文章编号：1003-0077（2011）00-0000-00

基于双向 LSTM 和两阶段方法的触发词识别*

何馨宇，李丽双

（大连理工大学 计算机科学与技术学院,辽宁 大连 116023）

摘要: 生物事件抽取是生物文本挖掘领域的一个重要分支，而触发词识别作为事件抽取的重要子过程，已经吸引了众多的关注。现有的触发词识别方法多为浅层的一阶段方法，训练代价较大，且需要丰富的领域知识抽取大量特征，人工成本较高。因此，本文提出了一种基于两阶段和双向 LSTM 神经网络的触发词识别方法。首先，将触发词识别分为识别和分类两个阶段，有效的缓解了训练过程中存在的类不平衡问题。其次，在两个阶段中均采用目前性能较好的双向 LSTM 神经网络来完成二分类任务和多分类任务，避免了浅层机器学习方法抽取人工特征时的代价。此外，利用 PubMed 数据库下载大规模语料训练带有依存关系的词向量，获得了更加丰富的语义信息，从而有效的提高了触发词的识别性能。本文方法在生物事件抽取通用语料 MLEE 上已获得目前最好抽取性能，F 值为 78.46%。

关键词：触发词识别；两阶段方法；双向 LSTM；依存词向量

中图分类号：TP391　　**文献标识码：**A

Trigger Detection Based on Bidirectional LSTM and Two-stage Method

HE Xinyu, LI Lishuang

(School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116023, China)

Abstract: Extracting biomedical events from biomedical literature plays an important role in the field of biomedical text mining, and the trigger detection has attracted much attention as a key sub-process in the biomedical event extraction. The existing trigger detection methods are almost one-stage methods based on shallow machine learning, the cost of training time is large; and rich domain knowledge and amount of artificial features are rather important in the system construction. In this paper, we propose a trigger detection method based on two-stage and Bidirectional Long Short Term Memory (BLSTM), which divides trigger detection into recognition stage and classification stage. Firstly, in the two-stage method, it can relieve the problem of class imbalance effectively. Secondly, we adopt Bidirectional LSTM Neural Networks to finish the binary class and multi-class tasks in the two stages, which can avoid the cost of exact features manually, and the generalization ability will be improved. In addition, to obtain more semantic information, we use the large-scale corpus downloaded from the PubMed database to train the dependency word embeddings, which can effectively improve the recognition performance of trigger detection. On the multi-level event extraction (MLEE) corpus test dataset, our method achieves an F-score of 78.46%, which outperforms the state-of-the-art systems.

Keywords: trigger detection; two-stage method; bidirectional LSTM; dependency word embeddings

---

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金项目(no.61672126)。

作者简介：何馨宇（1983—），女，博士，主要研究领域为自然语言处理。李丽双（1967—），女，教授，博士生导师，本文通讯作者，主要研究领域为自然语言处理、信息抽取与文本挖掘。

1 引言

近年来，网络与信息技术不断发展，生物研究者对医学领域持续关注，生物研究方向的相关文献呈指数级数量增长，这使得相关研究人员从海量的医学文献中快速获取有益的知识变得相当困难，因此生物医学信息抽取技术应运而生。

生物医学领域信息抽取的最终目的是将研究者感兴趣的非结构化数据以结构化的形式表示与呈现，方便研究。生物领域信息抽取经历了从生物医学命名实体识别到二元关系抽取，再到生物医学事件抽取的发展过程。其中，生物事件抽取属于复杂的关系抽取，是为描述更为复杂的、更为详细的分子变化的过程而提出的。一个生物事件由一个触发词和一个或者多个要素组成，如在句子片段“Prevented induction of 1L-10 production by gp41 in monocytes”中，参与事件的蛋白质为 IL-10 和 gp41，包含三个事件，分别为 Gene_expression 事件，Positive_regulation 事件和 Negative_regulation 事件，该片段中存在的事件如图 1 所示，三个事件的事件结构如下：

