面向新冠新闻的三阶段篇章级事件抽取方法

郭 鑫[1],高彩翔[1],陈 千[1,2],王素格[1,2],王雪婧[1]
1.山西大学 计算机与信息技术学院,太原 030006
2.山西大学 计算智能与中文信息处理教育部重点实验室,太原 030006

摘 要:事件抽取是信息抽取领域的一个研究热点。在新冠肺炎疫情常态化下,利用事件抽取技术可以筛选出有价值的信息。然而事件抽取领域缺乏精标注的新冠新闻训练数据集,且因部分事件的复杂性,论元不只存在于一句话中,需要多个句子才能完整描述一个事件。因此,首先构建新冠肺炎新闻数据集,接着提出一种三阶段的管道方法实现从篇章中抽取新冠肺炎事件。该方法对数据集进行事件类型分类;进行事件句的抽取;实现篇章级论元抽取。实验结果表明提出的方法能够减少事件分类时间,抽取两个事件句的条件下,对数据通报类论元识别效果最好,准确率、召回率和F1值达到75.0%、73.0%,和74.0%,证明方法能有效抽取新冠肺炎相关篇章级事件。

关键词:新冠肺炎;信息抽取;事件句抽取;篇章级事件抽取
文献标志码:A **中图分类号:**TP391 **doi:**10.3778/j.issn.1002-8331.2108-0353

Three-Stage Document-Level Event Extraction for COVID-19 News

GUO Xin[1], GAO Caixiang[1], CHEN Qian[1,2], WANG Suge[1,2], WANG Xuejing[1]
1.School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
2.Key Laboratory of Ministry of Education for Computation Intelligence and Chinese Information Processing of Shanxi University, Taiyuan 030006, China

Abstract: Event extraction is a hot research in the field of information extraction. In the face of COVID-19, event extraction technology can filter out valuable information. However, there is a lack of well-labeled news data set for COVID-19 in event extraction. Moreover, due to the complexity of some events, arguments do not only exist in one sentence, and multiple sentences are needed to fully describe an event. Therefore, an COVID-19 news events data set is constructed, and a three-stage pipeline method is proposed. It classifies event types, then it extracts event sentences. Finally, the document-level event arguments extraction is realized. The experimental results show that the model can reduce event classification time. When extracting two event sentences, the recognition effect for notification type event argument is the best. The accuracy rate, recall rate and F1 value reaches 75.0%, 73.0% and 74.0%, which proves that proposed method can effectively extract document-level COVID-19 events.

Key words: COVID-19; information extraction; event sentence extraction; document-level event extraction

新冠疫情于2019年底爆发,面对该公共卫生突发事件,各国政府积极采取应对措施,新闻媒体实时聚焦疫情事件报道,世界各国学者迅速投身新冠病毒及防治领域的研究。如何从海量的新闻中梳理出疫情发展的脉络,成为科研人员研究的热点问题。

作为信息抽取的一个子任务,事件抽取[1]旨在从非结构化数据中快速获取关键的结构化事件信息。事件抽取主要分为两个任务:事件类型检测[2]、事件论元抽取[3]。事件类型检测是识别句子中的触发词,接着对触发词分类,即对这句话所包含的事件进行分类;事件论元抽取是基于已经获取的事件触发词及事件类型,去识别事件中其余的事件相关论元。

事件检测中的触发词是指句子中能让一个事件发生的核心词语,触发词所对应的类别就是该句子当中所

基金项目:山西省应用基础研究计划(201901D111032);国家自然科学基金(61502288,61403238)。
作者简介:郭鑫(1982—),女,博士,副教授,CCF会员,主要研究方向为自然语言处理、特征学习;高彩翔(1994—),女,硕士研究生,主要研究方向为事件抽取;陈千(1983—),通信作者,男,博士,博士后,副教授,CCF会员,主要研究方向为自然语言处理、文本挖掘,E-mail:chenqian@sxu.edu.cn;王素格(1964—),女,博士,教授,主要研究方向为自然语言处理、情感分析;王雪婧(1997—),女,硕士研究生,主要研究方向为信息抽取。
收稿日期:2021-08-20 **修回日期:**2021-10-08 **文章编号:**1002-8331(2023)03-0150-08

包含的事件类别;事件论元则通常指一个事件的参与者,事件论元角色则是该参与者在事件当中所代表的具体含义[4]。事件抽取任务中存在一些挑战,对于简单事件可以直接从一句话中抽取出事件相关信息。但是对于部分复杂事件而言,句子级抽取不能涵盖事件的全部论元,需要从多个句子中才能完整地抽取出整个事件。如图1所示,s_1中"捐赠"触发一个爱心捐赠事件,"中国政府""加拿大""11日"在该事件中分别扮演捐赠方、接收方、时间的事件角色。但是只抽取出部分事件论元,在s_2中补充了具体的捐赠物,即"医用防护服""护目镜""口罩""隔离衣"。s_1和s_2组合抽取出该事件中所包含的所有事件论元以及在该事件中所扮演的角色,从而组成一个完整的事件:

