融合句义特征的元事件抽取方法

摘要：在句义结构分析的基础上，利用句义特征，提出一种元事件抽取方法。以词作为序列标注的基本单元，先采用条件随机场模型对触发词进行识别，并采用后处理规则对识别结果进行优化；然后采用双层条件随机场模型并结合语义格、谓词与一般项的关系识别事件元素；最后利用浅层次位置信息对前两部分识别结果进行共指判别，并在事件映射模式的指导下利用深层次句义信息提取事件对应的句义成分，完成元事件的抽取。在 CEC 语料上进行测试，元事件抽取的平均 F 值达到 59.9%。实验表明，该方法能有效抽取句中元事件，为事件抽取研究提供一个新的思路。

关键词：元事件抽取；触发词识别；事件元素识别；条件随机场；句义信息

中图分类号：TP391

Approach of Meta Event Extraction combing Sentential Semantic feature

Abstract: On the basis of sentential semantic structure analysis, and using sentential semantic feature, a new method is proposed for meta event extraction in this paper. The method which takes words as the sequence labeling units, the conditional random field model is trained to recognize the trigger word, a rule-based post-processing is applied to correct the triggers; then dual-layer conditional random field model and sentential semantic features are used to identify the event argument; finally shallow layer location position information is used for distinguishing which of the first two parts recognition results represent same meta event, and in the guidance of event mapping mode, sentential semantic components are extracted by using sentential semantic information,. The proposed method is evaluated using CEC corpus, the F-score of overall result is 59.9%. The experimental result show that this method is effective on event extraction and providing a new thought for event extraction.

Key words: meta event extraction; trigger recognition; event argument recognition; conditional random field; sentential semantic information

基金项目：国家 242 项目（2005C48）；北京理工大学科技创新计划重大项目培育专项（2011CX0101）

1 引言

作为信息抽取研究最具挑战性的任务之一，事件抽取旨在将无结构化文本中人们感兴趣的事件以结构化的形式表现出来。在自动文摘[1]、自动问答、信息检索[2]等领域有着广泛的应用。2005年，ACE评测内容中加入了事件抽取任务，认为事件是由触发词和事件元素两部分组成，触发词是句子中最能清晰表达事件发生的词语，事件元素主要指事件的参与者和属性（时间、地点等）[3]。

事件抽取的主要研究方法有模式匹配和机器学习。早期的信息抽取系统中，大量使用模式，这种方法准确率较高，在特定领域内可以取得比较好的效果。例如姜吉发提出了一种基于领域无关概念知识库的事件抽取模式学习方法GenPAM[4]；吴刚通过定义会见和访问元事件的事件模板完成这两类事件的抽取[5]；Yankova借助领域知识和制定模板实现足球领域的事件抽取系统[6]。但是这种方法可移植性差，从一个领域移植到另一个领域时，需要重新构建模式。基于机器学习的方法把事件抽取任务看作分类问题，核心在于分类器的构建和特征的发现、选择上。例如Chieu和Ng首次在事件抽取中引入最大熵分类器[7]，从分类的角度对事件元素进行抽取，在讨论发表会事件和工作交接事件中取得了很好的结果；David Ahn创新性的结合MegaM和Timbl两种机器学习方法[2]，分别实现了事件抽取任务中事件触发词及类别识别和事件元素识别；赵妍妍等采用了一种基于触发词扩展和二元分类相结合的方法[8]，较好的解决了事件抽取中训练实例正反例不平衡以及数据稀疏问题,在ACE的中文语料上取得较好的效果；Zheng Chen率先将事件抽取视为序列标注问题，通过使用MEMM模型，选取常规的特征进行试验，在ACE2005年的评测中，取得了很好的效果[9]；H.Llorens等通过条件随机场(CRF)模型进行语义角色标注[10]，并应用于TimeML的事件抽取，提升了系统的性能；Tian将扩展触发词表和机器学习方法相结合，提出一种触发词自动识别方法[18]。相对而言，这种方法较为客观，不需要太多的人工干预和领域知识，因此机器学习的方法成为目前主流的事件抽取方法。

近年来，在ACE英文语料上也出现了较多使用推理来提高事件抽取性能的研究,将特征选择由句子级别向段落、篇章甚至文档级扩展，以克服句子级信息在某些情况下不足的问题。例如Heng Ji基于平行语料库和跨语言信息，提出一个归纳学习框架，改善了事件抽取的性能[11]；Liao发现一类事件可以辅助识别其他类型的事件，提出了跨事件（Cross-Event）抽取辅助进行事件元素识别和事件类型识别推理的方法[12]，取得了不错的效果。

