基于依存句法和短语结构句法结合的金融领域事件元素抽取

孟 雷，丁 效，秦 兵，刘 挺

哈尔滨工业大学 计算机学院 信息检索研究室，哈尔滨 150001

E-mail: {lmeng, xding, bqin, tliu}@ir.hit.edu.cn

摘 要：事件抽取是信息抽取领域一个重要的研究方向。针对金融领域特定事件的事件元素抽取，本文提出了基于依存句法分析的事件元素核心词抽取方法，并结合短语结构句法分析进行事件元素完整边界的识别。实验表明，依存句法结合规则可以有效地抽取事件元素核心词，再结合短语结构句法则可以比较准确的识别出完整的事件元素。

关键词：依存句法；短语结构句法；事件抽取；事件元素

Financial Event Argument Extraction Based on Dependency Parsing and Noun Phrase Parsing

Meng Lei, Ding Xiao, Qin Bing, Liu Ting

Information Retrieval Laboratory of Computer Science & Technology School, Harbin Institute of Technology, Harbin 150001

E-mail: {lmeng, xding, bqin, tliu}@ir.hit.edu.cn

Abstract: Event extraction is an important research area in information extraction field. Aiming to extracting the event arguments of particular event in financial field, this paper presents the extraction method of the event argument's key word based on dependency parsing; and we have also applied the Noun Phrase parsing to the recognition of the noun phrase where the event argument locates. The experiment shows that the dependency parsing and the rules could efficiently extract the key word of the event argument, and the Noun Phrase parsing could accurately recognize the whole event argument.

Keywords: dependency parsing; noun phrase parsing; event extraction; event argument extraction

1 前言

事件抽取是将含有事件信息的非结构化文本以结构化的形式呈现出来，在多文档文摘，自动文摘，自动问答[1]和信息检索领域有着广泛的应用。ACE(Automatic Content Extraction)于 2005 年引入了事件抽取（Event Detection and Recognition、Event Mention Detection）评测任务。事件抽取主要分为事件类型识别和事件元素抽取两个关键任务。其中事件元素抽取对充分理解事件起着至关重要的作用。近些年来进行事件元素抽取的方法主要可以分为两类，一类是基于机器学习的方法，另一类是基于模式挖掘和匹配的方法。

基于机器学习的事件元素抽取方法需要首先给出候选的事件元素，然后通过如 Maxent，SVM (Support Vector Machine)等分类器来确定该候选事件元素是否为该事件的事件元素。这里的候选事件元素往往是固定的一类具体的实体，比如公司，人，项目名，演唱会名等。候选事件元素自身和其上下文的信息可以作为分类器的特征。该方法能够取得较好的召回率和准确率，但是如果要抽取的事件元素不是固定的一类实体，而是很泛化的内容，这种方法就很难被应用于事件元素的抽取过程中。

基于句法树进行模式挖掘和匹配的方法需要预先将句子中的事件元素和触发词进行标注，然后通过对事件元素和触发词在句法树中的关系进行模式的挖掘和构建。在模式的挖掘和构建过程中，非常重要的就是要找到高质量的模式，使得挖掘回来的模式既能准确的召回事件所涉及的事件元素，又不过多的引入噪声。该方法如需要在抽取的准确率和召回率之间平衡。

在某些特定域的事件抽取中，事件元素的实体类型比较容易确定；比如音乐领域中的演唱会类事件，事件元素类型为歌手和演唱会名，都是非常具体的实体。而在对金融语料的分析过程中发

现，金融领域事件词所涉及的事件元素范围非常广泛，所涉及的事件元素也很难被归类为一类具体的实体，无法确定具体的实体类型。其次机器学习方法要求标注大量的语料，工作量巨大。因此基于机器学习的事件元素抽取方法在金融领域上的事件元素抽取上无法应用。本文采用了基于模式匹配的抽取方法，手动构建候选规则用于提高事件元素抽取的召回率，并与短语结构相结合以提高事件元素抽取的准确率。