s_1:中国政府向加拿大捐赠的医疗物资11日晚运抵加拿大。

s_2:此次捐赠的医疗物资包括医用防护服、护目镜、口罩和隔离衣等。

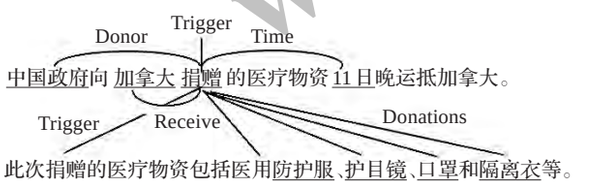

图1　爱心捐赠类篇章级事件抽取实例

Fig.1　Doc-level event extraction sample of caring donation

从已有研究近况来看,面向新冠肺炎新闻的篇章级事件抽取还存在如下问题:(1)事件抽取主要还是集中在从单一句子中抽取事件;(2)大部分已有的事件联合抽取工作都是基于ACE 2005数据集进行实验,该数据集仅在句子范围内标记事件联合模型,而且不包含面向新冠肺炎新闻数据集。

针对以上问题,本文的贡献在于以下三点:

(1)本文通过爬虫技术,构建了近6 644条基于篇章级的面向新冠肺炎的精标注新闻数据集以及15万左右的未标注新冠肺炎新闻数据集。

(2)通过改进的TextRank算法抽取关键的事件句,接着利用序列标注从篇章级角度进行事件抽取,进而获取到更加全面的事件信息。

(3)提出一种三阶段的管道方法,结合有监督和无监督模型,在降低人力成本的同时,将句子级事件抽取任务扩展到篇章级,实验证明该方法采用的篇章级事件抽取技术在新冠新闻数据集上的F1指标达到74%,从而验证了方法的有效性。

1　相关工作

已有的事件抽取方法大体上分为三大类:模式匹配方法、机器学习方法、深度学习方法。早期事件抽取方法采用模式匹配技术,首先构造特定事件模板,然后通过模板匹配从文本中提取事件。Riloff等人[5]通过建立触发词词典和事件匹配模式进行事件识别与抽取,但手动标注事件模式耗时费力,需要领域专家的指导。近年来基于机器学习的事件抽取技术得到迅速发展。Li等在2013年[6]和2014年[7]提出基于结构预测的事件抽取联合模型。Liu等人[8]研究了事件与事件关联和主题与事件关联两种全局信息。机器学习方法不仅需要人工设计特征,还需要借助外部NLP工具抽取特征,特征抽取过程中会产生误差。随着深度学习技术的兴起,端对端的神经网络模型被广泛应用于事件抽取。Zeng[9]和Liu[10]分别结合CNN和BiLSTM来进行事件触发器检测。Wu等人[11]应用参数信息训练BiLSTM网络的注意力来进行事件抽取。Chen等人[12]提出了一个HBTNGMA模型用于提取和融合句内和句间上下文信息,增强事件检测。

事件抽取也可以分为句子级和篇章级。现阶段事件抽取的研究主要基于ACE 2005数据的句子级事件抽取任务上,Chen等人[13]提出了一种动态多池卷积神经网络来评估句子的每个部分,捕获句子最重要的信息。Feng等人[14]基于递归神经网络对输入句子进行序列建模来获取整个句子的上下文信息。Nguyen等人[15]提出一种基于RNN的事件识别和角色分类联合学习模型。Miao等人[16]提出CNN-BiGRU模型,通过CNN获取词级别特征,BiGRU获取句子级特征。Ding等人[17]提出分层语义融合模型。Wu等人[18]提出FB-Latiice-BiLSTM模型,对仅能捕获字粒度语义信息的BiLSTM-CRF模型进行词语和实体维度的信息增强,但句子级的事件抽取会造成论元角色的缺失,忽略重要事件信息。近年来篇章级事件抽取也有所突破。Huang等人[19]利用基于管道的方法进行篇章级事件抽取。仲伟峰等人[20]提出基于自注意力机制的实体事件联合标注模型。Yang等人[21]基于句子抽取结果,利用上下文元素补齐策略得到篇章事件结构化信息。Du等人[22]将文档级事件角色填充符提取形式化为端到端序列标记问题。关于新冠事件抽取,Dimitrov等人[23]构建了COVID-19的语义标注Tweets语料库。Wang等人[24]提出从Twitter中抽取COVID-19事件。

综上,现有的新冠数据集大多数是基于英文语料库,且篇章级事件抽取任务存在输入篇幅过长的问题。本文构造中文新冠新闻数据集来扩充语料库,并提出一种事件句抽取的三阶段篇章级事件抽取方法。

