

# 汉语否定焦点识别研究：数据集与基线系统

盛佳璇<sup>1</sup>, 邹博伟<sup>1,2\*</sup>, 沈龙骧<sup>1</sup>, 叶静<sup>1</sup>, 洪宇<sup>1</sup>

苏州大学计算机科学与技术学院, 苏州, 215000<sup>1</sup>

新加坡资讯通信研究院, 新加坡, 138632<sup>2</sup>

shengjiaxuan1996@gmail.com, zou\_bowei@i2r.a-star.edu.sg,

{lxshen.scu,jye.scu,tianxianer}@gmail.com

## 摘要

自然语言文本中存在大量否定语义表达, 否定焦点识别任务作为更细粒度的否定语义分析, 近年来开始受到自然语言处理学者的关注。该任务旨在识别句子中被否定词修饰和强调的文本片段, 其对自然语言处理的下游任务, 如情感分析、观点挖掘等具有重要意义。与英语相比, 目前面向汉语的否定焦点识别研究开展缓慢, 其主要原因是尚未有中文数据集为模型提供训练和测试数据。为解决上述问题, 本文在汉语否定与不确定语料库上进行了否定焦点的标注工作, 初步探索了否定焦点在汉语上的语言现象, 并构建了一个包含5,762个样本的数据集。同时, 本文还提出了一个基于神经网络模型的基线系统, 为后续相关研究提供参照。

**关键词:** 否定焦点; 数据集; 人工标注

## Research on Chinese Negative Focus Identification: Dataset and Baseline

Jiaxuan Sheng<sup>1</sup>, Bowei Zou<sup>1,2\*</sup>, Longxiang Shen<sup>1</sup>, Jing Ye<sup>1</sup>, Yu Hong<sup>1</sup>

School of Computer Science and Technology, Soochow University, Suzhou, 215000<sup>1</sup>

Institute for Infocomm Research, Singapore, 138632<sup>2</sup>

shengjiaxuan1996@gmail.com, zou\_bowei@i2r.a-star.edu.sg,

{lxshen.scu,jye.scu,tianxianer}@gmail.com

## Abstract

There are a large number of negative expressions in natural language texts. As a more fine-grained negative semantic analysis task, negative focus identification has begun to attract the attention of natural language processing (NLP) researchers in recent years. The task aims to identify the text fragments modified and emphasized by negative cues in the sentence, and it is of great significance to the downstream tasks of NLP, such as sentiment analysis and opinion mining. Compared with English, the study on negative focus identification for Chinese is currently slow, the main reason is that there is no Chinese dataset to provide training and test data for the models. To solve the above issue, this paper carried out the manual annotation of negative focuses on the Chinese Negative and Speculation corpus (CNeSp), initially explored the language phenomena of negative focus on Chinese, and constructed a dataset containing 5,762 samples. Besides, we also come up with a baseline system based on neural network model to provide a reference for subsequent studies.

**Keywords:** Negative focus, Dataset, Manual annotation

\* 通讯作者

## 1 引言

自然语言文本中存在大量包含否定语义的表达，通常由表示否定语义的词(如“不”、“没有”等)对断言或其某一方面的含义进行反转，这类词被定义为否定线索词(Negative cue)，而被否定线索词强调的文本片段为否定焦点(Negative focus)。更确切地说，否定焦点指表述中最显著被否定的文本片段，从更细粒度上对文本中的肯定含义与否定含义进行了区分 (Blanco and Moldovan, 2011)。在例1中，否定断言为“酒店不提供24小时热水”，并包含否定线索词“不”，而从下文中“下午5点以后才有”来看，该断言中被否定的含义是“提供热水的时限”，而并非“提供热水”本身，因此，根据定义，其对应的否定焦点为“24小时”。需要注意的是，该断言隐含“酒店提供热水”的肯定含义。

**例1** 酒店不提供24小时热水，问了前台说要下午5点以后才有。

否定焦点的另一个特点是：同一个否定断言，在不同的上下文语境中，其否定焦点可能不相同。例如，在例2的三个句子中，根据不同的上下文，否定线索词所强调的语义，即否定焦点，发生了变化。

**例2.1** 酒店不提供24小时热水，但出门左边有浴池提供。<sup>1</sup>

**例2.2** 酒店不提供24小时热水，问了前台说要下午5点以后才有。

**例2.3** 酒店不提供24小时热水，仅能保证冷水供应。

否定焦点这一语言现象最早由Rodney和Pullum提出 (2002)，而Blanco和Moldovan (2011)在自然语言处理领域首次将否定焦点的自动识别作为任务提出，并基于PropBank (Palmer et al., 2005)语料库标注了否定焦点数据集。借助该数据集，\*SEM'2012 (Morante and Blanco, 2012)将否定焦点识别作为其评测任务之一，之后该任务开始受到自然语言处理领域学者的关注 (Zou et al., 2014; Zou et al., 2015b; Shen et al., 2019)。作为细粒度的否定语义解析任务，否定焦点的自动识别对用户意图识别或基于属性的情感分析等下游应用具有重要作用。如例2.1中，用户负面评价的对象是酒店；而在例2.2中，用户对酒店的负面评价则更具体，针对的是服务时限。

