	-3	-2	-1	0	1	2	3	
语块标记		B-VP	B-NP	I-NP	B-VP	I-VP	I-VP	B-NP	I-NP	I-NP	
依存句法节点		root	amod	nsubj	aux	neg	ccomp	det	nn	dobj	
语义角色		*	A0	A0	MOD	NEG	V	A1	A1	A1	
标签		O	I	I	O	O	O	O	O	O	

图 3 标记规则与特征表示

[24-26]。为比较各类特征在本文模型上的有效性，我们将词性、相对位置、句法、语义角色等特征加入模型。图 3 给出了各类型特征的示例，其向量化表示如下。

词性：向量矩阵 $\boldsymbol{M}_{\boldsymbol{E}_1}$ 将每个词的词性映射为一个维度为 d_{pos} 的实值向量，其中 $\boldsymbol{M}_{\boldsymbol{E}_1} \in P^{d_{pos}\times|V_{pos}|}$，$V_{pos}$ 表示词性的集合，采用随机初始化；

相对位置：向量矩阵 $\boldsymbol{M}_{\boldsymbol{E}_2}$ 将每个词和动词触发词之间的相对距离映射为一个维度为 d_{loc} 的实值向量，其中 $\boldsymbol{M}_{\boldsymbol{E}_2} \in P^{d_{loc}\times|V_{loc}|}$，$V_{loc}$ 表示相对距离的集合，采用随机初始化；

语块标签：向量矩阵 $\boldsymbol{M}_{\boldsymbol{E}_3}$ 将每个词在成分句法树中的语块标签映射为一个维度为 d_{con} 的实值向量，其中 $\boldsymbol{M}_{\boldsymbol{E}_3} \in P^{d_{con}\times|V_{con}|}$，$V_{con}$ 表示语块标签的集合，采用随机初始化；

依存句法节点：向量矩阵 $\boldsymbol{M}_{\boldsymbol{E}_4}$ 将每个词在依存句法树中的父节点映射为一个维度为 d_{dep} 的实值向量，其中 $\boldsymbol{M}_{\boldsymbol{E}_4} \in P^{d_{dep}\times|V_{dep}|}$，$V_{dep}$ 表示依存句法节点的集合，采用随机初始化；

语义角色：向量矩阵 $\boldsymbol{M}_{\boldsymbol{E}_5}$ 将每个词在句子中的语义角色映射为一个维度为 d_{sr} 的实值向量，其中 $\boldsymbol{M}_{\boldsymbol{E}_5} \in P^{d_{sr}\times|V_{sr}|}$，$V_{sr}$ 表示语义角色的集合，采用随机初始化；

4 实验

4.1 实验设置

本文的实验数据采用*SEM2012 评测任务数据集，其基于 PropBank 语料库[2]进行标注，共包含 3544 个否定聚焦点的实例，其中，2302 个实例作为训练集，530 个实例作为开发集，712 个实例作为测试集。*SEM2012 数据集中不仅人工标注了否定聚焦点，还给出了词性、命名实体、语块、成分句法、依存句法、语义角色等信息。表 1 给出了该数据集中训练集、开发集、测试集的实例数的统计，以及否定聚焦点对应的语义角色类型的统计数据。

表 1 *SEM2012 数据集中否定聚焦点语义角色类型统计（实例数）

语义角色	训练集	开发集	测试集
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
C-A1	46	6	16
AM-PNC	33	8	12
AM-LOC	25	4	10
A4	11	2	5
R-A1	10	2	2
Other	40	8	16
None	88	19	35
总计	2302	530	712

从表 1 中可以看出，否定聚焦点对应名称 A1 和 AM-NEG 两种类型的语义角色的情况较多。在大多数实例中，否定聚焦点只对应单一语义角色，而在一小部分实例中，否定聚焦点对应多个语义角色或不对应语义角色（表 1 中语义角色为“None”的数据）。

本实验采用划分好的 2302 个实例作为训练集，530 个实例作为开发集，712 个实例作为测试集。实验采用预训练好的 Senna 词向量[3]，维度为 50[27]。此外，我们同时尝试了其它不同的向量集，包括 Glove 的 100

