doi:10.3772/j.issn.2095-915x.2021.05.002

开放科学
（资源服务）
标识码
（OSID）

中医文献的句子级联合事件抽取

高甦[1] 陶浒[2] 蒋彦钊[2] 贾麒[2] 张德政[2] 谢永红[2]

1. 北京师范大学医院 北京 100875；
2. 北京科技大学计算机与通信工程学院 北京 100083

摘要：[目的 / 意义] 中医领域存在大量的文献，这些文献中含有大量中医诊疗的知识。但这些知识往往存在于非结构化文本中，通过信息抽取技术将其转化成结构化文本，不仅能够提高效率，还可以进一步推进中医智能辅助诊疗的发展。[方法 / 过程] 本文使用了联合事件抽取模型，利用 BERT 对中医文献进行字向量表示，并在此基础上加入 Self-Attention 层，最后与 CRF 模型结合，实现了对中医文献的句子级事件的初步抽取。[结果 / 结论] 通过实验与 Pipeline 模型进行对比，结果显示，本文使用的模型 F1 值较 Pipeline 模型提高了 14.2%。

关键词：事件抽取；信息抽取；中医文献；联合事件抽取模型

中图分类号：G35

Sentence-level Joint Event Extraction of Traditional Chinese Medical Literature

GAO Su[1] TAO Hu[2] JIANG Yanzhao[2] JIA Qi[2] ZHANG Dezheng[2] XIE Yonghong[2]

1. Beijing Normal University Hospital, Beijing 100875, China;
2. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

基金项目 国家重点研发计划云计算和大数据专项 “大数据驱动的中医智能辅助诊断服务系统”（2017YFB1002300）。

作者简介 高甦（1978-），硕士，研究方向为中西医结合临床专业；陶浒（1998-），硕士，研究方向为知识图谱、信息抽取；蒋彦钊（1995-），硕士，研究方向为实体识别、事件抽取、文本分类；贾麒（1989-），博士，研究方向为自然语言处理、知识图谱；张德政（1964-），博士，教授，研究方向为数据挖掘、知识工程，E-mail: zdzchina@ustb.edu.cn；谢永红（1970-），博士，副教授，研究方向为数据库技术、知识工程。

引用格式 高甦，陶浒，蒋彦钊，等. 中医文献的句子级联合事件抽取 [J]. 情报工程，2021, 7(5): 15-29.

doi:10.3772/j.issn.2095-915x.2021.05.002

Abstract: [Objective/Significance] There are many literatures in the field of Traditional Chinese Medicine (TCM), which contain massive knowledge of TCM diagnosis and treatment. However, the knowledge often exists in unstructured text. Transforming it into structured text through information extraction technology can not only improve efficiency, but also further promote the development of intelligent diagnosis and treatment with TCM. [Methods/Process] We use the joint event extraction model, representing the word vector of the TCM literatures with BERT. We add a Self-Attention layer on this basis, and finally combine with the CRF model to achieve the sentence-level events extraction from the TCM literature. [Results/Conclusions] The experiment is compared with the Pipeline model, and the results show that the F1 value of our model is 14.2% higher than that of the Pipeline model.
Keywords: Event extraction; information extraction; Traditional Chinese Medicine literatures; joint event extraction model

引言

中医领域存在大量的文献，这些文献中含有大量中医诊疗的知识。但这些知识大多都存在于非结构化的文本中，通过知识抽取将其转化成结构化数据，可以帮助进行中医辅助诊疗。但是由于文献众多，依赖人力进行抽取耗时耗力，为了提高中医临床诊疗的效率，利用现有的神经网络技术，可以对非结构化的文本进行有效的抽取，使之成为现代计算机可以处理的结构化数据。现有的中医诊断方法中大多包含多个要素并且每个要素之间是有相互联系的，三元组信息又不足以将信息完整表示，所以本文借鉴事件抽取技术，定义中医事件类型，对中医文献中的知识进行结构化处理进而实现信息抽取。

中医事件抽取旨在将中医文本中的一些诊断方法按一定的模式抽取出来。通过中医事件抽取，机器可以获取到文献中包含的各种诊断方法，为初步理解中医文献内容提供帮助，并为之后的机器辅助诊疗提供数据基础，从而为进一步的中医临床智能辅助诊断打下基础。

由于中医领域特性较为突出，传统事件类型无法准确地描述中医事件。中医语言表达简练的特点，一个句子中往往包含更多的中医事件要素。中医文本中有现代白话文也有古文，表达方式比较多样，因此语法层面存在一定差异，是异构的。另外，不同的中医事件之间存在语义接近导致难以区分的情况，例如“证候概念”和“证候分析”都属于描述中医证候的事件类型，但它们本质是不同的，这些因素导致中医事件抽取难度较大、挑战性更强。

通常一个事件是由一个句子描述的，这句话通常会包含事件的触发词以及事件的各种元素。采用管道模型（Pipeline）的方法对事件进行抽取的时候通常分为两步：首先进行事件触发词的识别，根据事件触发词定义事件类型，其次根据事件的类型进行事件论元的识别与分类。例如在“血脱证是指突然大量出血或长期反复出血，致使血液亡脱，以面色苍白、心悸、

脉微或芤为主要表现的证，又称脱血证。”这句话中，首先利用事件触发词识别模型识别出“指”这个触发词并可以判断其事件类型为证候概念；然后根据证候概念得知其中包含的论元为证候名称、概念以及其对应的别名；最后会利用事件论元识别模型识别事件所包含的论元并对其进行角色分类，例如在这句话中包含的事件论元证名为血脱证，而别名为脱血证。

