文章编号：2095-4980(2017)02-0279-06

基于语义角色分析的事件抽取技术

章顺瑞，骆　陈

(无锡工程兵科研 1 所 第 5 研究室，江苏 无锡 214035)

摘　要：利用语义角色分析的方法对动态新闻进行事件抽取研究。通过对句子进行论元结构标注，抽取出句子中以谓语动词为中心的论元结构，将其转化为具体的语义角色，并与事件要素进行匹配，完成事件抽取工作。论文提出并重点研究了如何利用 VerbNet 和 SemLink 资源对动词的论元结构与事件要素进行匹配。抽取系统基于该方法对 1 000 篇新闻语料进行了事件抽取实验，结果表明该方法的 *F* 值达到 70.6%，具有一定的应用价值。

关键词：事件抽取；论元结构；语义角色；事件要素

中图分类号：TN957.52；TP391　　**文献标志码：**A　　**doi：**10.11805/TKYDA201702.0279

Event extraction technology by semantic role analysis

ZHANG Shunrui，LUO Chen

(The Fifth Research Laboratory，The First Engineering Scientific Research Institute of Wuxi，Wuxi Jiangsu 214035，China)

Abstract：The event extraction from news on the internet is performed by a method of semantic role analysis. The sentence in the news is annotated with arguments labeler; the argument structures centered with verbs are extracted and converted to a specific semantic role of the verb; and then the semantic roles are matched to the event elements. How to use VerbNet and SemLink resources to match the verb's arguments and event elements is put forward and studied in detail. The experiment is carried out on 1 000 news corpus crawled from the web, and the results show that the *F* value is up to 70.6% and the proposed method has certain application value.

Keywords：event extraction；arguments structure；semantic role；event elements

网络大数据时代，浩瀚的互联网信息中隐含着大量武器装备的相关性能、技术战术指标，以及包括研制、试验、订购与生产在内的诸多事件信息。这些信息对于我国武器装备的研制与发展都有重要作用。然而网络中信息资源具有数量巨大，领域宽泛，资源庞杂等特点，使得以往通过人工的方式查找与提取武器装备的相关重要事件变得不再可行，因而亟需一种方法将网络大数据中关于诸多武器装备的相关事件信息自动提取出来，为我国的科研人员提供重要技术信息支撑。事件抽取技术是解决这一问题的有效方法。事件抽取是信息抽取领域一个重要的研究方向，事件抽取把含有事件信息的非结构化文本以结构化的形式呈现出来，在自动文摘、自动问答、情报分析等领域有着广泛的应用。事件抽取的研究主要有 2 种方法：模式匹配的方法和机器学习的方法。机器学习的方法把事件抽取任务看作分类问题，把主要的精力放在分类器的构建和特征的构建上[1]。这类方法不需要太多的人工干预和领域知识，移植性也较好，但是由于很难找到通用的特征，并且该类方法对语料的依赖性很高，总体上的识别准确率不高。模式匹配的方法是指对于某类事件的识别和抽取是在一些模式的指导下进行的，采用各种模式匹配算法将待抽取的句子和已经抽取的模板匹配[1]。这种方法准确性相对较高，但往往依赖于特定领域，可移植性较差。

本文从动词语义角色的角度出发，将事件抽取看作是对参与该谓语动词活动的中心实体所表示的语义角色进行识别的过程。通过对句子进行论元结构标注，识别出谓语动词的论元结构，利用动词语义库将其转化为具体的语义角色，最后将语义角色与事件模板中的元素进行匹配，完成事件抽取任务。

语义角色分析是当前自然语言处理领域的研究热点。对于给定句子中的每个谓词，语义角色分析目的是识别出其在句中的相应语义成分，并作相应的语义标记，如施事、受事、工具或附加语等。一般来说，这些语义成分表达了参与谓词活动的中心实体，谓词的施事指做某件事情的人或物，谓词的受事指接受某件事物的人或

