基于双向 LSTM 的维吾尔语事件因果关系抽取

田生伟[①] 周兴发[①] 禹 龙[*②] 冯冠军[③] 艾山·吾买尔[④] 李 圃[⑤]

[①](新疆大学软件学院 乌鲁木齐 830046)

[②](新疆大学网络中心 乌鲁木齐 830046)

[③](新疆大学人文学院 乌鲁木齐 830046)

[④](新疆大学信息科学与工程学院 乌鲁木齐 830046)

[⑤](新疆大学语言学院 乌鲁木齐 830046)

摘 要：针对传统方法不能有效抽取维吾尔语事件因果关系的问题，该文提出一种基于双向 LSTM(Bidirectional Long Short-Term Memory, BiLSTM)的维吾尔语事件因果关系抽取方法。通过对维吾尔语语言以及事件因果关系特点的研究，提取出 10 项基于事件内部结构信息的特征；同时为充分利用事件语义信息，引入词嵌入作为 BiLSTM 的输入，提取事件句隐含的深层语义特征并利用批样规范化(Batch Normalization, BN)算法加速 BiLSTM 的收敛；最后融合这两类特征作为 softmax 分类器的输入进而完成维吾尔语事件因果关系抽取。实验结果表明，该方法用于维吾尔语事件因果关系的抽取准确率为 89.19%, 召回率为 83.19%, F 值为 86.09%，证明了该文提出的方法在维吾尔语事件因果关系抽取上的有效性。

关键词：语言信号处理；事件因果关系；维吾尔语；双向 LSTM；词嵌入；批样规范化

中图分类号： TP391 **文献标识码**： A **文章编号**：1009-5896(2018)01-0200-09

DOI: 10.11999/JEIT170402

Causal Relation Extraction of Uyghur Events Based on Bidirectional Long Short-term Memory Model

TIAN Shengwei[①] ZHOU Xingfa[①] YU Long[②]

FENG Guanjun[③] Aishan WUMAIER[④] LI Pu[⑤]

[①](*School of Software*, *Xinjiang University*, *Urumqi* 830046, *China*)

[②](*Net Center, Xinjiang University*, *Urumqi* 830046, *China*)

[③](*College of Humanities*, *Xinjiang University*, *Urumqi* 830046, *China*)

[④](*School of Software*, *Xinjiang University*, *Urumqi* 830046, *China*)

[⑤](*School of Languages*, *Xinjiang University*, *Urumqi* 830046, *China*)

Abstract: Since the traditional events causal relation has the disadvantages of small recognition coverage, a method for causal relation extraction of Uyghur events is presented based on Bidirectional Long Short-Term Memory (BiLSTM) model. In order to make full use of the event structure information, 10 characteristics of the Uyghur events structure information are extracted based on the study of the events causal relationship and Uyghur language features; At the same time, the word embedding is introduced as the input of BiLSTM to extract the deep semantic features of the Uyghur events and Batch Normalization (BN) algorithm is usded to accelerate the convergence of BiLSTM. Finally, concatenating these two kinds of features as the input of the softmax classifier to extract the Uyghur events causal relations. This method is used in the causal relation extraction of Uyghur events, and the results show that the precision rate, the recall rate and F value can reach 89.19 %, 83.19% and 86.09 %, indicating the effectiveness and practicability of the method of causal relation extraction of Uyghur events.

收稿日期：2017-05-02；改回日期：2017-07-19；网络出版：2017-09-14

*通信作者：禹龙 yul_xju@163.com

基金项目：国家自然科学基金(61662074, 61563051, 61262064)，国家自然科学基金重点项目(61331011)，新疆自治区科技人才培养项目(QN2016YX0051)

Foundation Items: The National Natural Science Foundation of China (61662074, 61563051, 61262064), The Key Project of National Natural Science Foundation of China (61331011), Xinjiang Uygur Autonomous Region Scientific and Technological Personnel Training Project (QN2016YX0051)

Key words: Language signal processing; Event causal relation; Uyghur; Bidirectional Long Short-Term Memory (BiLSTM); Word embedding; Batch Normalization (BN)

1　引言

事件作为信息表示的一种重要形式，近年来在自然语言处理领域受到越来越多的关注。人们通过研究事件内部组成结构(参与者，时间，地点等事件论元)和外部关联(因果，共指，时序等语义关系)对文本进行抽取，进而以支持基于事件的问题回答，信息抽取，自动文摘等自然语言处理技术，实现对文本的深层理解。

因果关系作为事件外部关联中的一种语义关系，在文本中既常见又非常重要，有着广泛的应用前景。它反映了事件间的先后相继、由因及果的一种关系。因果关系的识别对文本事件抽取，深层语义理解有着重要意义，有助于获取事件演变的过程，对事件的发生进一步认识，从而为决策者提供重要的信息来预判事件后期的发展。事件因果关系的抽取主要分为两部分：(1)事件原因的抽取；(2)事件结果的抽取。通常情况下，根据原因和结果在文本上下文中是否同时出现，可以将因果关系分为显式因果关系(文本中原因和结果同时出现)和隐式因果关系(文本中只出现了原因或者结果)两种。隐式因果关系的抽取需要根据文本上下文知识进行推理和判断来确定事件的原因或者结果。例如，对于事件描述“جاڭ سەن لى سى ئۆلتۈردى(张三杀了李四)”，其结果事件“لى سى ئۆلۈپ كەتتى(李四死了)”被省略，需要根据上下文推理确定。对于显式因果关系，根据原因和结果之间的对应关系可以将其细分为一因一果、一因多果、多因一果和多因多果，另外，根据文本上下文中是否出现“因为、由于、造成”等带标记的因果连接词，显式因果关系又可以分为带标记因果关系和无标记因果关系。