然而，管道模型容易产生分级误差传递，使得上游任务的错误传播给下游任务，而且下游任务无法将信息反馈给上游任务，辅助上游任务提升识别的准确率，所以本文使用联合模型的方法对句子中的事件以及事件论元进行识别。在联合抽取模型中，由于让事件触发词检测和事件论元识别同时进行，所以不仅能够解决分级误差传递的问题，而且能够很好的捕捉到事件触发词和事件论元的相互依赖关系。本文利用中医中最具代表性的基础性文献《中医诊断学》、《中医内科学》以及多种疾病诊疗方案作为语料，针对中医文献探索句子级中医事件抽取的方法，其中主要的任务有事件触发词的识别以及事件论元的识别。本文设计了一个句子级别的事件抽取模型，利用预训练模型 BERT 作为文本的特征解码器，使用 BiLSTM 对上下文特征进行学习，之后通过 Self-Attention 机制强化事件触发词和事件论元之间的表达依赖，最后结合 CRF 对中医文献的事件进行抽取。我们在自己构建的中医事件抽取数据集上进行实验，最终我们的联合事件抽取模型在事件触发词识别任务上平均 F1 值达到 0.7696，在事件论元识别任务上平均 F1 值达到 0.7576，整体性能超过了主流的管道模型。

1 相关工作

事件抽取是信息抽取领域的一个重要研究方向。事件抽取旨在把句子中描述的事件，以结构化的形式表现出来。事件抽取作为信息处理领域的关键技术，在信息检索、自动问答、自动摘要、数据挖掘、文本挖掘等领域有着广泛的应用。

1.1 事件定义

事件作为信息的一种表现形式，其定义为特定的人、物在特定时间和地点相互作用的客观事实，通常事件的载体为句子 [1]。组成事件的各元素包括：事件触发词、事件论元、论元角色及事件类型。 其具体意义如下：

事件触发词：触发事件的核心词，通常是动词或者名词；

事件论元：事件所包含的用户感兴趣的语义对象，通常是一个名词；

论元角色：论元在事件中所扮演的角色，即为事件的参与对象；

事件类型：事件所属的事件类型，事件触发词和事件角色共同决定了事件的类别。

事件抽取任务可分解为 4 个子任务 [2]: 事件触发词识别、事件类型分类、论元识别和论元角色分类任务。其中，事件触发词识别和事件类型分类可合并成事件类别分类任务。事件论元识别和论元角色分类可合并成事件论元识别任务 [3]。

目前事件抽取的方法可以分为三类：模式匹配法、机器学习的方法和深度学习方法。其中模式匹配方法，指对于某类事件的识别和抽

doi:10.3772/j.issn.2095-915x.2021.05.002

取是在一些模式的指导下进行的[4]。模式主要用于指明构成目标信息的上下文约束环，集中体现了领域知识和语言知识的融合[1]。抽取时可利用模式匹配算法找出符合模式约束条件的信息。一个采用模式匹配法事件抽取系统的基本组成如图 1 所示。接下来主要介绍基于传统机器学习与深度学习的事件抽取方法。

图 1　模式匹配法事件抽取的基本组成

1.2　基于传统机器学习的事件抽取

采用机器学习的方法识别事件，就是借鉴文本分类的思想，将事件类别的分类及事件论元的识别转化成为分类问题，其核心在于分类器的构造和特征的选择。

对事件的识别主要包括事件类别分类以及事件论元识别两大核心任务。Hai 等[5]在处理讲座通告和人事管理事件的抽取问题中，运用最大熵模型用于事件论元的识别。H.Llorens 等[6]则利用条件随机场实现语义角色标注，有效地解决了 TimeML 的事件抽取问题。而 D.Ahn[7]结合 MegaM 和 TiMBL 两种机器学习方法分别完成了事件类别分类和事件论元识别两大任务，并通过实验证明二者结合的方法优于采用单一算法。

然而，以上提到的方法大多是基于触发词来进行事件的探测，这种方法的准确性十分依赖触发词识别的准确率，但实际运用中触发词只占所有词的小部分，这样会给训练过程引入大量反例从而使正反例不平衡，另外因为需要对每个词进行判断，这样无疑增加了额外计算。为了解决上述问题，赵妍妍[8]结合触发词扩展和二元分类的思路来识别事件类别，即在训练中将触发词收录在词典中并利用同义词林进行扩展，有效地处理了训练实例正反例不平衡以及数据稀疏造成的影响，在 ACE 的中文语料上取得较好的效果。而许红磊[9]与 Naughton[10]则基于事件实例，完成事件探测与抽取，这种方法面向句子，将句子作为识别实例，从而克服了上述基于触发词方法造成的正反例失衡和数据稀疏问题。

1.3　基于深度学习的事件抽取

传统基于机器学习的事件抽取方法均采用人工构建特征的方法表示每个候选触发词或每个触发词 – 实体对，然后借助基于统计的分类模型进行分类。以往方法重点在于研究有效的特征来捕捉候选触发词、候选论元的词法、句法等信息，并分别能够获取触发词 – 触发词之间的依赖关系、和同一触发词相关的多个实体之间的依赖关系、和同一实体相关的多个触发词之间的关系。此外，还有若干工作研究跨句子和跨文档的事件一致性特征[11]、跨事件的一致性特征[12]和跨实体的事件一致性特征[13]来提高事件抽取的性能。

近年来兴起的神经网络方法，能够自动学习有效特征。不同于传统的离散型特征，神经网络特征多是连续型向量特征。Chen 等[3]和 Nguyen 等[14]率先尝试将神经网络方法应用于事件抽取和识别任务中，通过实验结果验证了神经网络构建特征的有效性。更进一步地，