收稿日期：2015-11-06；修回日期：2016-01-08

物，其他的附加语(如时间、地点、方式)则是谓词的修饰成分。利用动词的语义角色或者论元结构的方法进行事件抽取近年来逐渐受到学者的关注。

事件元素通常可以直接与事件触发词的语义角色相对应。袁毓林[2]、肖升[3]等以职务变动领域为研究对象，根据职务变更动词的有关句法和语义特点，手工建立了从动词论元角色与事件模板的匹配关系，从而完成事件抽取工作。他们的工作揭示了动词对文本筛选和合并都有导向作用，从而也说明了由动词驱动的事件抽取方法的可行性。Surdeanu 等[4]则通过类似的方法实现了英文语料的事件抽取工作。与上述方法不同的是，于江德等[5]另辟蹊径，提出了一种基于条件随机场的语义角色标注方法，并使用该方法对“职务变动”和“会见”2 类事件的要素及其语义角色进行标注。于江德的方法优点在于无需构建动词的论元角色与事件要素之间的匹配关系，但是其却需要构建一个特定领域的语义角色标注器。

上述方法无论是构造动词论元角色与事件模板的匹配关系，还是构建特定领域的语义角色标注器，均需要具有专业技能的领域专家耗费大量的劳动和时间，人力负担较重，并且移植较为困难。因此，吴刚等[6]将目光转向了基于机器学习的方法，其利用人工标注的动词论元角色与事件模板的匹配关系，训练出了一个匹配模型，并将其用于指导其他动词论元角色与事件模板的匹配任务。与人工制定规则的方法相比，基于机器学习的方法只需要一个标注好的语料库进行训练，便可自动完成大量人工工作。但是该方法对大规模训练语料较为依赖，吴刚等人在实验中仅人工标注了一个小规模的训练语料，使得其效果并不理想，并且该方法由于缺乏人工参与的因素，使得其尚未达到实用的程度。

1 研究框架与方法

本文中研究的事件是以动词为驱动的，待抽取的事件元素是参与该谓语动词活动的中心实体。这样就可以将事件抽取任务转化为以谓语动词为中心的论元结构识别以及其到事件元素的匹配任务。与前人研究工作不同的是，本文在将动词的论元结构与事件要素的匹配过程中，研究利用 VerbNet 等语义资源库实现了动词的论元结构到具体的语义角色的转化，最终利用语义角色与事件要素的直接匹配完成了事件抽取工作。

本文进行事件抽取的过程如图 1 所示，主要包括如下几个步骤：

1) 文本预处理。对网页内容进行网页去噪、正文提取、分段、分句、分词等工作。

2) 论元结构标注。利用论元结构标注器对预处理后的句子进行处理，标注出谓语动词及其对应论元结构。

3) 事件元素抽取模块。首先，针对动词论元结构标注的结果，抽取出各谓语论元；然后利用 VerbNet 和 SimLink 资源将其映射到具体的语义角色；最后根据匹配算法，将动词的语义角色匹配到事件模板中的事件要素，生成候选事件。

4) 事件过滤。建立事件要素约束集合，并利用其对候选事件进行过滤，提高事件抽取的准确性。

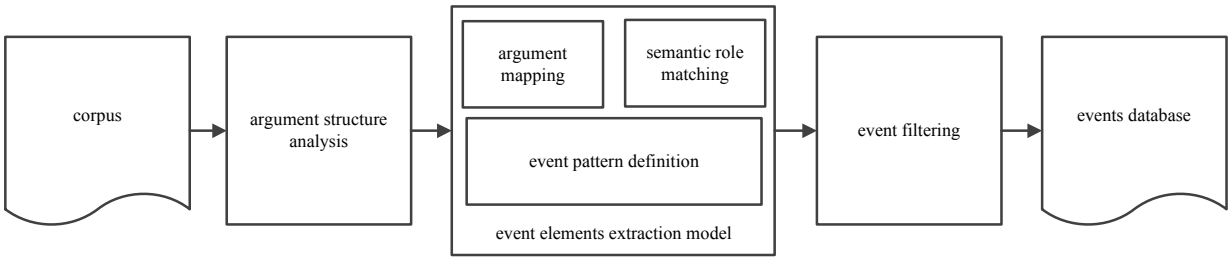

Fig.1 Process of event extraction
图 1 事件抽取流程图

1.1 论元结构分析

论元结构标注器是本文的一个重要的文本处理工具，近年来，论元结构标注的研究取得了显著的进步，一些学者和科研院所也相继开发了一些论元结构标注工具。本文使用了一个高性能的论元结构标注器[7]对文本进行处理，其在 CONLL-2009 的语义角色标注测评任务中取得了第 1 名。尽管对自然文本进行处理时，论元结构标注工具对论元的识别并不是 100%准确，尤其在句子中包含专有名词和修饰语的时候准确率会进一步降低，但是本文依然认为论元结构分析是事件抽取的一种有效方法。

