北京大学学报(自然科学版) 第 53 卷 第 3 期 2017 年 5 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 53, No. 3 (May 2017)
doi: 10.13209/j.0479-8023.2017.005

融合多模型与高置信度词典的事件线索检测

陈亚东 洪宇† 王潇斌 杨雪蓉 姚建民 朱巧明

苏州大学江苏省计算机信息处理重点实验室, 苏州 215006; † 通信作者, E-mail: tianxianer@gmail.com

摘要 提出一种融合多模型和高置信度词典的事件线索识别方法, 将高置信度词典特征分别加入最大熵模型和条件随机场模型, 然后融合两个模型的结果, 旨在提高触发词识别的召回率和整体性能。针对事件真伪性识别任务, 进一步考察否定词或不确定词与触发词的物理位置距离和依存路径距离等特征, 提高事件真伪性识别的性能。实验结果显示, 针对触发词识别和事件真伪性识别任务, 与仅使用最大熵模型相比, 所提出的融合多模型与高置信度词典的方法能够提高触发词识别的性能6.43%, 提高事件真伪性识别的性能1.69%。

关键词 事件线索检测; 最大熵模型; 条件随机模型; 高置信度词典

中图分类号 TP391

Combining Multiple Models and High-Confidence Dictionary for Event Nugget Detection

CHEN Yadong, HONG Yu†, WANG Xiaobin, YANG Xuerong, YAO Jianmin, ZHU Qiaoming

Provincial Key Laboratory of Computer Information Processing Technology, Soochow University, Suzhou 215006;
† Corresponding author, E-mail: tianxianer@gmail.com

Abstract This paper proposes a method that combines multiple models and high-confidence dictionary for event nugget detection. This method introduces dictionary features into maximum entropy model and conditional random fields model respectively, then combines the results of two models. In addition, the lexical length and the length of the dependency path between the trigger and negation or speculation in event realis recognition are considered to improve the accuracy of event realis detection. Compared to the method based on maximum entropy model, the experiment results show that proposed method can get 6.43% gain of F1 in event nugget recognition and 1.69% gain of F1 in event realis recognition.

Key words event nugget detection; Maximum Entropy; Conditional Random Fields; high-confidence dictionary

事件是一种描述特定人、物、事在特定时间和特定地点相互作用的客观事实。事件线索检测(event nugget detection)作为信息抽取的一个重要研究方向, 旨在从包含事件信息的自由文本中自动抽取触发事件的文本片段, 并辨别所抽取事件的真伪性。根据 KBP2015[①](knowledge base population)事件线索检测的任务定义, 事件线索检测包含两个子任务: 触发词识别(event nugget recognition)和事件真伪性识别(event realis recognition)。

触发词识别需要抽取触发事件的词或短语, 并正确识别出事件的类型。事件真伪性识别要求在触发词被正确识别的基础上, 进一步辨别事件的真伪性。例 1 给出一个完整的事件线索的结构表述, 其中, 触发词“born (译文: 出生)”表示事件类型Be-Born (表示出生事件), 其事件真伪性为Actual (表示真实发生的事件)。

例 1 Jack was **born** in England on July 20, 1984.
译文: 杰克于1984年7月20日在英格兰出生。

国家自然科学基金(61373097, 61272259, 61272260)资助
收稿日期: 2015–11–26; 修回日期: 2016–03–23; 网络出版日期: 2017–04–24

① http://www.nist.gov/tac/2015/KBP/Event/index.html

触发词→born, 对应事件类型“Be-Born”: born→Be-Born, 对应事件真伪性“Actual”: born→Actual。

触发词识别的方法主要沿用事件抽取中的方法。Ahn[1]针对触发词识别任务, 整合MegaM和Timbl两种机器学习方法, 首先使用MegaM分类器对当前词进行二元分类, 判断其是否是触发词, 然后使用多元分类器Timbl指定当前词所属的事件类别。Ji等[2]沿用Yarowsky[3]的“单片断单语义”假设, 提出跨篇章的事件抽取方法, 将事件抽取的范围从单文档引申到话题相关的文档集合中, 并且使用基于规则的方法, 解决了跨句子和跨文档的触发词识别问题。Liao等[4]提出文档级别的跨事件推理方法, 充分利用相关事件的内容信息和事件类型一致性等特征, 在触发词识别和解决事件歧义性方面得到很好的效果。Hong等[5]提出利用跨实体推理进行事件抽取, 其核心是充分利用实体类型的一致性特征, 进一步提升了触发词识别的性能。Chen等[6]针对中文事件抽取使用联合模型, 将事件触发词识别和类型判断以及事件元素识别和角色判断作为两个整体任务, 进而防止错误传递的情况。Li 等[7]针对中文事件抽取存在大量未登录触发词问题, 提出一个结合中文词语的形态结构和义原来推测未知触发词的方法, 旨在提高中文事件抽取系统的召回率。Li等[8]采用基于结构化感知机的联合模型, 将触发词识别和事件元素识别视为一个任务, 分析并检验了多种局部和全局特征, 达到目前事件抽取性能的最高值。针对事件真伪性识别任务的研究较少, Chen等[9]在事件共指任务中, 通过考察触发词所处上下文的词汇特征、实体特征、否定词和不确定词等特征, 考察事件的真伪性。

