基于动词论元结构的中文事件抽取方法

肖 升[1,2] 何炎祥[1]

(武汉大学计算机学院 武汉 430072)[1] (湖南第一师范学院信息科学与工程系 长沙 410205)[2]

摘 要 为将动词与其论元间的约束规则应用于事件抽取,在事件模型中引入动词论元结构形成模型变体,围绕模型变体提出基于动词论元结构的中文事件抽取方法。此方法首先对待抽取文本进行预处理和句法分析,得出其语法结构;然后将所得结构与动词论元结构属性进行比较,找出每个动词支配的论元;最后利用论元的语义属性确定与之对应的事件特征并由此完成事件抽取。实验结果显示,此方法能有效提高抽取系统的性能和效率。

关键词 动词,论元结构,事件特征,触发词,事件模型,中文事件,信息抽取

中图法分类号 TP391 **文献标识码** A

Approach of Chinese Event IE Based on Verb Argument Structure

XIAO Sheng[1,2] HE Yan-xiang[1]

(School of Computer Science, Wuhan University, Wuhan 430072, China)[1]

(Department of Information Science and Engineering, Hunan First Normal College, Changsha 410205, China)[2]

Abstract For the purpose of using the constraintrules between verbs and their arguments in the event extraction, an approach of Chinese event IE based on verb argument structure was proposed around the model variant, which was formed by introducing the verb argument structure into an event model. First pre-treating and syntactic analysis of text were made to get its grammatical structure. Then the structure was compared with the properties of verb argument structure to find out the disposal argument of each verb. Finally the features of the corresponding events were determined by the semantic properties of argument and the event extraction was completed. Experimental results show that this method can improve the extraction system performance, increase the efficiency of extraction.

Keywords Verb, Argument structure, Event feature, Trigger, Event model, Chinese event, Information extraction

1 引言

"事件"(Event)源自认知科学,被认为是人类感知和认识世界的一种基本单位。中文事件抽取是中文信息抽取重要的研究方向,它对非结构化文本进行结构化处理,并将结果应用于自动文摘、机器问答、情报检测等领域[1]。

事件模型是事件抽取的基础,其核心是触发词和事件特征。由于动词通常包含陈述事件的关键信息,因此基于动词的事件模型备受关注。此类模型中,触发词是动词,事件特征则是分布在动词周围的名词或名词短语[2]。例如在出生模型中,触发词是"出生"、"生"等动词,事件特征则是分布在其周围表人物、时间、地点的名词或名词短语。

目前,基于动词事件模型的研究焦点集中在动词本身,即从动词自身的词汇意义和概念意义来考察其所描述的事件,这种视角便于事件的识别和分类[3]。但与此同时,它忽视了动词对周围名词或名词短语的约束。近20年国内外配价语法、格语法,特别是论结构理论的研究成果表明,分布在动词周围的名词或名词短语很可能就是动词支配的论元,这些论元的结构属性是相对固定且可穷尽的,如果把这些属性应用于事件特征提取,定将大受其益[4,5]。比如,在提取"邓小平1904年生于四川广安"的事件特征时,首先对其进行预处理,结果为"邓小平$^{NE}_{N}$+1904年$^{TE}_{NP}$+生$_{V}$+于$_{P}$+四川广安$^{PE}_{NP}$"(上标NE、TE、AE分别表示人名实体、时间实体、地址实体,下标N、NP、V、P分别表示名词、名词短语、动词、介词);然后对其进行句法分析,结果为"邓小平$_{Sub}$+1904年$_{Adv}$+生于$_{Pre}$+四川广安$_{Adv}$"(下标Sub、Adv、Pre分别表示主语、状语、谓语);此时,若将动词"生"的论元结构属性,如共现方式"Ex+T+生+于(/在)+P"(Ex、T、P分别表示经事论元、时间论元、地点论元)、句法成分"Ex可充当Sub,T、P可充当Adv"和范畴特征"Ex、T、P可用N(/NP)表示"作为判断规则引入,显然可以确定"生"的Ex为"邓小平",T为"1904年",P为"四川广安";再利用Ex、T、P分别求解人物、时间、地点就可明确论元代表的事件特征,从而完成事件特征提取。动词的论元结构属性之所以有助于事件特征提取,是因为其内容本质上就是事件模型中触发词和事件特征之间的约束规则,这些规则可以帮助动词定位其论元,并确定论元代表的事件