目前，否定焦点识别的相关研究均面向英语，尚未有文献针对汉语中相关语言现象进行研究。主要归结为以下两方面原因：首先，语料库建设是开展自然语言处理相关研究的基础，由Blanco和Moldovan (2011)标注的数据集为面向英语的否定焦点识别提供了实验基础，而该研究领域尚未有面向汉语的语料资源。另一方面，由于句子结构和表述方式等诸多差异，使得英语中基于规则或特征的否定焦点识别模型和方法难以直接迁移到汉语中。

为了解决上述问题，本文首先在汉语否定与不确定数据集(Chinese Negation and Speculation dataset, CNeSp) (Zou et al., 2015a)的基础上人工标注并构建了汉语否定焦点识别数据集，该数据集共标注4,039个样本，其规模与英语否定焦点识别数据集相当，包括科技文献、酒店评论、金融新闻三个不同领域的文本，能够客观反映汉语表述中否定性语言现象，为相关研究的开展提供基准。其次，本文提出一个基于BiLSTM-CRF网络的否定焦点识别模型，在本文构建的数据集上的准确率为57.8%，与其在英文数据集上的性能相比，性能下降约为13%，表明了汉语否定焦点识别任务的难度，该模型能够作为基线系统为相关研究提供参照。

本文结构组织如下：第二章介绍否定焦点识别的相关工作；第三章介绍本文构建的否定焦点识别数据集，包括标注方法、数据统计等；第四章介绍本文提出的否定焦点识别基线系统；第五章介绍实验设置，给出实验结果并进行分析；第六章给出本文的结论。

## 2 相关工作

否定焦点识别的相关任务涉及句子的语义研究，属于自然语言处理技术中较难的问题；此外，与句法分析、机器翻译等领域相比，该问题很难标注大规模数据集。因此，针对否定焦点的工作主要集中在利用传统机器学习模型，尝试各种词法、句法和语义等特征。识别句子中否定词及其对应的否定含义，可以划分为否定线索词识别任务和否定焦点识别任务。

否定线索词识别研究主要包含基于词表、基于统计和基于序列标注的方法。基于词表的方法主要依赖于构建和扩充线索词词表算法，在识别线索词时，需严格匹配命中词表或词典中的词项，因此最终的识别性能主要受词表或词典质量的影响。例如，Kilicoglu和Bergler (2008)将

<sup>1</sup>本文中，否定线索词采用粗体表示，否定焦点采用下划线表示。

从生物学领域的专业词表及WordNet中提取的事实性和不确定性概念之间的词法以及语义关系加入词表。基于统计的方法的关键在于如何提取各类有效的句法、词法、语义等特征，同时将它们组合或者筛选。Light等(2004)最先使用此类方法，他们利用词特征，并采用了支持向量机辨别医学论文摘要相关句子内是否具有不确定性信息；之后，Georgescu(2012)利用基于高斯径向基核函数对Light的方法作出改善，并对类别的权重进行调节，以缓解训练数据不平衡的问题。Øvrelid等(2010)将线索词识别任务视为二元分类任务，并利用了句法、词性等特征进行线索词识别。鉴于线索词可能由多个连续的词组成，因此也有相关研究采用序列标注的方法识别线索词。Tang等(2010)利用条件随机场模型的序列标注方法和大规模基于边界模型分类器进行数据训练，利用了命名实体、词性等特征。Zhang等(2010)提出基于标准化特征的最大熵马尔可夫模型的线索词识别方法。Vinczel(2014)搭建了面向匈牙利语的不确定性语料库并把线索词识别作为序列标注任务，引入了语义、语用、词法、句法等特征并提出一个有监督的机器学习方法来识别线索词。总的来说，否定线索词识别任务相对简单，目前该任务的性能达到95%以上的准确率。

否定焦点识别研究最初由Blanco和Moldovan(2011)提出，他们从语义关系角度描述和定义否定焦点，并构建了否定焦点识别数据集，然后提出一种基于决策树的模型对否定焦点进行识别。\*SEM'2012(Morante and Blanco, 2012)将否定焦点识别作为其评测任务之一，然而由于任务难度较大，仅有一家机构提交了实验结果(Rosenberg and Bergler, 2012)，他们根据不同触发词之间词性的差异，提出了基于启发式规则的方法对否定焦点进行识别，其实验性能达到58.40%(F值)。之后，Zou等(2014; 2015b)融合上下文特征，提出“词-主题”双层图模型识别否定焦点。随着深度学习的发展，Shen等(2019)将否定焦点识别作为序列标注问题，提出了一种基于BiLSTM-CRF的否定焦点识别模型，并利用词级别和主题级别注意力机制的方法来更好地捕获句子间的上下文信息，一方面利用基于词级别上下文注意力机制来捕获当前句子中候选否定焦点和上下文句子的关联度，另一方面运用基于主题级别的上下文注意力机制计算当前句子中的每个候选否定焦点和上下文句子在主题上的分布相似性，他们提出的方法在英文否定焦点识别数据集上的性能达到了70.51%(准确率)。