图 2 是使用论元结构标注器对句子“In 2013, NASA and Boeing finished tests of 757 vertical tail with advanced technology.”进行标注后的可视化展现结果。可以看出通过标注，可以快速识别出句子中的谓语动词及其对应的论元结构。图 2 的测试句子中，对于谓语“finish”，可以准确地识别出该动作的触发方为“NASA and Boeing”，

接受方为“tests of 757 vertical tail”，动词的时间修饰语和方式修饰语为“in 2013”和“with advanced technology”，并且这些论元要素均使用相应的标签进行了标注。

	In	2013	,	NASA	and	Boeing	finished	tests	of	757	vertical	tail	with	advanced	technology
finish.02	AM-TMP			A0				A1					AM-MNR		
test.01									A1						

Fig.2 Annotation results

图 2 使用标注器对例句进行标注的结果

1.2 事件元素抽取

事件元素抽取模块主要用于将语句标注后生成的论元结构映射到预先定义的事件模板之中。针对这一问题，前人的研究工作主要基于 2 种方法：一是通过领域专家人工构建映射规则，二是通过机器学习的方法实现映射模式的自动归纳。前文对这 2 种方法均进行了分析，它们都存在着自身的不足。本文从语义分析的角度出发，借助动词语义资源 VerbNet[8]实现了论元要素向事件要素的转换。其过程如图 3 所示，首先将谓语动词的论元要素转换为更易于处理和辨别的 VerbNet 中的语义角色，然后再将该语义角色与预先定义的事件模板进行匹配。

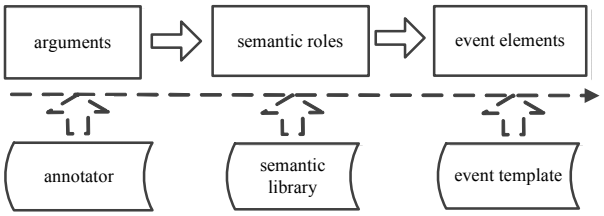

Fig.3 Event elements extraction model

图 3 事件抽取模块示意图

1) 事件模板相关定义

事件抽取就是相关动词的论元结构与事件元素的映射过程，本文做如下定义。

定义 1：事件模型定义为 *EM*={*verb*,*subject*,*object*,*place*,*time*}。其中，*verb* 指事件中产生的动作；*subject* 指事件中动作产生的主体，譬如人物、机构或组织；*object* 指事件中动作所施加的目标，譬如人物或事物；*place* 指事件发生的地点；*time* 指事件发生的时间。事件模型 *EM* 定义了构成本文中的事件所需要的基本要素。EM 的定义参照了事件模型 LODE[9]的基本思想，能够独立于事件的领域，具有较为通用的优点。

定义 2：事件模式定义为 *EP*={*EMP*,*D*}。其中，*EMP* 为事件的映射模式，定义了动词的语义角色与事件元素之间的映射关系，用于在标注过的句子中匹配事件。*D* 定义了构成一个事件所必须的事件要素集合。

定义 3：事件映射模式 *EMP*={M1(ArgTAG1,EventTAG1),M2(ArgTAG2,EventTAG2),⋯,Mi(ArgTAG*i*, EventTAGi),⋯} 是相关动词的论元要素与事件元素的映射关系集合。Mi(ArgTAGi, EventTAGi)表示论元要素 ArgTAGi 与事件元素 EventARGi 的对应，其中，ArgTAGi 为论元标签，EventARGi 为事件模型中的事件元素标签。

定义 4：事件要素约束集合 *D*={EventTAG1,EventTAG2,⋯,EventTAGi,⋯}。*D* 为一个约束集合，定义了事件的必要元素集合，即规定了必须出现的事件元素，*D* 在事件抽取过程中规定了构成事件的一个必要条件。

2) 论元映射

论元结构标注器使用的是 Propbank 的论元标注集[10]，其与常用的事件模型中的事件要素有着相似的元素结构，但二者并不完全对应。主要区别在于，Propbank 中的论元标注集面向具体的谓语动词语义场景，即在不同的谓语动词所表述的语义场景下，相同的论元符号所表述的内涵并不完全相同。然而在事件模型中，事件要素则通常可以描述为事件的主体、客体、时间和地点等几个主要要素。因此，针对 Propbank 中论元要素较为分散的特点，需要构建一个映射模型，将标注后的论元结构映射到事件模型中的元素之中。