事件 1 (Type:Gene_expression, Trigger: production, Theme:IL-10)；

事件 2 (Type:Positive_regulation, Trigger:induction,Theme:事件 1, Cause:gp41)；

事件 3 (Negative_regulation, Trigger:Prevented, Theme:事件 2)。

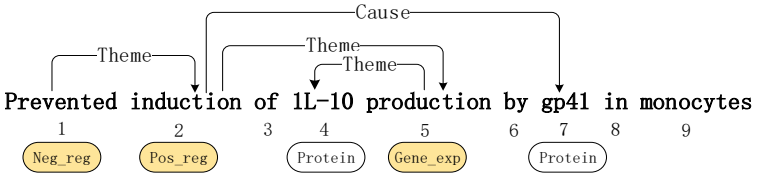


图 1　文本片段生物事件示例图

生物事件抽取方法主要分为两类：分阶段的 pipeline 方法和联合事件抽取。其中分阶段的 pipeline 方法为目前较为主流的事件抽取方法，该方法将事件抽取分为触发词识别、要素识别和后处理三个阶段，即先识别触发词，再根据触发词识别结果进行要素识别，最后通过后处理将触发词和要素构成整个事件。由于触发词识别错误很有可能被传播到要素识别阶段，从而影响整个生物医学事件抽取的性能，所以在分阶段的事件抽取过程中，生物医学事件触发词识别起到了至关重要的作用。研究显示，有超过 60%的抽取错误要归因于触发词识别阶段[5]。因此，本文将触发词识别任务作为研究重点。目前，触发词识别方法大体可以分成两大类：

基于浅层机器学习的方法：这类方法一般将触发词识别视为一个多分类任务。通常需要人工的总结、抽取特征，代价较大，且系统的泛化能力较差。Bjorne[8]等人使用 SVM 作为分类器，抽取了触发词的形态学特征、句子特征、词性、词干特征以及依存链上的信息等，在 BioNLP’09 Shared Task 取得了最好的结果。Pyysalo 等[9]总结了上下文、依存关系等丰富特征，并通过 SVM 进行分类，在 MLEE 语料[9]上的 F 值为 75.84%；Zhou 等[10]使用了半监督的学习模型，通过引入未标注的语料和事件抽取中的隐藏话题来识别触发词，在 MLEE 语料上的 F 值为 76.89%；Zhou[11]等将领域知识中学习到的特征与人工特征进行融合，通过 SVM 进行触发词分类，在 MLEE 语料上的 F 值为 78.32。

基于神经网络和词向量的方法：为了解决生物医学事件触发词提取过程中人工设计特征较为复杂以及缺乏语义信息等问题，基于词向量和神经网络的深度学习方法最近相继被提出。神经网络通常以词向量作为模型的输入，用于获取词与词之间的语义信息；同时，网络模型可以自动地学习一些抽象的特征，避免了机器学习模型人工设计复杂特征带来的问题。Wang[12]等通过词向量得到词之间的句法和语义功能信息，然后将生成的特征向量送到神经网络中进行分类。Nie[13]等将 Skip-gram 模型得到的词向量转化成特征矩阵，用于初始化神经网络的权重，以解决神经网络模型在训练时只得到局部最优解的问题。

上文提及的方法均为一阶段的触发词识别方法。该类方法直接对触发词进行分类，即一次性识别触发词类型和非触发词。一阶段方法训练代价较大，且对于生物医学领域语料中存在的常见问题数据不平衡问题也没有很好的解决。因此，本文提出了一种基于两阶段的触发词识别方法：将触发词识别分为识别和分类两个阶段。第一阶段，仅识别文本中的触发词正例，但不区分这些触发词的类型，分类任务中涉及到的类型仅为正例和负例；第二阶段，仅针对第一阶段识别出的触发词正例进行分类，分类任务中涉及到的类型全部为正例，所以，两个阶段中均对类不平衡有所缓解。此外，两阶段方法可以有效的避免过多的负例对正例分类造成的干扰。同时，在训练时间上，两阶段方法时间也更短，更高效。

深度学习采用词向量作为输入，可以避免由于人工特征向量稀疏而造成的维度灾难问题[14]，并且能够避免浅层机器学习方法中人工总结设计特征费时费力的不足。而深层神经网络能够对原始数据逐层进行表示优化，使得数据表示对分类更有利，从而提升系统性能。因此，本文在两个阶段的不同分类任务中均采用了目前较为流行的长短时记忆递归神经网络(LSTM)。此外，本文利用大规模的生物医学文献训练了一种基于依存关系的词向量，与传统的 Skip-gram 模型相比，基于依存关系的词向量可以获得更加丰富的语义信息，有助于提升触发词的识别性能。而双向 LSTM 和句子向量对于 LSTM 性能的提升也具有一定的作用，是本文触发词识别方法能够取得较好效果的关键因素。