到稿日期:2011-07-27 修返日期:2011-11-02 本文受国家自然科学基金项目(60703008),湖南省教育厅科学研究项目(10C0527),湖南省高校科技创新团队支持计划(湘教通[2010]212号),湖南省科技厅高新计划项目(2010GK3049)资助。

肖 升(1980—),男,博士后,副教授,主要研究方向为中文信息处理,E-mail:xiaosheng@mail.ccnu.edu.cn;何炎祥(1952—),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为可信软件、自然语言处理。

特征。

完成事件特征提取事实上就完成了事件抽取，加之论元结构是论元结构属性的载体，因此基于动词论元结构的事件抽取方法无疑值得研究。有鉴于此，本文首先在事件模型中引入动词论元结构，形成基于动词论元结构的事件模型，然后在此基础上探讨相关的事件抽取方法，并将其应用于中文职务变更事件，最后借助实验数据求证该方法的有效性。

2 动词论元结构及其属性

定义 1(论旨角色 TR, Thematic Role) 由动词根据与其相关的名词或名词短语之间的语义关系而指派给(assign)这些名词或名词短语的语义角色。动词有其固定的论旨角色，这些角色表示动词所涉及的主体、客体、动作、行为、状态、处所等。由于厘定标准有所不同，因此所得论旨角色的数目及名称理应存在差别。目前公认的论旨角色包括施事(agent)、感受者(experiencer)、受惠者(benefactive)、使役者(causer)、客体(theme)等，受影响的客体通常被称作受事(patient)。本文遵从文献[6]的标准，认定与汉语动词相关的论旨角色共17种，其具体的名称和句法语义特征请参考相关文献。

定义 2(论元 A, Argument) 带有论旨角色的名词或名词短语。

定义 3(论元结构 AS, Argument Structure) 动词支配的论元及其间关系。论元结构具有下列属性：

定义 4(论元数目 AA, Argument Amount) 动词所能支配的论元个数以及每个论元所具有的必要或可用属性。

定义 5(论元角色 AR, Argument Role) 论元的论旨角色。

定义 6(论元属性 AP, Argument Property) $AP=(AA, AR)$，即论元数目和论元角色合称论元属性，其内容可以参考配价语法、格语法和论元结构理论的研究成果。

定义 7(共现方式 AC, Argument Co-occurrence) 动词的论元在可能句式中同时或选择性出现所受到的约束。

定义 8(句法成分 SC, Syntactic Component) 论元在句中充当的句法成分。比如，施事、受事通常充当主、宾语。

定义 9(范畴特征 CF, Categorical Feature) 表达论元的词类范畴。比如，施事、受事通常由名词性成分来表达，致事通常由名词或动词性成分来表达等。

定义 10(语法特征 GF, Grammatical Feature) $GF=(AC, SC, CF)$，即论元的语法特征包括共现方式、句法功能和范畴特征，其内容可参考论元结构理论的研究成果。

定义 11(静态语义特征 SSF, Static Semantic Feature) 论元在述谓结构中表现出来的施动性、受动性等语义特征。

定义 12(动态语义特征 DSF, Dynamic Semantic Feature) 表达不同论旨角色的词语在语义上受到的约束。比如，施事、与事通常由指人名词来实现，受事则既可以由指人又可以由指物名词来实现。

定义 13(语义特征 SF, Semantic Feature) $SF=(SSF, DSF)$，即论元的语义特征包括静态语义特征和动态语义特征，其内容可参考词汇语义学和论元结构理论的研究成果。

定义 14(论元结构属性 ASP, Argument Structure Property) $ASP=(AP, GF, SF)$，即根据中文信息处理的应用需要，动词的论元结构属性包括论元属性、语法特征和语义特征3大类共7个指标。