VerbNet 是基于 Levin 动词分类标准[11]，可提供明确的句法和语义信息的动词词汇库。VerbNet(版本 3.2)在所有动词词汇中只使用了 36 个语义角色标签，并且使用了选择约束对每一个谓语角色进行了限制。Propbank 中包含的众多论元要素几乎均可以与 VerbNet 中的语义角色相对应。如果使用 VerbNet 中的语义角色代替 Propbank 中的论元要素，将会较大程度地提高事件抽取工作的可操作性。因此，本文试图将谓语动词的论元要素映射到 VerbNet 中的语义角色，为下一阶段的事件要素抽取提供重要的语义依据。

为了在 Propbank 的动词论元要素与 VerbNet 中的语义角色之间建立关联映射，本文使用 SemLink 资源[12]对 Propbank 中的论元要素进行了转换。SemLink 是一个映射资源库，针对每个谓语动词，SemLink 构建了 Propbank 中的论元要素和 VerbNet 中的语义角色的完整映射规则。使用 SemLink 资源，能够快速准确地实现 Propbank 中谓语动词的论元要素与 VerbNet 中的语义角色相转换。

3) 语义角色与事件要素匹配

前文将谓语动词的论元要素转换成 VerbNet 的语义角色，接下来的工作就是将这些有限的语义角色标签与事件中的元素标签进行匹配，生成候选事件。其中事件要素映射模型 *EMP* 的匹配规则如下：

输入：谓语动词 v 及其论元集合 $A=\{arg_1,arg_2,\cdots,arg_n\}$，对应的论元标签集合为 AL，其中 $AL=\{al_1,al_2,\cdots,al_n\}$
输出：事件 $E=\{s,v,o,p,t\}$
过程：
$\boldsymbol{E}\leftarrow\emptyset$
定义可生成事件要素*subject*,*object*,*place*,*time*的语义角色集合分别为S,O,P,T。其中
$S=\{Actor,Agent,Beneficiary, Experiencer, Recipient\}$，$O=\{Theme\}$，$P=\{Location, Source\}$，$T=\{Time\}$
通过SemLink资源，将动词verb的论元标签集合AL转换成VerbNet的语义角色集合$R=\{r_1,r_2,\ldots,r_n\}$，v←verb
for each r_i in R
{
if $r_i\in$Subject then $s\leftarrow r_i$
else if $r_i\in$Object then $o\leftarrow r_i$
else if $r_i\in$Place then $p\leftarrow r_i$
else if $r_i\in$Time then $t\leftarrow r_i$
}
if s is null, then $s\leftarrow$A0
if o is null, then $o\leftarrow$A1
if p is null, then $p\leftarrow$AM-LOC
if t is null, then $t\leftarrow$AM-TMP
$E=\{s,v,o,p,t\}$
return E

1.3 事件过滤

通过事件抽取模块中论元映射、语义角色与事件要素匹配 2 个阶段的分析和处理，一个句子中的各论元要素被合理地转换成事件要素。部分句子能够生成本文定义的事件 E 的全部要素，部分句子则仅能够生成部分事件要素。通常来说，一个句子产生的事件要素愈全面，那么其是一个事件的可能性则愈大。

为了更为准确地获取事件集合，过滤大量噪音信息，本文定义事件要素约束集合 D 用来对元素映射模块生成的候选事件进行过滤，生成事件集合。D 在事件生成的过程中定义了构成事件的必要要素集合。对于任意的候选事件 e，如果其全部包含集合 D 中的事件要素，则认为 e 为一个有效事件。本文的实验结果表明，事件约束集合 D 的定义能够过滤掉大量噪音信息，有效地提高了最终的事件抽取的准确性。

2 实验设计与结果分析

2.1 实验数据及预处理

本文面向武器装备科技信息领域，试图借助于前文研究方法，从武器装备科技信息领域的最新新闻动态中抽取出重要事件信息。为实现上述目的，本文从国外众多军工企业、重要国防机构以及政府部门的网站中采集了大量的动态新闻网页，均为英文网页。本文从采集下来的网页中随机抽取了部分网页作为实验语料。

对实验语料，本文进行了细致的预处理工作，预处理工作流程如图 4 所示。首先，识别并剔除原始网页中的噪音信息；然后提取网页正文，过滤掉正文中含有的各类网页标签；最后利用自然语言处理工具报对处理后的网页正文进行分段、分句和分词等处理工作。