目前，尚未有面向汉语的否定焦点识别研究，其主要原因是缺乏人工标注的数据集，而由于该任务难度较大，无监督的自动标注方法难于实现；另一方面，本文在汉语否定焦点识别数据集的标注过程中发现，该语言现象在英汉两种语言上有一定差别，主要集中在当文本中包含省略现象时，否定焦点可能发生转移。以上原因也说明了汉语否定焦点识别数据集标注的必要性。此外，相对否定线索词识别任务，汉语否定焦点识别研究更具一定挑战性，因此，本文专注于汉语否定焦点识别研究，实验中采用标准否定线索词。

### 3 汉语否定焦点识别数据集

自然语言处理任务的研究通常依赖于相关数据集，面向英语的否定焦点识别研究已经发布了较为成熟的数据集(Blanco and Moldovan, 2011)和评测任务(Morante and Blanco, 2012)。然而汉语否定焦点识别数据集研究较为匮乏，导致了面向汉语的相关研究受到极大限制。针对该问题，本文标注并构建了汉语否定焦点识别语料库。

考虑到不同领域或体裁的语料资源在语言特点上具有一定差别，本文根据以下三个方面选择标注文本：

- 尽可能基于现有的汉语否定数据集进行标注，因为其提供了很多与否定语义相关的标注信息，供标注人员参考，在降低标注任务难度<sup>2</sup>的同时，提高标注效率。
- 不同领域或语境下的文本在语言特点上具有差异性，本文希望通过构建该数据集，能够较全面地涵盖汉语中否定焦点的各种表述方式。
- 标注文本的所属领域中，相同类型的文本数量足够大，以增强该数据集的可扩展性；同时，也为将来能够开展半自动甚至自动标注研究提供可能。

基于此，本文选择现有汉语否定与不确定性语料库(CNeSp)<sup>3</sup>中包含否定线索词的文本作为基础，在其上人工标注否定焦点。目前，该语料库包含：1) 19篇《计算机学报》科技论文，该类型文本在语言表达上相对严谨，歧义性较小，但句子长度一般较长，且句法结构复杂；2)

<sup>2</sup>Blanco和Moldovan标注英文否定焦点识别数据集时，标注一致性仅为0.72。

<sup>3</sup><http://nlp.suda.edu.cn/corpus/CNeSp/>



821篇携程网酒店点评文章，由于大多文本由用户撰写，该类型文本语言表达方式多样化，句式简单，但修辞及省略现象较常见；3) 1,311篇新浪金融板块文章，该文本由专业编辑撰写，语言特性介于前两种类型之间。

### 3.1 标注规范

汉语否定焦点识别数据集标注方法部分参照了Blanco和Moldovan (2011)的标注规范，由标注者根据自己对文本的理解进行标注。由于否定焦点标注涉及较多语言现象，标注难度较大，为此本文为标注者制订了一套标注规范。否定焦点的基本标注概念和准则如下：

- 否定焦点：否定线索词明显强调的句子片段。
- 否定焦点是一段连续文本。

**标注最大化原则** 当片段中存在多个否定解释时，否定焦点应包含尽可能多的句法成分。如例3所示，否定线索词“不如”对应的否定焦点可能为：a) “好的”，解释为“这个星级酒店不如好的招待所，但差一些的招待所可以”，否定词强调“好的”；b) “招待所”，解释为“这个星级酒店只是不如招待所”，否定词强调“招待所”。根据上下文难以判断哪种解释更合理，此时，本文规定将其合并标注为否定焦点。

**例3** 这个星级酒店还不如好的招待所，只有表面功夫，位置有点喧闹...

**上下文最优原则** 否定焦点标注首先依赖上下文中的信息和证据。从例2.1-2.3可以看出，当给定不同的上下文信息时，所强调的部分(否定论元)可能不同，因此在标注句子中否定论元时，上下文信息是标注者首先需要参考的。在例4中，若不考虑下文中的信息，“全天”或“暖气”均可以作为否定焦点，而根据下文中“只有晚上提供”和“小旅社还提供24小时暖气呢”就能够判断作者强调的否定含义是“酒店供应暖气的不是全天”而非“不供应暖气”，据此，标注者将该句子的否定焦点标注为“全天”。

**例4** 酒店居然没有全天的暖气，问前台说只有晚上提供，就算小旅社还提供24小时暖气呢。

**积极意义优先原则** 标注者标注否定焦点时，优先考虑否定表述中是否存在潜在积极意义。如果将“这家酒店”或“酒店”作为否定焦点，则失去了该表述的潜在积极意义，即“下次去郑州会住酒店”，因此，应仅标注“这家”为否定焦点。注意，该原则与标注最大化原则在不同理解下存在冲突，当无法判断时优先采用标注最大化原则。