目前, 事件线索检测的方法主要沿用 ACE2005① (automatic content extraction)事件抽取的方法。根据ACE2005的任务定义, 事件抽取主要包含两个子任务: 触发词识别(trigger identification)和事件成员识别(argument identification)。事件线索检测的触发词识别沿用事件抽取的触发词识别定义, 但事件抽取的触发词识别方法并非直接移植到事件线索检测中的触发词识别任务。原因在于, 现有的事件抽取的触发词识别方法往往忽视短语作为触发词的情况, 而事件抽取英文语料中短语作为触发词的比例远小于事件线索检测中的比例。例如, 事件抽取中触发短语占所有触发词的比例为4.12%, 而事件线索检测中的比例为12.95%。此外, 现有的基于最大熵模型(Maximum Entropy)的触发词识别方法往往精确率较高而召回率较低。

本文提出融合最大熵模型和条件随机场模型(Conditional Random Fields, CRF)结果的方法, 旨在提高触发词识别的召回率和整体性能。CRF模型作为序列标注模型, 可以良好地识别出短语作为触发词的情况, 且融合两个模型的结果可以利用两个模型的差异性召回更多的触发词, 进而提高触发词识别的召回率和整体性能(F1值)。

高置信度词典特征指待测词在训练集中具有最大概率的非空事件类型, 例如待测词“strike” (译文“击打”)的事件类型分布为“Demonstrate (表示破坏事件) 63.64%, Attack (表示攻击事件) 27.27%, O (表示空事件) 9.09%”。本文将置信度大于阈值的事件类型Demonstrate作为该待测词的高置信度词典特征, 原因在于, 事件线索检测语料规模较小, 易造成特征稀疏, 而现有的基于监督学习的触发词识别方法虽然引入各种层面的特征来提高模型的泛化能力, 但也造成待测词先验事件分布对结果的影响减小。反之, 如果仅考虑待测词的先验事件分布而忽视其他层面的特征, 又难以提高触发词识别的泛化能力。本文将高置信度词典特征引入最大熵模型和条件随机场中, 分别针对两个模型构建不同的词典特征表示, 以期进一步提高触发词识别的性能。

针对事件真伪性识别的研究尚处于初步阶段, Chen等[9]在事件共指任务中对事件真伪性进行研究, 分析不同层面特征对事件真伪性识别的影响, 但没有考虑伪事件中否定词和不确定词的作用范围, 仅考虑当前句中是否包含否定词或不确定词。例如句子“恐怖分子可能袭击商场导致商家离开”中, 待测词“袭击”和“离开”分别触发Attack事件和Transport事件, 不确定词“可能”只作用于Attack事件, 而不作用于Transport事件。因此, 本文通过考察否定词或不确定词与触发词的物理位置之间的距离和依存路径距离等特征, 表征否定词或不确定词的作用范围, 以期进一步提高事件真伪性识别的性能。

1　任务描述

事件线索检测是知识库构建(knowledge base

① https://www.ldc.upenn.edu/collaborations/past-projects/ace

population, KBP)的子任务。事件线索检测旨在从含有事件信息的非结构化源文本中, 自动抽取出触发事件的词或短语, 并对事件的真伪性进行识别。下面给出事件线索检测中相关术语的定义。

事件触发词(event nugget) 触发事件的词或短语(常为名词性或动词性的词或短语)。

事件类型(event type) 描述事件的类型, 共包含8种事件类别以及33种子类别。本文针对33种事件子类型进行考察, 不考虑类别之间的层次关系。

事件真伪性(event realis) 描述事件发生的真伪性, 包含3种类别: Realis, Generic和Other。其中Realis表示在特定时间和地点确实发生的事件; Generic 表示泛化事件, 即不要求在特定时间和地点发生, 例如句子"Use of the death penalty is rare in Indonesia." (译文: 印度尼西亚的死刑很少)的触发词"death"表示的Die事件(表示死亡事件)即为泛化事件; Other表示没有发生的事件, 通常与否定词和不确定词共同出现。

事件描述(event mention) 包含事件触发词的短语或者句子。

事件线索检测需要正确抽取一个事件的描述信息, 包含触发词、事件类型和事件真伪性。下面用例 2 予以说明。

例2 The terrorists **killed** 7 people and **injured** 20 people in Baghdad.