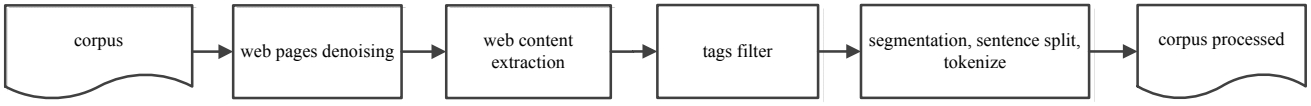

Fig.4 Process of the corpus preprocessing
图 4 语料预处理过程

2.2 评测方法

对于实验中所使用的所有语料，本文并没有一个完整的正确事件结果集合。为了能够良好地评估本文方法的有效性，实验决定对已经抽取出的事件信息进行结果评测。实验中随机选取了部分事件结果，找出每个事件所对应的原文句子，人工标注出准确的事件要素信息，并与本文自动抽取出的结果进行比对。抽取实验利用准确率 P 和召回率 R 两个指标进行评测，P 和 R 的计算公式如下：

$$P=\frac{\text{正确抽取出的事件个数}}{\text{抽取出的事件总数}},\quad R=\frac{\text{正确抽取出的事件个数}}{\text{实际的事件总数}}$$

在计算出 P 和 R 值后，抽取实验系统采用 F 度量作为最终的评价指标。F 度量方法计算如下：

$$F=\frac{2PR}{P+R}$$

除了对整个事件进行效果评测，实验还对事件中每个要素的准确性进行了评测，以分析影响事件抽取准确性的相关原因。

2.3 结果及分析

实验从国外众多军工企业、重要国防机构以及政府部门的网站中采集了 1000 篇动态新闻语料，针对本文的研究方法进行了抽取实验。实验具体的语料信息统计如表 1 所示。语料经过语义角色标注并抽取出论元结构后，通过论元映射和角色匹配，共抽取出 133 308 条候选事件。实验中定义了事件要素约束集合 D={*subject*, *object*, *time*}，经过 D 的过滤后，共获得 2 445 条事件信息。从实验中，发现事件要素约束集合 D 的定义过滤了大量的无意义的事件信息，较大程度地提高了事件抽取质量。

表 1 语料信息统计表

Table1 Statistical information of the corpus

item	value
total amount of corpus	1 000
total amount of candidate events	133 308
total amount of filtered events	2 445

实验从最终获取的 2 445 条事件信息中随机抽取了 100 条事件作为评测对象。为评判本文方法的优劣，基于相同语料和任务，对基于角色匹配的事件抽取(Event Extraction Based on Role Matching，EEBRM)[6]这种基于语义角色匹配的代表性方法进行了实验。该方法在标注好句子的论元结构后，利用标注样本自动训练出了一个映射模型，用于指导动词论元角色与事件模板的匹配任务，该方法详细过程参看文献[6]。

实验的评测结果如表 2 所示。从实验结果可以看出：a) 通过对句子进行论元结构标注，抽取出句子中以谓语动词为中心的论元结构，利用 VerbNet 和 SemLink 语义资源将论元结构转化为具体的语义角色并与事件要素进行匹配的方法要优于利用标注样本训练生成映射模型的方法；b) EEBRM 方法的实验结果中召回率 R 值较低，拉低了 F 指标。

分析认为：EEBRM 方法在文献[6]中发布类型的事件抽取实验中取得了不错的效果，其主要原因是其针对发布类型这一事件类型训练出了一个较好的事件映射模式，但是如果事件类型没有被严格指定或者较为模糊，那么其训练的事件映射模式则不够全面，导致抽取结果的召回率较低，使得最终的 F 指标较差；本文在构造映射模式时，使用 VerbNet 和 SemLink 语义资源构造映射规则，该方法使用了人工构造的较为全面丰富的语义资源，受事件类型的影响较小，具有灵活的可移植性。

表 2 事件抽取评测结果

Table2 Evaluation results of event extraction

	P/%	*R*/%	*F*/%
EEBRM	72.5	61.7	66.7
proposed method	69.0	71.1	70.6

表 3 错误源统计

Table3 Statistical information of the errors

error element	value/%
subject	38.9
object	33.3
time	15.0
verb	12.7

针对选取评测的事件信息，本文对抽取的错误事件要素进行了人工统计。表 3 是各错误的事件要素占全部错误事件要素的百分比。由于事件中的 Place 元素并不是必要元素，因此并没有对其进行评测。从统计结果可以看出，导致事件抽取误差的主要因素是 subject 和 object 2 个要素。通过对抽取结果数据的详细分析，发现导致这些错误的因素主要有：