**例5** 下次去郑州一定不会入住这家酒店。

**动词性否定(Verbal Negation)**中，否定线索词作用在动词上，但通常作者并非否定该动词表示的动作或事件，因此，需根据其具体含义标注否定焦点。例6.1中，“修”这一动作已经发生，否定线索词“没”否定了动作产生的结果，因此，其否定焦点标注为“好”。如果忽略上下文，通常情况标注者会将例6.2中的否定焦点标注为“正常工作”，而从上文来看，作者强调的是只有卫浴设备不能工作，其否定焦点应为“卫浴设备”。因此，标注否定焦点时，首先考虑上下文最优原则。针对例6.3中的转述动词(Reporting Verbs)或引述动词(Introducer)，如“知道”、“认为”、“说”等，否定焦点通常不是动词本身，而应从其转述内容中进行标注。

**例6.1** 门会发出报警，修了几次也没修好。

**例6.2** 其它设施可以使用，只有卫浴设备不能正常工作。

**例6.3** 这是一家相当差的酒店，不知道其他人是如何评价它的。

**副词标注** 汉语中副词的用法多种多样，在早期标注过程中，关于副词是否被标注有较大争议。除了相对明显的由副词本身作为否定焦点(例7.1)，本文对其它情况进行了规定：1) 程度副词具有实际含义时，是否将其标注为否定焦点主要取决于被其修饰的词。如例7.2所示，“太”作为程度副词，说明“近”的程度，因此将“太近”作为一个整体，标注为否定线索词“不”对应的否定焦点。2) 汉语中副词有时在习惯搭配或口语中并不具有实际意义，该情况下，不应将其标注在否定焦点内。如例7.3所示，“太”仅表示委婉语气，而并非修饰“好”，否则其含义为“周围环境不是非常好”，与作者原意“周围环境不好”不一致，因此该情况下，本文认为否定焦点应标注为“好”。

**例7.1** 酒店装修大多集中在下午，对客人的休息没有产生严重影响。

**例7.2** 到公园的路程不算太近。

**例7.3** 周围环境不太好，被噪声干扰到很晚。

### 3.2 标注中的特殊情况

**不存在明显否定焦点** 一般来说，本文规定否定焦点不应包含其对应的否定线索词。而在一些样本中，不存在明显的否定焦点，但确实包含否定语义。例8中，“没有”具有否定含义，但可能是由于评论者情绪较差，导致表达中省略了关键成分，若补全可能为“没有好感”或者“没有好的看法”，而该文本中并没有合适作为否定焦点的片段，因此该情况下，我们将否定线索词直接作为否定焦点。

**例8** 别问我对这家酒店怎么看，没有，感觉很差！

**否定词不表示否定含义** 特殊语境下，否定词并不表示否定含义。如例9.1中，否定词“不是”在反问句中仅具有强调语气的功能；同样，例9.2中，评论者采用双重否定的修辞方式“不是没有”来强调其中的肯定含义。本文认为这些情况中并不存在否定语义，因此也不存在否定焦点。在标注过程中，我们发现科技论文和金融文章文本中，由于表述相对严谨，极少出现反问或多重否定等修辞，而在酒店评论这类口语化程度较高的文本中存在较多。

**例9.1** 本人问服务员怎么没有荤菜肉菜，答复：鸡蛋不是荤菜？

**例9.2** 酒店管理混乱，外来人员都可以进电梯，看来丢东西这种事不是没有可能。

以上涵盖了否定焦点标注主要规则，以及针对标注中出现的特殊情况的处理，由于篇幅原因，更详细的标注规则将随同数据集一起发布。

### 3.3 标注过程

汉语否定焦点识别数据集标注工作由三位在本领域具有两年以上经验的硕士研究生(标注者)完成标注，并邀请一位从事该领域研究多年的计算语言学专家(指导者)参与数据集标注规则的制定、定期讨论标注结果、裁判标注结果等工作。标注过程大致分为以下两个阶段。

- 第一阶段：经过调研现有否定焦点识别工作及英文数据集，由指导者与标注者共同讨论和制定初步标注规范；之后，三位标注者两两分组，针对三个不同领域的文本，各抽取30%的样本进行标注，期间只允许标注者与指导者进行交流，而标注者之间禁止讨论；标注完成后，指导者与标注者针对标注结果不一致的样本进行讨论，对标注规范进行调整和修改，并将修改后的该30%的数据集作为最终标注结果。
- 第二阶段：仍由三位标注者两两分组，对三个领域剩余70%的样本进行标注；该期间同样只允许标注者与指导者进行交流；标注完成后，由另外一位标注者对其他两位标注者不同的标注结果进行重新审核，如果其与其中一位标注者的标注结果相同，则该结果作为最终标注结果，如果三位标注者结果均不同，则提交给指导者进行判断，给出最终标注结果；同时也会对标注规范进行修订。