3 基于动词论元结构的事件模型

定义 15(事件特征模块 EFM, Event Feature Module) $EFM=\{KI_1:SF_1(C_{11}, C_{12}, \cdots); KI_2:SF_2(C_{21}, C_{22}, \cdots); \cdots; KI_n:SF_n(C_{n1}, C_{n2}, \cdots)\}$，其中，$KI_1, KI_2, \cdots, KI_n$ 是事件所包含的关键信息(Key Information)；$SF_1, SF_2, \cdots, SF_n$ 是关键信息的语义特征(Semantic Feature)；$C_{n1}, C_{n2}, \cdots$ 是诠释这些语义特征的概念(Concept)。EFM 的功能一方面是描述事件特征，另一方面是用概念来诠释事件特征的语义特征[7]。

定义 16(触发词模块 TM, Trigger Module) $TM=\{KI_1(T_1:TP_1); KI_2(T_2:TP_2); \cdots; KI_n(T_n:TP_n)\}$，其中，$KI_1, KI_2, \cdots, KI_n$ 同定义 15；$T_1, T_2, \cdots, T_n$ 是触发关键信息的词语，即触发词；$TP_1, TP_2, \cdots, TP_n$ 是触发词属性(Trigger Property)。TM 的功能是描述触发词属性及其与关键信息之间的触发关系。

定义 17(分析模块 AM, Analysis Module) $AM=S_1(CW_1, GC_1, SC_1) \wedge S_2(CW_2, GC_2, SC_2) \wedge \cdots \wedge S_n(CW_n, GC_n, SC_n)$。其中，$S_1, S_2, \cdots, S_n$ 是抽取文本经分句处理后得出的小句(Sentence)；$CW_1, CW_2, \cdots, CW_n$ 是小句 S 的核心词(Core Word)；$GC_1, GC_2, \cdots, GC_n$ 和 $SC_1, SC_2, \cdots, SC_n$ 是对小句 S 进行语法分析后得到的语法表征(Grammar Characterization)和语义表征(Semantic Characterization)。AM 的功能是对抽取文本中切分出来的语句进行语法、语义分析，所得结果将与 EFM 和 TM 的分析结果进行比较，并由此确定待抽取信息的具体位置和内容[8,9]。

定义 18(事件模型 EM, Event Model) $EM=(EFM, TM, AM)$，即事件模型由事件特征模块、触发词模块和分析模块组成。

基于动词论元结构的事件模型是事件模型的变体，形成变体的替代规则有如下4条(":="表示替代)：

1)事件关键信息由动词论元替代，即 $KI:=A$；

2)诠释事件特征的语义特征及概念集合由诠释动词论元的语义特征和概念集合替代，即 $SF:=SF_V, C:=C_V$；

3)触发词由动词替代，触发词属性由动词论元结构属性替代，即 $T:=V, TP:=ASP$；

4)核心词由动词替代，即 $CW:=V$。

依据上述规则，基于动词论元结构的事件模型可定义为：

定义 19(基于动词论元结构的事件模型 EM_{AS}, Event Model Based on Argument Structure) $EM_{AS}=(EFM_{AS}, TM_{AS}, AM_{AS})$。其中，$EFM_{AS}=\{A_1:SF_{V1}(C_{V11}, C_{V12}, \cdots); A_2:SF_{V2}(C_{V21}, C_{V22}, \cdots); \cdots; A_n:S_{Vn}(C_{Vn1}, C_{Vn2}, \cdots)\}$，$TM_{AS}=\{A_1(V_1:ASP_1); A_2(V_2:ASP_2); \cdots; A_n(V_n:ASP_n)\}$，$AM_{AS}=S_1(V_1, GC_1, SC_1) \wedge S_2(V_2, GC_2, SC_2) \wedge \cdots \wedge S_n(V_n, GC_n, SC_n)$。

4 事件特征抽取步骤及实例

下文引入 EM_{AS} 的模块功能图(见图1)，参照此图并通过职务变更事件的实例来说明基于 EM_{AS} 的事件特征抽取步骤。

1)事件模板设定：根据需求确定待抽取的事件特征，形成事件模板。由于 EFM_{AS} 中的 A(论元)替代了 EFM 中的 KI

(关键信息),因此明确 A 的个数及其必要或可用属性(即 AA)就可设定事件模板。原则上,必要论元肯定要被设定为模板元素,而可用论元则需视情况而定。职务变更动词有变更对象(H)、变更职务(P)2 个必要论元和变更组织(O)、变更时间(T)2 个可用论元,由于 O 和 T 也是标识职务变更的重要特征,因此上述 4 者将组成事件模板。