事件抽取任务可分为元事件（Meta Event）抽取和主题事件（Topic Event）抽取[5]。元事件表示一个动作的发生或状态的变化，由表示动作的动词或名词来驱动，包括参与该动作行为的主要成分（如时间、地点、人物等）[3]。目前对元事件抽取研究集中在两项子任务上：事件触发词识别和事件元素识别。其中触发词识别精度在70%左右，事件元素识别精度在60%左右，距实用还有一定的差距。由于句子结构的复杂性，一个事件句往往包含多个元事件，仅依靠识别触发词和事件元素我们无法对发生的事件有直观的认识。

针对上述问题，本文提出一种融合句义特征的元事件抽取方法。通过采用 CRF 模型识别触发词和事件元素，并在此基础上结合位置信息与句义信息抽取句中元事件。文中定义元事件是由触发词，参与者，时间，地点这四个要素构成，其中触发词是元事件的核心，某些情况下后两个元素可以缺省。

2 汉语句义结构模型

句义结构模型[13]是对句义中的成分以及成分之间组合关系的结构化、形式化表示，将抽象的句义表示成计算机可处理的结构化数据。句义结构模型的基本形式如图 1。根据句义结构模型理论，课题组自动构建了一套句义结构分析系统ACSM①。后面用到的句义特征都是根据该系统提取出来的。

① http://www.isclab.org/csa/bfs-csa.php

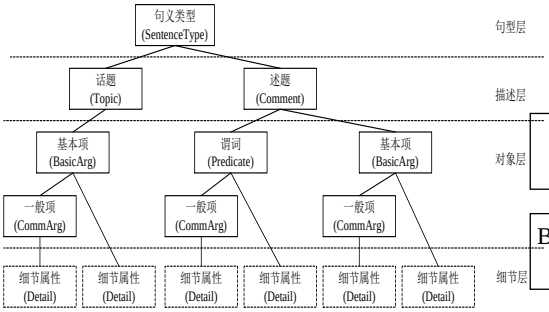

图 1句义结构模型的基本形式

Fig.1 Basic form of sentential semantic model

描述层反映的是句子的描述对象（话题）以及对话题的描述（述题）；对象层包含谓词、基本项和一般项，构成话题和述题的成分， 谓词是句义中说明话题的成分，是整个句义结构的核心，项是句义中表对象的成分。

句义结构模型是以谓词为核心，基本项担任谓词所描述的句义角色（施事、受事等）。而事件是以触发词为核心，对于动词驱动的事件，参与者是触发词所触发的对象。通过类比不难发现，事件与句义结构模型是存在一定映射关系的，可以将触发词映射到谓词上，相应的事件参与者映射到谓词所对应的基本格（施事格/受事格/遭遇格），事件的地点映射到谓词所对应的空间格或范围格，事件发生的时间映射到谓词对应的时间格。以“恐怖分子劫持了十名儿童”为例，采用 ASCM 进行句义分析，得到的句义结构树如图 2 所示。