2　三阶段COVID-19事件抽取方法

2.1　整体框架

本文提出一种基于三阶段的管道方法来实现篇章级的事件抽取。图2描述了事件抽取模型的总体架构。模型主要包含3个阶段:(1)事件类型识别,利用无

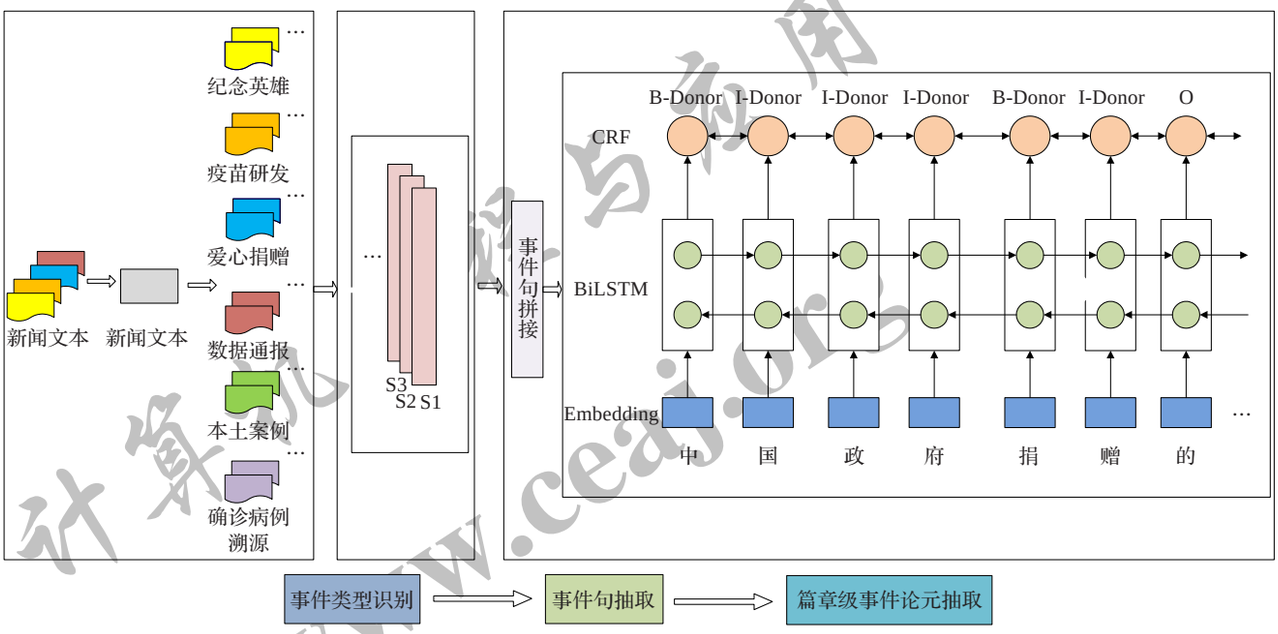

图2　篇章级事件抽取模型框架

Fig.2　Framework of document-level event extraction model

表1　新冠肺炎新闻事件类型

Table 1　COVID-19 news event types

事件类型	触发词	论元角色
确诊病例溯源	确诊、就诊、发现、新增、报告、确认……	确诊时间、发现案例城市、确诊病例人数、是否死亡
数据通报	新增、增加、确诊……	时间、发现案例城市、新增确诊人数、新增疑似人数、新增死亡人数
本土案例	确诊、就诊……	报道时间、确诊国家、确诊地点、确诊数量、状况
纪念英雄	奋战、奋斗、坚守、会诊、奔赴、驰援……	英雄姓名、职业、单位、是否殉职
爱心捐赠	提供、捐赠、募捐、支援、收到、捐助……	时间、捐赠方、接收方、受捐方、捐赠总数
疫苗研发	研发、完成、研制、启动、采用、开展……	时间、机构名、阶段、项目名称

监督算法进行事件类型分类;(2)事件句抽取,基于改进的TextRank算法进行含有论元的事件句抽取;(3)篇章级事件论元抽取,利用BiLSTM-CRF的序列标注模型,对事件句进行预测标注,采用拼接技术完成篇章级事件抽取。最终利用论元补充得到完整事件信息。

2.2　事件类型定义

已抓取的新冠肺炎新闻数据集根据预定义事件类型采用多人协同标注方式。根据卡帕值将事件大概分为六类,分别是确诊病例溯源、数据通报、本土案例、爱心捐赠、疫苗研发和纪念英雄,在每个大类下建立不同的触发词和与之对应的论元角色。如表1所示。

2.3　事件类型识别

事件类型识别是发现事件的触发词并为其分配预定义的事件类型。只有识别出事件类型,才能指导事件句的抽取,并进行相应事件的要素抽取。

输入是一个文档的集合 $D=\{d_1,d_2,\cdots,d_l\}$,同时还需要聚类的类别个数为 t;然后算法会将每一篇文档 d_i 在所有的主题上分布一个概率值 p;这样每篇文档都会得到一个概率的集合 $d_i=\{d_{p_1},d_{p_2},\cdots,d_{p_t}\}$,通过概率值来对每篇文档进行聚类。

为了在大量的新闻中快速区分各种事件类型,采用无监督聚类算法来分类。LDA和KMeans模型被广泛应用于文本分类。

2.4　基于衰减机制的TextRank事件句抽取

2.4.1　事件句分布统计

一般新闻类的文章,事件句出现在段首或者段尾的情况相当普遍,大多数都采用先总后分的方式,并且段落之间存在联系可能不是那么紧密,但是段落本身结构更紧凑的情况。通过统计事件句在篇章中的分布,得到如图3的汇总情况。