2 方法

图 2 为本文触发词识别的整体框架，主要由两部分构成：数据的向量表示，基于双向 LSTM 和两阶段方法的触发词识别。在数据的向量表示部分，本文先按照预训练的依存词向量，通过查表的方式将原始文本转换为词向量作为输入；此外，在训练的过程中不断对预训练的词向量进行微调，得到微调后的词向量，再通过对两套词向量进行相关运算获得句子级的向量特征信息，对输入进行优化。在基于双向 LSTM 和两阶段方法的触发词识别部分，本文在深度学习框架 Theano 的基础上，分别通过双向 LSTM 神经网络构建了两阶段方法中的二分类和多分类模型对输入数据进行触发词类型预测，相关内容将在后文详细阐述。

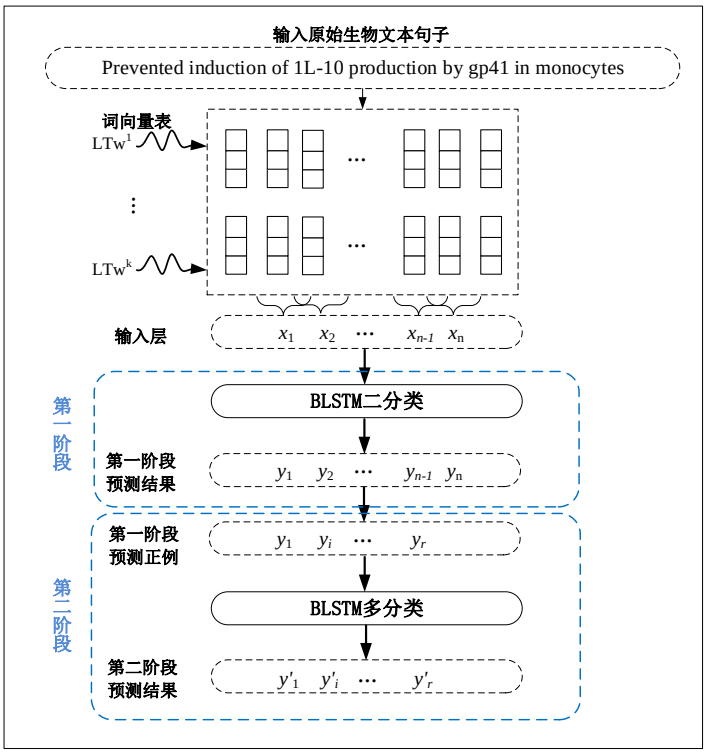


图 2 基于两阶段方法的触发词识别框架

2.1 数据的向量表示

2.1.1 依存词向量

词向量也称为词嵌入或词表达，使用词向量替代传统的 one-hot 方式用于词汇表示，解决了 one-hot 表示带来的维数灾难问题。近年来，随着深度学习在文本挖掘领域的不断发展，词向量也得到了更为广泛的应用。将词向量作为额外特征或者直接作为学习算法的输入，已经对许多文本挖掘系统性能起到了提升作用。

目前较为常用的词向量训练工具是由 Mikolov 等于 2013 年发布的 word2vec[15]，Mikolov 等提供了 Continuous Bag-of-word（CBOW）和 Skip-gram 两种常用的词向量训练模型，分别利用周围词预测目标词和利用目标词预测周围词。然而，与其他利用线性上下文信息来训练词向量的其他自然语言处理任务不同，生物医学触发词识别需要更多来自依存上下文的信息。为此，本文在 word2vec 的传统模型基础上，利用句子中词语的依存关系，通过依存上下文来代替传统 word2vec 模型中线性上下文训练得到依存词向量，使得触发词识别过程获得了更多的语义信息支持。

在本文中，我们从 PubMed 数据库中下载了 5.7G 的摘要内容，对摘要原文进行分句分词处理后，将其送至 Gdep 解析工具得到依存解析结果。最后，利用 word2vecf[16]将得到的依存上下文信息用于训练所需要的依存词向量。Gdep 是一个专门针对生物文本的依存句法分析工具，能够以较高的准确率对生物文本进行句法分析。如下图 3 所示，依存上下文可以捕获使用小窗口线性上下文难以获得的长距离词间的关系，例如，在线性上下文中，当线性窗口大小为 1 和 2 时，“discovers”和“star”、“telescope”的关系很难确定。因此，相较于线性上下文训练得到的词向量（例如 Skip-gram，CBOW 等模型），依存词向量可以获得更多的语义信息，从而提高触发词识别性能。