近年来，事件抽取在自然语言处理领域受到越来越多的关注，从基于模式匹配的方法[1]，到支持向量机(Support Vector Machine, SVM)[2]的浅层机器学习方法，再到基于深度学习的方法。如 Chen 等人[3]使用动态池化卷积神经网络(Convolutional Neural Network, CNN)在 ACE 2005 语料上抽取事件本体；Zhang 等人[4]通过基于深度信念网络的事件识别模型对中文文本进行事件识别；Chang 等人[5]在双向 LSTM(Bidirectional Long Short-Term Memory, BiLSTM)基础上，对隐藏状态进行最大、最小和平均池化操作，进而提取英文文本中的事件本体。然而，对于事件因果关系抽取研究还处于基于模式匹配和浅层机器学习方法研究阶段。在有关基于模式匹配方法研究中，Gan 等人[6]提出一种事件因果关系结构分析方法；Girju 等人[7]通过 Internet 和 WordNet 搜寻因果关系动词，建立 Lexico-syntactic 模式，实现了特定事件因果关系的自动识别。以上事件因果关系抽取的研究具备很强的领域性，需要完备的领域知识，因此近年来的研究更多地利用统计概率方法从文本中抽取显式因果关系(原因和结果同时在文本中出现)。Marcu 等人[8]采用朴素贝叶斯模型，通过分析相邻句子间的词对概率来提取因果关系；Sorgent 等人[9]首先通过制定的规则抽取事件因果关系，然后通过贝叶斯推理优化结果。以上基于统计概率的方法取得了较好的结果，但仅限于抽取带有标记的 (形如因为、导致等)、句内的或相邻语句间的因果关系。 研究发现，文本中存在大量不带标记和跨语句、跨段落的因果关系，因此学者们开始尝试事件序列标注方法来抽取语料中显式因果关系。如付剑锋等人[10]使用事件触发词、事件类型等事件内部结构特征作为输入特征，采用条件随机场识别事件间多种带标记或无标记因果关系。钟军等人[11]等在 CRFs 基础上采用双层模型对事件序列进行标记，从而抽取事件显式因果关系。

作为一种序列化模型，长短时记忆网络(Long Short-Term Memory, LSTM)将文本视为有序词汇序列，充分考虑文本的有序性和词汇间的关联性，更加符合自然语言规律。Tang 等人[12]针对语义关系分类问题，构建基于 LSTM 的学习模型，并发现 BiLSTM 与 LSTM 相比，能挖掘更丰富的语义信息且具有充分利用上下文信息的能力；Zhou 等人[13]借助 BiLSTM 从词嵌入中获取到高层面的语义信息特征，在 SemEval-2010 task 8 数据集上完成了句子级的关系分类问题。

以上研究表明，采用模式匹配方法抽取事件间因果关系，面临着通用性不强，需要大量领域知识等问题；而采用浅层机器学习方式则面临着特征提取困难，未考虑事件所在上下文的深层语义信息等问题，并且抽取结果只能覆盖文本中一因一果、一因多果的因果关系；同时现有的研究语种主要针对汉语和英语等大语种，而对维吾尔语(后面简称维语)这种使用范围相对较小的语种研究还不够深入。

针对上述存在的问题，本文提出一种基于 BiLSTM 的维吾尔语事件因果关系抽取方法。与付剑锋等人[10]，钟军等人[11]以往研究方法将因果关系的抽取问题转化为对事件序列的两次模式识别标注问题不同，本文将因果关系的抽取问题转化为对事

件对分类的三分类问题，然后用词嵌入表示事件句中的词汇作为 BiLSTM 的输入，并引入批样规范化(Batch Normalization, BN)算法加速 BiLSTM 的收敛，从而提取事件句的深层语义特征，最后融合对维吾尔语语言和事件因果关系特点研究后提取出的 10 项基于事件内部结构信息的特征，作为 softmax 分类器的输入，最终完成维吾尔语事件因果关系抽取。

2 预备知识

为了明确本文对于维语事件因果关系抽取研究的方法，先介绍一些基本概念。

定义 1 事件(event)：指在某个特定时间片段和地域范围内发生的由一个或多个角色参与，由一个或多个动作组成的一件事情，是关于某一主题的一组相关描述。如例 1 所示，共描述了 3 个事件：海地北部山区发生了交通事故；20 人受伤；6 人死亡(维语的书写习惯为从右向左)。

هايتىنىڭ شىمالىدىكى تاغلىق رايوندا 8-يانۋار ئېغىر قاتناش ۋەقەسى يۈز بېرىپ، 6 ئادەم قازا قىلدى، 20 ئادەم يارىلاندى.

例 1 海地北部山区 1 月 8 号发生一起严重交通事故，6 人死亡 20 人受伤。

定义 2 事件触发词(event trigger)：指在事件描述中能清晰表达事件所发生事情的词汇。如例 1 中的قاتناش ۋەقەسى(交通事故)、يارىلاندى (受伤)和قازا قىلدى(死亡)。

定义 3 事件论元(event argument)：指描述事件具体信息的文本短语，包括参与者、事件时间和发生地点等。如例 1 中的هايتىنىڭ شىمالىدىكى تاغلىق رايوندا(海地北部山区)，8-يانۋار(1 月 8 日)，20 ئادەم(20 人)，6 ئادەم(6 人)。

定义 4 事件对(event pair)：指对篇章中事件按照组对规则组对后的两个事件。

بۈگۈن چۈشتىن بۇرۇن ، نەنجىن يولى ئەتراپىدا قاتناش ۋەقەسى يۈز بېرىپ، بىر ئادەمنىڭ ئۆلۈشىنى، بەش ئادەمنىڭ يارلىنىشىنى كەلتۈرۈپ چىقارغان. ساقچى تەرەپنىڭ دەسلەپكى قەدەملىك تەكشۈرۈشىدە، بۇ قاتناش ۋەقەسى يولنىڭ مۇز تۇتۇش سەۋەبىدىن يۈز بەرگەن.