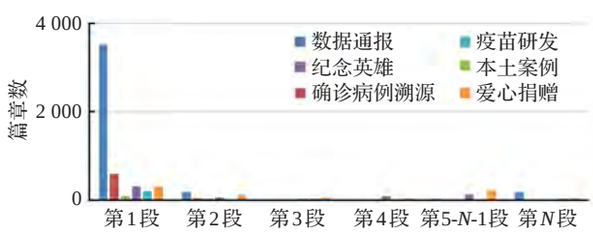

图3　新冠新闻事件句分布情况

Fig.3　Distribution of COVID-19 news event sentences

2.4.2　事件句抽取

事件句是包含事件触发词和事件论元的句子。由

于事件句基本上是一篇新闻中较为重要的摘要句,为了从大量新闻中快速找到事件句,采用TextRank算法[25]来抽取。TextRank算法是一种基于图的用于关键词抽取和候选句抽取的排序算法,通过把文本分割成若干个句子,构建节点连接图,用句子之间的相似度作为边的权重,通过循环迭代计算句子的TextRank值,给新闻正文的每个句子进行打分,之后选取排名靠前的 k 个句子最后抽取排名高的句子,作为文本候选句。然而原始的TextRank算法仅利用文本自身的信息进行抽取,不能充分利用词语之间的语义相关等信息。基于此,提出基于衰减机制的词嵌入TextRank算法。根据图3统计可知,事件句往往会在段落开始位置提及。设计加权因子decay_rate,即对每个段落的段首做适当的加权,往后逐渐衰减,呈梯度模型。decay_rate取值范围为(0,1],值越小,倾斜越大,默认为1代表无任何倾斜。

词语间语义信息可以通过预训练的词向量来实现。首先对给定的篇章级文本 d 进行了断句处理,利用规则的方法进行断句 $d_i=\{s_1,s_2,\cdots,s_m\}$;然后利用jieba分词技术将句子进行分词,得到每个句子的词集合 $s_i=\{w_1,w_2,\cdots,w_n\}$;接着利用Word2vec将文档集中所有词汇进行向量表征,在生成词向量之后,基于Word2vec模型实现句子中每个词语相似度的计算进而计算句子相似度。即根据 s_1 的句子,找到 s_1 中第一个词语在 s_2 所有词语中最大相似值的词语,再依次找到 s_1 中第二个,第三个,直到第 n 个词语在 s_2 所有词语中最大相似值的词语,取平均值作为 s_1 和 s_2 句子的单项匹配pipei_reverse (s_1,s_2);接着,同理反过来计算 s_2 和 s_1 句子的单项匹配pipei_reverse (s_2,s_1);最后取双向匹配的平均值作为 s_1 和 s_2 的句子相似度。

2.5　基于BiLSTM-CRF的新冠新闻事件抽取

为了实现对新冠肺炎新闻事件的抽取,本文构建基于BiLSTM-CRF的事件抽取模型。双向长短期记忆网络(BiLSTM)具有捕获数据的时序性和解决长序列信息依赖问题的优点,能主动学习新冠肺炎新闻事件的抽象特征和提高检测性能。条件随机场层(CRF)使用条件随机场模型对全连接层的输出进行解码,能有效地考虑了序列前后的标签信息,通过学习标签间的约束条件提升标签预测的准确性,得到最终的预测标签序列。事件抽取模型的具体步骤如下:

在预处理阶段,本文采用word-embedding将文本的每个字符映射成一个字符向量,即输入向量 $s=\{x_1,x_2,\cdots,x_n\}$,其中 n 表示该句中字符个数,x_i 表示文本中每一个维度的数据。

首先,将 s 作为神经网络的输入,得到输入层输出向量 $O_i=\{o_1,o_2,\cdots,o_n\}$。其次,将 O_i 输入到BiLSTM层前向的LSTML,通过前向学习输出特征向量 $q=\{q_1,q_2,\cdots,q_n\}$,q_n 为经过BiLSTM层后每一维度的数据;将 O_i 输入到BiLSTM层后向的LSTMR,通过后向学习输出特征向量 $\boldsymbol{h}=\{h_1,h_2,\cdots,h_n\}$;将前向特征 $\boldsymbol{q}$ 和后向特征 $\boldsymbol{h}$ 进行拼接,得到BiLSTM提取出的抽象特征 $\boldsymbol{b}=[\boldsymbol{q}:\boldsymbol{h}]=\{q_1,q_2,\cdots,q_n,h_1,h_2,\cdots,h_n\}$。然后,经过softmax层,得到网络输出结果,做论元角色类别的分类处理。最后,加上CRF层融合。CRF层的作用在于加入一些约束来保证最终预测结果是有效的。目标是让真实序列的概率在整个序列所有概率中最大。最终得到预测标签序列 $y=\{y_1,y_2,\cdots,y_n\}$。当前序列得分为:

$$score(x,y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n+1}A_{y_{i-1},y_i} \tag{1}$$

式中,y_i 是第 i 个位置的标签值;P_{i,y_i} 是第 i 个位置softmax输出为 y_i 的概率;A_{y_{i-1},y_i} 为 y_{i-1} 到 y_i 的转移概率。最后利用softmax计算归一化后的概率,公式为:

$$P(y|x)=\frac{\exp(score(x,y))}{\sum\exp(score(x,y'))} \tag{2}$$

采用最大化对数似然函数优化目标函数,训练样本 (x,y) 的对数似然为:

$$\ln(y|x)=score(x,y)-\ln(\sum\exp(score(x,y'))) \tag{3}$$

在预测时,使用动态规划的Viterbi算法求解最优路径,得到序列标注任务中每个字对应的标签概率,最大概率对应的标签即为正确标签,概率公式为:

$$y^*=\arg\max_{y'}score(x,y') \tag{4}$$

采用BIO标注法对事件进行标注。B_label代表字符为触发词或论元的开始位置,I_label代表字符为触发词或论元的中间位置,O_label代表字符为非触发词或论元。