译文: 在巴格达, 恐怖分子杀死7人同时伤害20人。

事件线索检测应该正确识别例2中待测词"killed"和"injured"分别触发Die事件(表示死亡时间)和Injure事件(表示伤害事件), 其中Die事件和Injure事件的真伪性都为Realis (表示事件真实发生), 例 2 的抽取结果如表 1 所示。

表 1 事件线索检测结果

Table 1 Result of event nugget detection

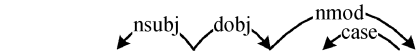
序号	触发词	事件类型	真伪性
1	killed	Die	Realis
2	injured	Injure	Realis

2 融合多模型和高置信度词典的事件线索检测方法

本文集中于事件线索检测的两个子任务: 触发词识别和事件真伪性识别。我们提出的融合多模型和高置信度词典的方法主要用于触发词识别任务。

2.1 触发词识别

触发词识别任务要求不仅正确识别出触发事件的词或短语, 并且能够正确判断出所触发的事件类型。表 2 描述触发词识别各个层面的特征, 主要包含词汇层面、句法层面和高置信度词典层面特征, 下面以例 3 中的"killed"为待测词, 举例说明各个层面特征的用法。



例 3	Jack	killed	Tom	in	Wednesday.
Pos :	NNP	VBD	NNP	IN	NNP
Chunk :	B-NP	B-VP	B-NP	B-PP	B-NP
Label :	O	Die	O	O	O

例 3 中的Pos, Chunk和Label分别表示该句词性标记、语块标记以及事件类型标记。表 2 中的词汇层面主要考察待测词的词形、词性、上下文以及同义词特征。其中, 待测词的同义词集指该词在WordNet[10]中具有最大概率的同义词集合, 该特征可以减小由于语料规模较小导致的特征稀疏情况。句法层面考察待测词的语块特征、与待测词具有依存关系的词及词性等特征, 以便在句法层面更好地表征触发词。本文将高置信度词典特征引入到最大熵模型和CRF模型中, 旨在提高待测词的先验事件分布对结果的影响。原因在于, 事件线索检测语料规模较小, 易产生特征稀疏的情况, 其他层面特征的引入在一定程度上可以提高模型的泛化能力。相应地, 待测词在训练集中的先验事件分布对结果的影响会减小。反之, 如果不考虑其他层面的特征, 又会导致模型的泛化能力较弱。因此, 本文在传统层面特征的基础上引入高置信度词典层面特征。在最大熵模型和CRF模型中使用高置信度词典特征的方式不同, 下面分别予以介绍。

现有的基于最大熵模型的触发词识别方法往往只考虑单个词作为触发词的情况, 即只对单个词进行分类, 得到该词的事件类型。本文采用MALLET①的最大熵模型。在最大熵模型中的高置信度词典特征指该词在训练集上的先验事件分布中, 具有最大概率的非空事件类型。例如, 待测词"killed"的先验事件分布为"Die 95.36%, Injure 2.32%, O 2.32%", 事件类型Die的概率大于置信度阈值, 因此将事件