[2] PropBank 语料库对谓语动词和 20 多种语义角色进行了标注。

[3] 以维基百科和 Reuters RCV-1 语料库为训练数据，http://ronan.collobert.com/senna/

维词向量[4]以及 Google 预训练好的 300 维新闻语料[5]的词向量[28-29]。

在实验中，我们将特征维度设置为 50，LSTM 隐藏层的维度设置为 150，mini-batch 大小设置为 3，dropout 设置为 0.3。参数更新时采用随机梯度下降（stochastic gradient descent，SGD）算法，其中学习率设置为 0.015，动量（momentum）设置为 0.9。此外，我们还尝试了其它优化算法，包括 Adadelta[30]和 Adam[31]，这些方法虽然使得模型收敛速度加快，但是最终性能均不如 SGD。本文采用准确率（Accuracy，以下简写为 Acc）作为系统性能评价指标，以句子为单位计算，即仅当一个句子中预测的标签序列全部正确时，才被判为正确。

4.2 不同特征对否定聚焦点识别性能影响

表 2 给出了不同模型的性能比较，以及使用各种特征的 BiLSTM-CRF 模型的性能。其中，PoS 表示词性特征，Chunk 表示语块标签特征，Dep 表示依存句法节点特征，RP 表示相对位置特征，SR 表示语义角色特征，ALL 表示以上五种特征的组合。

表 2 不同模型及特征组合的否定聚焦点识别系统性能比较

系统	Acc
LSTM	58.71
BiLSTM	60.85
BiLSTM-CRF	64.10
BiLSTM-CRF+PoS	65.30
BiLSTM-CRF+Chunk	64.71
BiLSTM-CRF+Dep	64.95
BiLSTM-CRF+RP	65.52
BiLSTM-CRF+SR	**69.58**
BiLSTM-CRF+SR+PoS	69.06
BiLSTM-CRF+SR+Chunk	68.88
BiLSTM-CRF+SR+Dep	69.05
BiLSTM-CRF+SR+RP	69.01
BiLSTM-CRF+ALL	69.10

首先，我们比较了不同序列标注网络结构在否定聚焦点识别任务上的性能（表 2：2-4 行）。实验结果表明：1）BiLSTM 模型的准确率比 LSTM 模型高 2.14%，主要原因是 BiLSTM 模型考虑了前向和后向两个方向的信息，比单向的 LSTM 模型能够更加充分地利用上下文特征。2）BiLSTM-CRF 模型的准确率达到 64.10%，比单使用 BiLSTM 模型的性能提升了 3.25%，其原因是否定聚焦点通常由连续文本片段构成，甚至是一个完整的语义角色或句法结构，其中相邻词之间具有较强的依赖关系，仅采用 LSTM 或 BiLSTM 模型无法有效学习此类特征，而增加 CRF 层后，通过对转移概率的训练和学习，我们的否定聚焦点识别方法能够捕捉这些信息。

为验证不同特征的有效性，我们在 BiLSTM-CRF 模型中使用不同类型的特征，并比较其性能（表 2：5-9 行）。结果显示，添加词性、语块标签、依存句法节点和相对位置特征后，系统性能均有微弱提升。单独添加语义角色特征后，系统性能提升了 5.48%。由此可见，语义角色特征对否定聚焦点识别任务较为有效。根据表 1 的语料统计，*SEM2012 数据集中，大多数否定聚焦点对应单一的语义角色，因此，语义角色是该任务的一个重要特征。

为进一步验证以上结论，本文基于添加了语义角色特征的 BiLSTM-CRF 系统，分别加入其它四类特征（表 2：10-14 行）。实验结果表明，分别增加这些特征后，系统性能并没有获得明显提升。这说明，在否定聚焦点识别任务上，语义角色特征很可能包含了以上各类特征提供的信息，其它特征对识别否定聚焦点贡献不明显。

4.3 超参数设置与分析

本文对添加语义角色特征的 BiLSTM-CRF 模型（表 2 中 BiLSTM-CRF+SR 系统）尝试了不同的参数设置，包括语义角色特征的维度、mini-batch 大小、LSTM 隐藏层维度、不同的预训练词向量以及梯度下降算法。在观察某一超参数值对模型性能影响时，其它参数值固定为 4.1 节中给出的值。