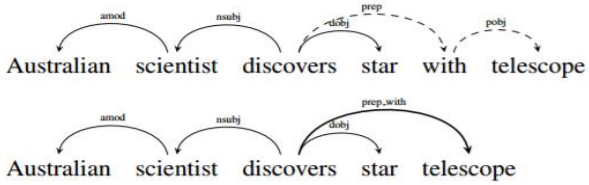


图3 依存关系示例

2.1.2 句子向量

在原始的 LSTM 框架中，所有的输入都是基于词级的向量特征信息，并且需要通过输入门控制其读入到记忆单元中。但是单纯的词级向量容易忽视句子本身潜在的特征信息，而把句子信息作为一种补充输入，有助于在隐层抽象出更加精确的特征表示。因此，为了能够建立起单词与句子之间的潜在关系。本文在 BLSTM 框架中融入句子级的向量特征信息，从而将句子信息通过读入门输入到记忆单元中，获得更加丰富的文本信息。

Li[17]验证了在生物命名实体识别任务上句子向量对于 LSTM 的性能提升具有一定作用。本文采用了类似的句子向量构成方式，即在训练过程中提供两种词向量，一种是预训练的词向量 x_t，另一种是在训练过程中微调后的词向量 x_t'。而句子向量 d_0 为句子中所有单词对应的两种词向量的差值进行加和求平均，如公式(1)所示。预训练的词向量包含了训练语料中无法捕捉到的特征信息，不断微调的词向量包含了更丰富更具有针对性的文本信息，句子向量建立了单词与句子之间的潜在关系。实验证明，句子向量在生物触发词识别任务上也具有一定的提升性能。

$$d_0 = \frac{1}{n}\left(\sum_{t=1}^{T}(x_t' - x_t)\right) \tag{1}$$

2.2 基于双向 LSTM 和两阶段方法的触发词识别

2.2.1 LSTM 神经网络

传统的递归神经网络在有监督的训练过程中的误差传播会随着神经网络递归深度的增加而不断的减小或夸大，这种影响被称之为梯度弥散[18]。为了解决梯度弥散问题，90 年代的研究人员进行了各种各样的尝试，目前 Hochreiter 和 Schmidhuber[19]提出的长短时记忆（Long Short-Term Memory，LSTM）结构是目前最受研究者青睐也是最有效的用来解决递归神经网络梯度弥散问题的方法。LSTM 的基本构成单位是一个记忆存贮块，其主要包括一个记忆单元和三组具有自适应性的元素乘法门，即输入门、忘记门和输出门。这三个门是非线性的求和单元，旨在收集存贮块内外的激活信息，并且通过乘法运算控制记忆单元中的激活值。正是这种有选择的读写上下文信息的优势极大的弥补了梯度弥散的缺陷。

双向 LSTM 神经网络[20]的基本思想是对每一句话分别采用顺序（从第一个词开始，从左往右递归）和逆序（从最后一个词开始，从右向左递归）递归神经网络计算得到两套不同的隐层表示，如图 4 所示。然后通过向量求和或拼接的方式计算得到最终的隐层表示。这样，文本序列中的每个单词的隐层都包含完整的前后上下文信息。相较于单向 LSTM 而言，双向 LSTM 可以提供更为全面的语义信息，从而提高系统性能。

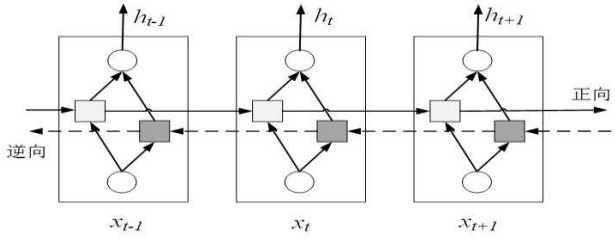


图 4 双向递归神经网络的一般结构

2.2.2 两阶段方法

在本文中，我们采用了两阶段的触发词识别方法。将触发词的识别过程分为识别和分类两个阶段。

(1) 识别阶段

在这个阶段中，生物医学文献中的触发词和非触发词被区别开来，但不对识别出的触发词进行分类，即此阶段为触发词二分类任务。在此阶段，我们通过双向 LSTM 构建触发词二分类模型，并对预测出来的触发词正例进行筛选，作为第二阶段的输入。

(2) 分类阶段

在这个阶段中，为识别阶段得到的触发词确定其具体类型，此阶段为触发词的多分类任务。在此阶段，我们通过双向 LSTM 构建触发词多分类模型，并将第一阶段识别出来的正例按照预定义的 19 种触发词类型进行分类，之后在预测结果中加回第一阶段过滤掉的预测负例，从而得到 MLEE 语料测试集的完整预测结果。