最终以第二阶段修订后的标准作为数据集标注规范。标注过程中，尽管标注人员遵循统一的标注规范，但由于每个人对上下文语境理解存在差异且受限于标注者语言知识水平，导致数据集的标注结果存在不一致现象。标注一致性是用来评价数据集质量的重要指标，能够验证标注者在同一标注规范下对标注结果的主观性差异程度以及问题本身的难易程度。本文采用Kappa值 (Mchugh, 2012)评价标注的一致性，计算公式如下：

$$Kappa = \frac{P_0 - P_c}{1 - P_c} \quad (1)$$

其中， $P_0$ 和 $P_c$ 分别表示观察一致率和期望率， $Kappa \in [-1, 1]$ 。通常认为Kappa值大于0.6表示标注具有较好的一致性，而Kappa值小于0.4表示标注一致性较差。

本文以否定焦点样本为单位，当两位标注者标注的否定焦点完全相同时，认为该样本的标注结果一致。表1给出了标注第二阶段三个子数据集上各自的Kappa值。可以看出，三个子数据集的标注一致性均在0.65-0.7之间，一方面，表明该语料的标注结果是可靠的，另一方面，也

子数据集	Kappa值
科技论文	0.69
酒店评论	0.65
金融文章	0.68

表 1: 汉语否定焦点识别数据集标注一致性

表明否定焦点识别任务具有较大的挑战性。此外，科技论文子数据集上的标注一致性最高，而酒店评论子数据集的标注一致性最低，这说明在表述相对严谨的文本上，否定焦点比较容易判断，而在口语化表达较多的酒店评论子数据集上，则存在较大的争议，其否定焦点的识别难度也较高。

### 3.4 数据集统计与分析

汉语否定焦点识别数据集包含科技论文、酒店评论、金融文章三个子语料库，共4039句，5762个样本。表2展示了该数据集的相关数据统计。比较三种不同领域的数据集，可以看出，酒店评论数据与另外两种类型的数据相比，其句子长度更短，这是因为该语料来源于携程网1500家酒店的评论页面，造成该语料表达方式较为口语化，写作风格相对自由。此外，从统计数据可以看出，酒店评论语料中否定句子数、否定句子占比、以及否定样本个数也远高于其它两类数据，造成这一现象的原因是酒店评论中包含了大量评论者的负面观点或意见，这也表明正确识别这些评论或意见中的否定焦点对情感分析(Jansen et al., 2009)、观点挖掘(Dundar et al., 2018)等领域具有重要意义。

统计项	科技论文	酒店评论	金融文章
文档数	19	821	1311
句子数	4626	4997	7213
词数	140900	120436	221265
句子平均长度	30.4	24.1	30.6
否定句子数	132	2644	1263
否定句子占比%	2.9	52.9	17.5
否定样本个数	161	3985	1616
否定线索词集合	25	121	200
否定线索词平均长度	1.58	1.47	1.86
否定焦点平均长度	6.25	2.86	4.34

表 2: 汉语否定焦点识别数据集统计

科技论文数据集中，否定句子数/占比和否定线索词集合大小均远低于其它两种类型的数据集，可以看出，在科技论文中，作者在使用否定或负面表述时较为谨慎。此外，科技论文类型的否定焦点平均长度高于其它类型的数据集，也表明了科技论文具有叙述更完整、论证更严谨的语言特点。

金融文章数据集来源于由经济分析师或财经记者撰写的股市评论类文章，其否定线索词的平均长度大于其它类型的数据集，这是由于作者在表达否定观点时，倾向于使用更长的否定词使表述更加精确或完整，例如“不容乐观”、“并不是”、“并没有”等。

## 4 否定焦点识别任务基准模型

否定焦点是一段连续的文本。因此，本文将否定焦点识别作为序列标注任务，采用基于BiLSTM-CRF的神经网络模型作为基准模型。如图1所示，其中，Bi-LSTM网络能够同时有效利用句子的上下文信息并抓取全局特征，而CRF层能够充分学习输出标签序列间的前后依赖关系。本节将按照Embedding层、Bi-LSTM层、CRF层和输出层的方式详细介绍否定焦点识别方法。

**Embedding层** 给定句子 $S = (w_1, w_2, \dots, w_n)$ ， $w_i$ 表示句子 $S$ 中的词，其中 $i \in [1, n]$ 。本文利用向量矩阵 $W_E$ 将 $S$ 中的每个词转换成维度为 $d_w$ 的实值向量，其中 $W_E \in R^{d_w \times |v|}$ ， $|v|$ 为词表大

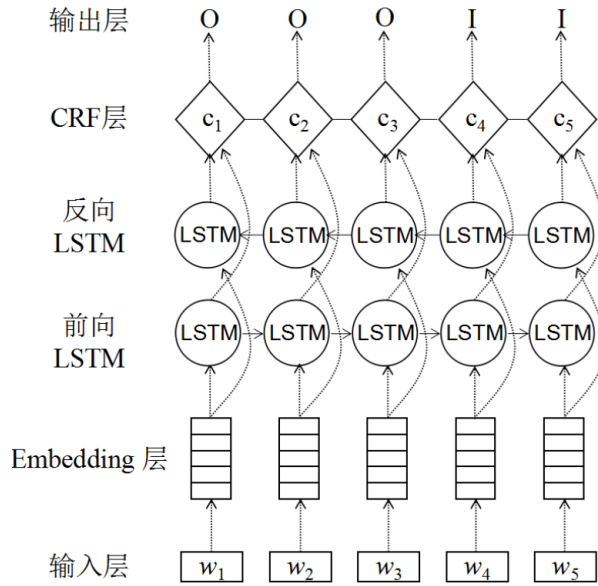


图 1: BiLSTM-CRF模型框架.