① http://mallet.cs.umass.edu

表 2　触发词识别的特征

Table 2　Features of trigger identification

特征层面	特征标记	特征描述	待测词“killed”的特征
词汇层面	W_0	待测词本身	W_0=killed
	Lem	待测词的词形	Lem=kill
	Pos	待测词的词性	Pos=VBD
	Synonym	待测词的同义词集(WordNet)	Synonym=kill
	W_i (i=–2, –1, 1, 2)	待测词左/右2个词	W_{-1}=Jack, W_1=Tom, W_2=in
	Pos_i (i=–2, –1, 1, 2)	待测词左/右2个词的词性	Pos_{-1}=NNP, Pos_1=NNP, Pos_2=IN
句法层面	Chunk	待测词的语块特征	Chunk=*B*-VP
	DepW	与待测词具有依存关系的词	DepW=nsubj_Jack, DepW=dobj_Tom
	DepPos	与待测词具有依存关系的词的词性	DepPos=nsubj_NNP, DepPos=dobj_NNP
高置信度词典层面(最大熵模型)	Dict	待测词在高置信度词典中对应的事件类型	Dict=Die
高置信度词典层面(条件随机场模型)	B_i (i=–1, 0, 1)	在当前位置, 根据高置信度词典, 采用最大匹配获得的词或短语对应的事件类型	B_0=*B*-Die
	M_i (i=–1, 0, 1)	根据高置信度词典, 采用正向最大匹配获得的经过当前位置但不以当前位置结尾的短语对应的事件类型	空特征
	E_i (i=–1, 0, 1)	在当前位置, 根据高置信度词典, 采用逆向最大匹配获得的词或短语对应的事件类型	E_0=*B*-Die
	$B_{-1}B_0B_1$, $M_{-1}M_0M_1$, $E_{-1}E_0E_1$	B_i, M_i以及E_i的Tri-gram表现形式	$B_{-1}B_0B_1$=0|*B*-Die|0, $E_{-1}E_0E_1$=0|*B*-Die|0

类型Die作为待测词“killed”的高置信度词典特征。

与最大熵模型将触发词识别看做单个词的分类任务不同, CRF模型将触发词识别看做序列标注任务, 这表示触发词不再局限于单个词, 也有可能是短语。因此, 本文利用MALLET的CRF模型, 根据传统序列标注模型中的BIO模式[11]对触发词进行训练, 例如“The murder went to prison last week.”中短语“went to prison”为触发词并且触发Die事件, 则短语“went to prison”的标记为“B-Arrest-Jail I-Arrest-Jail I-Arrest-Jail” (Arrest-Jail表示逮捕入狱事件), 而非触发词的标记为“O” (表示空事件)。在CRF模型中高置信度词典特征也考虑短语作为触发词的情况, 其形式化定义如下所示, 给定句子 $x=c_1...c_n$ 以及高置信度词典 D, 考虑其中第 j 个单词 $c_j(1\leqslant j\leqslant n)$, w 表示触发事件的词或短语。

$$B(x,j,D)=\underset{\text{event}}{\text{argmax}}\ l,$$
$$\text{s.t. } w=c_j\cdots c_{j+l-1}\in D,\ j+l-1\leqslant n,$$
$$M(x,j,D)=\underset{\text{event}}{\text{argmax}}\ l,$$
$$\text{s.t. } w=c_s\cdots c_{s+l-1}\in D,\ j<s+l-1\leqslant n,1\leqslant s<j,$$
$$E(x,j,D)=\underset{\text{event}}{\text{argmax}}\ l,$$
$$\text{s.t. } w=c_{j-l+1}\cdots c_j\in D,1\leqslant j-l+1。$$

$B(x,j,D)$ 表示句子 x 在 j 位置, 根据高置信度词典 D, 采用正向最大匹配获得的词或短语对应的事件类型; $M(x,j,D)$ 表示句子 x 在 j 前面的某个位置, 根据高置信度词典D, 采用正向最大匹配获得的经过 j 位置但不以 j 结尾的短语对应的事件类型; $E(x,j,D)$ 表示句子 x 在 j 位置, 根据高置信度词典 D, 采用逆向最大匹配获得的词或短语对应的事件类型。

现有的基于监督学习的触发词识别方法召回率较低。融合最大熵模型和CRF模型的结果, 可以通过CRF模型识别短语类型触发词的优势以及利用两个模型的差异性, 提高触发词识别的召回率和整体性能。融合两个模型结果的方法主要依赖以下规则。