本文内容组织为：第二部分介绍系统框架；第三部分介绍金融领域的事件抽取方法；第四部分为实验部分；第五部分为结论及进一步工作。

2 系统框架

事件元素抽取的流程主要分三个部分：事件类型确定，获取事件元素核心词，识别事件元素名词短语（见图 1）；下面简要介绍一下各个部分所负责的功能和它们之间的关系。

（1）事件类型确定

该部分负责进行触发词的识别，从而可以使得事件元素抽取系统依据匹配出的触发词类别进行后续的抽取工作。本文的触发词是经过动词细分类进行过滤，再通过基于的 HowNet 进行动词聚类和人工的调整从而形成了触发词的类别。在后面的事件抽取系统中，我们对不同的触发词类别采取不同的抽取模式和候选规则进行抽取。

（2）获取事件元素核心词

本文采用了本实验室开发的依存句法分析结合候选规则进行事件元素核心词的获取。在系统中依存句法分析负责对有触发词的句子进行预处理，具体包括分词，词性标注，句法分析。事件元素抽取系统结合具体的抽取模式和手工设计的候选规则对进行过预处理后的句子进行事件元素核心词的抽取。

（3）识别事件元素名词短语

本文采用了 MIT 开发的 DBParser 短语结构句法分析器进行名词短语的识别。对于依存句法分析后的事件元素核心词，短语结构句法分析器负责对其所在的名词短语进行名词短语的识别，从而形成最终的完整的事件元素。需要指出的是，该短语结构句法分析器的输入是分词后的句子，然后输出名词短语分析结果，这样就可以在短语结构句法分析结果中定位事件元素核心词。因此和依存句法分析器可以很好的结合在一起，而不会产生冲突。

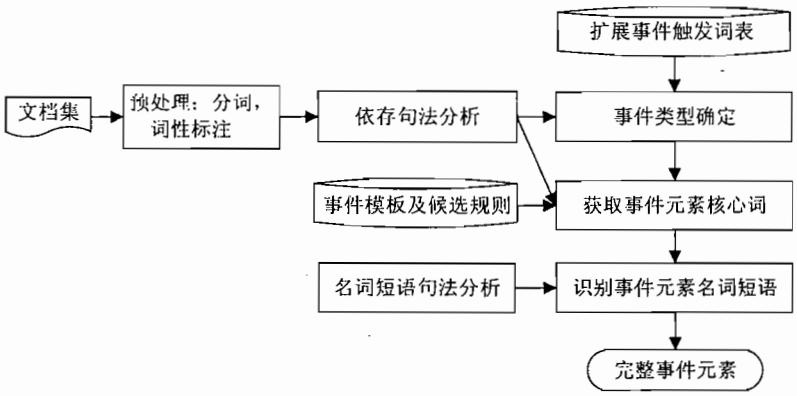


图 1 事件元素抽取系统框架图

3 金融领域事件抽取

在金融领域的事件抽取中，首先需要确定要抽取的事件类型，通过动词细分类进行动词过滤

来确定金融领域具有实际意义的动词，再对这些实义动词所涉及的事件元素进行抽取。通过进一步的研究发现，这些事件词在金融语料中大部分都是以谓语动词形式出现，只要应用主谓宾模板召回这些事件词的主语和宾语，基本上就召回了这个事件词所涉及的事件元素，而不需要引入其他过多的模式进行抽取。

在实际的事件元素抽取过程中，在一些情况下依存句法分析无法明确给出一个事件的主语事件元素或者宾语事件元素。为了提高事件元素抽取的召回率，本文在依存句法分析的基础上给出了候选事件元素抽取方法。对于依存句法分析器和候选规则给出事件主语关键词或者宾语关键词，再结合短语结构句法分析器给出的名词短语结构识别出主语或者宾语所在的名词短语，从而给出完整的主语事件元素和宾语事件元素。在后面的实验证明，本文给出的候选事件元素抽取方法抽取出的事件元素占正确抽取出来的事件元素 35%左右，对于提高事件元素抽取的召回率起了很重要的作用，对于单纯基于依存句法分析结果是一个有益的补充。短语结构句法分析器用来进行事件元素核心词所在的名词短语的识别，可以有效地提升要抽取的事件元素的准确率。

3.1 事件元素核心词抽取

在事件元素的抽取中，对于大部分的触发词都是采用主谓宾模式进行抽取。但是在某些情况下主语和宾语无法由依存句法分析器给出，通过对语料的分析，手动构建了三条主语事件元素抽取候选规则和三条宾语事件元素候选抽取规则，分别是：直接抽取触发词左边（宾语为右边）的名词短语，抽取考虑句法分句边界的触发词左边（宾语为右边）的名词短语，抽取前面（宾语为后面）分句的主语。