语义角色特征维度

表2中验证了语义角色特征对否定聚焦

[4] 以维基百科和网页文本 60 亿个词为训练数据，http://nlp.stanford.edu/projects/glove/

[5] 以谷歌新闻语料 1000 亿个词为训练数据，https://code.google.com/archive/p/word2vec/

点识别的有效性，因此本文尝试采用不同的维度对语义角色特征进行向量化表示。实验结果如图 4 所示。

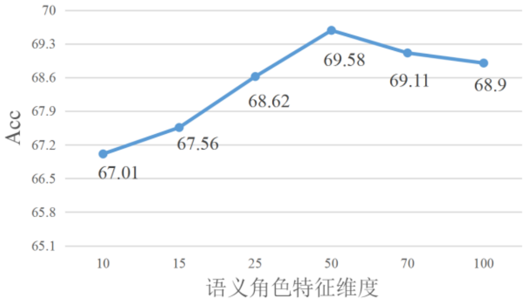

图 4 语义角色特征维度对否定聚焦点识别的影响

可以看出，提升语义角色特征的维度后，系统性能有比较明显的提升，当特征维度为 50 时，系统性能达到最高值 69.58%。然而，当继续增加特征维度时，系统性能开始出现下降。其原因可能是语义角色表征能力随着维度的增加而变强，直到维度超过某个阈值，其表示的信息开始变得稀疏或饱和，表征能力下降。

Mini-batch 大小

考虑到如果仅以单个实例来更新模型参数可能会使实验结果具有偶然性，在随机梯度下降过程中可能会越过全局最小值而仅收敛于局部最小值，我们探索了不同的 mini-batch 大小对模型性能的影响，实验结果如图 5 所示。

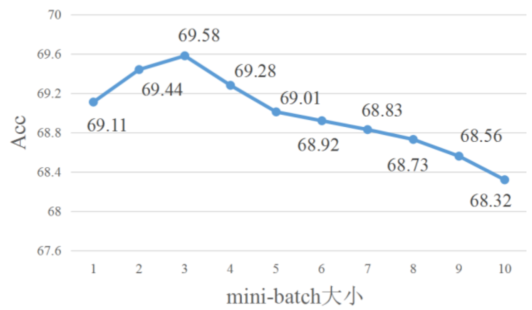

图 5 mini-batch 大小对否定聚焦点识别的影响

可以看出，改变 mini-batch 的大小能够使模型的性能得到进一步提升，当 mini-batch 的大小为 3 时，系统性能达到最高值 69.58%。从图 5 中还可以看出，mini-batch 的值过大时，系统性能下降，可能是由于模型的泛化能力下降所致。

LSTM 隐藏层维度

LSTM 隐藏层维度和输入维度可能有着一定的联系和相互影响：隐藏层维度偏大会使得模型更为复杂，泛化能力下降；隐藏层维度偏小会导致神经网络学习不充分，丢失一些重要特征。因此，本文验证了 LSTM 隐藏层维度对模型性能的影响，实验结果如图 6 所示。结果表明，隐藏层维度为 150 时，系统性能达到最高值 69.58%

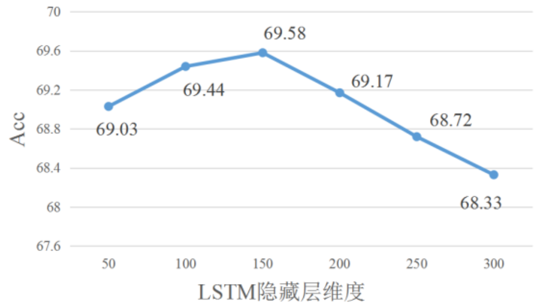

图 6 LSTM 隐藏层维度对否定聚焦点识别的影响

预训练词向量

为验证预训练词向量对模型性能的影响，本文对不同的公开词向量以及随机初始化的词向量进行了对比实验，实验结果如表 3 所示。

表 3 不同预训练词向量对否定聚焦点识别的影响

词向量类型	词向量维度	Acc
Random_uniform	50	64.58
Random_normal	50	66.39
Senna	50	**69.58**
Glove	100	68.50
Word2Vec	300	68.12

其中，Random_uniform 指采用范围在 $[-\sqrt{\frac{3}{dim}},+\sqrt{\frac{3}{dim}}]$ 的均匀分布对词向量值进行随机初始化，dim 是指词向量的维度；Random_normal 指服从 $normal(0,\frac{1}{\sqrt{dim}})$ 的高斯分布。

比较两种随机初始化词向量的方法，采用高斯分布的方法比均匀分布高，但两种方法均低于公开的预训练词向量，这表明预训练词向量在本任务中的重要性。在三种不同的预训练词向量中，Senna 的 50 维词向量获得了最高系统性能，达到 69.58%，使用斯坦福 Glove 的 100 维词向量的性能比 Senna 低 1.08%，而 Google 的 300 维词向量略逊于 Glove，也低于 Senna。