例 2 今天上午南京路附近发生一起车祸(e_1)，导致 1 死亡(e_2)，5 人受伤(e_3)，经警方初步调查(e_4)，车祸(e_1)是由道路结冰(e_5)引起的。

基于 BiLSTM 的维语事件因果关系抽取的核心是通过把因果关系抽取问题转换为对事件对分类的问题。假设语篇中的事件集合为：$E=\{e_1,e_2,\cdots,e_n\}$，通过事件组对算法，对事件集合 E 中事件进行组对从而构成事件对 $<e_i,e_j;y>$，其中 $y\in\{0,1,2\}$ 为事件对的标签，0 表示 e_i 与 e_j 不具有因果关系；1 表示 e_i 是 e_j 的原因事件；2 表示 e_i 是 e_j 的结果事件。如例 2 所示，共描述了 5 个事件，“قاتناش ۋەقەسى(车祸)”、“ئۆلۈشىنى(死亡)”、“يارلىنىشىنى(受伤)”、“تەكشۈرۈشىدە(调查)”和“مۇز تۇتۇش(结冰)”作为触发词清晰地表达了 5 个事件所属类别。其中事件 e_5 是 e_1 的原因事件，事件 e_3 和 e_4 不存在语义上的因果关系，因而将获得 $<e_5,e_1;1>$、$<e_1,e_5;2>$ 和 $<e_3,e_4;0>$ 等事件对。

3 维吾尔语因果关系抽取模型

与付剑锋等人[10]，钟军等人[11]等以往研究方法将因果关系的抽取问题转化为对事件序列的两次模式识别标注问题不同，本文将因果关系的抽取问题转化为对事件对分类的三分类问题，进而判断事件对之间是否存在因果关系，若存在因果关系，并指出原因事件和结果事件，整个抽取过程如图 1 所示。接下来将详细叙述图中的每个部分。

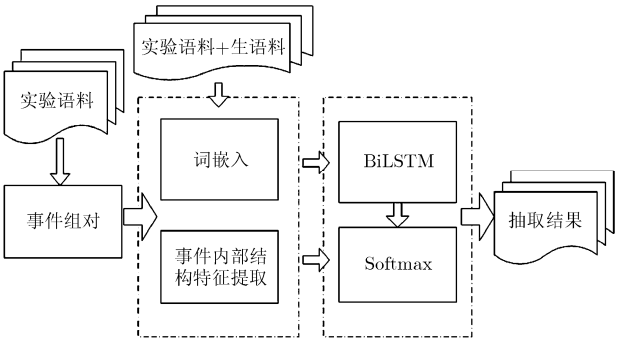

图 1 维语事件因果关系抽取框架

3.1 事件组对

本文首先将已标注事件的维语实验语料进行事件提取，并按照一定的规则两两组对，通过标注语料中的事件因果关系对应链，判断组对的事件对是否具有因果关系，然后按照第 2 节的叙述贴上对应的标签。具体的算法步骤如下：

步骤 1 提取语料中的事件，存入事件集合{Events Set}中；提取语料中的因果关系链，存入因果关系链集合{Chains Set}中。

步骤 2 对于每一个事件 $e_i\in$ {Events Set}，判断是否满足 $e_i\in$ {Chains Set}，是则存入{Causal Events Set}中否则存入{NonCausalEvents Set}。

步骤 3 对事件集合{CausalEvents Set}中的所有事件两两组对，构成事件对 $<e_i,e_j>$，然后根据{Chains Set}中的因果关系链为 $<e_i,e_j>$ 贴上对应的标签构成样本实例 $<e_i,e_j,\text{label}>$，其中 label $\in\{0,1,2\}$ (0 表示 e_i 和 e_j 间不具有因果关系；1 表示 e_i 是原因事件，e_j 是结果事件；2 表示 e_j 是原因事件，e_i 是结果事件)。

步骤 4 对事件集合{NonCausalEvents Set}

中的所有事件两两组对，构成事件对$<e_i,e_j>$，然后贴上类别标签 0，构成样本实例$<e_i,e_j,0>$，表示事件e_i和e_j间不具有因果关系。

步骤 5　合并步骤 3 和步骤 4 中的样本实例构成最终的样本实例。

3.2 维语事件内部结构特征提取

数据和特征决定了机器学习的上限，特征的选择在机器学习中占有相当重要的地位。提取的特征是否有效直接影响模型的预测性能，使用准确的特征对数据进行描述，实验结果会相应地提高。维语事件的内部结构信息(事件类别、触发词、事件极性等)详细反映了事件描述的内容，因而本文结合实验组维语语言专家关于维语语言和事件内部结构特点的意见选取了以下 10 个方面的特征作为维语事件内部结构特征：

(1)触发词：事件触发词能反映一个事件的大部分信息，具有很强的事件指向性。因此本文选择维语事件对中两个事件的触发词，然后将触发词去掉词尾，判断剩下部分是否相同，若相同该特征值标记为 1，否则标记为 0。例如：“يارىلانماق(受伤)”和“يارىلانغان(受过伤)”去掉词尾后具有相同的词干يارىلان，因而该特征值取 1。

(2)触发词词性：结合触发词的词性能很好地反映事件的信息。若两个触发词的词性相同，则该特征值标记为 1，否则标记为 0；同时根据语料统计，事件触发词中 90%为名词和动词，所以若触发词词性为名词取值为 0，若为动词取值为 1，否则取值为 2。

(3)触发词语义类别：同词性一样，语义类别能很好地反映事件的信息。若两个触发词的语义类别相同，则该特征值标记为 1，否则标记为 0。

(4) 触发词句法结构：触发词的句法结构反映了触发词在事件句中所作的成分。若两个触发词的句法结构相同，则该特征值标记为 1，否则标记为 0；在维语事件中，统计发现 85%的触发词作为谓语或宾语，所以若触发词句法结构为谓语取值为 0，若为宾语取值为 1，否则取值为 2。

(5)事件类别：事件类别不同，由原因事件引发的结果事件也不同。本文参照国际事件标注体系 ACE(标注了 Arabic, English, Chinese 等 3 种语言)和实验组维语专家给出的维语事件结构特点划分了 8 大类、33 小类事件类别。根据事件对中的两个事件类别，若相同则为 1，否则为 0。