3　实验设计与结果分析

3.1　数据集

实验使用的面向新冠肺炎新闻数据集来自于网络信息,借助网络爬虫技术从信息门户网站、论坛等地获取新冠疫情相关的新闻语料,包括山西省人民政府网、CCTV新闻网、各省卫健委网站、中国新闻网等。其中中国新闻网上爬取的新闻占80%(约5 500条新闻数据)。这些网站上的新闻舆情通常都是紧跟时事热点、内容完整度较高、主题较明确的高质量文本信息,对事件抽取模型具有较好训练作用。作为实验的数据,根据定义的事件模型,对所有语料进行标注。其中,标注内容包括:事件类型、事件触发词、事件论元、事件角色,以及事件触发词和事件论元在文本中的位置信息。

在进行标注前,首先对篇章级文本进行了断句处理,包含“。”“?”“;”“!”符号的位置进行断句处理。然后多人一起进行标注,标注的时候同时进行交叉检验,保

表2 事件类型分类检验

Table 2 Event type classification test

R_1	R_2						
	确诊病例溯源	数据通报	本土案例	爱心捐赠	疫苗研发	纪念英雄	求和
确诊病例溯源	698	5	7	0	0	0	710
数据通报	3	3 935	2	4	4	20	3 968
本土案例	3	13	170	3	2	0	191
爱心捐赠	0	7	1	789	1	2	800
疫苗研发	0	3	0	2	323	2	330
纪念英雄	0	2	0	2	0	641	645
总计	704	3 965	180	800	330	665	6 644

证了标注数据的质量。具体标注过后最终的生成格式为.ann格式,其中每列分别对应:标号、论元角色、起始位置、结束位置、具体论元。如图4所示。

```
T1    Time 0 3   28日
T2    Donor 5 9 中国政府
T3    Donations 12 16    抗疫物资
T4    Recipient 20 27    马来西亚吉隆坡
T5    Donate 9 11    捐赠
T6    Donations 36 47    10万人份核酸检测试剂
T7    Donations 48 59    10万只N95医用口罩
T8    Donations 60 70    50万只医用外科口罩
T9    Donations 71 79    5万件医用防护服
T10   Donations 80 90    200台便携式呼吸机
```

图4 新闻数据集.ann格式

Fig.4 News dataset .ann format

卡帕值用于计算标注者之间标注结果的吻合程度。如表2所示,表2中统计了标注者的标注情况,第一行第一个数“698”表示R_1和R_2都判断为确诊病例溯源类行的个数,第一行第二个数表示“R_1判断为确诊病例溯源类而R_2判断为数据通报类”的个数。基于混淆矩阵的kappa系数计算公式如下:

$$kappa=\frac{p_o-p_e}{1-p_e} \tag{5}$$

$$p_o=\frac{\text{对角线元素之和}}{\text{整个矩阵元素之和}} \tag{6}$$

$$p_e=\frac{\sum_i \text{第}i\text{行元素之和}\times\text{第}i\text{列元素之和}}{\left(\text{矩阵所有元素之和}\right)^2} \tag{7}$$

p_e表示所有类别分别对应的“实际与预测数量的乘积”总和除以“样本总数的平方”。计算得到kappa值为0.978,因此将新冠肺炎新闻数据集分为6类,分别为确诊病例溯源、数据通报、本土案例、爱心捐赠、疫苗研发和纪念英雄。

其中关于各类新闻事件的数据分布中数据通报类数据占多数,本土案例和疫苗研发占比较少。采用无监督方法对新闻数据进行分类。按照8:2划分为训练集和测试集,数据集的统计基本情况如表3所示。

3.2 实验设置

事件类型分类阶段本文采用无监督聚类算法。判断一个LDA模型是否合理的标准一般有两个,一个是

表3 新冠肺炎新闻数据集统计情况

Table 3 Statistics on COVID-19 news dataset

事件类型	训练集	测试集	总计
确诊病例溯源	568	142	710(10.7%)
数据通报	3 174	794	3 968(59.8%)
本土案例	153	38	191(2.7%)
爱心捐赠	641	159	800(12.1%)
疫苗研发	264	66	330(5.0%)
纪念英雄	516	129	645(9.7%)

一致性(coherence),另一个是困惑度(perplexity)。对于LDA主题模型中的困惑度用于在语料库中确定合理的主题个数。

$$perplexity(D)=\exp\left(-\frac{\sum\ln p(w)}{\sum_{d=1}^{M}N_d}\right) \tag{8}$$