Chen 等 [3] 改进了传统的 CNN 模型，增加了动态多池机制 (dynamic multi-pooling) 来提高事件抽取性能，而秦美越等 [15] 在生物医学事件抽取任务上利用并行多池化 CNN 来提高性能。研究者认为，与最大池机制相比，将候选触发词和候选实体位置分割成 3 部分的动态多池机制，有助于提取更有效的特征用于角色分类。Nguyen 等 [16] 则使用一种基于离散短语的卷积神经网络模型来辅助事件识别，该模型能够获取更丰富的非连续短语的特征，而不局限于局部连续短语特征。另外，不少研究者采用循环神经网络 [17,18](Recurrent Neural Network，RNN) 建模文本序列，以支撑下游的事件抽取任务。Feng 等 [19] 基于 RNN 获取文本序列信息，利用一个卷积层来获取文本中的短语块信息，通过结合上述两种信息实现事件触发词识别，从模型特点来看，该混合模型是语言无关模型，因此在多种语言的事件识别任务中取得了良好效果。然而上述提到的事件抽取方法多数是将触发词的相关特征用于辅助论元角色分类任务，而忽略了论元信息对触发词识别任务的作用。借助注意力机制，Liu 等 [20] 采用有监督的论元注意力模型将论元信息融入事件识别模型中，通过实验证明论元信息能够提升事件触发词识别任务性能。

1.4 事件抽取模型分类

根据学习方式的不同，当前主要分为管道（Pipeline）模型和联合模型两种事件抽取框架 [21]。管道模型主要是将事件类别分类任务和事件论元识别任务按照流水线的方式进行；而联合模型将两个任务同时进行。

在管道模型中，后者输入依赖于前者输出。首先需要先经过触发词识别模型识别出句子中所包含的事件触发词并且对其进行事件类别分类，论元角色分类任务的输入是前一步识别出的触发词和整个句子。之前大部分事件抽取方法都是管道模型学习方式，先进行事件识别模型的学习，再进入论元角色分类模型的学习。这种学习方式会存在以下两个问题：（1）管道模型在进行触发词识别这个上游任务的过程中由于样本不平衡问题而导致识别准确率偏低；（2）管道模型容易产生级联错误，所以上游任务产生的误差会在下游任务进一步扩大，而且处于下游的任务无法将信息反馈至上游任务，辅助提高上游任务的精度。

为了减少上述两个问题所带来的影响，研究者们后续提出了联合事件抽取模型。在联合抽取模型中，由于事件类别分类任务与事件论元识别任务是同时进行的，所以能够解决管道模型产生的级联误差问题。由于联合抽取模型的训练集同时包含事件触发词标签和事件论元标签，所以在模型训练的时候能有效的捕捉事件触发词和事件论元相互的依赖关系，并且当一句话中带有更多有角色标签的实体，这样就解决了触发词识别中样本不平衡问题而导致识别准确率偏低问题。所以联合抽取模型对事件抽取两个子任务的准确率都有提升。

2 任务定义与数据说明

2.1 任务定义

根据 ACE 模型设计，事件的模型框架应该

doi:10.3772/j.issn.2095-915x.2021.05.002

包含事件触发词、事件类型、事件论元、论元角色。通常事件抽取是通过事件触发词进行事件类型的检测，再根据不同的事件类型来进行事件论元的识别。

事件定义是事件抽取的首要步骤，而事件粒度的大小对之后的抽取是非常重要的。我们的中医专家确定并定义了两种基本的中医主题事件，为了便于抽取与整合，又将其划分出六种元事件类型，并且规定了每种元事件所对应的事件论元，具体结构如下表 1 所示。

表 1　中医事件类型及其对应的事件论元

中医主题事件	元事件类型	事件论元	部分事件触发词
诊断主题	证候概念	证名	称、指、根据、俗称、为
		别名	
	临床表现	症状	表现为、描述为、特征是
		舌象	
		苔象	
		脉象	
	证候分析	病因病机	说明、提示、多因、因、所致
	辩证要点	常见病证	可见于、主、由于、多属
治疗主题	发病信息	发病日期	见于、分为
		发病分期	
		易发病人群	
	治法信息	治法名称	（方剂）为、(服法）为
		方药信息	
		用法信息	
		加减信息	

例如在“失神，又称无神”这句事件句中，其对应的事件类型以及事件论元分别如表 2 所示。

表 2　事件句对应其元素

类型	取值
事件句	失神，又称无神
触发词	称
事件类型	证候概念
证名	失神
别名	无神

本文研究的任务目标即从各类中医文本中，对以上两大类中医主题事件的六种事件类型进行抽取，达到中医知识结构化的目的。

2.2　数据说明及预处理

本文使用各类中医临床教材作为基础语料。《中医诊断学》中的内容包括如何通过望、闻、问、切四诊收集病情，也介绍了在收集病情的基础上如何对其进行病情分析、融合、判断，进而判断出疾病的种类和该种类下对应的证候。《中医内科学》则是主要讲述内科疾病的病因病机、症候特点、辨证论治及预防、康复的一门学科，是研究中医其他临床学科的基础教材。除此之外，本文使用的数据还包括新冠肺炎疫情期间，全国各地发布的各种肺炎诊疗方案。

利用 OCR 技术将获取的中医文献语料转化为文本，同时对结果进行去噪，包括删除乱码、修改错字等步骤，得到数据清洗后的语料。

针对不同的事件类型分别对《中医诊断学》《中医内科学》以及各种诊疗方案进行筛选，其中《中医诊断学》总共包含句子 4286 句，《中医内科学》包含句子 8296 条以及各种诊疗方案 2435 条。然后我们对这些文本进行了标注，数据集标注方式如下：