(6)事件子类别：事件子类别进一步定义了事件所属类别。与事件类别类似，若事件子类别相同则为 1，否则为 0。

(7)事件极性：描述了事件是肯定的事件还是否定的事件。当事件对的极性相同时该特征值取 1，否则取 0。例如对于例 3 中事件“ئوققا تۇتۇش ۋەقەسى (枪击事件)”和“ئۆلۈپ(死亡)”都显式地表示已经发生，所以对于事件对$<e_6,e_7>$该特征值取 1。

امېرىكا فىلورىدا ئىشتاتىدىكى بىر ئايروپورتتا ئوققا تۇتۇش ۋەقەسى يۈز بېرىپ، 5 ئادەم ئۆلۈپ 8 ئادەم يارىلاندى

例 3　美国佛罗里达州一个机场发生枪击事件(e_6), 5 人死亡(e_7), 8 人受伤(e_8)。

(8)事件时态：描述事件是过去发生的事件还是正在发生的事件或者将来发生的事件。正在发生的事件不可能是过去发生的事件的原因事件，将来发生的事件也不可能是过去和正在发生的事件的原因事件，因而事件时态对因果关系的抽取有着很大的影响。若事件发生在过去则该特征值取 0，事件正在发生则取 1，若发生在将来则取 2。

(9)事件间的相对位置：事件间距离越近，具有因果关系的概率越大，因此本文计算了事件间的相对距离作为特征之一。首先按照事件在文本中出现的先后顺序给每一个事件一个递增的 id，然后计算 id 间的差值绝对值 distance，若$0 \le \text{distance} \le 3$，则取值为 0，若$4 \le \text{distance} \le 6$，取值为 1，否则取值为 2。

(10)事件对间相同论元的个数：若事件间相同的论元越多，事件间的相对距离则越近，事件间具有因果关系的概率就越大，因而本文计算了事件对间相同的论元个数，把它作为样例的特征之一。

3.3 词嵌入

事件间的因果关系是一种语义关系，3.2 节提取的特征虽然能较好地解决因果关系抽取问题，却缺少对整个事件句的语义考虑，因此本文引入词嵌入提取整个事件句的深层语义特征。不同于传统的 one-hot 模型，基于神经网络训练的词嵌入包含丰富的上下文信息，可以很好地表现目标词在当前文本中的语义规则，同时也避免了维数灾难[14]。本文使用 Mikolov 等人[15]提出的 word2vec 工具进行词嵌入的训练，选择 Skip-gram+HS 模型作为训练框架。为了准确地获取每个词在低维空间中语义的分布情况，在原有实验语料的基础上进行了扩充。选取天山网和人民网等维语版网页作为语料来源，利用网络爬虫下载网页，进行去重、去噪处理之后获取约 7000 篇不限题材且未标注任何内容的文本作为生成词嵌入的生语料。

3.4 双向 LSTM

LSTM 网络是 RNN 的扩展，可以看作是同一神经网络的多次复制，这些神经网络模块共享权值，每一个神经网络模块把信息传递给下一个模块，从

而达到信息传递，如图 2 所示。与 RNN 网络结构不同，LSTM 通过特别设计的门控机制来避免长期依赖问题，包含输入门(input gates)、遗忘门(forget gates)和输出门(output gates)3 种门结构，以保持和更新细胞状态。以下公式中 $\boldsymbol{i}_t$，$\boldsymbol{f}_t$，$\boldsymbol{o}_t$ 和 $\boldsymbol{C}_t$ 分别表示 t 时刻对应的 3 种门结构和细胞状态。

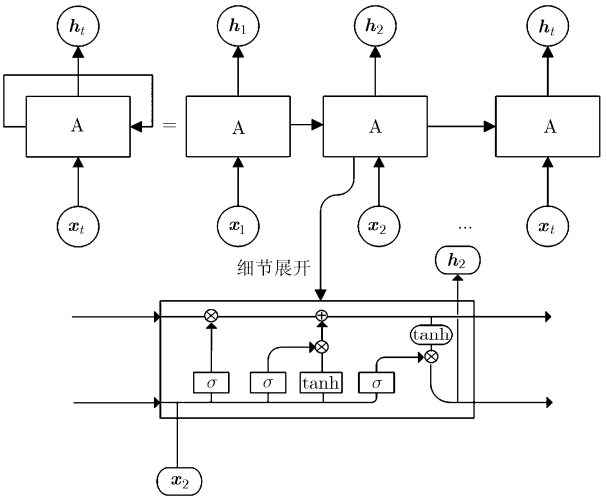

图 2 LSTM 展开图

(1)LSTM 的第 1 步决定从细胞状态中丢弃什么信息。由忘记门的 sigmoid 层决定，以当前层的输入 $\boldsymbol{x}_t$ 和上一层的输出 $\boldsymbol{h}_{t-1}$ 作为输入。

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_f\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_f\right) \tag{1}$$

(2)下一步是确定什么样的新信息被存放在细胞状态中。这里包含两个部分：(1)sigmoid 层“输入门”决定什么值将要更新；(2)tanh 层创建一个新的候选值向量 $\widetilde{\boldsymbol{C}}_t$ 加入到状态中。

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_t \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_i\right) \tag{2}$$

$$\widetilde{\boldsymbol{C}}_t = \tanh\left(\boldsymbol{W}_c \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_c\right) \tag{3}$$

(3)通过式(1)、式(2)、式(3)的计算，最终来更新细胞状态。首先将旧的细胞状态 $\boldsymbol{C}_{t-1}$ 乘以 $\boldsymbol{f}_t$，丢弃掉确定需要丢弃的信息；然后加上 $\boldsymbol{i}_t \odot \widetilde{\boldsymbol{C}}_t$，生成细胞状态的更新值。$\odot$ 表示逐点乘积。

$$\boldsymbol{C}_t = \boldsymbol{f}_t \odot \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \odot \widetilde{\boldsymbol{C}}_t \tag{4}$$

(4)最终由输出门决定信息的输出。先使用 sigmoid 层来决定要输出细胞状态的部分信息，接着用 tanh 处理细胞状态，两部分信息的乘积得到输出的值。

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_o \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_o\right) \tag{5}$$