1) 如果两个模型对于同一触发词判断的事件类型不同, 则优先选取置信度大的且大于50%的非空事件类型作为最终结果, 否则判定事件类型为空

(CRF模型的置信度可以通过边缘概率得到)。

2) 如果CRF模型的触发短语包含最大熵模型中的触发词, 例如“pass away”包含“pass”, 则只保留条件随机场模型的结果。

3) 在符合规则 1 和 2 的情况下, 将两个模型结果的并集作为最终结果。

2.2 事件真伪性识别

事件真伪性识别要求在正确识别触发词及其事件类型的基础上, 进一步对事件的真伪性进行识别。本文采用MALLET①的最大熵模型来实现事件真伪性识别。表 3 给出事件真伪性识别的特征, 其中, 基本层面参考Chen等[9]实现, 扩展层面为本文提出的对现有特征进行的扩充。基本层面中仅考察触发词前面是否包含否定词或不确定词, 这会导致以下情况无法识别。例如“恐怖分子可能袭击商场导致商家离开”中待测词“袭击”和“撤离”分别触发Attack事件(表示攻击事件)和Transport事件(表示转移事件), 仅判断当前句中触发词前面是否包含否定词或不确定词, 会导致Transport事件被错误地判断为伪事件。因此, 根据位于触发词前面最近的否定词或不确定词与触发词之间的物理位置距离或依存路径距离, 对否定词或不确定词的作用范围进行考察。例 3 中“killed”到“Wednesday”的物理位置距离为 3, 依存路径距离为 2 (依存路径为dobj-nmod)。

根据Zou等[12]针对否定域和不确定域识别任务总结的情感词典, 在Chen等[9]所用特征的基础上, 分别扩充否定词和不确定词, 旨在提高否定词和不确定词的覆盖范围。其中, 扩充不确定词 27 个(例如if, whether和perhaps等), 扩充否定词 13 个(例如refute, fail 和 never等)。由于Chen等[9]没有在事件真伪性识别中考虑句法层面特征, 因此本文也对与触发词具有依存关系的词的词形、词性和实体特征进行考察。

3 实验

3.1 实验数据与评测方法

本文针对事件线索检测中的触发词识别和事件真伪性识别任务, 选取 KBP2015 的 151 篇英文文档作为实验语料, 共包含 5425 个句子, 其中有2197句包含事件描述。为防止训练过拟合, 采用 4 倍交叉验证, 即每次选取 121 篇作为训练集, 剩余的 30篇作为测试集。

表 3 事件真伪性识别的特征

Table 3 Features of event realis identification

特征层面	特征标记	特征描述
基本层面	Lem	触发词的词形
	Pos	触发词的词性
	Event	触发词的事件类型
	Lem_i (i=–2, –1, 1, 2)	触发词左/右 2 个词的词形
	Pos_i (i=–2, –1, 1, 2)	触发词左/右 2 个词的词性
	FirstVerbLem	当前句中第一个动词的词形
	FirstVerbPos	当前句中第一个动词的词性
	Time	当前句中描述时间的词
	NegWord	当前句触发词前面是否含有否定词
	ModalWord	当前局触发词前面是否含有情态动词
扩展层面	PhyDistance	触发词与前面最近的否定词或不确定词的物理位置的距离
	DepDistance	触发词与前面最近的否定词或不确定词的依存路径的距离
	NegWords	根据情感词典扩充表示否定的词, 例如 refute, never 和 fail 等
	Uncertain	根据情感词典扩充表示可能性的词, 例如 if, whether 和 perhaps 等
	Dep	与触发词具有依存关系的词的词形、词性和实体

① http://mallet.cs.umass.edu

触发词识别任务不仅要求抽取触发词，并且要求触发词表示的事件类型被正确识别。在触发词任务中，系统采用准确率(*P*)、召回率以及F1值作为评价指标。事件真伪性识别任务要求在触发词被正确识别的基础上，判断事件的真伪性。由于语料中没有给出被错误识别的触发词的真伪性，因此，参照KBP2015评测的方法，本文仅对在触发词识别任务中正确识别出的触发词进行真伪性判断，即在事件真伪性识别中采用精确率(*A*)。

3.2　触发词识别实验系统设置

为验证本文提出的融合多模型和高置信度词典方法对触发词识别任务的有效性，本文给出以下实验对比系统。

ME　使用最大熵模型，特征为表 2 中的词汇层面和句法层面特征。

CRF　使用CRF模型，特征为表 2 中的词汇层面和句法层面特征，并在特征表示中设置当前词的事件状态仅依赖于前一个词的事件状态。

Combine　采用 2.1 节中的规则融合最大熵模型和CRF模型的结果。

ME_Dict　在ME系统的基础上，引入表 2 中适用于最大熵模型的高置信度词典层面特征。

CRF_Dict　在CRF系统的基础上，引入表 2 中适用于CRF模型的高置信度词典层面特征。

Combine_Dict　根据2.1节中的规则融合ME_Dict系统和CRF_Dict系统的结果。

实验系统Combine旨在考察本文提出的融合最大模型和CRF模型结果的方法对触发词识别性能的影响，实验系统ME_Dict和CRF_Dict是为了验证高置信度词典特征分别对最大熵模型和CRF模型性能的影响，实验系统Combine_Dict在两个模型中分别引入高置信度词典特征，然后融合两个模型的结果，考察高置信度词典特征与融合结果方法的兼容性。