（1）触发词标注标准

在筛选出完整事件句之后，就要对其进行标注。不同的事件类型是根据不同的事件触发词的分类结果来确定的，由于本文使用的是根据字的序列标注模型来分类的，所采用的标注方法为 BIOES 的标注方法。因此，首先需要规定每类事件触发词的标签符号，不同于其他序列标注任务如实体识别，为了区分事件触发词和事件论元，采用 X-T-XX 和 X-A-XX 来进行区分。例如在“里证是指病变部位在内，脏腑、气血、骨髓等受病，以脏腑受损或功能失调症状为主要表现的证”中，“指”应标注为 S-T-ZHGN，而“里证”的标签则为：里（B-A-ZM）证（E-A-ZM）。在六个不同类别的事件中，ZHGN 表示证候概念、LCBX 表示临床表现、ZHFX 表示证候分析、BZYD 表示辩证要点、FBXX 表示发病信息、ZFXX 表示治法信息，其事件类型及其对应的事件触发词标签如下表 3 所示。

表 3　事件类型及其对应的事件触发词标签表

事件类型	单字触发词	多字触发词（首字）	多字触发词（中间）	多字触发词（尾字）
证候概念	S-T-ZHGN	B-T-ZHGN	I-T-ZHGN	E-T-ZHGN
临床表现	S-T-LCBX	B-T-LCBX	I-T-LCBX	E-T-LCBX
证候分析	S-T-ZHFX	B-T-ZHFX	I-T-ZHFX	E-T-ZHFX
辩证要点	S-T-BZYD	B-T-BZYD	I-T-BZYD	E-T-BZYD
发病信息	S-T-FBXX	B-T-FBXX	I-T-FBXX	E-T-FBXX
治法信息	S-T-ZFXX	B-T-ZFXX	I-T-ZFXX	E-T-ZFXX

（2）事件论元标注标准

我们依据定义的事件论元，为每个元事件类型划分论元角色。以一句中医事件句“表证通常表现为新起恶风寒，或恶寒发热，头身疼痛，喷嚏，鼻塞，流涕，咽喉痒痛，微有咳嗽、气喘，舌淡红，苔薄，脉浮”为例，根据“表现为”可以判断出该句事件句所属事件为临床表现，论元角色包含症状、舌象、苔象、脉象，症状角色的实体为恶风寒、恶寒发热、头身疼痛、喷嚏、鼻塞、流涕、咽喉痒痛、微有咳嗽、气喘；舌象角色对应的实体为舌淡红；苔象角色对应的实体为苔薄；脉象角色对应的实体为脉浮。而每一种事件论元的角色都有其对应的标签，同样采用 BIOES 的标注方法，事件论元角色标签表如下表 4 所示。

表 4　事件论元角色标签对照表

事件类型	事件论元角色	论元角色标签
证候概念	证名	A-ZM
	别名	A-BM
临床表现	症状	A-ZZ
	舌象	A-SX
	苔象	A-TX
	脉象	A-MX
证候分析	病因病机	A-BJ
辨证要点	常见病证	A-CJBZ
发病信息	发病日期	A-FBRQ
	发病分期	A-FBFQ
	易发病人群	A-FBRQ
治法信息	治法名称	A-ZFMC
	方药信息	A-FYXX
	用法信息	A-YFXX
	加减信息	A-JJXX

doi:10.3772/j.issn.2095-915x.2021.05.002

（3）基于规则构建标注训练集

依据约定的标注标准，我们采用词表匹配的方式对《中医诊断学》、《中医内科学》进行初步的标注，然后结合人工校对的方式对标注结果进行校正。利用前期工作从《中华古代名老中医医案》和《中医证候学》中抽取的实体，构建中医事件的规则，通过正则的方法实现词表的抽取，抽取规则如下：

1）总结中医事件类型的触发词，标记其对应的事件类型。

2）判断待标注语料句子中是否含有事件触发词，若含有触发词，则根据触发词对应的事件类型来匹配对应的事件论元。

3）根据事件触发词识别出的事件类型，利用词表，对该事件对应的事件论元进行填充。如识别出一句话中包含“表现为”这个事件触发词，那么将利用总结的证名、症状、舌象、苔象、脉象词表去发现其中对应的实体并将其标注出来。

4）若一句话中含有多种类型的事件触发词，则分别去匹配该种事件对应的事件论元；若一句话中存在同一种事件类型的多个事件触发词，则同时匹配出来。

5）若一句话中只可以匹配出事件触发词而没有发现事件论元，则可以不对该事件触发词进行标注。

初步标注完成之后，则可得到一个初步标注的数据训练集，然后对其进行人工校对，最终得到有标注的训练集。下图 2 为部分标注的数据示例。

证候概念　(1)得神　得神，又称有神。
证候分析　得神说明精气充盛，体健神旺，是健康的表现；若病而有神，则表明脏腑功能不衰，正气未伤，病多轻浅，预后良好。
证候概念　(2)少神　少神，又称神气不足。
辨证要点,证候分析　少神多由正气不足，精气轻度损伤，脏腑功能减退所致，多见于轻病或疾病恢复期的患者；素体虚弱者，平时亦多出现少神。
证候概念　(3)失神　失神，又称无神。
辨证要点　可见于久病虚衰（常见病症）或邪实神乱的重病患者。
证候分析　提示人体精气大伤（病机），脏腑功能严重受损，机能衰竭，预后不良。
临床表现　表现为本已神识不清，却突然精神转佳，语言不休，想见亲人；本已目光晦暗，却突然目似有光而浮露；本已面色晦暗枯槁，却突然颧赤如妆；本已久病卧床不起，却忽思下床活动；本来毫无食欲或久不能食，而突然食欲大增或主动索食。
证候概念　表证（证名）是指（触发词）六淫、疫疠等邪气，经皮毛、口鼻侵入机体的初期阶段，正气抗邪于肌表，以新起恶寒发热为主要表现的证。
临床表现　表证的主要表现为新起恶风寒（症状），或恶寒发热，头身疼痛，喷嚏，鼻塞，流涕，咽喉痒痛，微有咳嗽、气喘，舌淡红（舌象），苔薄（苔象），脉浮（脉象）。
证候概念　里证是指病变部位在内，脏腑、气血、骨髓等受病，以脏腑受损或功能失调症状为主要表现的证。
None　里证的范围极为广泛，其表现多种多样，概而言之，凡非表证（及半表半里证)的特定证，一般都属里证的范畴。其表现特征是无新起恶寒发热并见，以脏腑症状为主要表现的证。
标题　国家卫健委新型冠状病毒肺炎诊疗方案
发病信息　该病见于 2020 年 2 月 18 日（发病日期），临床分为轻型（发病分期），临床症状轻微，影像学未见肺炎表现；普通型，具有发热、呼吸道等症状，影像学可见肺炎表现；重型，出现急促；危重型，出现呼吸衰竭，且需要机械通气，出现休克。
临床表现　寒湿郁肺证临床表现为发热，乏力，周身酸痛，咳嗽，咯痰，胸紧憋气，纳呆，恶心，呕吐，大便粘腻不爽。舌质淡胖齿痕或淡红，苔白厚腐腻或白腻，脉濡或滑。
方药信息　治疗其基础方剂为生麻黄 6g（方药信息）、生石膏 15g、杏仁 9g、羌活 15g、葶苈子 15g，分 3 次服用（用法信息），早中晚各 1 次，饭前服用。