在具体的主语事件元素抽取过程中，优先采用句法分析器直接给出的主语事件元素，毕竟依存句法分析器给出的结果在大部分情况下还是最准确的，对于其他 3 种候选抽取方法，采用对 3 种抽取方法进行优先级排序，通过测试语料来看哪一种排序方法能够在保证准确率的同时比较大幅度的提高主语事件元素的召回率，从而决定最终抽取该类事件触发词的候选抽取方法排序。下面对本文所具体设计的主语事件元素候选规则进行介绍。宾语部分由于与其非常类似，基本上和主语候选规则是对称的关系，因此不具体展开介绍。

（1）直接抽取依存句法树中触发词的主语

该方法即是直接抽取依存句法分析器给出的主语成分，比如对于句子“二手房价格如果跌到2007 年年初水平。”，依存句法分析器给出触发词"跌"的主语即是“价格”，通过依存弧上关系将“二手房价格”全部召回作为主语。

（2）候选规则 1 前置分句主语作为主语

具体来讲就是抽取事件触发词前面那个分句的第一个谓语动词的主语成分（后面简称为 PrevSub）。对于句子“成交量放大至 1743.5 亿元，创下新高。”，无法直接找出触发词“创下”的主语，而通过找到前面谓语动词“放大”的主语“成交量”，该主语也是“创下”的主语，从而找到了事件触发词的主语。

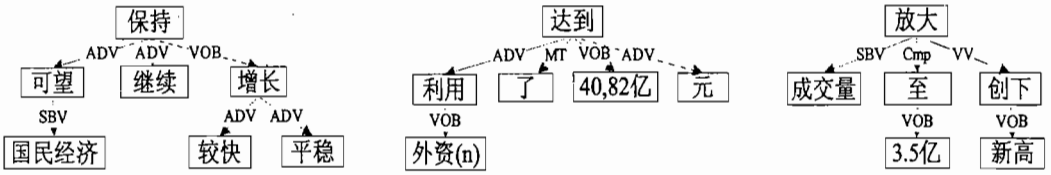
（3）候选规则 2 考虑句法边界的前置名词作为主语

定义 1 触发词句法关系的最左儿子：在同一个分句中，依存于触发词（依存句法弧从触发词触发指向该词）且位于触发词最左侧的那个词称为触发词句法关系的最左儿子。

比如对于句子“国民经济可望继续保持平稳较快增长。”，“可望”和触发词“保持”有 ADV 句法关系，那么就将“可望”就是触发词句法关系的最左儿子（后面简称为 PrevIC）。但本句直接提取主语是提取不出来的，因为没有和触发词“保持”具有“SBV”关系的词。但是这个句子中确实存在主语 “国民经济”，在图 2 中可以清晰地看出“国民经济”是离触发词“保持”句法关系最左儿子“可望”最近的名词短语，将其作为候选事件主语元素。

（4）候选规则3 不考虑句法规则边界的前置名词作为主语

抽取触发词左边的名词短语作为主语事件元素（后面简称为 PrevNC）。比如句子“利用外资达到了 40.82 亿元。”，“外资”是触发词“达到”左边最近的那个名词短语，因此将其抽取作为候选主语事件元素。



候选规则1例句　　候选规则2例句　　候选规则3例句

图2　短语结构句法分析图

3.2　事件元素名词短语识别

在事件元素抽取中，可以通过依存句法分析器给出的依存句法弧识别出主语或者宾语所在的名词短语，从而给出完整的主语事件元素和宾语事件元素。在采取这种方法时，通过遍历依存句法树中在核心词为树根的子树下面的所有具有修饰关系的子节点进行名词短语的识别。但通过测试发现，该方法识别的事件元素的精确评价准确率低。经过分析，这是由于依存句法树是基于词与词的依存关系，在名词短语结构中有一个依存关系分析错误，就会造成整个名词短语的抽取错误，因此对于依存分析器的准确率的要求非常高；为了能够抽取出更加准确的事件元素，本文采用了短语结构句法分析器来代替依存句法分析器来进行主语或者宾语所在名词短语的识别。