事件及其映射模式如

表 1 所示。其中，1 表示必须有，0 表示不一定有。

表 1事件及其映射模式

Tab.1 Definition of event and its mapping model

事件	事件映射模式	是否必要
触发词	谓词	1
参与者	施事格/受事格/遭遇格	1
时间	时间格	0
地点	范围格/空间格	0

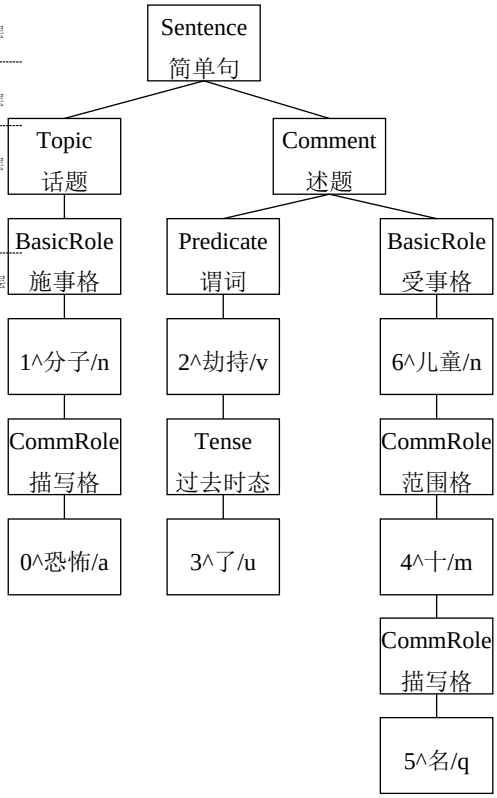

图 2“恐怖分子劫持了十名儿童”句义结构树

Fig.2 Example of sentential semantic structure of a Chinese sentence

3 算法原理

本文提出的融合句义特征的元事件抽取是在文本预处理的基础上，以词作为序列标注的基本单元，选取合适的词法，上下文等特征，分别采用CRF和双层CRF模型识别触发词和事件元素，对识别结果采用后处理规则进行处理，得到触发词集合和事件元素集合；然后根据位置信息对前两部分识别结果进行共指判别，根据事件映射模式从句义结构中抽取事件对应的各个句义成分，完成元事件的抽取。

系统主要包括：预处理、触发词识别、事件元素识别、元事件抽取四个模块。算法

原　　理　　如

图 3 所示，各模块具体内容将在下文详细介绍。

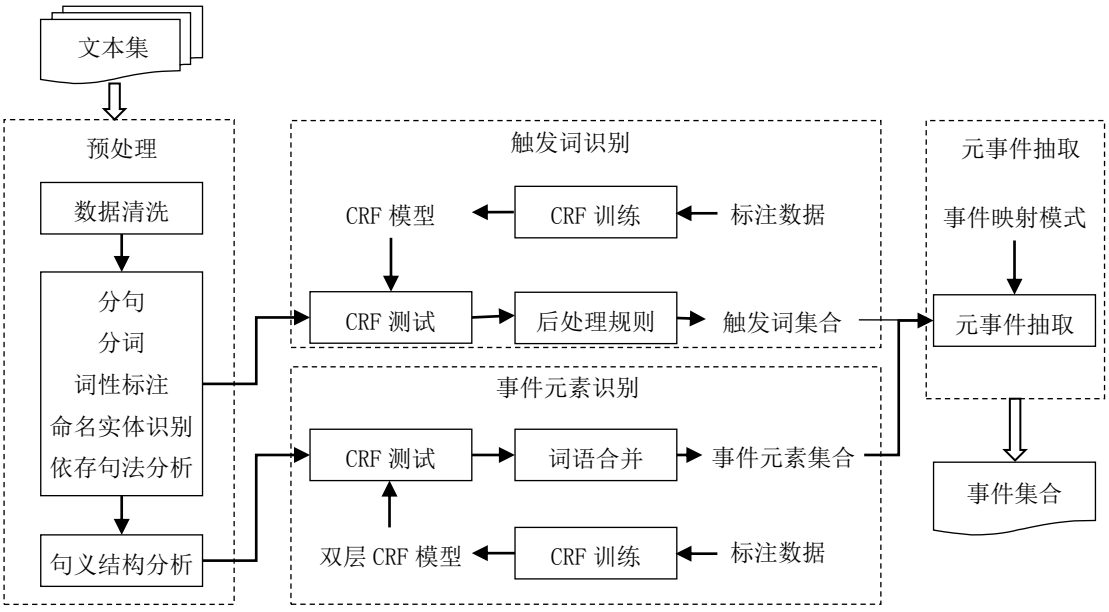

图 3 事件抽取原理图

Fig.3 Processes of event extraction

3.1 预处理

预处理模块是对输入的文本集进行数据清洗、分句、分词、词性标注、命名实体识别、依存句法分析和句义分析。数据清洗阶段去除 XML 格式实验数据中的 XML 标记，恢复原始数据（不添加任何标记）以及保留事件要素的各个标记和编号。然后采用哈工大信息检索研究室开发的 LTP 对原始文档进行分句、分词、词性标注等操作，将得到的 XML 文档进行处理分别得到带有编号的句子、词性标注、命名实体和依存句法分析文件。最后利用 ASCM 系统对得到的词性标注文件进行句义结构分析，方便后面获取句义特征。

3.2 触发词识别

触发词标志事件的发生。本文将触发词识别看成序列标注问题，采用条件随机场(Conditional Random Field，CRF)模型进行识别。

条件随机场(Conditional Random Field，CRF)是由John Lafferty等提出的一种基于统计的序列标注模型[14]，其核心思想是利用无向图理论使序列标注的结果达到在整个序列上全局最优，其特点是假设输出随机变量构成马尔科夫随机场。CRF模型克服了传统的隐Markov模型(HMM)和最大熵Markov模型(MEMM)的标记偏置等问题，已被应用到自然语言处理的多个领域，如中文分词、命名实体识别[15]等等。