为了更好的与一阶段方法比较，本文在触发词识别的两阶段实验中均采用了与一阶段方法相同的双向 LSTM 神经网络构建触发词二分类和多分类的模型，同时采用了依存词向量捕捉词语语义信息，增加了句子向量信息建立词级特征和句子级特征之间的联系，丰富上下文信息，得到更加精确的隐层表示。

3 实验与分析

3.1 语料及评价方法

为了对本文提出的触发词识别方法进行评价，本文在生物信息抽取领域的通用语料 MLEE[12]语料上进行了触发词识别实验。MLEE 语料由 Pyysalo[9]组织标注，不仅仅抽取分子

级别的事件，还面向细胞、组织、器官等更多的生物实体相关的事件，共包含了 19 种事件类型，涵盖了从分子到器官水平的大多数事件类型。这些事件类型按照功能可以被分为“Anatomical”,“Molecular”,“General”,“Planned”四大类，具体类型包括“Cell proliferation”，“Regulation”，“Blood vessel development”等。此外，MLEE 语料为了丰富事件表示还引入了更加精确的事件描述。该数据集的数据静态分布如下表 1 所示。

表 1　MLEE 数据集的统计数据

项目	训练集	测试集	总计
文档数	175	87	262
句子数	1728	880	2608
事件	4471	2206	6677

本文使用三个性能评价指标评价每类触发词的性能，分别是准确率 *Precision(P)*、召回率 *Recall(R)*和 *F* 值 *F-score(F)*。其定义为公式：

$$P=\frac{TP}{TP+FP} \tag{2}$$

$$R=\frac{TP}{TP+FN} \tag{3}$$

$$F=\frac{2*P*R}{P+R} \tag{4}$$

其中 *TP*（True Positives）表示正例中判断正确的样本数，*FP*（False Positives）表示负例中判断错误的样本数，*FN*（False Negatives）表示正例中判断错误的样本数。

3.2 实验结果及分析

为了能够更好的比较不同神经网络结构之间触发词识别的性能差异，本文对所有实验采用了统一的参数标准。在一阶段方法和两阶段方法中的不同阶段训练模型时，梯度下降的学习率均设置为 0.001,最大迭代次数设置为 200 次，隐层节点数为 200，输入层的上下文窗口规定为 5。

3.2.1 依存词向量 Vs Skip-gram 模型词向量

如前文 2.1 所述，基于依存关系的词向量可以获得更多的语义信息，从而提高触发词识别性能。为了验证依存词向量对于触发词识别性能的影响，针对基于单向 LSTM 的一阶段触发词识别方法，本文分别采用了两种不同的词向量，即通过 word2vec 词向量训练工具，Skip-gram 模型训练的普通词向量和使用 word2vecf[16]词向量训练工具训练的基于依存关系的词向量，词向量维度均为 200 维。触发词识别性能如下表 2 所示(第 1 行&第 2 行)，基于依存词向量的触发词识别 F 值较基于 Skip-gram 模型训练的词向量的触发词识别 F 值提高 1.8%。

表 2 不同方法的触发词识别性能对比

方法	准确率(%)	召回率(%)	F 值(%)
Skip-gram 模型词向量+LSTM(一阶段)	68.69	72.08	70.35
依存词向量+LSTM（一阶段）	74.66	69.80	72.15
依存词向量+BLSTM（一阶段）	77.65	69.99	73.62
依存词向量+BLSTM+句子向量（一阶段）	78.36	75.93	77.13
依存词向量+BLSTM+句子向量（两阶段）	78.46	78.46	78.46

3.2.2 基于 LSTM 的触发词识别 Vs 基于 BLSTM 的触发词识别结果

为了能够在单向的基础上，进一步探究双向递归神经网络的识别性能，实验采用了对正向和逆向 LSTM 的隐层相加的方式表示新的隐层，识别效果为 73.62%，比单向的 LSTM 提高了 1.47%，如上表 2（第 2 行&第 3 行）所示。从性能上来看，无论是召回率还是准确率，双向的 LSTM 递归神经网络明显优于单向的网络。这主要是因为双向的递归神经网络可以访问更加丰富的上下文信息。

3.2.3 句子向量对触发词识别性能的影响

为了验证句子向量对于双向 LSTM 性能的影响，本文在上述实验的基础上增加了句子向量。如前文所述，本文句子向量的计算采取的是句子中所有单词对应的预训练词向量和微调后词向量的差值加和求平均的方式。如上表 2 所示，从性能上来看，增加句子向量后的 F 值（第 4 行）较不加句子向量的 F 值（第 3 行）提升了 3.51%，可见句子级的向量特征信息可以通过获取丰富的文本信息从而提升系统的识别性能。