图 2　中医文献事件展示

上图中淡红色字体表示触发词，蓝色字体表示的是证名，绿色表示别名，浅黄色表示舌象，浅绿色表示苔象，浅蓝色表示脉象，金黄色字体表示常见病证，橘黄色字体表示病因，红色字体表示发病日期，灰色字体表示发病分期，浅粉红色字体表示方药信息，深绿色字体表示用法信息，事件类型为每句话前面所打的标签，例如方框内的这句事件类型为证候概念，事件触发词为“称”。我们统计了各类事件句在数据集中的分布情况，如下表 5 所示。

表 5　事件类型占比数

事件类型	所占比例
证候概念	20.1%
临床表现	17.3%
证候分析	16.3%
辩证要点	16.6%
发病信息	15.7%
治法信息	14.1%

3　联合句子级事件抽取模型

事件抽取的本质就是对句子中的每个字进行分类，所以其本质是一个序列标注模型。绝大多数序列标注模型都是以词作为输入，这样会产生一个问题即分词所带来的误差会传递到下游任务中，而由于中医语言文字语法结构和表达方式的特殊性，现有的针对中文的分词器并不能很好的对中医文献进行分词，而且中医文献中单字表示一个完整语义的情况较多，例如：浮沉分表里，浮，是面色浮显于皮肤之外，主表证；沉，是面色沉隐于皮肤之内，主里证。这里面的“浮”“沉”“主”“表”“里”都具有单独的语义。因此本文使用基于字的序列标注模型，省去分词步骤而减少误差传递。结合事件抽取任务的特性，为了让事件触发词与事件论元之间可以相互辅助从而提升事件抽取任务的精度，本文将 Self-Attention 层融入到序列标注任务中，通过权重强化语义表达，从而能够更深层的捕获事件触发词与事件论元之间的语义联系，提升对两个任务的识别分类效果。

模型一共包含四层，分别为 Embedding 输入层，BiLSTM 层，Self-Attention 层，CRF 层。模型架构图如图 3 所示。

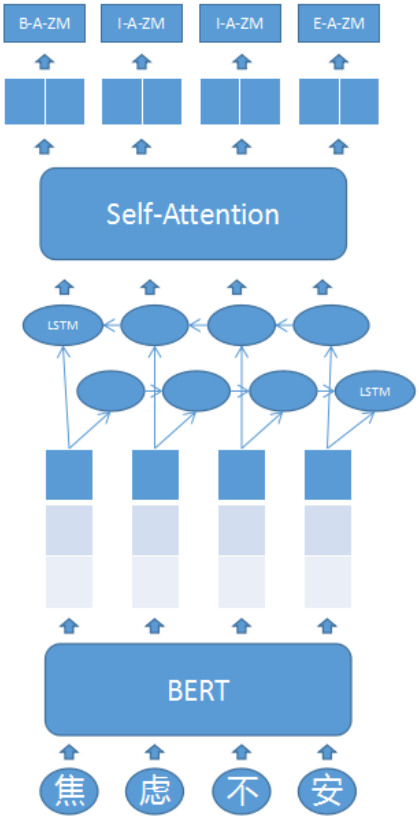

图 3　模型结构图

1）Embedding 输入层——BERT

我们利用 BERT 进行中医文本的特征提

doi:10.3772/j.issn.2095-915x.2021.05.002

取，BERT 在大规模语料上通过 MLM（Masked Language Model）和 NSP（Next Sentence Prediction）两个预训练任务来更好地学习每个字的上下文相关的向量表示。在此基础上，我们利用中医文本进行微调，由此能够更好地捕捉中医文本中的语义信息和抽象特征，这对后续进行中医事件抽取是有益的。在输入层将 BERT 所生成的字向量作为输入，假设 BERT 生成的字向量为 h_1，则该字符输入模型的向量为 $[h_1]$。

2）BiLSTM 层

BiLSTM 层主要是对输入向量进行特征学习。该结构由于有三个门单元控制着上下文信息的遗忘、保留和输出，所以该结构处理序列标注任务具有得天独厚的优势。该层的门控制机制能够充分保留上文信息以及当前信息中对任务有用的信息。在我们的中医事件抽取任务中，利用 BiLSTM 可以让每个字的特征中包含上下文特征。对文本进行前向和后向的双向计算，对同一个字符得到两个输出$\overrightarrow{h_t}$和$\overleftarrow{h_t}$，通过计算得到当前字符的特征向量 h，这里使用的计算方法同样为两个向量的拼接。