LSTM 通过记忆单元来学习从细胞状态中忘记信息，更新细胞状态的信息，能有效利用序列数据中长距离依赖信息，因而被广泛地应用在机器翻译[12]、事件抽取[16]、情感分类[17]等自然语言处理领域。

从图 2 中 LSTM 模型结构可以看出，序列信息从前往后依次传播，这种单向机制仅包含序列当前词的前文信息，而对后文信息并未涉及，在此基础上提出的 BiLSTM 很好地解决了这个问题。BiLSTM 在隐层同时有一个正向 LSTM 和一个反向 LSTM，正向 LSTM 捕获了上文的特征信息，而反向 LSTM 捕获了下文的特征信息，然后通过融合捕获的上文特征信息和下文特征信息最终获得全局的上下文信息。

3.5 批样规范化

机器学习领域的一个重要假设是训练数据和测试数据满足相同的分布，这是通过训练数据得到的模型在测试数据上获得理想结果的重要前提条件。LSTM 学习过程的本质就是为了学习数据的这种分布，然而在训练过程中，LSTM 在做非线性变换前的激活输入值随着网络深度加深或者在训练过程中各层参数的不断更新导致其分布逐渐发生偏移或者变动，一旦网络某一层输入数据的分布发生改变，那么这一层网络就需要学习这个新的数据分布。训练过程中，训练数据的每一层分布一直在发生变化，并且每一层所需要的学习率不一样，通常需要使用最小的那个学习率才能保证损失函数有效下降，因此随着网络深度加深模型的收敛越来越慢。为了解决这个问题，Ioffe 等人[18]提出了 BN 算法。BN 算法使每一层的数据都归一化均值为 0、标准差为 1 来保证数据分布稳定，从而可以使用较大的学习率进行训练，使网络加快收敛，提高训练速度。BN 算法主要使用式(6)进行归一化：

$$\mathrm{BN}(\boldsymbol{h}; \gamma, \beta) = \beta + \gamma \frac{\boldsymbol{h} - E(\boldsymbol{h})}{\sqrt{\mathrm{Var}(\boldsymbol{h}) + \varepsilon}} \tag{6}$$

其中，$\boldsymbol{h} \in R^d$ 为当前神经网络层激活函数的输入值，$\gamma \in R^d$，$\beta \in R^d$ 用来保持模型的表达力，$\varepsilon \in R$ 为模型的正则化超参数。$E(\boldsymbol{h})$ 和 $\mathrm{Var}(\boldsymbol{h})$ 的值在训练阶段时基于当前批量的样本值计算得到，而在测试阶段则基于整个数据集计算得到。为了解决 LSTM 收敛速度慢问题，本文在式(1)、式(2)、式(3)、式(5)中引入了 BN 操作，如式(7)所示。

$$\begin{pmatrix} \boldsymbol{f}_t \\ \boldsymbol{i}_t \\ \widetilde{\boldsymbol{C}}_t \\ \boldsymbol{o}_t \end{pmatrix} = \mathrm{BN}\left(\boldsymbol{W}_h \boldsymbol{h}_{t-1}; \gamma_h, \beta_h\right) + \mathrm{BN}\left(\boldsymbol{W}_x \boldsymbol{x}_t; \gamma_x, \beta_x\right) + \boldsymbol{b} \tag{7}$$

3.6 基于 BiLSTM 的维语事件因果关系抽取

图 3 描述了整个维语事件因果关系抽取的模型结构。通过将事件因果关系抽取转换为对事件对的分类问题，整个过程可以分为 3 个部分：(1)事件句

语义特征抽取部分；(2)特征融合部分；(3)softmax 分类部分。其核心思想是将事件因果关系抽取问题转换为对事件对的分类问题。模型引入词嵌入作为事件句维语词汇序列的特征，然后利用 BiLSTM 能有效利用序列数据中长距离依赖信息特效，挖掘事件句上下文隐含的语义特征，并引入 BN 操作加速模型的收敛速度；此外基于维语事件因果关系特性，提取出 10 项经验表示的事件内部结构特征；最后将两类特征融合，作为 softmax 分类器的输入，最终完成事件因果关系抽取任务。

4　实验与分析

4.1 语料准备

基于机器学习的方法抽取事件因果关系需要相应的语料库来进行训练和测试。目前，国际上有两个会议机构提供常用的事件信息抽取评测语料库，即消息理解会议 (Message Understanding Conference, MUC；仅有 English 语料)和自动内容抽取会议(Automatic Content Extraction, ACE；有 Arabic, Chinese, English 3 种语料)。然而上述会议主要关注事件要素识别和模板填充，尚未发现有可供公开评测的维语事件因果关系抽取语料库，因此，本文针对维语事件因果关系抽取对语料进行了筛选和标注。

新闻报道是事件信息抽取评测典型的语料。因此, 实验中选取天山网、人民网和博客以及论坛等网站维语版网页作为语料来源。利用网络爬虫下载这些网页，然后经过去重、去噪筛选出以新闻报道为题材且包含事件描述的新闻报道文本作为实验语料。参照 ACE 事件标注体系，在实验组维语专家指导下对语料进行标注。

本次实验共标注了经过上述处理后的 250 篇语料，统计发现其中包含 349 条因果关系，然后按照 3.1 节样本构建方法，共生成了 1680 条样本数据。

4.2 实验结果与分析

theano 是目前常用的深度学习框架，本文所有模型基于于 theano 实现，同时采用批量梯度下降法更新模型参数。为保证实验结果的稳定性，所有实验均首先将样本数据随机打乱，然后采用 5 折交叉验证，取 5 次结果的平均值作为最终结果。通过网格搜索算法反复实验不同的参数组合，确定各个模型的最优参数。后续实验均采用各自最优参数进行对比实验。图 3 架构的最优参数如表 1 所示。

表 1 模型最优参数

ε	0.001	bs	10
α	0.7	eds	100
es	90	hus	100

其中，ε 表示 BN 算法的正则化参数；α 表示模型训练过程中的学习率；es 表示模型训练达到最优的迭代次数；bs 表示每一次迭代训练，批量处理样本个数；eds 表示词向量的维度；hus 表示 LSTM 隐藏层的节点数。