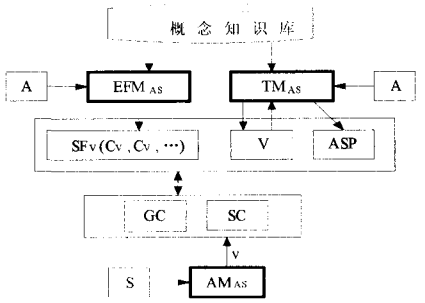

图 1 EM_{AS}模块功能图

2)待抽取文本预处理:

①分句:承载动词及其论元的句子是小句(包括单句和复句中的分句)而不是整句,因此在使用分句工具时,除"。"、"?"、"!"外,","和";"也将被设定为句末标识符。

②命名实体识别:选取一组命名实体识别软件对表达事件特征的命名实体进行判别并标注。职务变更事件需要 4 个功能模块分别对人名实体 NE(Names Entity)、职务实体 PE(Position Entity)、组织实体 OE(Organization Entity)和时间实体 TE(Time Entity)进行识别。除人名实体外,前 3 种命名实体都是可穷尽的,除少数不规则形式外,这 3 种命名实体也是可列举的,比如,职务实体可以用"一长、代一、副一"等列举,组织实体可用"一党、一国/州/省/地区/直辖市/市/县/区、一部/厅/局/处、一厂/公司、一委员会/协会/理事会/董事会、一校/院/系/所/室"等列举,时间实体可用"一年一月一日"等列举。

③命名实体组合:命名实体除单独存在外,还可能形成组合体。组合体将被视为具有最后一级实体特征的整体标识。职务变更事件中的命名实体的组合规则如下:

(1)$E_{OE}+E_{PE}=E_{PE}$。例如,湖北省委$_{OE}$+〈的〉书记$_{PE}$=湖北省委〈的〉书记$_{PE}$。

(2)$E_{PE}+E_{NE}=E_{NE}$。例如,原总经理$_{PE}$+周鹏$_{NE}$=原总经理周鹏$_{NE}$。

(3)$E_{OE}+E_{NE}=E_{NE}$。例如,政治部$_{OE}$+〈的〉邓凯$_{NE}$=政治部〈的〉邓凯$_{NE}$。

(4)$E_{OE}+E_{PE}+E_{NE}=E_{NE}$;例如,总部$_{OE}$+〈的〉部长$_{PE}$+宋文$_{NE}$=总部〈的〉部长宋文$_{NE}$。

④分词及词性标注处理。分词及词性标注需要在命名实体识别和组合之后进行,当一个文本片段被标识为一个命名实体后,其内部不再进行分词及词性标注处理。经预处理后,用例形如:

2006 年 8 月 31 日/$_{TE}$詹夏来/$_{NE}$辞去/$_{V}$芜湖市委书记/$_{PE}$一职/$_{NP}$,/$_{W}$调任/$_{V}$安徽省政协副主席/$_{PE}$,/$_{W}$兼任/$_{V}$奇瑞公司董事长/$_{PE}$。/$_{W}$(怀宁论坛,2006-8-31)其中,下标 W 表示标点符号。

3)待抽取文本句法分析:对预处理后的文本进行分析,确定起中词和短语的句法成分(由 AM_{AS}完成)。经此步骤后,用例形如:

2006 年 8 月 31 日$_{Adv}$詹夏来$_{Sub}$辞去$_{Pre}$芜湖市委书记$_{Obj}$一职$_{App}$,调任$_{Pre}$安徽省政协副主席$_{Obj}$,兼任$_{Pre}$奇瑞公司董事长$_{Obj}$。下标 Adv、Sub、Pre、Obj、App 分别表示状语、主语、谓语、宾语、同位语。

4)事件特征语义分析:此步骤在概念知识库的支持下由 EFM_{AS}完成,它用相关概念对事件特征(论元)的语义特征进行描述,得出 $SF_V(C_{V1},C_{V2},\cdots)$。职务变更事件中,事件特征的语义分析结果如下:

①变更对象:{+human}(人名、人名代指概念)。变更对象具有人物属性,描述它的概念包括人名("文强"、"埃文·特纳"等)和人名代指概念("驻港特派员"、"他"等)。