小。此外，本文将相对位置特征矩阵与矩阵 $W_E$ 拼接，作为Embedding层中的输入。即：向量矩阵 $F_E \in R^{d_{loc} \times |V_{loc}|}$ ，其中 $d_{loc}$ 为 $S$ 中的每个词与否定线索词间的相对距离映射的一个实值向量， $V_{loc}$ 是相对距离的集合，采用随机初始化。

**Bi-LSTM层** 传统循环神经网络 (Recurrent Neural Networks, RNN) 适合为序列化数据建模，然而，在实际应用中，其受限于梯度消失和梯度爆炸问题 (Bengio et al., 1994; Razvan et al., 2013)，之后为缓解该问题，Hochreiter和Schmidhuber (1997)提出了长短期记忆网络 (Long Short-Term Memory, LSTM)，该网络能够有效利用长距离依赖关系，缓解冗余上下文信息带来的影响。图2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，它们控制着当前时刻信息传递到下一时刻的比例。本文将经过Embedding层编码后的矩阵输入Bi-LSTM层，分别从两个相反的方向并行计算，然后将通过前向LSTM网络的矩阵 $\vec{H}$ 和通过后向LSTM网络的矩阵 $\overleftarrow{H}$ 进行拼接得到矩阵 $H$ ，如公式(2)所示。该矩阵能够有效的捕获了正向和反向的双向上下文信息。

$$H = \vec{H} \oplus \overleftarrow{H} \quad (2)$$

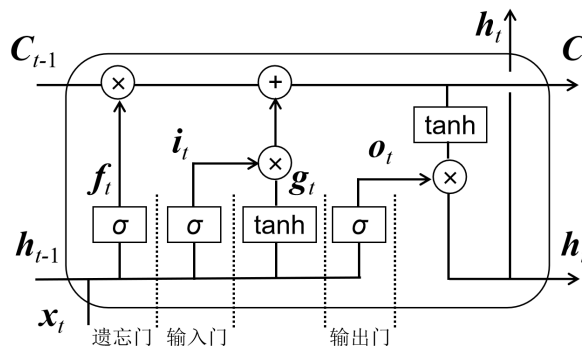


图 2: LSTM记忆单元结构.

**CRF层** 本文将否定焦点识别作为序列标注任务，通常情况下，一个词的标签生成与其周围词存在一定关联。而条件随机场 (Conditional Random Fields, CRF) 能够将句子中当前词与周围词的标签关系考虑在内，从而解码出全局最优标签序列。因此，本文将经过Bi-LSTM层



得到的矩阵 $H$ 输入CRF层解码出最优标签序列。对于给定句子 $S = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ ，其预测标签序列为 $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)$ ，定义其得分为：

$$G(\mathbf{S}, \mathbf{y}) = \sum_{i=0}^n T_{\mathbf{y}_i, \mathbf{y}_{i+1}} + \sum_{i=1}^n E_{i, \mathbf{y}_i}, \quad (3)$$

其中， $\mathbf{T}$ 表示转移得分矩阵， $T_{i,j}$ 表示从标签 $i$ 到标签 $j$ 的转移得分， $\mathbf{y} = 0$ 与 $\mathbf{y} = n$ 是句子起始标签和终止标签， $\mathbf{T}$ 的维度为 $(k+2) \times (k+2)$ ； $\mathbf{E}$ 是Bi-LSTM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， $E_{i,j}$ 表示句子中第 $i$ 个词的第 $j$ 个标签的得分。本文采用softmax对所有可能出现的预测序列得分进行归一化表示：

$$p(\mathbf{y}|\mathbf{S}) = \frac{e^{G(\mathbf{S}, \mathbf{y})}}{\sum_{\tilde{\mathbf{y}} \in Y_S} e^{G(\mathbf{S}, \tilde{\mathbf{y}})}} \quad (4)$$

然后，对正确的否定焦点标签预测序列对数概率进行最大化：

$$\log(p(\mathbf{y}|\mathbf{S})) = G(\mathbf{S}, \mathbf{y}) - \log\left(\sum_{\tilde{\mathbf{y}} \in Y_S} e^{G(\mathbf{S}, \tilde{\mathbf{y}})}\right) = G(\mathbf{S}, \mathbf{y}) - \log \text{add}_{\tilde{\mathbf{y}} \in Y_S} G(\mathbf{S}, \tilde{\mathbf{y}}) \quad (5)$$