给定输入序列X后，标记序列Y的条件概率被定义为：

$$p(Y \mid X)=\frac{1}{Z(X)}\exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k\left(Y_{i-1},Y_i,X,i\right)\right). \quad (1)$$

其中，f_k 是特征函数，$k\in[1,K]$，K 为特征函数的数量。λ 是权值向量，通过训练获得相应的估计值。$Z(X)$ 是归一化因子，是所有可能的标签序列 S 的总和：

$$Z(X)=\sum_{S}\exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k\left(Y_{i-1},Y_i,X,i\right)\right). \quad (2)$$

对于输入序列X，最有可能的输出标记序列Y为：

$$Y=\arg\max_{Y} p(Y \mid X) \quad (3)$$

一般情况下，触发词是句子中的主要动词，有些情况下也为名词。例如“美国军队炮继续轰法鲁耶”，句中“炮轰”是触发词，其为动词；“双方领导人在北京举行了会谈”，句中“会谈”是触发词，其为名词。由于句中存在大量噪声词语，导致正反例严重失衡。因此，本文首先对词性标注文件进行噪声词语过滤，提取出每句话对应的名词和动词作为候选触发词；然后将触发词识别看成是一个二元分类问题，通过选取合适的特征，此处用到的义元和义类特征分别是通过 HowNet[②]和哈工大同义词林[③]获得，采用 CRF 模型进行序列标注，识别出正确的触发词。本文选取了词法特征、实体特征、上下文特征、词典信息、句法特征来对候选触发词进行描述，如表 2 所示。

表 2事件触发词识别特征描述

Tab.2 Description of features for trigger recognition

特征	描述
Trigger	候选触发词
TriggerPos	候选触发词词性
NerType-1 NerType+1	候选触发词左右侧命名实体类型
Word-1 Word-2 Word+1 Word+2	候选触发词左右侧两个词语
WordPos-1 WordPos-2 WordPos+1 WordPos+2	候选触发词左右侧两词语词性
WordCodeT	候选触发词的义类代码
WordCodeH	候选触发词的义元代码
WordDR	候选触发词与父节点的依存关系

由 CRF 模型识别出的触发词对应的是单个词语，其中大部分能很清晰地代表触发词，但也有少部分由于分词过程中被分开，作为触发词不能完整的表达事件的发生。

[②]http://www.keenage.com

[③]http://ir.hit.edu.cn/

例如：“大货车翻倒在樟树湾的转弯处”。

词性标注结果为：0^大/a 1^货车/n 2^翻/v 3^倒/v 4^在/p 5^樟树湾/ns 6^的/u 7^转弯/v 8^处/n 9^。/wp

CRF 模型识别出的触发词为“倒”，相比之下，“翻倒”能更好的表示整个事件的发生。对于这种情况，采取比较简单的后处理规则进行修正。根据已识别触发词相邻词语的词性，采取表 3 所述规则。

表 3后处理规则算法

Tab.3 Rule-based post-processing algorithm

Input：CRF 模型识别的词语序列集合 wordSet；
Output：规则修正后的触发词集合 triggers；
变量：词语标记 flag，词性 pos，词语包含字数 len，当前词语 word，当前词语前一个词语 word-1，当前词语后一个词语 word+1。

```
for ∀ wordSet do
    for ∀flag(word) do
        if flag(word)＝“Y” and len(word)＝1
            if flag(word+1) ＝ “Y”， 则 merge(wordword+1)∈trigger；
            else if pos(word-1) ＝ “v” or pos(word-1)＝“f” and len(word-1)＝1，则 merge(wordword+1)∈triggers；
            else if pos(word+1) ＝ “v” and len(word+1) ＝ 1 ， 则 merge(wordword+1) ∈triggers；
            else word∈triggers；
            end if
        else if flag(word)＝“Y” and len(word)≠1，则 word∈triggers；
            else word ∉ triggers.
            end if
    end for
end for
```

3.3 事件元素识别

事件元素识别是事件抽取系统的关键部分。本文参考文献[16]中方法，通过选取合适的特征，采用 CRF 模型进行双层训练。第一层训练进行词语融合，选取词法特征、上下文特征、词性组合特征和句义特征（语义格类型），利用 CRF 模型初步识别候选事件元素对应的短语，采用（B，I，E）三种标签标记识别结果。第二层训练再次使用 CRF 模型对第一层序列标注结果进行分类，在第一层训练特征的基础上添加了句义特征(谓词与一般格间的关系)和词语融合标记，采用（P，L，T，O）四种标签来标记事件元素和非事件元素。最后对每句话的序列标注结果进行简单的处理，将一句话中相邻的同一标记的词语进行合并，得到每句话对应的事件元素及其类型。两层训练中使用的标签含义如表 4 所示，事件元素识别选用的特征如表 5 所示。