3.2.4 一阶段方法的触发词识别 Vs 两阶段方法的触发词识别结果

在两阶段的触发词识别方法中，触发词识别被分为识别和分类两个阶段。在识别阶段，候选实例仅被识别为触发词和非触发词两类，语料中的类型数目比例为：所有正例总数：负例总数，而一阶段方法中类型的比例为：每个子类数目：负例总数，显然对于两阶段方法中的第一阶段而言，这个比例将大于一阶段方法的相应比例，从而缓解了类不平衡的问题；在分类阶段，由于只对识别阶段筛选出来的预测正例进行分类，数据集中类的不平衡问题也得到了很好的缓解。此外，这种方式也可以有效的避免过多的负例对正例分类造成的干扰。同时，实验表明，在训练时间上，两阶段方法时间也更短，更高效。综上，针对触发词识别任务而言，两阶段方法是一个比较有效的方法。为了验证两阶段方法的有效性，本文在基于依存词向量，双向 LSTM 和句子向量的实验基础上结合了两阶段方法，实验结果如上表 2（第 5 行）所示，两阶段方法的触发词识别相较于一阶段方法 F 值提高了 1.33%。

3.2.5 本文系统与其他系统整体性能的比较

为了更好的评价本文提出方法的性能，我们选取了 MLEE 语料上，生物触发词识别的现有参考文献与本文进行了整体性能比较，结果如下表 3 所示。本文所提出的方法在总体性能上，分别比 Pyysalo 等[9]基于丰富特征的 SVM 的分类方法 F 值高 2.62%；比 Zhou 等[10]基于半监督的学习模型，通过引入未标注的语料和事件抽取中的隐藏话题来识别触发词的方法 F 值高 1.57%；比 Wang[12]基于依存关系的词向量的触发词识别的方法 F 值高 1.36%；比 Nie[13]等神经网络和词向量的方法 F 值高 1.23%；此外，Zhou[11]等将大规模语料中学习到的领域知识与人工总结的语义、句法等特征进行融合，然后通过 SVM 进行触发词分类，取得了目前 MLEE 语料上触发词识别的最好性能，本文方法 F 值比 Zhou 高 0.14%。相较而言，本文方法通过深层神经网络自动学习特征，避免了抽取人工特征时的代价。

表 3 本文系统与其他系统的性能比较†

方法	**准确率(%)**	**召回率(%)**	**F 值 (%)**
Pyysaloet al.[9]	70.79	81.69	75.84
Zhou et al.[10]	72.17	82.26	76.89
Wang et al.[12]	73.27	81.35	77.10
Nie et al.[13]	71.04	84.60	77.23
Zhou et al.[11]	75.35	81.60	78.32
Proposed	78.46	78.46	**78.46**

†注：Nie et al.[13]原文中触发词识别结果为 14 分类结果，本文按照 19 分类对其进行了换算。

3.2.6 本文系统与其他系统在 19 个子类上的性能比较

MLEE 语料上共有 19 种预定义事件类型，为了更好的分析本文所提出方法的触发词识别性能，我们分别在 19 个子类上与 Pyysalo[9], Zhou[11], Nie[13]的触发词识别性能进行了比较，具体如下表 4 所示。可以看出，相较于 Pyysalo 提出的基于特征的 SVM 分类方法，本文提出的方法在 10 种类型的触发词识别性能上优于文献[9]的方法，例如在“Regulation”，“Positive regulation”和“Binding”等类型上 F 值分别提升了 3.18%, 4.05%和 0.39%。相较于 Zhou 等[11]的机器学习方法，本文提出的方法在 6 种类型的触发词识别性能上优于文献[11]的方法，例如在“Breakdown”，“Positive regulation”, “Regulation”和“Binding”等类型上 F 值分别提升了 18.19%, 1.89%, 0.45%和 0.76%。与 Nie 等[13]基于神经网络和触发词的方法相比，本文提出的方法有 7 种类型的触发词抽取结果好于文献[13]方法。其中“Positive regulation”和“Remodeling”等类型上 F 值分别提升了 2.91%和 50%。总体来看，本文方法在处理“Negative regulation”、“Positive regulation”、“Regulation”以及“Binding”等复杂事件类型时具有一定优势，而这些复杂事件触发词的抽取往往更需要语义信息的支持。