4.2.1 语义特征对事件因果关系抽取的影响　事件间的因果关系是一种语义关系，3.2 节提取的规则特征主要围绕事件触发词、事件论元和事件类别等事件内部结构，缺少对整个事件句的语义考虑。本节探讨了基于词嵌入和 BiLSTM 生成的事件句语义特征对事件因果关系抽取的影响，分别将表 2 中的两类特征集作为 softmax 分类器的输入来验证语义特征对事件因果关系抽取的影响。实验结果如表 2 所示。

结果表明，在去掉语义特征仅包含规则特征条件下的因果关系抽取性能与包含全部特征的性能相比有所下降，反映模型性能的准确率 P 与包含语义特征的准确率 P 相比降低了 7.94%，召回率 R 下降了 6.72%，衡量模型整体性能的 F 值下降了 7.3%。实验结果说明了对语义特征引入的有效性，这是因为因果关系是一种语义关系，基于规则的特征仅仅考虑了事件的内部结构特点，缺乏对整个事件句的语义信息的考虑，而语义特征则通过词嵌入对事件

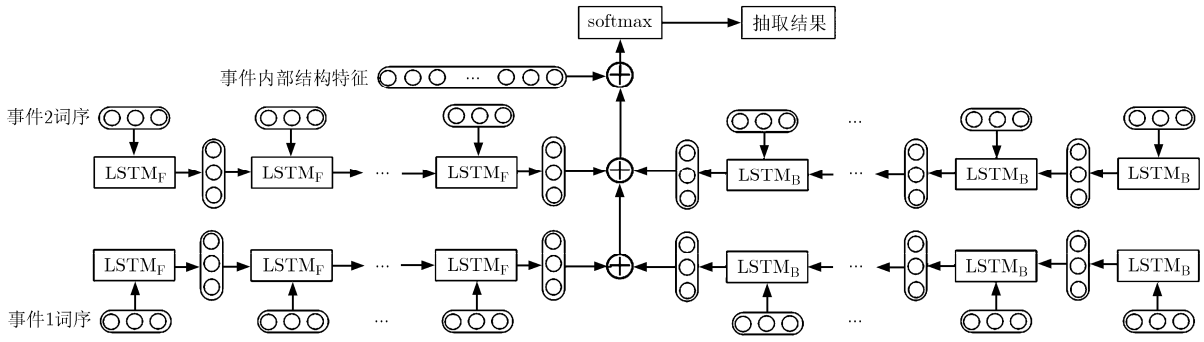

图 3 基于 BiLSTM 的维语事件因果关系抽取模型结构，其中每个 LSTM 单元都经过了 BN 算法处理

表 2 语义特征对因果关系抽取的影响(%)

特征类型	P	R	F
规则特征+语义特征	89.19	83.19	86.09
规则特征	81.25	76.47	78.79

句的词汇序列建模，然后由 BiLSTM 自动学习了整个事件句的隐含语义信息。

4.2.2 词嵌入维度对因果关系抽取的影响 训练词嵌入时，可以选择不同维度大小作为参数。不同维度的词嵌入蕴含的语义信息不同，理论上维度越大词嵌入蕴含的语义信息越多，所以本文分别选择 50, 100, 150, 200 和 250 维作为词嵌入的维度，在图 3 架构下验证维度对因果关系抽取性能的影响。实验结果如表 3 所示。

表 3 词向量维度对因果关系抽取性能影响对比实验(%)

维度	P	R	F
50	87.39	81.51	84.35
100	89.19	83.19	86.09
150	90.38	79.69	84.30
200	84.96	80.67	82.76
250	87.72	84.03	85.84

从表 3 可以看出，准确率，召回率和 F 值分别在 150 维，250 维和 100 维时达到最高，为 90.38%，84.03%和 86.09%。实验结果表明，词向量维度对因果关系抽取的性能并没有明显的规律，实验结果也比较不稳定在本实验数据集下，100 维的词向量结果相对比较均衡，因此，在后续实验中均采用 100 维词向量作为词汇特征验证后续实验。

4.2.3 BiLSTM 与 LSTM 因果关系抽取性能对比

在维语事件因果关系抽取上，为验证 BiLSTM 是否比 LSTM 性能更优越，将图 3 中的模型与去掉两个后向传播的 LSTM 模型作比较。实验结果如表 4 所示，为确保对比实验结果的有效性，实验结果均在两个模型各自最优参数下获得。

从表 4 实验结果可知，BiLSTM 的准确率，召回率和 F 值分别比 LSTM 高出了 4.79%, 5.88%和 5.39%。BiLSTM 的性能优于 LSTM，这是因为 LSTM 的序列信息从前往后依次传播，并不包含从后向前的传播过程，这种单向机制仅包含事件句词汇序列当前词的前文信息，而对后文信息并未涉及；而 BiLSTM 在 LSTM 基础上增加了一个反向 LSTM，正向 LSTM 用于捕获上文的特征信息，反向 LSTM 用于捕获下文的特征信息，然后通过融合捕获的上文特征信息和下文特征信息最终获得全局的上下文信息。因而与 LSTM 相比，维语事件因果关系抽取使用 BiLSTM 可更利于挖掘事件句隐含的上下文语义信息，更适用于本任务。

表 4 BiLSTM 与 LSTM 因果关系抽取性能对比(%)

模型	P	R	F
BiLSTM	89.19	83.19	86.09
LSTM	84.40	77.31	80.70

4.2.4 LSTM 层数对因果关系抽取的影响 图 3 架构下的前向 LSTM 和后向 LSTM 均采用单层 LSTM 模型。和其它深度学习模型一样，BiLSTM 可以通过堆叠 LSTM 形成更深层次的模型，从而挖掘不同抽象层度的语义信息。为此，本文在图 3 架构的基础上分别增加前向 LSTM 和后向 LSTM 的层数验证 LSTM 层数对因果关系抽取的影响。实验结果如图 4 所示。