表 4标签含义

Tab.4 Meaning of label

第一层 CRF 识别标签		第二层 CRF 识别标签	
标签	含义	标签	含义
B	短语对应的第一个词语或单个词语	P	参与者
		L	时间
I	短语对应的中间词语	T	地点
E	短语对应的最后一个词语	O	非事件元素

表 5事件元素识别特征描述

Tab.5 Description of features for event argument recognition

特征	描述
Word	当前词语
WordPos	当前词语的词性
WordTag	当前词语标记
Word-1	当前词语的左右侧两个词语
Word-2	
Word+1	
Word+2	
WordPos-1	当前词语左右侧两个词语的词性组合
WordPos-2	
WordPos+1	
WordPos+2	
WordPos-2WordPos-1WordPos	
WordPos-1WordPosWordPos+1	
WordPosWordPos+1WordPos+2	
WordTag-1	当前词语左右侧两个词语的标记
WordTag-2	
WordTag+1	
WordTag+2	

B&PrelationSemanticCase	一般格与谓词的关系和语义格类型的组合

以“恐怖分子劫持了十名儿童”为例，图 4 显示了双层 CRF 模型的工作原理。

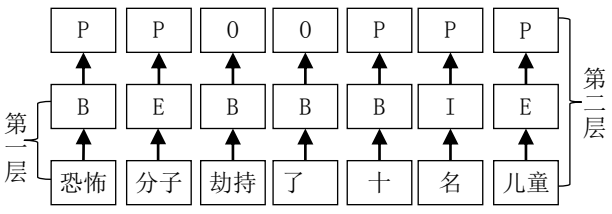

图 4 双层 CRF 事件元素识别模型

Fig.4 Dual-layer CRF model for event argument recognition

3.4 元事件抽取

元事件抽取是从句子层面抽取描述某一事件的触发词和事件元素，一个句子可能不存在元事件，也可能存在一个或多个元事件，触发词的出现代表元事件的产生。由于根据位置信息能锁定小句（小句以“，”和“；”为分隔符）元事件；而句义结构能综合分析各个句义成分及它们之间的关系，能在一定程度上克服跨小句元事件抽取的困难。因此，本文采用位置信息和句义结构模型相结合的方法完成元事件的抽取。输入是触发词集合、事件元素集合和句义分析结果文件，输出是结构化的元事件集合。

基于位置信息抽取元事件是指对 CRF 模型识别出的触发词和事件元素进行共指判别，即判断哪些触发词和事件元素描述同一个元事件。判别方式是：如果触发词和事件元素在同一个小句中，则它们对应同一个元事件，否则不是。

例如：瑞丽市民族医院等医疗卫生部门积极开展受伤人员救治，各受伤人员已得到妥善医治。

其中“医疗卫生部门”、“受伤人员”和“各受伤人员”分别被识别为参与者，“救治”和“医治”被识别为触发词。整个句子包含两个小句，根据位置信息可以很容易得到完整的“救治”事件和“医治”事件。

基于句义结构模型抽取元事件是指根据事件映射模式，从句义结构树中提取句子所描述的对象以及谓词所涉及的时间、地点属性，作为事件的各个元素。具体的抽取过程如下：

step1：搜索句义结构树，找到当前触发词在句义结构树中对应的谓词；

step2：查找谓词对应的话题和述题，如果话题和述题都存在，转向 step5，退出；如果仅存在话题，转向 step3；如果仅存在述题，转向 step4；如果二者不存在，先后转向 step3 和 step4；

step3：沿着句义结构树向下遍历，找到谓词的孩子节点，如果孩子节点是谓词，转向 step5，否则退出；

step4：沿着句义结构树向上遍历，找到谓词的父节点，如果父节点是谓词，转向 step5，否则退出；

step5：如果存在描述当前谓词的时间和地点信息，判断时间格、地点格、话题或述题是否是已识别出事件元素的子集，如果是就将该事件元素直接提取，否则根据事件映射模式提取相应的句义成分。

通过上面步骤得到的句义成分是事件元素对应的核心词语，由于在某些情况下仅根据这些核心词语我们无法对与事件相关的元素有清晰的认识，这就需要对识别出来的核心词语添加适当的修饰成分（描写格、范围格等）来丰富事件元素。通过观察，得到离核心词语最近和最远的修饰词语之间间隔的词语不超过 3 个时比较合适。