本文采用了 MIT 开发的 DBParser 短语结构句法分析器。对于已经抽取出事件核心词的句子，通过短语结构句法分析器分析出句子的名词短语树结构，然后通过定位事件核心词在名词短语树中的叶子节点位置，逆向的找到其祖先节点中为 NP 节点中辈分最高的那一个节点，然后通过遍历该节点的子树识别出了事件元素核心词所在的名词短语，从而得到完整的事件元素。如句子“金三银四的楼市行情增强了开发商的信心。”，可以得到.“行情”为增强的主语核心词，在短语结构句法树中其辈分最高的节点为“NP”（加粗部分），通过遍历该节点的子树即可得到整个名词短语“金三银四的楼市行情”；而增强的宾语为“信心”，同样可以得到完整的宾语事件元素核心词为“开发商的信心”。

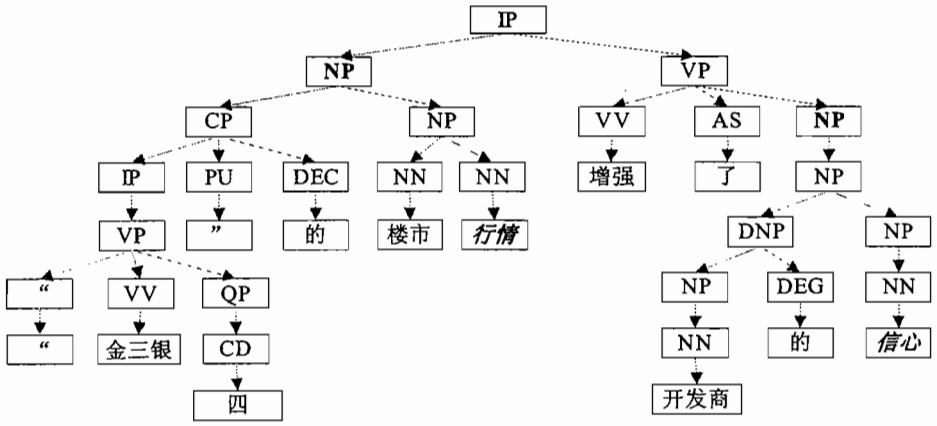


图3　短语结构句法分析图

4 实验结果与分析

4.1 准确率和召回率测试

测试针对要抽取的事件类型的主语事件元素和宾语事件元素进行准确率和召回率的测试。本部分测试结果是基于依存句法抽取的主语和宾语所组成的名词短语，由于基于依存短语句法识别的短语结构的精确准确性较低，因此除了一般的基于精确匹配（标注与抽取结果完全一致）的准确率和召回率测试之外，还进行了基于模糊匹配（标注是抽取结果的子串或抽取结果是标注结果的子串）的准确率和召回率测试，从而评估设计的候选规则定位是否准确。表 1 是部分触发词的测试结果。宾语的精确匹配的准确率和召回率大概在 30%-80%之间，模糊匹配的准确率和召回率在 70%-90%之间，这是由于大部分事件触发词的宾语相对简单，比较容易抽取。而主语的精确匹配的准确率和召回率都相对较低，在 20%-40%之间，模糊匹配的准确率和召回率在 60%-80%之间。综合来看，本文设计的候选规则基本上达到了定位的作用。

表 1 部分触发词测试结果

抽取类别	精确准确率（%）	精确召回率（%）	模糊准确率（%）	模糊召回率（%）	语料规模（句）
主语（创下）	36.07	36.07	72.15	72.15	94
宾语（创下）	67.21	65.78	90.96	88.06	
主语（达）	20.08	20.08	67.87	67.74	702
宾语（达）	73.57	71.47	94.52	91.69	
主语（累计）	29.20	29.20	67.43	67.43	189
宾语（累计）	55.81	55.22	88.55	87.61	
主语（暴跌）	36.44	36.44	71.94	71.94	334
宾语（暴跌）	36.44	31.50	88.41	75.94	

4.2 候选事件元素抽取排序测试

以触发词“*达*”为例，应用不同的候选主语事件元素抽取方法排序有着不同的（主语模糊）结果。抽取元素方法排序 1 为“SBV→PreSub→PrevNC→PrevIC”，抽取元素方法排序 2 为“SBV→PreNC→PrevSub→PrevIC”。