$Y_S$ 表示句子 $S$ 所有可能预测出的否定焦点标签序列。解码时，将得分最高的标签序列作为句子 $S$ 最终对应的否定焦点标签序列，如公式(6)所示：

$$\hat{\mathbf{y}} = \arg \max_{\tilde{\mathbf{y}} \in Y_S} G(\mathbf{S}, \tilde{\mathbf{y}}). \quad (6)$$

**输出层** 本文采用I/O标记方案来定义CRF层解码出的标签序列，即：标签I表示句子中的词位于否定焦点内；标签O表示句子中的词不属于否定焦点。如例10所示，该句子的否定焦点为“到位”，因此这个词的标记为I，其余词标记为O。

**例10** 这/O 是/O 服务/O 态度/O 没有/O 做/O 到位/I 。/O

## 5 实验

### 5.1 实验设置

本文实验数据采用第三章介绍的汉语否定焦点识别数据集。实验中，本文采用随机初始化的50维词向量和相对位置特征向量作为Embedding层的输入。模型超参数设置如下：LSTM隐藏层维度为150，dropout设置为0.3，并采用随机梯度下降算法(Stochastic Gradient Descent, SGD)对参数进行更新，其中，学习率为0.015，动量为0.9。本文采用准确率(Accuracy, Acc)作为实验性能评价指标，当一个样本中预测的标签序列全部正确时，则否定焦点正确。

考虑到每个领域的数据集中样本数量较少，本文采用两种不同的训练方式来进行模型的训练，分别为单领域训练方式以及多领域训练方式：

- 单领域训练：对于科技论文、酒店评论和金融文章这三个领域的语料，分别按照70%、10%、20%的比例划分为训练集、开发集和测试集，仅使用单领域的训练集进行训练。
- 多领域训练：将所有单领域的训练集合并，利用合并后的训练集进行训练。

### 5.2 不同系统对否定焦点识别性能影响

表3给出了基于不同神经网络模型的汉语否定焦点识别性能，同时考虑到虽然三个语料来源于不同领域，但是在汉语语言共性上仍然存在一定关联。因此，本文在汉语否定焦点语料上分别尝试了单领域和多领域两种不同的训练方式。同时，本节对不同语料间的差异及不同系统在三个语料上的实验进行了详细分析。

**不同语料之间的对比。**表3中2-4行实验结果可以看出：1) 不同领域之间的否定焦点识别性能存在着一定的差异性，酒店评论语料上的实验性能最好，而科技文献语料上的实验性能最低，造成这一现象的主要原因是酒店评论表达方式较为口语化，其写作风格相对自由，因此句



语料	单领域训练			多领域训练
	LSTM	BiLSTM	BiLSTM-CRF	BiLSTM-CRF
科技论文	24.24	27.27	30.30	<b>39.39</b>
酒店评论	49.25	54.38	59.27	<b>62.53</b>
金融文章	40.43	48.15	57.10	<b>60.18</b>
全部	46.07	51.87	57.84	<b>61.21</b>

表 3: 汉语否定焦点识别模型性能比较

子平均长度以及否定焦点长度更短，模型更容易识别出否定焦点。而科技论文写作方式更为严谨，该语料中句子长度和否定焦点长度在三个语料上最长，导致模型在学习时更容易遗忘长距离信息，所以系统识别性能较低；2) 面向汉语的否定焦点识别难度较大，挑战性比较高，虽然酒店评论和金融文章识别性能相对较好，最好性能达到60%以上，但是整体识别性能均没有超过65%，这也验证了本文对否定焦点识别任务难易程度所作的判断。

**不同模型之间的对比。**为验证不同模型对识别性能的影响（表3中第2列的3个子列），本章对LSTM、BiLSTM和BiLSTM-CRF三个模型进行性能上的比较，从实验结果可以看出，1) BiLSTM模型的性能均比LSTM模型性能高，主要原因是BiLSTM模型考虑了前后两个方向的信息，相较于单向的LSTM模型，它能更充分的利用上下文信息；2) BiLSTM-CRF模型相较于BiLSTM模型，在三个语料上实验性能分别提高了3.03%，4.89%，8.95%，这表明CRF层对于否定焦点识别的有效性，因为否定焦点通常是一段连续的文本，相邻词之间具有较强的依赖关系，而CRF层能够有效捕捉当前词标签与周围词标签之间存在的关联，然后解码出全局最优的标签序列。

同时，本文发现，相较科技论文和酒店评论语料，模型在金融文章上的性能提升尤为显著，其原因可能是：1) 科技论文语料中没有足够的否定样本，所以性能上存在着偶然性和不确定性，且科技论文的否定焦点长度普遍较长（平均长度为6.25），因此识别难度较大，性能提升不明显；2) 酒店评论语料中的否定焦点长度较短（平均长度为2.86），因此识别较为容易，所以仅用LSTM模型即可获得较高的识别正确率，而金融文章的否定焦点长度介于其它两个语料之间（平均长度为4.34），该语料对上下文信息和相邻标签之间联系的依赖程度较酒店评论语料更大，因此引入了BiLSTM和BiLSTM-CRF模型后性能提升更为显著。