3.3　高置信度词典特征的阈值选择实验

为确定在最大熵模型和CRF模型中分别引入高置信度词典特征的最优阈值，本文首先进行置信度阈值选择实验。图 1 为两个模型根据不同置信度阈值引入词典特征后的结果。阈值为0表示待测词只要在训练集中触发过事件，则将其最大概率的非空事件类型作为高置信度词典特征。例如，待测词“started”的先验事件分布为“O 92.31%，Start-Org 7.69%” (O表示空事件，Start-Org表示创建公司事件)，当阈值取0时，该词的词典特征为事件类型Start-Org。从图 1 可以看出，当置信度阈值取0.4时，最大熵模型和CRF模型的性能均达到最高，分别为65.11%和62.57%。如果置信度阈值取值过小，会导致非触发词产生更多噪音特征。例如当待测词“started”所在的上下文环境并没有表示创建公司事件时，强制加入词典特征Start-Org会对结果产生错误影响。如果置信度阈值取值过大，又会造成词典特征的覆盖范围较小。例如当阈值取0.9时，满足条件的待测词仅有323个，占所有触发过事件的待测词数量的35.33%。因此，在触发词识别实验系统中，最大熵模型和CRF模型的词典特征的置信度阈值都设为0.4。

3.4　触发词识别实验结果与分析

本文针对触发词识别的实验结果如表 4 所示。实验系统Combine比实验系统ME和CRF的召回率分别提高2.69%和5.97%，F1值分别提高0.88%和2.94%，验证了融合最大熵模型和CRF模型结果的有效性。作为序列标注模型，CRF模型可以有效地

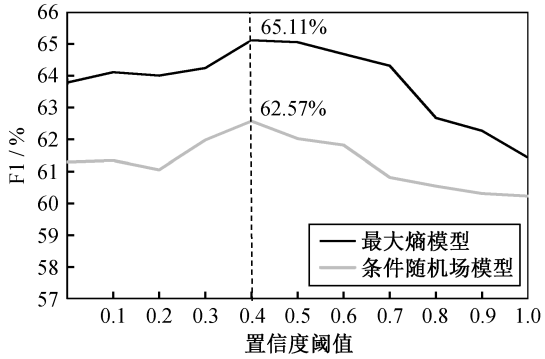


图 1　词典特征的置信度阈值选择
Fig. 1　Confidence threshold of dictionary features

表 4　触发词识别的实验结果
Table 4　Result of trigger identification

实验系统	触发词识别/%		
	P	*R*	F1
ME	75.34	48.52	59.03
CRF	76.90	45.24	56.97
Combine	72.19	51.21	59.91
ME_Dict	71.31	59.89	65.11
CRF_Dict	70.46	56.93	62.57
Combine_Dict	67.48	63.57	65.46

识别长文本的短语作为触发词的情况，而最大熵模型仅针对识别单个词作为触发词的情况。此外，最大熵模型和CRF模型的性能都受限于召回率较低的情况。因此，本文融合两个模型的结果，可以借助两个模型结果的差异性提高触发词识别的召回率和整体性能。然而，融合结果方法的准确率下降较大，与实验系统ME和CRF相比，准确率分别下降3.15%和4.71%。原因在于，融合两个模型的结果虽然可以提高召回率，但也会引入更多的错误结果，进而导致融合结果方法的准确率有所下降。

此外，实验系统ME比CRF的性能高2.06%。分析原因发现，CRF模型更易受到特征稀疏的影响：一方面因为触发词识别的语料规模较小，仅有151篇文档；另一方面由于CRF模型在特征训练时需要依赖前一个位置的状态，例如，针对当前待测词的特征集合，最大熵模型仅对应一个状态(比如，对应事件类型Attack)，而CRF模型不仅对应当前位置的状态，还需要考虑前一个位置的状态(比如，对应“O”和“Attack”，其中，O表示前一个位置是空事件，Attack表示当前位置的事件类型)。