通过对表 2 可以看出，不同的候选抽取排序对抽取结果的确有着较大影响，排序 2 要比排序 1 的召回率高 6%，因此“达”这个触发词的候选主语事件元素更适合优先用 PrevNC 方法进行抽取；在宾语候选事件触发词也有着类似的现象。从表 2 还可以看出，候选规则 PrevNC 准确率为 72.36%，依存句法直接给出的 SBV 为 60.65%，本文所给出的候选规则在定位的准确性上并不比依存句法直接给出的结果差；PrevNC 候选规则召回的元素占总召回约 1/3，可以说很好的补充了依存句法的抽取结果。

4.3 名词短语识别对比测试

本文采用短语结构句法替代依存句法进行事件元素所在名词短语的识别工作，这是由于短语结构句法对于句子的分析是基于句子分块的方式，因此粒度较粗，在进行名词短语识别时发生错误的概率要低于依存句法。本文选取了抽取效果较好的事件类别中约 300 句进行对比测试。实验证明，宾语事件元素的精确准确率有了大幅度的提高，在准确率上提升了 20%的准确率，主语上也有 10%的提高（表 3）。通过实验可以证明短语结构句法更适合进行事件元素所在的名词短语的识别。

表2 事件元素候选方法排序对比测试

抽取方法1	抽取（句）	准确率(%)	召回率（%）	抽取方法2	抽取（句）	准确率(%)	召回率（%）
SBV	521	60.65	65.29	*SBV*	521	60.65	65.29
PreSub	240	52.91	30.07	*PreNC*	240	72.36	34.46
PrevNC	36	88.89	4.51	*PrevSub*	——	——	——
PrevIC	1	100	0.13	*PrevIC*	1	100	0.25
总计	701	67.87	67.77	总计	702	73.61	73.61

表3 名词短语识别对比测试

	精确准确率	精确召回率	模糊准确率	模糊召回率
依存主语	30.68%	30.68%	58.80%	58.80%
名词短语主语	40.90%	40.90%	70.44%	**70.74%**
依存宾语	57.94%	62.21%	88.89%	88.89%
名词短语宾语	76.19%	**81.82%**	92.06%	92.06%

5 结论及未来工作

本文针对金融领域的事件抽取中的事件元素抽取相关工作展开研究。本文基于依存句法的主谓宾模板手动设计了6种候选事件元素抽取规则，该规则与句法分析共同给出了事件元素核心词的位置；并且结合短语结构句法分析进行了事件元素所在名词短语的识别。最终我们的实验结果表明，我们的候选规则结合句法分析给出的直接结果可以很好的定位事件元素核心词，再结合短语结构句法分析可以进一步提升抽取的事件元素的质量，从而达到比较好的抽取效果。

在未来的工作我们需要进一步去挖掘我们所设计的候选规则在何种句式中应用比较合适，提炼出更具有普适意义的候选规则，从而使得本文设计的候选规则能适应不同领域的需求。

参 考 文 献

[1] Ellen Riloff. Automatically Generating Extraction Patterns from Untagged Text. In Proceedings of the Thirteen National Conference on Artificial Intelligence. AAAI-96. pp. 1044-1049.

[2] Roman Yangarber Automatic acquisition of domain knowledge for Information Extraction. Proc. COLING 2000.

[3] 赵妍妍, 秦兵, 车万翔, 刘挺, 中文事件抽取技术研究, 中文信息学报 2008Vol.22 No.1 pp3-8.

[4] Shasha Liao. Using Document Level Cross-Event Inference to Improve Event Extraction. Proc. ACL 2010.

[5] David Ahn. The stages of event extraction[A]. In: Proceedings of the Workshop on Annotations and Reasoning about Time and Events[c]. 2006. 1-8.

[6] Heng Ji. Refining Extraction through Cross-document Inference. Proc. ACL 2008.

[7] K Sudo. An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition. Proc ACL 2003.

[8] Hai Leong Chieu. A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text. AAAI 2002.

[9] ACE Chinese Annotation Guidelines for Events. National Institute of Standards and Technology, 2005.