②变更职务:{+position}(行政职务,业务职务,职务代指概念)。变更职务具有职务属性,描述它的概念包括行政职务("代总统"、"现任董事长"等)、业务职务("教授"、"工程师"等)和职务代指概念("前任"、"原职"等)。

③变更组织:{+organization}(单位,机构、组织代指概念)。变更组织具有组织属性,描述它的概念包括单位("卫生部"、"中兴公司"等)、机构("市人大常委会"、"省作家协会"等)和组织代指概念("贵公司"、"本协会"等)。

④变更时间:{+time}(日期,时间代指概念)。变更时间具有时间属性,描述它的概念包括日期("2002 年 7 月"、"6 月 1 日"等)和时间代指概念("去年"、"此后"等)。

5)命名实体映射:将命名实体映射为相应的事件特征(论元),此步骤由 EFM_{AS}完成。职务变更事件中命名实体的映射规则包括 $NE\rightarrow H$、$PE\rightarrow P$、$OE\rightarrow O$、$TE\rightarrow T$("→"表示映射,下文同此)。经命名实体映射后,用例形如:

[2006 年 8 月 31 日]$_T$[詹夏来]$_H$ 辞去/$_V$[芜湖市委书记]$_P$一职/$_{NP}$,/$_W$调任/$_V$[安徽省政协副主席]$_P$,/$_W$兼任/$_V$[奇瑞公司董事长]$_P$。/$_W$

6)代指概念映射:确定代指概念所指内容(由 EFM_{AS}完成)。除有形代指概念外,事件特征(论元)还可用无形代指概念,即空语类 e(empty category)代指[10]。职务变更事件中 e 可以代指$[A]_O$、$[A]_P$ 和$[A]_H$。例如:

①[市人大]$_O$ 通过罢免案,[e]罢免陈启市人大代表职务。(晋江日报,2010-7-19,e 代指变更组织"市人大")

②[[岳麓区]$_O$[区长]$_P$]$_P$ 选举产生,陈中当选[e]。(星辰在线,2007-11-30,e 代指变更职务"岳麓区区长")

③[金滉植]$_H$ 当选韩国总理,[e]提名金星焕为外长。(新快网,2010-12-2,e 代指变更对象"金滉植")

如果存在先行句,则可利用先行句中的已现论元来求解后面代指概念的具体内容。经代指概念映射后,用例形如:

[2006 年 8 月 31 日]$_T$[詹夏来]$_H$ 辞去[芜湖市委书记]$_P$一职,[e]调任[安徽省政协副主席]$_P$,[e]兼任[奇瑞公司董事长]$_P$。

7)动词识别:构造词表并运用词形匹配确定文本中的动词(由 AM_{AS}完成)。由于动词数量有限,因此所得词表规模通常不大,这使词形匹配变得易于实现。职务变更动词有 3 大类 6 小类共 36 个:

①任命动词(共 5 个):任命、提名、选、选聘、选举;

②担任动词(共 14 个):任、担任、就任、出任、上任、现任、连任、历任、兼任、接任、继任、就职、当、当选〈为〉("〈〉"中的内

容表示实例中可能出现的从属成分)；

③免职动词(共 8 个)：免去、撤职、撤销、撤消、撤除、解除、罢免、免职；

④辞职动词(共 5 个)：辞去、辞职、离任、任满、下台；

⑤调遣动词(共 3 个)：调〈动〉…任…、升〈为〉、提升；

⑥受命动词(共 1 个)：调任。

其中，①、②合称任职动词，③、④合称离职动词，⑤、⑥合称调职动词；①、③、⑤具有主动性，②、④、⑥具有被动性。经动词识别，可知用例中包含“辞去”、“调任”、“兼任”3 个职务变更动词。

8)论元结构属性分析：此步骤在概念知识库的支持下由 TM_{AS} 完成，它分析步骤 7)中找到的 V，得出其 ASP。用例中“辞去”、“调任”、“兼任”的 ASP 如表 1 所列。