例如：恐怖分子劫持了十名儿童。

这句话的触发词为“劫持”，采用上述方法对 2.3 节中得到的句义结构树进行句义成分提取，得到的事件参与者为“分子”和“儿童”。修正后，事件参与者变为“恐怖分子”和“十名儿童”。

由于一个元事件包含的触发词和事件元素分布存在两种情况：分布在同一小句中；分布在所属的句子中。根据这个分布特点，制定以下规则：

1、事件触发词所在小句中含有事件参与者信息时，以基于位置信息的抽取结果为准。

2、事件触发词所在小句中没有识别出或没有事件参与者信息出现时，以基于句义结构模型的抽取结果为准。

4 实验及结果分析

为了获取模型最优参数以及验证算法各模块的识别准确率和识别效果，分别进行

了参数选择，触发词识别，事件元素识别以及元事件抽取四组实验。

4.1 实验数据源

实验数据来自上海大学计算机实验室中文事件语料(CEC)[17]，该语料包含地震、交通事故、恐怖袭击、食物中毒、火灾五大类突发公共事件新闻报道，共 203 篇。实验采用五折交叉验证测试方法，用语料的 4/5 作为训练集，1/5 作为测试集。在元事件抽取部分，将正确的元事件界定为至少包含触发词和事件参与者，通过对 CEC 中的语料进行整理，得到满足要求的事件总数为 2429。

4.2 评价方法

采用准确率（P）、召回率（R）、F 值（$F-score$）来评价触发词、事件元素和元事件的识别情况。其定义为：

准确率：

$$P=\frac{\text{正确识别的结果数}N_{right}}{\text{识别出的总的结果数}N_{recog}} \tag{4}$$

召回率：

$$R=\frac{\text{正确识别的结果数}N_{right}}{\text{标注语料中的总数目}N_{total}} \tag{5}$$

F 值：

$$F-score=\frac{2PR}{P+R} \tag{6}$$

在元事件抽取部分，如果一个元事件包含的触发词与参与者都正确，则认为是正确的元事件。

4.3 实验结果及分析

第一组实验是参数选择实验，目的为文中使用的两处 CRF 模型选取最优训练参数。本文在触发词和事件元素识别中都使用 CRF 模型，共涉及到两对参数选择。由于多模型的联合参数很难进行参数调优，本文通过贪心思想找到局部最优解，通过局部最优逼近全局最优。

下面以触发词识别中CRF模型参数选择实验为例。首先将 c 以步长 1 从 0.2 增长到 9.2， f 以步长 2 从 1 增长到 19，对应参数下的 F 值如图 5 所示。然后根据结果分布缩小 c 和 f 的取值范围和步长进行进一步实验，将 c 以步长 0.1 从 0.8 增长到 1.7，f 以步长 1 从 1 增长到 5，结果如图 6 所示。分析实验结果得到触发词识别中 CRF 模型最优参数 c 为 1.4，f 为 1。

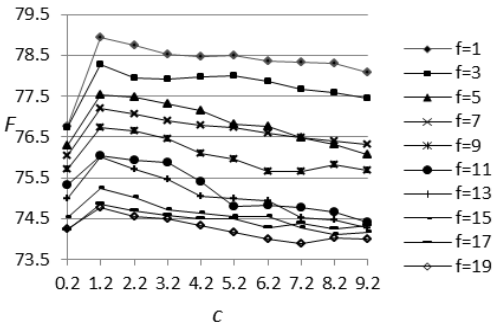

图 5 第一次参数选择实验结果

Fig.5 First selection results of parameter selection experiment

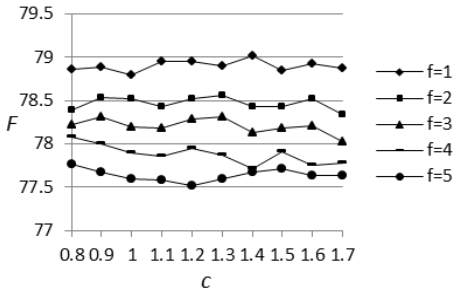

图 6 第二次参数选择实验结果

Fig.6 Second selection results of parameter selection experiment

事件元素模块中 CRF 模型参数选择方法同上述步骤，得到最优参数 c1 为 1.5，f1 为 2。

第二组实验是触发词识别实验，目的是验证本文提出的基于 CRF 模型识别触发词方法的有效性。将基于扩展触发词词表的识别方法和文献[18]中基于最大熵模型的识别方法，与本文方法进行对比，结果如表 6 所示。