3）Self-Attention 层

在 BiLSTM 层之后为 Self-Attention 层，这层的作用主要是捕捉词之间依赖关系。中医事件抽取任务中事件触发词和事件论元之间有很强的依赖关系，如触发词“表现为”和论元“症状”“舌象”之间。利用 Self-Attention 则能很好地捕捉到这样的关系，并且过滤掉无用信息的影响，从而提高任务的精度。假设 BiLSTM 层的输出为 $H=[h_1,h_2,h_3,\cdots,h_n]$，则接下来最终用于输入 Self-Attention 层的矩阵 P 可以经过公式 (1) 计算得到：

$$Attention(\mathrm{Q},\mathrm{K},\mathrm{V})=\mathrm{softmax}(\frac{QK_T}{\sqrt{d_k}})\mathrm{V} \quad (1)$$

其中 $Q=K=V=H$，此公式得到的结果是一个 $n\times d$ 的矩阵，n 表示句子中字符的个数，d 表示每个字符经过注意力层所对应的向量的维度。利用下面公式 (2) 则可以得出作为接下来 CRF 层的输入的向量 P。

$$P=soft\max(Attention(\mathrm{Q},\mathrm{K},\mathrm{V})\times \mathrm{W}^{\mathrm{d}\times\mathrm{k}}) \quad (2)$$

其中，W 为一个模型内置的参数矩阵，为的是使得注意力机制输出的 $n\times d$ 矩阵转换为一个 $n\times K$ 的矩阵，其中 K 为标签类别数。假设事件触发词以及事件论元总共有 n 类，则 $k=4\times n+1$。

4）CRF 层

为了使得模型可以从全局来进行规划，在 Self-Attention 层之后使用了 CRF 层，CRF 层可以为经过 softmax 计算后的类别数的标签添加一些约束，这样可以尽最大可能来保证标签的合法性。例如对于临床表现类事件的事件触发词标签 T-LCBX，CRF 层可以尽可能使 B-T-LCBX 后不会再有 O、S-T-LCBX 或者 E-T-LCBX。

假设存在一句包含 n 个字的输入序列 $C=(c_1,c_2,\cdots,c_n)$，经过 Self-Attention 层之后可以得到一个 $n\times k$ 的矩阵 P，那么，该序列的输出标签序列 $L=(l_1,l_2,\cdots,l_n)$ 的最终得分 s 为公式 (3)：

$$S(\mathrm{C},\mathrm{L})=\sum_{i=0}^{n}A_{l_i,l_{i+1}}+\sum_{i=1}^{n}P_{i,\mathrm{l}i} \quad (3)$$

在上式中，A 是一个大小为 $k\times k$ 的矩阵，称为转移矩阵，其中，$A_{i,j}$ 表示的是第 i 个标签到第 j 个标签的转移概率。$P_{i,li}$ 为第 i 个位字符标签为 l_i 的概率，每个字符可以根据转移

矩阵和 softmax 计算之后得到一个得分，而标签序列的整体得分为每个字符的得分之和。这就会使得标签序列的预测不仅仅是基于当前字符标签的最大概率。若当一个字符的预测出现 B-T-LCBX 的概率最高，下一个字符同样预测为 B-T-LCBX 的概率最高，但是在 CRF 层的转移矩阵中 B 后接 B 的概率是非常低的，所以出现 B-T-LCBX—>B-T-LCBX 的概率就不是最高的，因此 CRF 层对非法标签序列有一个纠正作用。

对于一句输入样本 C，每个字符对应的标签有 k 个，那就会产生种标签序列，而模型的最终的目标函数为公式 (4)：

$$p(\mathrm{L}\,|\,\mathrm{C}) = \frac{e^{s(\mathrm{C},\mathrm{L})}}{\sum_{L\in\bar{L}} e^{s(\mathrm{C},\mathrm{L})}} \quad \text{（公式 4）}$$

公式 (4) 中的$\bar{L}$指的是 k^n 种标签序列，而在训练的时候采用的目标函数为对上式取对数的形式，具体形式如公式 (5)，最终模型利用反向传播对模型参数进行优化。

$$\log(p(\mathrm{L}\,|\,\mathrm{C})) = s(c,l) - \log(\sum_{L\in\bar{L}} e^{s(\mathrm{C},\mathrm{L})}) \quad \text{（公式 5）}$$

4 实验与分析

4.1 模型评价指标与实验设置

本实验主要使用精准率、召回率和 F1 值作为评判指标，精准率表示预测为正的样本中真正为正的样本所占的比例，召回率表示样本中的正例有多少预测正确，F1 值为精准率和召回率的调和平均。

我们将标注的数据按照 8:1:1 的比例划分为训练集、验证集以及测试集，然后投入模型进行训练，下表 6 展示了模型部分超参数设置。

表 6　模型超参数设置

参数	参数值	含义
eval_batch_size	8	每个batch中验证集句子个数
predict_batch_size	8	每个batch中测试集句子个数
learning_rate	5e-5	模型的学习率
num_train_epochs	100	训练的轮数
droupout_rate	0.8	防止过拟合参数
lstm_size	128	BiLSTM层的LSTM单元数
num_layers	1	BiLSTM层的层数
max_seq_length	128	输入句子的最大长度

4.2 实验结果与对比分析

1）实验结果

本实验中，使用《中医诊断学》、《中医内科学》以及各种诊疗标准的标注语料进行实验分析，所采用的模型是添加注意力机制的序列标注模型。因为一次训练的结果具偶然性，所以本实验采用三次结果的均值来作为实验的最终结果，模型最终对触发词的准确率以及召回率进行计算。