表 1 “辞去”、“调任”、“兼任”的 ASP

ASP \ V		辞去	调任	兼任
AP	AA	2(必)	2(必)	2(必)
	AR	Ex,Re	Ex,Re	Ex,Re
GF	AC	①Ex+__+Re〈的〉职务 ②Ex+__+Re ③Ex+__+职务	①Ex+由(/从)Re1+__+为Re2 ②Ex+__+Re	①Ex+__+〈为〉Re ②由 Ex+__+Re ③Re+由 Ex+__
	SC	①Ex:Sub,Re:Obj ②Ex:Sub,Re:Obj ③Ex:Sub	①Ex:Sub,Re1:Obj 由(/从),Re2:Obj(为) ②Ex:Sub,Re:Obj	①Ex:Sub,Re:Obj ②Ex:Obj(由),Re:Obj ③Re:Sub,Ex:Obj(由)
	CF	①Ex,Re:N(/NP) ②Ex,Re:N(/NP) ③Ex:N(/NP)	①Ex,Re1,Re2:N(/NP) ②Ex,Re:N(/NP)	①Ex,Re:N(/NP) ②Ex,Re:N(/NP) ③Ex,Re:N(/NP)
SF	SSF	Ex:{+I,+C},Re:{+I,+T}	Ex:{+I,+C},Re:{+I,+T}	Ex:{+I,+C},Re:{+I,+T}
	DSF	Ex:NE,Re:PE	Ex:NE,Re:PE	Ex:NE,Re:PE

表 1 中，Ex 表示经事(Experiencer)，Re 表示系事(Relevant)，它们是 V 的论元角色；句式中的“__”代替同列第 1 行中出现的 V；Sub 表示主语(Subject)，Obj 表示宾语(Object)；N 表示名词(Noun)，NP 表示名词短语(Noun Phrase)；I 指自立性(Independent)，即所指对象先于动词所表示的事件独立存在；C 指变化性(Change)，即所指对象的状态在动词所表示事件中发生了变化；T 指类属性(Type)，即所指表示相应经事的类型、属性等。例如，表 1 中的第 3 列说明，“调任”带有经事、系事两个必要论元；“调任”和由名词(或名词短语)充当的论元角色之间可以形成两种句式。句式①中，经事充当主语，系事 1 充当“由(/从)”的宾语，系事 2 充当“为”的宾语；句式②中，经事、系事分别充当“调任”的主语和宾语，经事由人名实体表示且具有自立性和变化性，系事由职务实体表示且具有自立性和类属性。

9)论元映射：确定事件实例中每个动词支配的论元。此步骤由 TM_{AS} 和 AM_{AS} 协作完成，它针对每个被识别出的 V(步骤 8)，调用其 ASP(步骤 9)，找出与其对应的 A。下文以“辞去”为例说明论元映射的具体步骤。

①调用 ASP 中的 AP，确定 V 的论元数目和论旨角色。表 1 显示，“辞去”支配 Ex、Re 两个必要论元，因此求解这两个论元成为论元映射的主要任务。

②调用 ASP 中的 GF，确定 V 论元的合法句式及合法词性序列。依据表 1，将“辞去”的 SC 代入 AC，可知“Sub+__+Obj”为其合法句式，再将 CF 代入 AC，可知“N(/NP)+__+N(/NP)”为其合法词性序列。

③调用 ASP 中的 SF，确定 V 论元的合法命名实体序列。依据表 1，将“辞去”的 DSF 代入 AC，可知“NE+__+PE”为其合法命名实体序列。

④步骤 3)的结果显示，用例中的“辞去”与周围词语形成的句式为“Sub+__+Obj+App”，排除 App(同位语)等修饰成分的干扰，可知其具有合法句式。由此初步可以认定充当 Sub 的“詹夏来”和充当 Obj 的“芜湖市委书记”分别是“辞去”的 Ex 和 Re。

⑤步骤 2)的结果显示，用例中的“辞去”与周围词语形成的命名实体序列为“NE+__+PE”，这说明其同样具有合法的命名实体序列，因此又可以进一步判定，NE“詹夏来”和 PE“芜湖市委书记”分别是“辞去”的 Ex 和 Re。

⑥步骤 4)、5)的结果显示，NE→H、PE→P 且 H、P 均可用 N(/NP)表示，因此用例中的“辞去”与周围词语形成的词性序列同样合法，至此可以肯定，N“詹夏来”和 NP“芜湖市委书记”分别就是“辞去”的 Ex 和 Re。

