基于故事浅层理解与事件框架的语义建模

谢秋妹 高春鸣 王小兰

(湖南大学信息科学与工程学院 长沙 410082) (湖南大学数字媒体研究所 长沙 410082)

摘 要 针对故事文本的语义理解需要,采用开放式信息抽取方式对故事文本进行多元事实抽取,并将多元事实框架表示成事件语义模型。本方法提出了基于依存关系分析和正则表达式相结合的多元事实抽取方法,得到故事浅层语义的多元事实框架,然后将多元事实框架通过规则映射到具有定量时空描述的事件本体模型即Story-Oriented Semantic Description Language(SOSDL)本体。实验表明,多元事实抽取方法能抽取出较多的事实,具有较高的准确率,且SOSDL本体能有效地表示多元事实框架的事件、语义要素以及它们之间的关系。

关键词 开放式信息抽取,自然语言处理,故事文本,事件本体

中图法分类号 TP391 **文献标识码** A

Semantic Modeling for Story Based Shallow Text Understanding and Event Frame

XIE Qiu-mei GAO Chun-ming WANG Xiao-lan

(School of Information Science and Engineering, Hunan University, Changsha 410082, China)

(Institute of Digital Media, Hunan University, Changsha 410082, China)

Abstract For the semantic understanding task of story text, this paper used open information extraction method to capture N-ary facts from story, and then described N-ary facts frames as event semantic model. Our method proposes extraction rules for frame elements based on dependency parser and regular expressions, and an event semantic model SOSDL ontology for story text with representation of qualitative temporal and spatial relations. Our experiments indicate that this approach captures more facts per sentence, is greater completeness, and SOSDL can effectively model the semantic elements of N-ary facts frames and their relationship.

Keywords Open information extraction, Natural language processing, Story text, Event ontology

1 引言

文本内容理解是人工智能与计算机语言领域至今尚未完全解决的基础问题之一。浅层语义旨在研究主体、客体、时间、地点等关于谓词的细粒度知识,是文本语义理解任务的基础工作,能够对问答系统、类型推导和机器翻译等应用产生推动作用[1]。本文关注儿童故事文本的浅层语义理解。儿童故事既是幼儿教育的一种教育资源又是一种教育方法,将儿童故事中的情节和动作信息用于动画自动生成或者幼儿舞蹈创编,对幼儿可视化教育具有重要作用。

从20世纪80年代开始,在Message Understanding Conference(MUC)、Automatic Content Extraction(ACE)以及Text Analysis Conference(TAC)等评测会议的推动下,信息抽取技术研究得到蓬勃发展。目前有两种基于浅层语义的自然文本理解方法:机器阅读(Machine Reading)[2]和阅读学习(Learning by Reading)[3],它们都是获取文本浅层语义的无监督信息抽取方法。在机器阅读中,文本由关系(动词或动词短语)和两个参数构成的固定结构表示,是针对预定义的目标关系集合的信息进行抽取方法。机器阅读发展了被称为"开放式信息抽取"的信息抽取模式,该方法是从自由文本中获取所有关于动词的三元组或者断言集合[4]。但是在阅读学习中,文本的表示是一种更具灵活性的多元参数结构,其从句法依存关系中抽取得到关系,且关系不局限于动词或动词短语。国外的信息抽取已经从传统的限定类别、限定领域信息抽取任务发展到开放类别、开放领域信息抽取[5]。开放式信息抽取系统ReVerb[6]和WANDERLUST[7]分别利用句子的浅层语义和深层句法特征从标注语料中训练得到领域无关的抽取器,然后从文本中抽取出基于动词的二元关系(Arg1, Rel, Arg2)。KRAKEN[8]从句子的依存关系结果中得到关系短语及关系关联的参数的抽取规则,实现了针对英文网页文本内容的高阶多元事实的信息抽取,同时保证了较高的抽取信息的完整性。

国内的中文信息抽取研究方向目前主要集中使用基于机器学习的方法或基于规则匹配的方法针对领域事件的信息进行抽取。吴平博等人建立灾难领域的16个事件抽取框架并基于句法模板构造事件框架的抽取规则[9];北京大学王伟提出了从新闻故事的主题句抽取新闻事件语义元素5W1H[10],但这些方法针对的是感兴趣的信息抽取而不是全文信息抽

到稿日期:2013-01-03 返修日期:2013-05-23 本文受广东省教育部产学研结合项目(2011B090400002)资助。

谢秋妹(1986—),女,硕士生,主要研究方向为自然语言处理、语义Web应用;高春鸣(1957—),男,博士,教授,主要研究方向为数字媒体与服务计算。

取。刘耀华采用基于句法分析的事件抽取方法对新闻故事的动词二元关系进行抽取，并取得了66.11%的F值，摆脱了对语料库和事件模板的依赖，增强了通用性[11]，但该方法缺少对时空信息等多元组事实的抽取。

基于框架的事件表示结构难以实现机器的语义查询和推理，而本体事件模型具有很强的事件语义表达和推理能力。目前有多种事件模型用于不同领域的事件知识表示，Jain R在多媒体内容管理领域中提出了通用多媒体事件模型E[12]；Event-Model-F用于表达现实世界中的事件及事件间的整体-部分、因果和相关关系[13]；NOEM围绕新闻事件5W1H要素进行新闻事件本体建模[14]。但是故事驱动的动画自动生成立足于故事情节内容，其生成的动画必须与故事表达的情景高度一致，包括文本中涉及的事件、时空信息、有生命或无生命的参与者、参与者的固有属性和数量属性以及参与者间的空间方位关系等细粒度的动画信息，上述本体事件模型因其领域限制性，不适合于故事文本的自动理解任务。