由于事件抽取涉及到的实体种类较多，在本实验中，我们定义了 6 种中医事件类型，对应 6 种事件触发词实体类型以及 15 种事件论元，因此存在的标签类型有 4×(15+6)+1=85 种。在分类任务中，标签类型越多，对分类准确率影响越大。为了提高模型的准确率，本实验针对不同事件类型分别训练模型，将标注好的语料按照事件类型分开，对同一种事件类型只标注与其相关的事件论元并放入模型中，训练得到该种事件的事件抽取模型。但是这种方式会使得训练语料规模大幅度降低，结合表 1 可知。本实验将交集较多的划分为一组，将证候概念、临床表现、证候分析、辨证要点作为一个训练集，

doi:10.3772/j.issn.2095-915x.2021.05.002

发病信息与治法信息作为一个训练集，总共训练两个模型，模型的各项指标如下表 7 所示。

表 7　事件触发词评价指标结果

事件触发词类别	P(%)	R(%)	F(%)
证候概念	78.66	83.48	81.00
临床表现	80.70	78.67	79.67
证候分析	69.99	79.88	74.61
辨证要点	67.32	80.39	73.28
发病信息	76.28	73.52	74.88
治法信息	76.99	79.67	78.31
AVE	74.99	79.26	76.96

上表为六种中医事件触发词的评价指标，在六种中医事件类型中，模型整体的精确率达到 74.99%，召回率达到 79.26%，F1 值达到 76.96%。模型的触发词识别的总体效果能达到 70% 以上，但是证候分析和辨证要点的精确率相对较低。这是因为证候分析和辨证要点的触发词变化相对较大，如在证候分析中“多因”“因”“多为”“为”都是证候分析的触发词，那么识别过程中可能出现未识别到“多因”中的“多”的情况，图 4 展示了模型预测的部分结果。

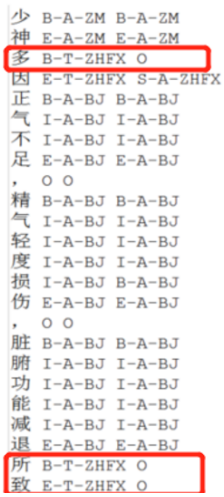

图 4　模型预测结果图

图 4 中左侧为人为标注的标签，右侧为模型预测的标签，可以看到“多因”中的“多”是没有预测出来的，因为通常语句中“多因”和“因”使用变化比较频繁而且变化无明显规律，也可以看到下方的“所致”也是没有预测出来的，这是因为在标注时通常只标注一个事件触发词，而人为校正的过程也会存在一些噪声，这些噪声也是精确率偏低的原因。表 8 为各类事件对应的事件论元各项指标。

表 8　事件论元评价指标结果

元事件类型	事件论元角色	P(%)	R(%)	F(%)
证候概念	证名	76.29	71.3	73.71
	别名	75.06	77.42	76.22
临床表现	症状	65.4	74.74	69.76
	舌象	79.29	85.16	82.13
	苔象	77.25	84.38	80.66
	脉象	78.75	84.15	81.36
证候分析	病因病机	65.44	79.19	71.66
辨证要点	常见病证	70.1	77.75	73.72
发病信息	发病日期	80.01	79.89	79.95
	发病分期	74.33	78.46	76.34
	易发病人群	76.03	79.15	77.56
治法信息	治法名称	73.83	75.5	74.66
	方药信息	75.24	73.32	74.26
	用法信息	68.18	70.29	69.22
	加减信息	77.27	73.14	75.15

从上表可以看出症状、病因病机和用法的精确率是偏低的。这是因为症状表达方式较为多样而且症状边界不明显，而病因病机的精确率较低是因为有些病因病机的描述较长，所以模型识别时会将从来没有遇到过的语句切分开来。图 5 分别展示了部分结果，其中

左边表示症状预测的结果，右边表示病因病机预测的结果。

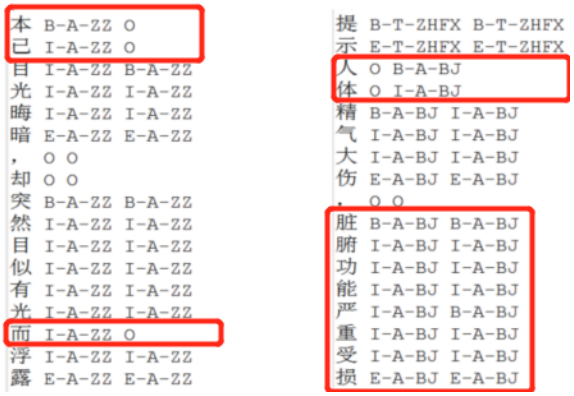

图 5　部分预测结果

2）对比分析

另外，本文还针对 Pipeline 模型和联合抽取模型做了对比分析实验，Pipeline 模式分为两个阶段：事件触发词识别阶段。我们利用一个事件触发词分类器 DMCNN[3] 对句子中每个词进行分类，判断是否为触发词；如果存在触发词，执行论元分类阶段，将事件论元分类过程看作一个多分类问题，使用一个分类器判定论元角色，同样使用 DMCNN[3] 帮助实现。因为第一阶段主要目的是为了确定事件类型，对事件信息的描述并没有什么实质性的影响，所以我们主要针对事件论元进行对比实验，实验结果如表 9 所示。