“调任”和“兼任”必要论元的求解步骤与“辞去”基本雷同，只不过它们还需完成空语类映射(步骤 6))。

⑦可用论元映射，即根据约束规则确定 V 的可用论元，这需另文详解，本文不做讨论，仅用例进行分析。用例中的变更时间没有和哪个 V 形成特定约束，因此“辞去”、“调任”、“兼任”都将支配此论元。用例中 V 的 Re 为 NP，这样基本可以判定 Re 是包含 OE 的 PE，因此可以逆用命名实体组合规则(步骤 2))并从中分解出 OE。

最终，“辞去”、“调任”、“兼任”的映射结果为：{辞去}→{詹夏来，芜湖市委书记，芜湖市委，2006 年 8 月 31 日}；{调任}→{詹夏来，安徽省政协副主席，安徽省政协，2006 年 8 月 31 日}；{兼任}→{詹夏来，奇瑞公司董事长，奇瑞公司，2006 年 8 月 31 日}。确定动词和其论元之间具体的映射关系事实上就完成了事件特征提取，也就完成了事件抽取。

5 实验及结果分析

为测定本文方法的效率，课题组实现了一个基于动词论元结构的职务变更事件抽取模块。当输入某类职务变更动词及其 ASP 后，此模块就可以以半自动化模式进行工作。由于仅是一个半自动化模块，因此在事件选取和结果评判时都需要进行人工干预。

本课题组从华中师范大学语言研究所开发的《人民日报》语料库(语料规模：15,429,593 ±364,1381 词)中选取 279 个职务变更事件(任职 173 个，离职 64 个，调职 42 个)进行实验，实验涉及的事件元素包括 H、P、O、T，实验结果如表 2 所列。

表 2 职务变更事件抽取结果

	任职		离职		调职	
	任命	担任	免职	辞职	调遣	受命
precision	77.8%	74.2%	79.6%	76.7%	82.9%	89.3%
Recall	74.1%	71.3%	70.1%	69.4%	81.2%	86.7%
F-Measure	75.9%	72.7%	74.5%	72.9%	82.0%	88.0%

(下转第 176 页)

did incremental principal component analysis[R]. MSU-CSE-01-23. East Lansing: Michigan State University, 2001

[8] Chung F. Spectral graph theory(Vol. 92 of the CBMS Regional Conference Series in Mathematics)[C]// Conference Board of the Mathematical Sciences. Washington, 1997

[9] Fiedler M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal, 1973, 23(98): 298-305

[10] 胡茂林. 矩阵计算与应用[M]. 北京: 科学出版社, 2008: 219-220

[11] Aitken A C. The evaluation of the latent roota and latent vectors of a matrix[C]// Proc. roy. Soc. Eedinb. 1937, 57: 269-304

[12] Wilkinson J H. 代数特征值问题[M]. 石钟慈, 邓健新, 译. 北京: 科学出版社, 2003: 588-597

[13] 蔡晓妍, 戴冠中, 杨黎斌. 谱聚类算法综述[J]. 计算机科学, 2008, 35(7)

(上接第 164 页)

表 2 中, precision(准确率)=正确抽取事件数目/抽取事件数目, recall(召回率)=正确抽取事件数目/实际事件数目, *F*-Measure(*F* 度量)=(2×recall×precision)/(recall+precision)。对表 2 及实验过程进行分析可知:

1)受命事件的抽取效率远高于其他事件,这主要是因为触发此类事件的动词只有"调任"一个,大大降低了后期处理难度。因此,在保证完整性的前提下尽可能缩减触发动词数目,是提高此类方法效率的重要环节。

2)除受命事件外,调遣事件的抽取效率要高于其他事件。这是因为和其他动词相比,调遣动词的 ASP 更加清晰规范。因此,进行全面深入的句法语义分析,寻找更加可靠的匹配规则,仍是提高此类方法效率的重中之重。

3)如果放弃人工干预,进行全开放实验,结果(平均准确率约为 55.2%、平均召回率约为 53.9%、平均 *F* 度量值为 54.5%)和现有结果比起来存在较大差距,这说明今后很长一段时间内,半自动化抽取方法仍将占据主流。