表 6触发词识别结果

Tab.6 Results of trigger recognition

	Precision	Recall	F-score
基于触发词表	0.3976	0.8655	0.5446
文献[18]方法	0.8095	0.6296	0.7083
本文方法	0.8107	0.7709	0.7902

由表 6 可见，基于扩展触发词表的方法

准确率较低，但是能够获得比基于机器学习的方法更高的召回率，这是由于基于扩展触发词词表的方法对触发词抽取限制比较少，抽取结果中返回触发词数目较多所致。两种机器学习方法中，本文的方法效果略好。分析原因如下：本文通过特征筛选，在文献[18]所用特征的基础上，进一步丰富了特征向量，由于实体会伴随触发词出现，选取实体特征对触发词识别起到指示作用；义元特征可以间接描述词语之间的相似程度，在一定程度上对触发词语义进行了较为深入的挖掘。

第三组实验是事件元素识别实验，目的是验证本文提出的基于双层 CRF 并结合句义特征识别事件元素方法的有效性。将文献[19]中基于最大熵模型识别事件元素的方法与本文的方法进行比较。结果如表 7 所示。

表 7事件元素识别结果

Tab.7 Results of event argument recognition

	Precision	Recall	F-score
文献[19]方法	0.5613	0.6101	0.5847
本文方法	0.6925	0.8148	0.7486

由表 7 可见，基于双层 CRF 进行的事件元素识别方法结果要比基于最大熵模型的方法效果好。分析原因有两点：1、由于一些复杂事件元素会嵌套包含简单事件元素，采用双层 CRF 训练，将第一层识别结果作为一维特征传至第二层能提供更多有用的信息，尽可能的消除词语跨越标记错误，同时对标注结果起到平滑作用；2、句义特征能够更加细致的描述词语在句子中担任的句义成分以及与其他句义成分之间的关系，融合句义特征，能够提取候选事件元素词语深层次的句义信息，增强了特征向量的表达能力；3、词语组合特征对语义上联系紧密的词语进行组合，能对事件元素识别界限起到指导作用。

第四组实验是元事件抽取实验，目的是验证本文提出的元事件抽取方法的有效性以及引入句义结构模型能提升元事件抽取效果。将仅基于位置信息的方法作为 Baseline，与融合句义结构模型后的方法（Our Method）进行比较，结果如表 8 所示。

表 8元事件抽取结果

Tab.8 Results of meta event extraction

	Precision	Recall	F-score
Baseline	0.686	0.4749	0.561
Our Method	0.6549	0.5528	0.5994

由表 8 可见，融合句义结构后元事件抽取的准确率有所下降，但是召回率和 F 值都有所提升，其中召回率平均提高 7.8%，F 值平均提高 3.8%，说明在综合性能方面融合句义结构后的元事件抽取策略比仅基于位置信息的元事件抽取策略要好，并且能很好的平衡准确率和召回率。分析原因，主要有以下两点：1、引入句义结构模型能够对句子进行深层语义分析，能够提取出小句中 CRF 未识别出来的事件元素；2、句义结构模型能够综合分析句义成分以及它们之间的关系，克服了基于位置信息跨小句抽取事件的局限性。

由表中数据可知，本文提出的元事件抽取方法平均 F 值为 59.9%，说明该方法能比较有效的抽取句中元事件。相比触发词和事件元素识别的平均 F 值都达到 74%以上，元事件抽取效果还存在一定的差距。分析原因主要有以下两点：1、从触发词和事件元素映射到元事件存在一定的级联错误；2、由于句子结构复杂，获取元事件对应句义成分的方法有一定的局限性，使得跨小句部分元事件的事件元素抽取不完整。

5 结论

本文对事件抽取的关键技术进行了研究，在触发词识别和事件元素识别的基础上，提出了融合句义特征的元事件抽取方法。针对触发词识别任务，选取合适的特征，采用CRF模型进行二元分类，并采用后处理规则对识别结果进行优化；针对事件元素识别任务，采用双层CRF模型并结合句义特征进行分类。在CEC语料上进行测试，平均F值分别达到79%和74.8%。实验表明，本文提出的触发词和事件元素识别方法可以有效提高事件识别的精度。针对元事件抽取问题，在前两部分识别结果的基础上，根据位置信息进行共指判别，并结合句义结构分析，在事件映射模式的指导下抽取相应的句义成分，完成元事件抽取。在CEC语料上进