**不同训练方式的对比。**多领域训练，即：将来源于三个不同领域的语料训练集统一为一个整体训练集，并利用不同神经网络模型在该训练集上进行了实验。如表3中2-3列，为验证多领域训练方式的有效性，本节将两种不同训练方式的识别性能进行对比，实验结果显示，采用多领域的训练方式后系统识别性能提升显著。这是因为一方面神经网络模型在学习过程中，语料的数量对系统识别性能具有重要影响，而多领域训练方式扩充了数据集，使得模型的鲁棒性得以提升；另一方面，三个不同领域语料之间虽然存在着语义、句法上的差异，但是仍然存在汉语上的语言共性，而神经网络可以有效捕获这部分特征。

## 6 结论

本文在汉语否定与不确定语料库(CNeSp)的基础上，人工标注了首个汉语否定焦点识别数据集，该数据共包含5,762个样本。此外，本文还提出了一个基于BiLSTM-CRF的基准系统。该数据集和基准系统为汉语否定焦点识别的后续研究提供了基础。未来工作中，如何针对汉语语言特点以及如何将英语中相关模型迁移到汉语否定焦点识别模型中，是需要探索的方向。另外，汉语否定焦点识别数据集和基准系统将同论文一起发布。

## 致谢

本文工作得到国家自然科学基金（基金号61703293，61672368，61672367），江苏省高校优势学科建设工程资助项目资助。

## 参考文献

- Yoshua Bengio, Simard Y. Patrice and Frasconi Paolo. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*. 5(2): 157-166.
- 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 518-589. Association for Computational Linguistics.
- Betül Dündar, Diyar Akay, Fatih Emre Boran and Suat Özdemir. 2018. Fuzzy Quantification and Opinion Mining on Qualitative Data using Feature Reduction. *International Journal of Intelligent Systems*. 33(9): 1840-1857.
- Maria Georgescu. 2012. A Hedgehop over a Max-Margin Framework Using Hedge Cues. *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task*.
- Sepp Hochreiter and Schmidhuber Jurgen. 1997. Long short-term memory. *Neural Computation*. 9(8): 1735-1780.
- Bernard J. Jansen, Mimi Zhang, Kate Sobel and Abdur Chowdur. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the Association for Information Science and Technology*. 60(11): 2169-2188.
- Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *Journal of BMC Bioinformatics*. 9(11):S10.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami and Chris Dyer. 2016. Neural architectures for named entity recognition. *North american chapter of the association for computational linguistics*. pages 260-270.
- Marc Light, Xin Ying Qiu and Padmini Srinivasan. 2004. The Language of Bioscience: Facts, Speculations, and Statements in between. *In Proceedings of the HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*. pages 17-24. Association for Computational Linguistics.
- Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1064-1074. Association for Computational Linguistics.
- Marry L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Media*. 22(3):276-282.
- 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 265-274. Association for Computational Linguistics.
- Xiaofeng Mu, Wei Wang and Aiping Xu. 2020. Incorporating token-level dictionary feature into neural model for named entity recognition. *Neurocomputing*, pages 43-50.
- Lilja Ovreliid, Erik Velldal and Stephan Oepen. 2010. Syntactic scope resolution in uncertainty analysis. *In Proceedings of the 23rd International Conference on Computational Linguistics*. pages 1379-1387.
- Martha Palmer, Daniel Gildea and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*. 31(1):71-106.
- Pascanu Razvan, Tomas Mikolov and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *International conference on machine learning*. pages 1310-1318.
- Huddleston Rodney and Geoffrey K. Pullum. 2003. The Cambridge Grammar of the English Language. *Modern Language Review*. 98.3.
- Sabine Rosenberg and Sabine Bergler. 2012. UConcordia: CLaC negation focus detection at \*SEM2012. *In proceedings of the First Joint Conference on Lexical and Computational Semantics (\*SEM)*, pages 294-300. Association for Computational Linguistics.
- Longxiang Shen, Bowei Zou, Yu Hong, Qiaoming Zhu, Guodong Zhou and Ai Ti Aw. 2019. Negative Focus Detection via Contextual Attention Mechanism. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2251-2261. Association for Computational Linguistics.

- Buzhou Tang, Xiaolong Wang, Xuan Wang, Bo Yuan and Shixi Fan 2010. A Cascade Method for Detecting Hedges and their Scope in Natural Language Text. *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning-Shared Task*. pages 13-17.
- Veronika Vincze1 2014. Uncertainty Detection in Hungarian Texts. *In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*. Association for Computational Linguistics.
- Shaodian Zhang, Hai Zhao, Guodong Zhou and Baoliang Lu 2010. Hedge detection and scope finding by sequence labeling with normalized feature selection. *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning-Shared Task*. pages 92-99.
- 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, Guodong Zhou and Qiaoming Zhu. 2016. Research on Chinese negation and speculation: corpus annotation and identification. *Frontiers of Computer Science in China* 10(6): 1039-1051.
- Bowei Zou, Qiaoming Zhu and Guodong Zhou. 2015. Negation and Speculation Identification in Chinese Language. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL)*, pages 656-665. 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.