4)此外,实验过程中出现的词法分析错误、短语识别错误和句型匹配错误对实验结果也造成了不同程度的影响。

为评判本文方法的优劣,课题组基于相同语料和任务,选取任职事件测定了另外两种代表性事件抽取方法的效率,这两种方法分别是事件框架的方法和 IEPAM 方法,其方法过程可参看文献[11,12]。之所以选择上述两者作为比较对象,一方面因为它们的效率得到了已有成果的肯定;另一方面因为它们分属模式匹配方法和机器学习方法,而本文方法属于模式匹配方法,因此,用本文方法与前者比较可以判定其在同类方法中的优劣,与后者比较则可评判其在异类方法中的优劣。实验结果如表 3 所列。

表 3 动词论元结构方法与其他方法的对比分析

	动词论元结构方法		事件框架方法		IEPAM 方法	
	任命	担任	任命	担任	任命	担任
precision	77.8%	74.2%	73.2%	70.6%	59.1%	57.3%
Recall	74.1%	71.3%	70.3%	67.7%	56.8%	54.9%
F-Measure	75.9%	72.7%	71.9%	68.4%	57.4%	55.5%

从表 3 可以看出:1)以动词论元结构为核心来完成事件分类和定位要优于抽象的事件框架机制,由此形成的模式匹配方法更具应用潜质。2)与模式匹配方法相比,机器学习方法在效率上仍存在较大差距,虽然不具有灵活的可移植性,但模式匹配方法在一定时期内仍将是主流。

综合表 2 和表 3 的实验结果可知,基于动词论元结构的事件抽取方法已具备基本应用价值,值得深入研究。

结束语 本文所提方法属于动词驱动的信息抽取方法,触发动词在其中一直起着驱动和引导作用,和事件驱动的信息抽取方法不同,此类方法主要依靠动词聚合来实现跨事件融合[13]。

利用动词的 ASP 来确定其句法成分(如主语、宾语等)和语义角色(如施事、受事等)之间的映射规则,是本文所提方法中最关键的技术。此技术虽然只能视为浅层的文本理解技术,但从应用层面看,由于事件抽取只关注有限的基本信息,不深究文本意义和写作意图等深层理解问题,因此,此技术完全能够满足中文事件抽取的应用要求。

为适应计算机系统,本课题组在实现上述技术时尽可能穷举了表征特定事件的动词及其 ASP。在此过程中,本课题组借鉴了大量汉语言本体研究成果,并从中找出了一些具备可计算性的句法语义规则。

当然,仅靠本文提及的动词论元结构知识来实现全面的事件信息抽取显然不切实际,进一步考虑利用基于论元结构的篇章和逻辑知识来确定事件信息类型和特征分布将是本课题组下一步的努力方向。

参 考 文 献

[1] 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

[2] 谭红叶. 中文事件抽取关键技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2008

[3] 邓擘, 郑彦宁, 樊孝忠. 汉语信息抽取中事件的定位与分类[J]. 情报理论与实践, 2009, 28(10): 104-107

[4] 顾阳. 论元结构介绍[J]. 国外语言学, 1994(1): 1-11

[5] 袁毓林. 用动词的论元结构跟事件模板相匹配——种由动词驱动的信息抽取方法[J]. 中文信息学报, 2002, 19(5): 37-43

[6] 袁毓林. 一套动词的论元角色的语法指标[J]. 世界汉语教学, 2003(3): 24-36

[7] Yangarber R, Grishman R, Tapanainen P, et al. Automatic Acquisition of Domain Knowledge for Information Extraction[C]// Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000). Saarbriicken, Germany, 2000: 412-416

[8] 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

[9] Kim J, Moldovan D. Acquisition of Linguistic Patterns for Knowledge-based Information Extraction[J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(5): 713-724

[10] 许荣华, 吴刚, 李培峰, 等. 基于指代消解的中文事件融合方法[J]. 计算机应用, 2009, 29(8): 2264-2267

[11] 梁晗, 陈群秀, 吴平博. 基于事件框架的信息抽取系统[J]. 中文信息学报, 2006, 20(2): 40-46

[12] 姜吉发. 一种事件信息抽取模式获取方法[J]. 计算机工程, 2005, 31(15): 27-29

[13] 许荣华, 吴刚, 李培峰, 等. 基于事件框架的主题事件融合研究[J]. 计算机应用研究, 2009, 26(12): 4542-4545