从图 4 可知，准确率 P 随着 LSTM 层数的增加微小下降后逐渐增加，在 LSTM 层数为 4 时达到 89.36%。然而召回率 R 和 F 值随着 LSTM 层数的增加逐渐下降，在 LSTM 层数为 4 时分别为 70.59%和 78.87%。实验结果表明，随着 LSTM 层数的增加，模型的整体性能低于图 3 架构下的模型。这是因为 LSTM 层数的增加，导致模型更加复杂，学习到的特征过度抽象；同时需要学习的参数更多，从而学习到的特征包含一些无用信息。

4.2.5 BiLSTM 与其它模型的因果关系抽取性能对比 在自然语言处理领域中，SVM 和 CNN 是两个常用的模型。为充分验证图 3 架构的有效性，本节将图 3 架构下的模型与 SVM 和 CNN 进行事件因果关系的抽象性能的对比。实验结果如表 5 所示。

从表 5 可知，SVM 与 BiLSTM 和 CNN 相比，反映模型性能的 P, R 和 F 值均有所下降，这是因为 SVM 挖掘数据中隐藏深层特征的能力与 BiLSTM 和 CNN 相比相对较弱，基于深度学习思想的 BiLSTM 和 CNN 能够捕捉数据中更加的复杂分布，

表 5 BiLSTM 与其它模型的因果关系抽取性能对比(%)

模型	特征集	P	R	F
BiLSTM	规则特征+语义特征	89.19	83.19	86.09
CNN	规则特征+语义特征	84.21	80.67	82.40
SVM	规则特征	82.20	79.51	81.86

学习到文本中隐含的更抽象的特征。此外，BiLSTM 和 CNN 能够利用词嵌入挖掘事件句隐含的深层语义信息，然后整合规则特征，从而整体性能优于 SVM。另外，BiLSTM 与 CNN 相比，P, R 和 F 值分别提高了 4.98%, 2.52%和 3.69%，这是因为从模型角度来看，CNN 利用局部卷积思想，只能捕获文本序列的局部特征，缺乏对全局上下文信息的捕获能力；而 BiLSTM 的前向 LSTM 的序列信息从前往后依次传播，用于捕获上文的特征信息，反向 LSTM 的序列信息从后往前依次传播，用于捕获下文的特征信息，然后通过融合捕获的上文特征信息和下文特征信息最终获得全局的上下文信息。因而 BiLSTM 的整体性能优于 CNN。

4.2.6 BN 算法对 BiLSTM 收敛速度的影响　为验证 BN 算法对 BiLSTM 收敛速度的影响，将图 3 中的模型与去掉 BN 处理后的模型作比较，观察两个模型迭代过程中的 F 值，实验结果如图 5 所示。

图 5 表明，未使用 BN 算法时，模型在迭代 133 次后收敛，F 值稳定为 83.04%；使用 BN 算法后，模型迭代 90 次后收敛，同时 F 值达到 86.09%。由此可知，在 LSTM 单元结构加入 BN 处理后，模型的收敛速度明显提高，且 F 值提升了 3.05%，原因在于 BN 算法对式(1)、式(2)、式(3)和式(5)的非线性变换的输入进行了归一化操作，使它们的均值为 0、方差为 1，防止了输入数据的分布发生改变，而重新学习这个新的数据分布，从而可以使用大的学习率对网络结构进行训练，加快模型的收敛速度。

4.2.7 本文实验结果和其他学者实验结果比较　为了全面衡量本文提出方法的有效性，我们与以往研究方法结果做了对比，如表 6 所示。虽然本文标注的实验语料语言和标注结构与表 6 中其他研究者的不一样，直接与他们的实验结果相比较不具有可比性，但将本文实验结果和他们在事件因果关系抽取上的最好实验结果做一个对照，虽可比性不强，却可以给读者提供一些有用的信息。

表 6 中所列的方法，除付剑锋等人[10]的方法和钟军等人[11]的方法能够理论上抽取所有类别的因果关系外，其余学者的方法都只能抽取特定的事件因果关系(带标记的或句内的等)。与钟军等人[11]将事件类别简单地分为行动类、知觉类、声明陈述类、紧急情况类和状态更改类 5 个类别相比，本文将事件类别划分得更详细，从而导致事件因果关系抽取更加复杂，但准确率、召回率和 F 值比钟军等人[11]的方法分别高出 3.8%, 5.66%和 4.82%。总体而言，本文提出的方法能够有效抽取包括无标记因果关系在内的各类维语事件因果关系，实验效果明显。

表 6 其他学者事件因果关系抽取实验结果列表(%)

研究者	语言	P	R	F
钟军[11]	维语	**85.39**	**77.53**	**81.27**
付剑锋[10]	汉语	89.20	81.70	85.30
Girju[7]	英语	73.91	**88.69**	80.63
本文	维语	**89.19**	**83.19**	**86.09**

5　结束语

事件因果关系是一类重要的语义关系，研究维语事件因果关系抽取有助于维语事件预测、评估和问题回答等自然语言处理技术的进一步发展。现有的研究主要关注汉语、英语等大语种，而对维语这种小语种的事件因果关系抽取还不够成熟，且现有的研究只能覆盖文本中的部分类型的因果关系，未考虑整个事件句隐含的语义信息。针对以上不足，本文提出了基于双向 LSTM 的维吾尔语事件因果关系抽取方法。与以往研究方法相比，该方法将维语因果关系的抽取问题转化为对维语事件对分类的三分类问题，然后提取出事件句中的维语词汇序列，引入词嵌入作为维语词汇的特征，代入双向 LSTM 提取事件句中隐含的深层语义特征；同时通过对维语事件因果关系以及维语词干词尾、语序和格语法等语言特征的研究，提取出 10 项维语事件内部结构特征；最后融合这两类特征作为 softmax 分类器的输入进而完成事件因果关系抽取。实验结果表明，本文提出的算法能够很好地抽取维语事件间各类因果关系，且抽取的范围覆盖全文，总体效果较好。