1) 自然语言处理技术不够成熟从而导致了级联错误，尤其是现有的自然语言处理工具在句法分析、命名实体识别、语义角色标注等任务中还存在一定的误差，影响了本文在论元结构识别工作中的准确性。

2) 文本中的复杂从句导致了句子中指代不明、指代歧义等现象的产生，增加了语义分析的难度。

3 总结与展望

本文在研究事件抽取工作时，从语义分析的角度，将事件抽取模式建设的主要精力放在动词论元结构与事件要素的对应关系上，从而从纷繁复杂的句法规则中解放出来。为了建立对应关系，本文运用了 VerbNet, SemLink 等语义资源库。从初步的实验结果来看，本文方法取得了不错的效果。从实验结果中发现，误差主要是由事件中主体(subject)和客体(object) 2 个要素产生，主要原因是由于自然语言处理技术还不十分成熟，尤其是在命名实体识别、语义角色标注等领域还存在一定的误差，导致产生了级联错误。此外，本文在利用 SemLink 资源将动词的论元要素转换为语义角色的过程中所使用的转换方法，主要是通过主观经验的方法，并没有进行比较分析。

下一步的研究工作将围绕上述问题展开，首先通过词表等方法提高自然语言处理中命名实体识别的准确性，以减少级联误差；其次在论元要素转换为语义角色过程中，重点研究如何制定合理的转换规则。

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

[2] 袁毓林. 用动词的论元结构跟事件模板相匹配——一种由动词驱动的信息抽取方法[J]. 中文信息学报, 2005,19(5):37-43. (YUAN Yulin. Matching even-template with argument structure of verbs: towards a verb-driven approach of information extraction[J]. Journal of Chinese Information Processing, 2005,19(5):37-43.)

[3] 肖升,何炎祥. 基于动词论元结构的中文事件抽取方法[J]. 计算机科学, 2012,39(5):161-164. (XIAO Sheng,HE Yanxiang. Approach of Chinese event IE based on verb argument structure[J]. Computer Science, 2012,39(5):161-164.)

[4] SURDEANU M,HARABAGIU S,WILLIAMS J,et al. Using predicate-argument structures for information extraction[C]// Proceedings of the 41th Annual Meeting of the ACL. Sapporo,Japan:ACL, 2003:8-15.

[5] 于江德,樊孝忠,庞文博. 事件信息抽取中语义角色标注研究[J]. 计算机科学, 2008,35(3):155-157. (YU Jiangde, FAN Xiaozhong,PANG Wenbo. Research on semantic role labeling for event information extraction[J]. Computer Science, 2008,35(3):155-157.)

[6] 吴刚,徐荣华,朱巧明,等. 一种基于角色匹配的事件抽取方法[J]. 微计算机信息, 2010,26(9):187-189. (WU Gang,XU Ronghua,ZHU Qiaoming,et al. An approach of event extraction based on role matching[J]. Control & Automation, 2010, 26(9):187-189.)

[7] BJÖRKELUND A,HAFDELL L,NUGUES P. Multilingual semantic role labeling[C]// Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task. Boulder,Colorado:ACL, 2009:43-48.

[8] SCHULER K. VerbNet: a broad-coverage, comprehensive verb lexicon[D]. Philadelphia:University of Pennsylvania, 2005.

[9] SHAW R B. Events and periods as concepts for organizing historical knowledge[D]. Berkeley:University of California, 2010.

[10] PALMER M,GILDEA D,KINGSBURY P. The proposition bank: an annotated corpus of semantic roles[J]. Computational Linguistics, 2005,31(1):71-105.

[11] LEVIN B. English Verb Classes and Alternations: a Preliminary Investigation[M]. Chicago:University of Chicago Press, 1993.

[12] BONIAL C,FEELY W,HWANG J D,et al. Empirically validating VerbNet using SemLink[C]// Joint ISA-7 Workshop on Interoperable Semantic Annotation SRSL-3 Workshop on Semantic Representation for Spoken Language I2MRT Workshop on Multimodal Resources and Tools. Istanbul,Turkey:European Language Resources Association, 2012:45-51.

作者简介：

章顺瑞(1988-)，男，江苏省淮安市人，硕士，工程师，主要研究方向为智能信息处理、指控信息化.email:ruiruisoldier@163.com.

骆　陈(1988-)，女，江苏省淮安市人，硕士，工程师，主要研究方向为智能信息处理.