表 4 本文系统与其他系统在 19 个子类上的性能比较

触发词类型	Pyysalo *et al.* (P/R/F, %)	Zhou *et al.* *(P/R/F, %)*	Nie *et al.* *(P/R/F, %)*	本文系统 *(P/R/F, %)*
Cell proliferation	63.83/69.77/66.67	78.38/ 67.44/ 72.50	81.40/ 89.74/ 85.37	82.35/65.12/72.73
Development	68.07/83.51/75.00	69.30/ 81.44/ 74.88	48.54/ 84.69/ 61.71	65.81/79.38/71.96
Blood vessel develop	95.70/96.33/96.01	98.65/ 97.33/ 97.99	96.60/ 93.11/ 94.82	97.77/ 93.59 /95.64
Growth	69.12/83.93/75.81	77.05/ 83.92/ 80.34	100.00/ 91.07/ 95.33	68.18/80.36/ 73.77
Death	56.90/94.29/70.97	72.09/ 88.57/ 79.49	62.75/ 88.89/ 73.56	73.68/ 77.78/ 75.68
Breakdown	80.00/34.78/48.48	80.00/ 34.78/ 48.48	84.21/ 69.57/ 76.19	76.47/ 59.09/ 66.67
Remodeling	85.71/54.54/66.66	85.71/ 60.00/ 70.59	16.67/ 10.00/ 12.50	83.33/ 50.00/ 62.50
Synthesis	33.33/50.00/40.00	40.00/ 50.00/ 44.44	75.00/ 75.00/ 75.00	60.00/ 75.00/ 66.67
Gene expression	83.78/93.94/88.57	84.72/ 92.42/ 88.41	85.21/ 91.67/ 88.32	85.61/ 90.40/ 87.94
Transcription	25.00/14.28/18.18	0.00/ 0.00/ 0.00	24.00/ 85.71/ 37.50	0.00/0.00/0.00
Catabolism	0.00/0.00/0.00	16.67/ 33.33/ 22.22	12.50/ 25.00/ 16.67	50.00/33.33/40.00
Phosphorylation	50.00/100/66.66	75.00/ 100/ 85.71	100.00/ 100.00/ 100	60.00/ 100/ 75.00
Dephosphorylation	0.00/0.00/0.00	100.00/ 100/ 100	100.00/ 100.00/ 100	0.00/0.00/0.00
Localization	79.86/83.46/81.62	80.85/ 85.71/ 83.21	65.50/ 84.21/ 73.68	86.18/ 79.70/ 82.81
Binding	84.00/76.36/80.00	81.13/ 78.18/ 79.63	81.82/ 80.36/ 81.08	89.13/73.21/80.39
Regulation	60.37/46.48/52.52	56.49/ 53.05/ 54.72	59.90/ 67.98/ 63.68	54.55/ 55.81/ 55.17
Positive regulation	67.85/86.73/76.14	71.58/ 86.41/ 78.30	67.14/ 91.03/ 77.28	79.55/ 80.84/ 80.19
Negative regulation	74.35/77.03/75.66	77.09/ 78.83/ 77.95	70.86/ 84.55/ 77.10	73.39/ 75.33/ 74.35
Planned process	53.92/75.00/62.73	56.46/ 75.64/ 64.66	64.53/ 74.86/ 69.31	60.12/ 56.65/ 58.33

4 结论

本文针对生物触发词识别任务，提出了一种基于双向 LSTM 和两阶段方法的触发词识别模型，在生物事件抽取通用语料 MLEE 语料上获得了较好的性能。主要结论如下：

首先，在生物触发词识别任务中，浅层的机器学习方法需要设计复杂的人工特征，丰富的专业领域知识，以及大量的实验选择特征。这一方面增加了系统的设计成本，另一方面也对系统的移植带来了困难。从而，本文采用了深层神经网络方式识别触发词。而 LSTM 神

经网络可以学习长距离依赖的信息，避免了传统递归神经网络在处理长句子时产生的梯度弥散问题。因此，本文提出了基于双向 LSTM 的触发词识别模型。

其次，为了获取更好的数据表示，本文针对 PubMed 数据库中下载的大规模语料训练了基于依存关系的词向量，该词向量可以捕获长距离词间的关系，从而获得更加丰富的语义信息；此外，本文在预训练词向量的基础上扩展了一套随着训练过程不断微调的词向量，进而通过计算得到句子向量，句子向量信息可以建立起词级特征和句子级特征之间的联系，丰富上下文信息，得到更加精确的隐层表示。