其中,M是测试语料库中的文本的数量,N_d是第d篇文本的单词数,$p(w)$代表文本的概率。

如图5、图6所示,通过困惑度和一致性这两个指标,确定最优的主题个数为7。

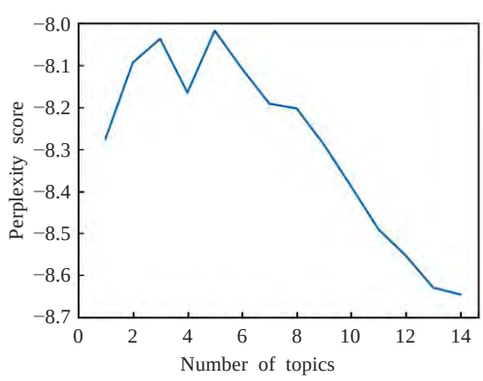

图5 困惑度随主题个数的变化情况

Fig.5 Perplexity varies with number of topics

事件句抽取阶段中本文采用jieba分词技术将句子进行分词,过滤掉文本中无意义的停用词,然后使用Gensim库中的Word2vec模块,设置维度为60、窗口大小为2对该数据集进行学习训练得到词向量模型文件。通过多次实验,选定的权重值decay_rate为0.1,达到效果最佳。

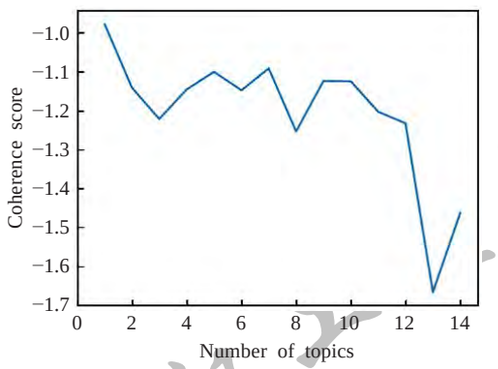

图6　一致性随主题个数的变化情况

Fig.6　Coherence varies with number of topics

事件论元抽取阶段本文实验基于TensorFlow框架,编程语言为Python 3.6。通过多次实验,选定的实验参数如下:优化器为Adam,learning-rate值为0.005,batch-size值为270,epoch值为20。

3.3　评价指标

本文改进的TextRank算法性能的评估中,算法为每篇文章都识别事件句,实验中使用准确率作为算法的评估指标。

$$准确率=\frac{系统正确识别的关键句数}{手工标注文档关键句数}\times 100\% \quad (9)$$

本文利用系统自动抽取信息,再对抽取结果进行评估。LDA、KMeans无监督聚类算法和论元抽取模型在实验中的评估度量方式一致,都是使用准确率(precision,P)、召回率(recall,R)、F1(F1-measure)值作为算法的评估指标。其中,论元抽取是基于句子级和篇章级的多分类任务,而论元是基于词级别的,所以词语的BIO标签预测都计入评估。精确率 P 是指分类结果预测为正确的数据占所有预测为正确的数据的比重,召回率 R 是指分类结果预测为正确的样本占所有真实为正确的样本的比重。其中,无监督聚类算法评估是看预测正确的篇章数;论元抽取模型算法评估是看预测正确的标签数。

3.4　无监督聚类实验

将已有的6 644条新闻数据利用LDA和KMeans模型进行聚类,定义主题个数 $num\_topic=7$,生成7个主题文档;将分类结果得到的7个主题文档和人工标注好的6类数据作比较,计算P、R、F1值,结果取每类占比最大的类别(Max)。实验结果如表4。

由表4可知,LDA能识别出5类数据。爱心捐赠、疫苗研发、纪念英雄、数据通报这四类新冠疫情新闻聚类效果要比其他两类好;确诊病例溯源、本土案例这两类新冠疫情新闻聚类效果相对较差一些。因为本土案例类数据较少,不易识别,而且本土案例和确诊病例溯源的事件论元较为相似,不易于区分,导致这两类易于混淆,聚类效果相对较差。

表4中,KMeans只能识别出4类数据,没有识别出易于混淆的本土案例和确诊病例溯源类数据。而且,从总的平均结果的评估指标来看,也是LDA算法优于KMeans算法。

由上述分析可知,选择利用LDA模型进行聚类,效果更佳。所以,爬取到的6 644条新冠新闻数据选择利用LDA模型进行聚类,可以更快速准确地获取到爱心捐赠、疫苗研发、纪念英雄、数据通报这四类新冠疫情新闻。部分易于混淆的本土案例类和确诊病例溯源类,可以加上人工干预,进行区分。

3.5　消融实验

为证明加入衰减机制的TextRank算法能提高抽取关键事件句的精度,将未加衰减机制和加入衰减机制的TextRank进行了消融实验。结果如表5所示。

表5中,2_Sents、3_Sents、5_Sents分别表示对每篇文档分别抽取了2、3、5个事件句进行比较。实验结果表明,利用TextRank算法抽取事件句和人工抽取的正确事件句进行论元识别相比,TextRank算法抽取事件句的效果欠佳。但确诊病例溯源类、数据通报类论元的识别效果相对较好,这是因为这两类事件构成相对简单,包含的修饰性词语较少,结构性较强,如许多数据通报类实体都包含“新增确诊病例”“新增疑似病例”“新增死亡病例”等词。实验结果表明,数据通报类事件在2句事件句抽取的论元识别效果最好。本土案