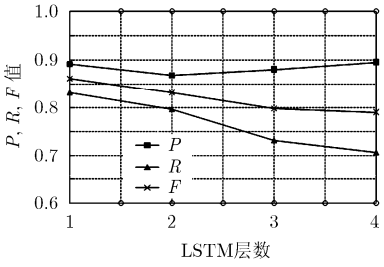

图 4　LSTM 层数对因果关系抽取的影响

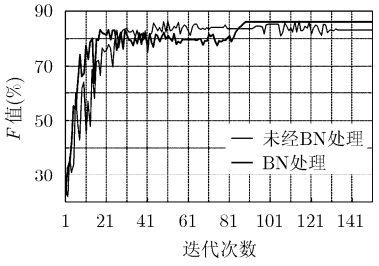

图 5　BN 算法对模型收敛速度的影响

参 考 文 献

[1] SHA L, LI S J, CHANG B B, *et al*. Joint learning teplates and slots for event schema induction[C]. Proceedings of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies, California, USA, 2016: 428–434. doi: 10.18653/v1/N16-1049.

[2] CHEN C and NG V. Joint modeling for chinese event extraction with rich linguistic features[C]. Proceedings of COLING 2012, Mumbai, India, 2012: 529–544.

[3] CHEN Y, XU L, LIU K, *et al*. Event extraction via dynamic multi-pooling convolutional neural networks[C]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 2015: 167–176. doi: 10.3115/v1/P15-1017.

[4] ZHANG Y J, LIU Z T, and ZHOU W, Event recognition based on deep learning in Chinese texts[J]. *Plos One*, 2016, 11(8): e0160147. doi: 10.1371/journal.pone.0160147.

[5] CHANG C Y, TENG Z Y and ZHANG Y. Expectation-regulated neural model for event mention extraction[C]. Proceedings of NAACL-HLT, California, USA, 2016: 400–410. doi: 10.18653/v1/N16-1045.

[6] 干红华, 潘云鹤. 一种基于事件的因果关系的结构分析方法[J]. 模式识别与人工智能, 2003, 16(1): 56–62. doi: 10.3969/j.issn.1003-6059.2003.01.011.
Gan Honghua and Pan Yunhe. A new analysis of the structure of event causation[J]. *Pattern Recognition and Artiflcial Intelligence*, 2003, 16(1): 56–62. doi: 10.3969/j.issn.1003-6059.2003.01.011.

[7] GIRJU R and MOLDOVAN D. Text mining for causal relations[C]. Proceedings of the 15th International Florida Artiflcial Intelligence Research Society Conference, Florida, USA, 2002: 360–364.

[8] MARCU D and ECHIHABI A. An unsupervised approach to recognizing discourse relations[C]. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, USA, 2002: 368–375. doi: 10.3115/1073083.1073145.

[9] SORGENTE A, VETTIGLI G, and MELE F. Automatic extraction of cause-effect relations in natural language text[C]. Proceedings of the 13th Conference of the Italian Association for Artificial Intellgence, Rome, Italian, 2013: 37–48.

[10] 付剑锋, 刘宗田, 刘炜, 等. 基于层叠条件随机场的事件因果关系抽取[J]. 模式识别与人工智能, 2011, 24(4): 567–573. doi: 10.3969/j.issn.1003-6059.2011.04.016.
FU Jianfeng, LIU Zongtian, LIU Wei, *et al*. Event causal relation extraction based on cascaded conditional random flelds[J]. *Pattern Recognition and Artiflcial Intelligence*, 2011, 24(4): 567–573. doi: 10.3969/j.issn.1003-6059.2011.04.016.

[11] 钟军, 禹龙, 田生伟, 等. 基于双层模型的维语突发事件因果关系抽取[J]. 自动化学报, 2014, 40(4): 771–779. doi: 10.3724/SP.J.1004.2013.00771.
Zhong Jun, Yu Long, Tian Shengwei, *et al*. Causal relation extraction of uyghur emergency events based on cascaded model[J]. *Acta Automatica Sinica*, 2014, 40(4): 771–779. doi: 10.3724/SP.J.1004.2013.00771.

[12] TANG D, QIN B, FENG X, *et al*. Effective LSTMs for target-dependent sentiment classification[C]. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 2016: 3298–3307.

[13] ZHOU P, SHI W, TIAN J, *et al*. Attention-based bidirectional long short-term memory networks for relation classification[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 207–212. doi: 10.18653/v1/P16–2034.

[14] 贺宇, 潘达, 付国宏. 基于自动编码特征的汉语解释性意见句识别[J]. 北京大学学报(自然科学版), 2015, 51(2): 235–240. doi: 10.13209/J.0479-8023.2015.041.
HE Yu, PAN Da and FU Guohong. Chinese explanatory opinionated sentence recognition based on autoEncoding features[J]. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 2015, 51(2): 235–240. doi: 10.13209/J.0479-8023.2015.041.

[15] MIKOLOV T, SUTSKEVER I, CHEN K, *et al*. Distributed representations of words and phrases and their compositionality[C]. Proceedings of Advances in Neural Information Processing Systems, Vancouve, Canada, 2013: 3111–3119.

[16] FENG X C, HUANG L F, TANG D Y, *et al*. A language-independent neural network for event detection[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 66–71. doi: 10.18653/v1/P16-2011.

[17] SUTSKEVER I, VINYALS O, and LE Q V. Sequence to sequence learning with neural networks[C]. Proceedings of Advances in Neural Information Processing Systems, Quebec, Canada, 2014: 3104–3112.

[18] IOFFE S and SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]. Proceedings of the 32th International Conference on Machine Learning, Lille, France, 2015: 448–456.

田生伟： 男，1973年生，教授，研究方向为自然语言处理、计算机网络技术.

周兴发： 男，1990 年生，硕士生，研究方向为自然语言处理.

禹 龙： 女，1974年生，教授，研究方向为计算机网络技术、人工智能.

冯冠军： 男，1972年生，副教授，研究方向为中国古典文学和维语研究.

艾山·吾买尔： 男，1981年生，副教授，研究方向为自然语言处理.

李 圃： 女，1972 年生，副教授，研究方向为语言学.