梯度下降算法

采用不同梯度下降算法对模型性能也有一定的影响，本文尝试了不同的优化算法。表 4 给出了各优化算法的性能，以及在 10 折交叉实验中完成迭代的实验平均轮数。

表 4 不同梯度下降算法对否定聚焦点识别的影响

梯度下降算法	Acc	收敛轮数
Adadelta	67.95	15
Adam	68.33	14
RMSprop	68.10	15
SGD	**69.58**	31

实验结果表明，相比 SGD 算法，其它优化方法，包括 Adadelta、Adam 和 RMSprop，均加快了模型的收敛速度，而从系统性能来看，这些算法性能比 SGD 算法的性能略低。

4.4 错误分析

我们选取 BiLSTM-CRF+SR 系统在测试集上的 50 条错误实例进行了分析。主要包含以下几种类型的错误：

否定聚焦点识别错误（27/50）

*SEM2012 数据集在标注聚焦点时，充分考虑了当前句子的上下文信息，即前一句和后一句，而我们的模型仅凭借当前句子，有时很难确定否定聚焦点。如下面句子所示：

*But <u>a majority of the Addison council</u> did **n't** buy those arguments .*

其否定聚焦点为 *a majority of the Addison council*，而仅凭当前句子的含义，很难确定其聚焦点，换一种角度理解，也可能是 *those arguments* 或 *n't*。因此，在未来工作中需要考虑引入上下文信息帮助识别否定聚焦点。

否定聚焦点对应多个语义角色（8/50）

由于模型并未约束否定聚焦点对应单一语义角色，因此如果模型分配给不同语义角色的分值都比较高时，便会造成此类型的错误。

标准答案不符合标注规则（13/50）

我们还发现部分错误实例由标注答案错误所致，而我们系统给出的结果符合标注规则。如下所示：

标注结果：*A panic on Wall Street does <u>**n't**</u> exactly inspire confidence .*

系统结果：*A panic on Wall Street does **n't** exactly inspire <u>confidence</u> .*

否定词 *n't* 聚焦点应为 *confidence*，而语料标注为 *n't* 本身。

此外，*SEM2012 评测任务在数据标注规则[14]中指出：否定聚焦点应为单一且完整的语义角色[6]。而我们发现数据集中标注的否定聚焦点并非严格对应单一且完整的语义角色。因此，我们对测试数据集进行了进一步分析，统计了标注结果与标注规则不一致的句子数目，结果如下：

- 否定聚焦点不对应语义角色：35 句
- 否定聚焦点对应多个语义角色：44 句
- 否定聚焦点对应不完整语义角色：34 句

该类型的实例共 113 个，占测试集的 15.9%，而这部分否定聚焦点识别难度较大。因此，未来工作可尝试修正此类标注不一致问题，同时需从理论层面考虑，是否存在否定聚焦点对应多个语义角色或不完整语义角色。

4.5 与现有方法的性能比较

本文将我们的方法与现有的否定聚焦点识别模型进行了比较，结果如表 5 所示。

表 5 否定聚焦点识别性能比较

系统	Acc
B&M(2011)	65.50
Zou et al.(2014)	67.14
BiGRU-CRF(ours)	68.47
BiLSTM-CRF(ours)	**69.58**

B&M[1]系统使用决策树模型，融合了包括词性、语义角色、句法节点、位置等 22 类特征；Zou 的系统[15]使用基于“词−主题”结构的双层图模型对否定聚焦点进行识别。本文提出的基于 BiLSTM-CRF 的否定聚焦点识别方法，准确率达到 69.58%，比目前的最好系统性能高 2.44%。此外，我们还尝试了另一种序列标注网络，双向门控循环网络（BiGRU）。在该网络上增加 CRF 层之后，其性能也达到了 68.47%。说明本文提出的“RNN 网络+CRF 层”结构能够有效地提升

[6] We only target verbal negations and focus is always the full text of a semantic role.