表4　LDA、KMeans模型的事件聚类

Table 4　Event clustering of LDA model and KMeans model

6类	LDA				KMeans			
	Max	P/%	R/%	F1/%	Max	P/%	R/%	F1/%
0	数据通报 2 257	93.8	52.8	67.6	数据通报 992	55.7	13.8	22.1
1	疫苗研发 301	92.4	84.5	88.3	数据通报 775	93.9	18.2	30.5
2	数据通报 1 713	85.5	36.6	51.2	爱心捐赠 842	85.7	91.2	88.4
3	爱心捐赠 1 042	80.2	92.4	85.7	数据通报 1 047	97.4	25.5	40.4
4	数据通报 124	94.4	2.90	5.70	纪念英雄 770	77.1	92.1	84.0
5	确诊病例 581	68.5	56.1	61.7	数据通报 1 951	85.9	41.8	56.3
6	纪念英雄 634	89.9	88.4	89.1	疫苗研发 275	96.4	80.5	87.7
汇总	数据通报 4 094	90.3	92.3	91.3	数据通报 4 765	83.4	99.3	90.7
平均	6类	86.4	59.1	64.2	6类	84.6	51.9	58.5

表5　抽取事件句2种算法准确率比较

Table 5　Comparison of accuracy of two algorithms for event sentences extraction

单位:%

事件类型	算法	2_Sents	3_Sents	5_Sents
本土案例	TextRank	36.1	44.2	60.7
	+decay_rate	60.7	77.9	78.2
纪念英雄	TextRank	12.1	16.5	23.4
	+decay_rate	16.5	17.8	29.5
疫苗研发	TextRank	35.5	43.4	58.7
	+decay_rate	49.7	51.5	51.9
确诊病例溯源	TextRank	36.6	48.0	58.7
	+decay_rate	37.7	38.1	39.0
爱心捐赠	TextRank	30.0	40.1	53.9
	+decay_rate	30.7	33.6	36.1
数据通报	TextRank	42.4	52.3	65.4
	+decay_rate	31.1	32.2	32.7
平均	TextRank	32.1	40.8	53.5
	+decay_rate	37.7	41.7	44.6

例、纪念英雄和疫苗研发类事件类型利用改进过的+decay_rate算法抽取事件句抽取,效果明显提升,且整体效果较好。确诊病例溯源和爱心捐赠类事件抽取2个事件句,改进过的+decay_rate算法效果更佳。总体来看,抽取2、3个事件句,使用改进过的+decay_rate算法效果更佳。

选择表5中抽取的事件句和人工精标注的正确事件句作为事件抽取的训练数据对六类新冠疫情事件进行实验。图7显示了不同标注类别的实体识别结果,及自动抽取事件句和人工抽取事件句对论元识别的影响。

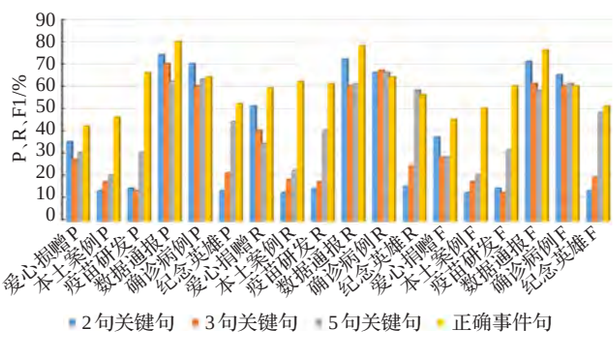

图7　关于不同事件类型的P、R、F1值

Fig.7　P,R,F1 values for different event types

实验结果表明,数据通报类事件在2句事件句抽取的论元识别效果最好,准确率(precision,用 P 表示)、召回率(recall,用 R 表示)和F1(用 F 表示)分别为75.0%,73.0%、74.0%。

3.6　篇章级事件句个数对比

为进一步确定篇章级事件抽取中关键事件句个数,分别比较了2、3、5句事件句对篇章级事件抽取的论元识别的F1指标效果影响,结果如图8所示。

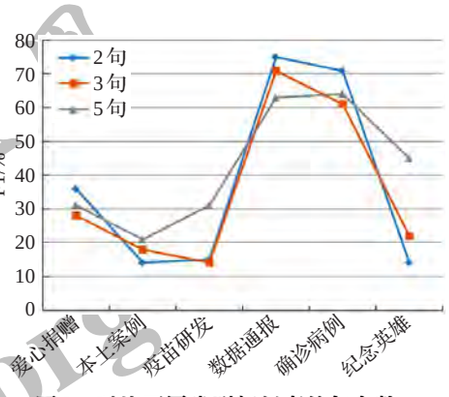

图8　对比不同类型新闻事件句个数

Fig.8　Number of sentences forvarious types of news events

实验结果表明,爱心捐赠、数据通报、确诊病例溯源类事件在2句事件句抽取中性能较好,抽取更多的事件句,会减低篇章级事件抽取的效果。

4　结论与展望

针对新冠肺炎领域的事件抽取任务中存在缺乏中文的新冠疫情新闻数据集和跨句抽取论元的问题,本文设计了三阶段的管道方法。并且通过人工参与的方式进行数据集标注。实验表明,该方法能够更快捷地进行事件抽取,对于数据通报、确诊病例溯源类事件在事件句抽取的论元识别效果较好。