行测试，平均F值达到59.9%。实验表明，本文提出的元事件抽取方法能有效提取句中元事件。同时整个元事件抽取过程是在机器学习和通用事件模式的指导下完成的，不涉及人为干预，能为现阶段事件抽取面临的领域受限问题提供一定的参考价值。

下一步可尝试根据已识别出的事件元素通过句义结构分析识别 CRF 模型未识别出的触发词。同时在触发词和事件元素识别上还有待寻找泛化能力更强的特征，可尝试添加适当的规则，来提高识别的准确率。对于长句子分析，可通过一定的规则进行分解，转换成短句分析。这些都是接下来要重点研究的方向。

考参文献:

[1] LIAO Tao, LIU Zongtian, WANG Xianchuan. Research and implementation on event-based method for automatic summarization. Advances in Intelligent and Soft Computing. Springer-Verlag. 2013.

[2] DAVID Ahn. The stages of event extraction[A]// Proceedings of the Workshop on Annotations and Reasoning about Time and Events[C]. 2006: 1-8.

[3] ACE(Automatic Content Extraction) Chinese Annotation Guidelines for Events. National Institute of Standards and Technology[R]. 2005.

[4] 姜吉发. 自由文本的信息抽取模式获取的研究[D]. 北京: 中国科学院, 2004.
JIANG Jiefa. A research about the pattern acquisition for free text IE[D]. Bei jing: Chinese Academy of Sciences, 2004.

[5] 吴刚. 基于主题的中文事件抽取技术研究及应用[D]. 苏州大学, 2009.
WU Gang. Research and application on Chinese topic event extraction[D]. Soochow University, 2009.

[6] YANKOVA M. Focusing on scenario recognition in information extraction[C]// Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2. Association for Computational Linguistics, 2003: 41-48.

[7] CHIEU H L, NG H T. A maximum entropy approach to information extraction from semi-structured and free text[J]. Proceedings of the18th National Conference on Artificial Intelligence, 2002: 786~791.

[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.
Zhao Yanyan, Qin Bing, Che Wanxiang, et al. Research on Chinese event extraction[J]. Journal of Chinese Information Processing, 2008, 22(1): 3-8.

[9] CHEN Z, JI H. Language specific issue and feature exploration in Chinese event extraction[C]// Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Association for Computational Linguistics, 2009: 209-212.

[10] LLORENS H, SAQUETE E, et al. TimeML events recognition and classification learning CRF models with semantic roles[C]// Proceedings of the 23rd International Conference on Computational, 2010.

[11] Heng Ji. Cross-lingual Predicate Cluster Acquisition to Improve Bilingual Event Extraction by Inductive Learning[C]. In Proceedings of the NAACL HLT Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics, pages 27-35. Boulder, Colorado, 2009: 27-35.

[12] LIAO Shasha, GRISHMAN Ralph. Using Document Level Cross-Event Inference to Improve Event Extraction[C]. In Porc. ACL-2010, Uppsala, Sweden, July, 2010: 789-797.

[13] 罗森林, 韩磊, 潘丽敏等. 汉语句义结

构模型及其验证[J]. 北京理工大学学报(自然科学版), 2013, 33(2): 166-171.
LUO Senlin, HAN Lei, PAN Limin, et al. Chinese sentential semantic mode and verification[J]. Beijing Institute of Technology (Natural Science), 2013, 33(2): 166-171.
[14] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the Eighteenth International Conference on Machine Learning, 2001: 282-289.
[15] CHEN A, PENG F, SHAN R, et al. Chinese named entity recognition with conditional probabilistic models[C]// 5th SIGHAN Workshop on Chinese Language Processing. 2006: 173-176.
[16] 黄德根, 焦世斗, 周惠巍. 基于子词的双层 CRFs 中文分词[J]. 计算机研究与发展, 2010 (5): 962-968.
HUANG Degen, JIAO Shidou, ZHOU Huiwei. Dual-layer CRFs based on sub-word for Chinese word segmentation[J]. Journal of Computer Research and Development, 2010(5): 962-968.
[17] 付剑锋. 面向事件的知识处理研究[D]. 上海大学, 2011.
FU Jianfeng. Research on event-oriented knowledge processing[D]. Shanghai University, 2011.
[18] TIAN L, MA W, ZHOU W. Automatic event trigger word extraction in Chinese event[J]. Journal of Software Engineering and Applications, 2013, 5: 208.
[19] 丁效. 句子级中文事件抽取关键技术研究[D]. 哈尔滨工业大学, 2011.
DING Xiao. Research on sentence level Chinese event extraction[D]. Harbin Institute of Technology, 2011.