实验系统 ME_Dict 比 ME 的性能提高 6.08%，相应地，实验系统 CRF_Dict 比 CRF 的性能提高 5.6%，说明了在最大熵模型和 CRF 模型中引入高置信度词典特征的有效性。分析原因发现，两个模型引入的各个层面的特征虽然极大地提高了模型的泛化能力(例如，可以根据不同待测词拥有相似的上下文环境，推断出两者触发相同事件)，但是更多层面特征的引入会造成触发词本身的先验事件分布对结果影响程度的下降。因此，将待测词在训练集中具有最大概率的非空事件作为高置信度词典特征，在一定程度上可以提高待测词先验事件分布对结果的影响。另一方面，如果仅考虑待测词在训练集中的事件分布，而不考虑各个层面的特征，又会导致模型没有泛化能力。为了进一步验证高置信度词典特征的有效性，本文通过表 5 和 6 分别给出在最大模型和 CRF 模型中特征权重为前 10 的特征。从表 5 可以发现，最大熵模型中特征权重占前 10 的特征有 9 个为高置信度词典特征(除第 8 个为待测词的词形特征外)。表 6 显示 CRF 模型中特征权重占前 10 的特征都 为高置信度词典特征，说明高置信词典特征无论在最大熵模型还是 CRF 模型中，都有很好的表征作用。

实验系统Combine_Dict的性能达到最优(F1=65.46%)，说明在两个模型中引入高置信度词典特征的基础上，进一步进行结果融合，仍然可以提高触发词识别的性能，验证了本文提出的高置信度词典特征与融合模型结果的方法之间具有良好的互适性。但是，实验系统Combine_Dict仅比ME_Dict的性能高0.35%，而实验系统Combine比ME高0.88%，实验系统Combine_Dict的性能提升较小。原因在于，引入高置信度词典特征后，虽然两个模型可以极大地提高召回率，但也会引入更多的噪音，造成准确率下降，进而导致融合结果的性能提升较小。

3.5 事件真伪性识别实验结果与分析

事件真伪性识别要求在触发词被正确识别的基础上，进一步判断事件的真伪性。由于在触发词识别任务中被错误识别的触发词没有标注真伪性，

表 5 最大熵模型中前 10 位的特征权重

Table 5 Top 10 features of Maximum Entropy Model

排名	特征名称	事件标记	特征权重
1	Dict=Attack	Attack	5.23
2	Dict=Transport-Person	Transport-Person	4.29
3	Dict=Transfer-Ownership	Transfer-Ownership	4.02
4	Dict=Transfer-Money	Transfer-Money	3.80
5	Dict=Die	Die	3.70
6	Dict=End-Position	End-Position	3.05
7	Dict=Injure	Injure	2.92
8	Lem=convict	Convict	2.88
9	Dict=Meet	Meet	2.78
10	Dict=Communicate	Communicate	2.60

表 6　CRF 模型中前 10 位的特征权重

Table 6　Top 10 features of CRF model

排名	特征名称	事件标记	特征权重
1	E_{-1}=*B*-Attack	*B*-Attack, O	1.95
2	B_{-1}=*B*-Attack	*B*-Attack, O	1.95
3	E_0=*B*-Attack	O, *B*-Attack	1.83
4	B_0=*B*-Attack=1	O, *B*-Attack	1.83
5	E_{-1}=*B*-Transfer-Money	*B*-Transfer-Money, O	1.77
6	B_{-1}=*B*-Transfer-Money	*B*-Transfer-Money, O	1.77
7	E_0=*B*-Transport-Person	O, *B*-Transport-Person	1.72
8	B_0=*B*-Transport-Person	O, *B*-Transport-Person	1.72
9	B_{-1}=*B*-Communicate	*B*-Communicate, O	1.70
10	E_{-1}=*B*-Communicate	*B*-Communicate, O	1.70

因此参照 KBP2015 的评测，本文仅对在触发词任务中被实验系统 Combine_Dict 正确识别的事件进行真伪性判断。本节专注于 2.2 节事件真伪性识别中引入扩展层面特征对性能的影响，表 7 中，序号 2~6 为在基本层面特征的基础上，引入扩展层面中的每一种特征对事件真伪性识别性能的影响，序号 2，7 和 8 为使用序列前向选择算法(sequential forward selection, SFS)寻找最优特征子集的中间结果性能。简单地说，SFS 算法就是每次都选择一种使分类性能达到最优的特征加入。结果显示，在基本层面的基础上引入特征 PhyDistance 时，性能提升 1.42%，说明触发词与前面最近的否定词或不确定词的物理位置的距离可以良好地表征否定词或不确定词的作用范围。例如“恐怖分子可能袭击商场导致商家离开”中触发词“袭击”和“撤离”分别表示 Attack 事件和 Transport 事件，由于触发词“撤离”与不确定词“可能”的物理位置的距离较大，可以避免判断 Transport 事件为伪事件。然后，继续加入特征 Dep，性能相较于基本层面提高1.62%。原因在于，与触发词具有依存关系的词的词形、词性和实体可以更好地表示该触发词所处的上下文环境。例如，与多个实体具有依存关系的触发词更有可能触发真实发生的事件。最后，加入特征Uncertain，性能达到最优(F1＝50.93%)，原因在于，原有事件真伪性识别的不确定词范围较小。本文参照Zou等[12]提供的情感词典，提高不确定词的覆盖范围，进一步提高性能。然而，特征 DepDistance和 NegWords 无法通过 SFS 算法继续提高事件真伪性识别的性能。原因在于，这两种特征单独加入到基本层面时对性能提高作用不大，如果强制加入到最终特征集合中，反而会引入较多噪音，因此在SFS 算法中继续引入这两种特征，性能并没有提高。