否定聚焦点识别性能。

5 结论

本文提出了基于 BiLSTM 网络和 CRF 结构相结合的否定聚焦点识别方法，该模型在*SEM2012 数据集上取得了目前最好的性能。以下是本文主要结论：

首先，凭借 BiLSTM 模型在捕获全局信息和长距离依赖关系的优势，有效地利用了上下文信息，使模型的性能得到提升。

其次，考虑到否定聚焦点通常由几个连续的词所构成，为了获取更准确的识别结果，我们将 CRF 融合到 BiLSTM 模型中，使得模型兼具了 CRF 在权衡相邻标签之间的联系与依赖关系的优点，从而预测全局最优的输出标签序列。

最后，通过实验比较了各种特征对否定聚焦点识别性能提升的效果。据我们所知，这是首次将深度学习方法应用于否定聚焦点识别任务，并取得该任务目前的最好性能，因此本方法可以作为基线系统为相关研究提供参考。

本文方法仅仅针对当前句子内容识别否定聚焦点，而正如 4.4 节分析，对部分实例而言，需要根据前后句子的信息判断聚焦点，这也与 Zou[15]所指出的相一致。因此，未来研究考虑将上下文信息引入模型中，以进一步提升否定聚焦点识别的性能。

参考文献

[1] Blanco E., Moldovan D. Semantic Representation of Negation Using Focus Detection[C]//In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL), 2011: 581-589.

[2] Rosenberg S, Bergler S. UConcordia: CLaC negation focus detection at *Sem 2012[C]//Joint Conferece on Lexical and Computational Semantics. Association for Computational Linguistics, 2013: 294-300.

[3] Cho K., Merrienboer B., Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.

[4] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.

[5] Santos C., Gattit M. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts[C]//In Proceedings of the International Conference on Computational Linguistics. 2014.

[6] Wang J., Yu L., Lai K., et al. Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model[C]//In Proceedings of the Meeting of the Association for Computational Linguistics. 2016: 225-230.

[7] Zeng D, Liu K, Chen Y, et al. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks[C]//In Proceedings of Conference on Empirical Methods in Natural Language Processing. 2015:1753-1762.

[8] Lin Y., Shen S., Liu Z., et al. Neural Relation Extraction with Selective Attention over Instances[C]//In Proceedings of the Meeting of the Association for Computational Linguistics. 2016:2124-2133.

[9] Goller C., Kuchler A. Learning Task-Dependent Distributed Representations by Backpropagation Through Structure[C]//IEEE International Conference on Neural Networks, 1996: 347–352.

[10] Hochreiter S., Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[11] Gers F., Schmidhuber J., Cummins F. Learning to Forget: Continual Prediction with LSTM. Neural Computation 12(10): 2451-2471[J]. Neural Computation, 2000, 12(10):2451-2471.

[12] Cho K., Merrienboer B., Bahdanau D., et al. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches[C]//In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014: 103-111

[13] Palmer M., Gildea D., Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles[J]. Computational Linguistics, 2005, 31(1):71-106.

[14] Morante R., Blanco E. *SEM 2012 Shared Task: Resolving the Scope and Focus of Negation[C]//First Joint Conference on Lexical and Computational Semantics (*SEM), 2012: 265-274.

[15] Zou B., Zhu Q., Zhou G. Negation Focus Identification with Contextual Discourse Information[C]//In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics(ACL), 2014: 522-530.

[16] Bengio Y., Simard P., Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 2002, 5(2):157-166.

[17] Pascanu R., Mikolov T., Bengio Y. On the difficulty of training recurrent neural networks[C]//In Proceedings of the International Conference on International Conference on Machine Learning. JMLR.org, 2013:III-1310.

[18] Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Netw, 2005, 18(5):602-610.

[19] Graves A, Mohamed A., Hinton G. Speech recognition with deep recurrent neural networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 2013: 6645-6649.

[20] Lafferty J., Mccallum A., Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//In Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2001: 282-289.

[21] Huang Z., Xu W., Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.

[22] Ma X., Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C]//In Proceedings of the Meeting of the Association for Computational Linguistics. 2016: 1064-1074.

[23] Lample G., Ballesteros M., Subramanian S., et al. Neural Architectures for Named Entity Recognition[C]//In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL-HLT). 2016: 260-270.

[24] Poon H., Domingos P. Unsupervised semantic parsing[C]//In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2009:1-10.

[25] Pradhan S., Ward W., Hacioglu K., et al. Shallow Semantic Parsing using Support Vector Machines[C]// North American Chapter of the Association for Computational Linguistics. 2003:233-240.

[26] Soricut R., Marcu D. Sentence level discourse parsing using syntactic and lexical information[C]//In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. 2003:149-156.

[27] Collobert R., Weston J., Bottou L., et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12:2493–2537.

[28] Pennington J., Socher R., Manning C. Glove: Global Vectors for Word Representation[C]//In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.

[29] Mikolov T., Sutskever I., Chen K., et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[30] Zeiler M. ADADELTA: An Adaptive Learning Rate Method[J]. Computer Science, 2012.

[31] Kingma D., Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.