最后，本文采用的两阶段方法可以缓解训练过程中存在的类不平衡问题。两阶段方法将触发词识别分成识别和分类两个阶段，通过将一次分类转换为两次分类，每个阶段数据不平衡的严重性低于一次分类，间接的缓解了数据不平衡问题。此外，在两阶段方法中，由于第二阶段只对预测的正例进行多分类，可以有效的避免过多的负例对正例分类造成的干扰。同时，实验表明，在训练时间上，两阶段方法时间也更短，更高效。综上，针对触发词识别任务而言，两阶段方法是一个比较有效的方法。

参考文献

[1] Jin-Dong K, Tomoko O, Sampo P, et al. Overview of BioNLP'09 Shared Task on Event Extraction[C]. Proceedings of the Workshop on BioNLP: Shared Task. 2009: 1-9.

[2] Jin-Dong K, Yue W, Toshishisa T, et al. Overview of the Genia Event Task in BioNLP Shared Task 2011[C]. Proceedings of the BioNLP 2011 Workshop. 2011: 7-15.

[3] Jin-Dong K, Yue W, Yamamoto Y. The Genia Event Extraction Shared Task, 2013 Edition-Overview[C]. Proceedings of BioNLP Shared Task 2013 Workshop. 2013: 8-15.

[4] Louise D, Robert B, Estelle C, et al. Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016[C]. Proceedings of 4th BioNLP Shared Task Workshop. 2016: 12-22.

[5] Björne J. Biomedical Event Extraction with Machine Learning[J]. TUCS Dissertations, 2014, 178: 1-121.

[6] Vlachos A. Two strong baselines for the BioNLP 2009 Event Extraction Task[C]. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics. 2010: 1-9.

[7] Bretonnel K, Karin V, Helen L, et al. High-precision Biological Event Extraction with a Concept Recognizer[C]. Proceedings of the Workshop on BioNLP: Shared Task. 2009: 50-58.

[8] Bjorne J, Juho H, Filip G, et al. Extracting Complex Biological Events with Rich Graph-Based Feature Sets[C]. Proceedings of the Workshop on BioNLP: Shared Task. 2009: 10-18.

[9] Pyysalo S, Ohta T, Miwa M, et al. Ananiadou S, Event Extraction across Multiple Levels of Biological Organization[J]. Bioinformatics, 2012, 28(18): 575–i581.

[10] Zhou D, Zhong D. A Semi-Supervised Learning Framework for Biomedical Event Extraction Based on Hidden Topics[J]. Artificial Intelligence in Medicine, 2015, 64(1): 51-58.

[11] Zhou D, Zhong D, He Y. Event Trigger Identification for Biomedical Events Extraction Using Domain Knowledge[J]. Bioinformatics, 2014, 30(11): 1587-1594.

[12] Jian W, Zhang J, Yuan A, et al. Biomedical Event Trigger Detection by Dependency-Based Word Embedding[C]//IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2015: 429-432.

[13] Nie Y, Rong W, Zhang Y, et al. Embedding Assisted Prediction Architecture for Event Trigger Identification[J]. Journal of Bioinformatics & Computational Biology, 2015, 13(3): i575-7.

[14] Yoshua B, Rejean D, Pascal V, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research (JMLR), 2003, 3: 1137-1155.

[15] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. 2013,

arXiv preprint arXiv: 1301.3781.

[16] Levy O, Goldberg Y. Dependency-Based Word Embeddings[C]. Meeting of the Association for Computational Linguistics. 2010: 302-308.

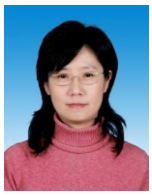
[17] Li L, Jin L, Jiang Y, et al. Recognizing Biomedical Named Entities Based on the Sentence Vector/Twin Word Embeddings Conditioned Bidirectional LSTM[C]//China National Conference on Chinese Computational Linguistics. Springer International Publishing, 2016: 165-176.

[18] Kolen J, Kremer S. Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-term Dependencies[M]. Wiley-IEEE Press, 2007, 28(2): 237-243.

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

[20] Schuster M, Paliwal K. Bidirectional Recurrent Neural Networks[J]. Signal Processing, 1997, 45(11): 2673-2681.

何馨宇（1983—），女，博士，主要研究领域为自然语言处理。Email: hexinyu@mail.dlut.edu.cn;



李丽双（1967—），女，教授，博士生导师，主要研究领域为自然语言处理、信息抽取与文本挖掘。本文通讯作者。Email: lilishuang314@163.com。