表 7　事件真伪性识别的实验结果

Table 7　Result of event realis identification

序号	实验特征及其组合	事件真伪性识别/%
1	基本层面	49.24
2	基本层面+PhyDistance	50.66 (+1.42)
3	基本层面+DepDistance	49.60 (+0.36)
4	基本层面+NegWords	49.41 (+0.17)
5	基本层面+Uncertain	49.81 (+0.57)
6	基本层面+Dep	50.31 (+1.07)
7	基本层面+PhyDistance+Dep	50.86 (+1.62)
8	基本层面+PhyDistance+Dep+Uncertain	50.93 (+1.69)

说明：括号中数字表示相较于基本层面提升的性能百分比。

4 总结

本文针对事件线索检测的触发词识别和事件真伪性识别任务, 提出融合多模型与高置信度词典的方法。该方法首先将高置信度词典特征分别引入最大熵模型和 CRF 模型中, 然后融合两个模型的结果, 通过 CRF 模型识别短语类型触发词的优势以及利用两个模型的差异性, 提高触发词识别的召回率和整体性能。针对事件真伪性识别任务, 提出引入否定词或不确定词与触发词的物理位置等特征, 考察否定词或不确定词的作用范围, 以期提高事件真伪性识别的性能。

然而, 融合最大熵模型和 CRF 模型的结果会引入较多的噪音, 导致触发词识别的准确率较低。在以后的工作中, 将尝试融合 3 个以上模型的集成学习方法, 并对每个模型使用关联性较小的特征集合, 以减小融合结果时准确率减小的情况。针对现有的事件线索检测语料规模较小的情况, 尝试引入主动学习或半监督学习的方法, 通过扩充实验语料, 减小特征稀疏的情况, 以期提高事件线索检测的性能。

参考文献

[1] Ahn D. The stages of event extraction // Proceedings of ACL 2006 Workshop on Annotating and Reasoning about Time and Events. Sydney, 2006: 1–8

[2] Ji H, Grishman R. Refining event extraction through cross-document inference // Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics(ACL). Colunbus, 2008: 254–262

[3] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods // Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL). Stroudsburg, 1995: 189–196

[4] Liao S, Grishman R. Using document level cross-event inference to improve event extraction // Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL). Uppsala, 2010: 789–797

[5] Hong Y, Zhang J, Ma B, et al. Using cross-entity inference to improve event extraction // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL). Portland, 2010: 1127–1136

[6] Chen C, Ng V. Joint modeling for chinese event extraction with rich linguistic features // Proceedings of COLING 2012. Mumbai, 2012: 529–544

[7] Li P, Zhou G. Employing morphological structures and sememes for Chinese event extraction // Proceedings of COLING 2012. Mumbai, 2012: 1619–1634

[8] Li Q, Ji H, Huang L. Joint event extraction via structured prediction with global features // Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL). Sofia, 2013: 73–82

[9] Chen Z, Ji H, Haralick R. A pairwise event coreference model, feature impact and evaluation for event coreference resolution // Proceedings of the Workshop on Events in Emerging Text Types. Borovets, 2009: 17–22

[10] Miller G A. WordNet: a lexical database for English. Communications of the ACM, 1995, 38(11): 39–41

[11] 张梅山, 邓知龙, 车万翔, 等. 统计与词典相结合的领域自适应中文分词. 中文信息学报, 2012, 26(2): 8–12

[12] Zou B, Zhou G, Zhu Q. Tree kernel-based negation and speculation scope detection with structured syntactic parse features // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing(EMNLP). Seattle, 2013: 968–976