本文主要聚焦面向新冠肺炎的篇章新闻数据,目前该领域未见公开数据集,因此未进行其他领域的公开数据集下的对比实验。在利用TextRank算法进行事件句抽取的过程中,发现爱心捐赠、纪念英雄等事件句的精度较低。对后续事件论元的抽取影响较大。在接下来的工作中,会考虑对TextRank算法进行改进,通过引入句子位置、句子相似度和论元词信息融合3个影响因素,以此计算句子之间的影响权重。进而提升事件句的抽取精度。

参考文献:

[1] ZHANG H,SONG H,WANG S,et al.A BERT-based end-to-end model for Chinese document-level event extraction[C]//Proceedings of the 19th Chinese National Conference on Computational Linguistics,2020:390-401.

[2] GUO H,WANG Z,LI P,et al.Semi-supervised method to cluster Chinese events on social streams[C]//Proceedings of the 19th Chinese National Conference on Computational Linguistics,2020:634-644.

[3] YU W,YI M,HUANG X,et al.Make it directly:event extraction based on tree-LSTM and Bi-GRU on tree-LSTM and Bi-GRU[J].IEEE Access,2020(8):14344-14354.

[4] 贺瑞芳,段绍杨.基于多任务学习的中文事件抽取联合模型[J].软件学报,2019,30(4):1015-1030.
HE R F,DUAN S Y.Joint Chinese event extraction based

on multi-task learning[J].Journal of Software,2019,30(4):1015-1030.

[5] RILOFF E.Automatically constructing a dictionary for information extraction tasks[C]//Proceedings of the 11th National Conference on Artificial Intelligence, 1993:811-816.

[6] LI P,ZHU Q,ZHOU G.Joint modeling of argument identification and role determination in Chinese event extraction with discourse-level information[C]//International Joint Conference on Artificial Intelligence,2013:2120-2126.

[7] LI Q,JI H,YU H,et al.Constructing information networks using one single model[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,2014:1846-1851.

[8] LIU S,LIU K,HE S,et al.A probabilistic soft logic based approach to exploiting latent and global information in event classification[C]//Thirtieth AAAI Conference on Artificial Intelligence,2016:2993-2999.

[9] ZENG Y,LUO B,FENG Y,et al.WIP event detection system at TAC KBP 2016 event nugget track[C]//Proceedings of the 2016 Text Analysis Conference,2016:1-5.

[10] LIU Y,LI Q,LIU X,et al.Document information assisted event trigger detection[C]//2018 IEEE International Conference on Big Data,2018:5383-5385.

[11] WU W,ZHU X,TAO J,et al.Event detection via recurrent neural network and argument prediction[C]//CCF International Conference on Natural Language Processing and Chinese Computing,2018:235-245.

[12] CHEN Y B,YANG H,LIU K,et al.Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,2018:1267-1276.

[13] CHEN Y B,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,2015:167-176.

[14] FENG X,QIN B,LIU T.A language-independent neural network for event detection[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,2016:66-71.

[15] NGUYEN T H,CHO K,GRISHMAN R.Joint event extraction via recurrent neural networks[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,2016:300-309.

[16] MIAO J,DUAN Y X,ZHANG Y Q,et al.Method for extracting event trigger words based on the CNN-BiGRU model[J].Computer Engineering,2021,47(9):69-74.

[17] DING L,XIANG Y.Chinese event detection with hierarchical and multi-granularity semantic fusion[J].Computer Science,2021,48(5):202-208.

[18] WU G L,XU J N.Chinese emergency event extraction method based on named entity recognition task feedback enhancement[J].Journal of Computer Applications,2021(7):1891-1896.

[19] HUANG E R.Peeling back the layers:detecting event role fillers in secondary contexts[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies,2011:1137-1147.

[20] 仲伟峰,杨航,陈玉博.基于联合标注和全局推理的篇章级事件抽取[J].中文信息学报,2019,33(9):88-95.
ZHONG W F,YANG H,CHEN Y B.Document-level event extraction based on joint annotation and global reasoning[J].Journal of Chinese Information Processing,2019,33(9):88-95.

[21] YANG H,CHEN Y B,LIU K,et al.DCFEE:a document-level Chinese financial event extraction system based on automatically labeled training data[C]//Proceedings of the Association for Computational Linguistics,2018:50-55.

[22] DU X,CARDIE C.Document-level event role filler extraction using multi-granularity contextualized encoding[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,2020:8010-8020.

[23] DIMITROV D,BARAN E,FAFALIOS P,et al.Tweets COV19-a knowledge base of semantically annotated tweets about the COVID-19 pandemic[C]//Proceedings of the 29th ACM International Conference on Information and Knowledge Management,2020:2991-2998.

[24] WANG C,LILLIS D.UCD-CS at W-NUT 2020 shared Task-3:a text to text approach for COVID-19 event extraction on social media[C]//Proceedings of the Sixth Workshop on Noisy User-Generated Text(W-NUT 2020),2020.

[25] MIHALCEA R,TARAU P.TextRank:bringing order into text[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,2004:404-411.