表 9　**Pipeline** 与联合抽取模型对比实验结果

事件类型	事件论元角色	P(%)		R(%)		F(%)	
		Pipeline	联合抽取	Pipeline	联合抽取	Pipeline	联合抽取
证候概念	证名	67.94	76.29	60.77	71.3	64.16	73.71
	别名	64.32	75.06	51.51	77.42	57.21	76.22
临床表现	症状	59.68	65.4	55.95	74.74	57.75	69.76
	舌象	69.99	79.29	54.42	85.16	61.24	82.13
	苔象	61.74	77.25	56.11	84.38	58.79	80.66
	脉象	61.63	78.75	56.99	84.15	59.22	81.36
证候分析辩证要点	病因病机	60.98	65.44	60.48	79.19	60.73	71.66
	常见病证	68.27	70.1	54.91	77.75	60.82	73.72
发病信息	发病日期	69.28	80.01	64.56	79.89	66.84	79.95
	发病分期	68.28	74.33	60.6	78.46	64.21	76.34
	易发病人群	63.44	76.03	64.31	79.15	63.87	77.56
治法信息	治法名称	61.79	73.83	60.65	75.5	61.22	74.66
	方药信息	66.62	75.24	68.09	73.32	67.35	74.26
	用法信息	60.12	68.18	53.59	70.29	56.67	69.22
	加减信息	61.08	77.27	65.82	73.14	63.36	75.15
	AVE	64.34	74.16	59.25	77.59	61.56	75.76

由上表可以看出，联合抽取模型效果普遍高于 Pipeline 抽取模型。这是由于联合抽取模型可以捕获到触发词与事件论元之间的关系，使得两个任务的准确率都有所提升，而且受益于 Self-Attention 层可以更精确地捕获这两种词之间的语义关系，并且 CRF 层可以学习到标签

doi:10.3772/j.issn.2095-915x.2021.05.002

之间的依赖关系。针对不同的事件论元，可以看出联合抽取模型 F1 值平均高于 Pipeline 模型 14.2 个百分点。因此实验证明在我们的数据集上联合抽取的方法确实较 Pipeline 模式的方法更为有效。

5 结论

本文主要介绍了针对中医文献的句子级事件抽取的任务、数据来源，又根据 ACE 事件模式定义了中医领域的六类事件类型及其对应的事件论元，对比当前流行的 Pipeline 事件抽取模型与联合事件抽取模型的优缺点，提出了一种针对中医文献的句子级联合事件抽取模型，并且结合事件抽取任务的特性，提出将 Self-Attention 机制融入模型进行改进的方法。通过对比实验以及分析结果可知，本文提出的方法是有效的。另外从数据准备与数据预处理的角度，本文还介绍了如何筛选包含中医事件类型的句子，定义数据标注规则并介绍如何利用规则和人工结合的方式对数据进行标注，从而更好地完成针对《中医诊断学》、《中医内科学》以及各种诊疗方案的初步事件抽取任务。

未来，我们将针对一个句子可能缺少事件触发词或同一个事件的事件论元可能分布在不同的句子中的情况开展进一步的事件抽取方法研究。

参 考 文 献

[1] 高强 , 游宏梁 . 事件抽取技术研究综述 [J]. 情报理论与实践 , 2013, 36(4): 114-117.

[2] Ouertatani A, Gasmi G, Latiri C. Argued opinion extraction from festivals and cultural events on Twitter[J]. Procedia Computer Science, 2018(126):205-213.

[3] Chen Y, Xu L, Liu K, et al. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks[C]. The 53rd Annual Meeting of the Association for Computational Linguistics. 2015.

[4] Sauri R, Verhagen M, Pustejovsky J. Annotating and recognizing Event Modality in text[C]. Nineteenth International Florida Artificial Intelligence Research Society Conference. DBLP, 2004.

[5] Hai L C, NG H T. A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text. American Association for Artificial Intelligence. 2002.

[6] Llorens H, Saquete E, Navarro-Colorado B . TimeML Events Recognition and Classification: learning CRF models with semantic roles[C]. International Conference on Computational Linguistics. Association for Computational Linguistics, 2010.

[7] Ahn D. The stages of Event Extraction. Proceedings of the Workshop on Annotating and Reasoning about Time and Events[C]. Association for Computational Linguistics, Sydney. Australia: ACL. 2006:1-8.

[8] 赵妍妍 , 秦兵 , 车万翔 , 等 . 中文事件抽取技术研究 [J]. 中文信息学报 , 2008, 22(1):3-8.

[9] 许红磊 , 陈锦秀 , 周昌乐 , 等 . 自动识别事件类别的中文事件抽取技术研究 [J]. 心智与计算 , 2010(1):34-44.

[10] Naughton M, Kushmerick N, Carthy J. Event Extraction from heterogeneous news sources[C]. In Proceedings of AAAI Workshop on Event Extraction and Synthesis. Boston, US: Amercation Association for Artifical Intelligence. 2006:7-13.

[11] Ji H, Grishman R. Refining Event Extraction through cross-documents infere[C]. Proceedings of 46th Annual Meeting of ACL, Ohio, USA: Association for Computational Linguistics. 2008:254-262.

[12] Liao S, Grishman R. Using document level Cross-Event inference to improve Event Extraction[C]. ACL, Meeting of the Association for Computational Linguistics, July, Uppsala, Sweden. DBLP, 2010.

[13] Hong Y, Zhang J, Ma B., et al. Using Cross-Entity inference to improve Event Extraction[C]. Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011.

[14] Nguyen T H, Grishman R, et al. Event Detection and domain adaptation with Convolutional Neural Networks[C]. The Meeting of the Association for Computational Linguistics. 2015:365-371.

[15] 秦美越 . 基于并行多池化 CNN 的生物医学事件抽取 [D]. 大连 : 大连理工大学 , 2017.

[16] Nguyen T H , Grishman R . Modeling Skip-Grams for Event Detection with Convolutional Neural Networks[C]. Conference on Empirical Methods in Natural Language Processing. 2016.

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

[18] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation[J]. Computer Science, 2014:1724-1734.

[19] Feng X, Qin B, Liu T. A language-independent neural network for Event Detection[J]. Science China-Information Sciences, 2018, 61(9):81-92.

[20] Liu S, Chen Y , Liu K , et al. Exploiting argument information to improve Event Detection via supervised Attention Mechanisms[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:1789-1789.

[21] 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.