本文提出多元事实抽取方法与基于事件机制的故事语义描述语言(Story-Oriented Semantic Description Language，简称SOSDL)。针对故事理解和动画脚本生成需求，采用开放式信息抽取方式对故事文本进行多元事实抽取，基于Stanford Chinese Parser[15]依赖分析和正则表达式相结合的多元事实抽取普适性方法，将故事浅层语义表示成层次多元事实框架，在主句框架结构内，将复合型句型中修饰从句表示为子框架结构，使得事件驱动的语义嵌套框架结构可以解决复杂句子中各种从句的浅层语义表达。多元事实框架通过规则映射到层次事件本体模型SOSDL，该模型包括事件类型、参与者、时间、地点、结构和媒体等多元特征，因而具有定量时空描述能力；通过继承多元事实嵌套框架结构，SOSDL有利于表达复杂故事语义，并易于转换为动画脚本。

2 故事语义框架元素抽取

本文的故事事实框架抽取系统如图1所示，系统主要由两部分组成：(1)原始语料数据获得和语料预处理；(2)基于依存关系分析和正则表达式规则相结合的多元事实抽取方法：从依存关系分析结果中得到一般句型和把字句的框架元素抽取规则，利用正则表达式实现双宾语、兼语类等部分特殊句型的框架元素抽取规则，并将抽取的框架元素表示为框架结构或嵌套框架结构。

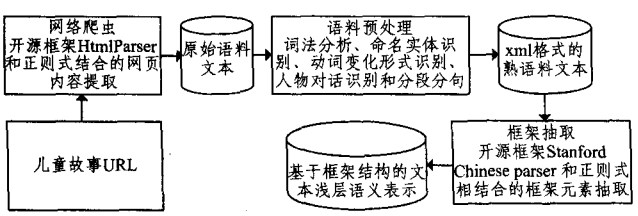

图1 故事事实框架抽取系统结构图

2.1 语料数据获得和预处理

2.1.1 原始语料获取

我们采用开源框架HtmpParser[16]和正则表达式编写网络爬虫。将网页标签title的内容作为文件名，标签p的内容作为文本内容，最后对抓取的数据进行清洗并保存为文本文件作为原始数据。

2.1.2 生语料加工

汉语句子由一串前后连续的汉字组成，词与词之间没有明显的分界标志，必须先对文本进行中文分词建立词的边界。另外，我们发现儿童故事文本具有以下几个特征：(1)文本中人物对话较多；(2)文本中的角色名大多采用拟人、借代等修辞方法，例如：棉花姑娘，小鹿弟弟，或者以身份、特征等指代，如：老奶奶，小儿子；(3)动词具有多种不同的变化形式，动词的变化形式主要有两大类，第一类是双音节动词，有两种形式，即AB→ABAB，如：研究→研究研究；AB→AAB，如：散步→散散步，第二类是单音节动词，有4种形式：即V→VV，如：看→看看；V→V一V，如：看一看；V→V了V，如：看了看；V→V了又V，如：看了又看等；(4)文本中的时间短语比较具有规律，多数时间短语的识别并不依赖于句法分析结果。时间词类型可分为时刻、时间段、事件时间3类时间子类型，表示时刻的有“同时”、“2012年12月6日中午12点”等；表示时间段的有“过了三年”、“三个星期以后”等；表示事件时间的有“快要开花的时候”、“下午放学的时候”等。

本文语料预处理过程包含下面4个步骤：(1)使用ICTCLAS[17]进行汉语分词和词性标注，973专家组测试结果显示ICTCLAS的分词正确率在97.58%，ICTCLAS具有以下的显著特点：能对原始语料进行分词、自动地识别人名、地名、机构名等未登录词、新词标注以及词性标注，并可在分析过程中导入用户自定义的词典用于指导分词；(2)根据从语料实例归纳出的儿童故事角色、动词变化形式和时间短语的主要构词规则，使用正则表达式实现上述短语边界的识别任务；(3)将人物对话内容视为整体出现，并对文本分段分句；(4)将上述步骤处理后的熟语料保存为XML格式的语料文件。

2.2 基于依存关系分析和正则表达式的多元框架元素抽取

汉语是非形态语言，其句子是由“话题.述评”构成的开放式框架，即“话题”铺排成包含若干“小句”的“话题链”，不一定显示地要求句子包含动词和形容词等候选谓词，而是强调句子意思的表达[18]。因此，本文提出基于依赖关系和正则表达式相结合的方法对文本进行多元事实抽取：对于故事文本的句子集合，若当前句子含谓词成分，则从Stanford Chinese Parser的依赖关系结果中抽取事实元素；若不含谓词成分，则使用正则表达式抽取时间、地点、角色等实体类型；最后将抽取出来的事实保存为框架结构，对于复合句型的从句，则使用嵌套的子框架结构表达事实框架。

定义1(事实) 定义事实为文本中出现的客观事实，由属性-值字符串对构成。本文定义事实属性集合为{“谓词”、“主体”、“客体”、“时间”、“地点”、“工具”、“与事”、“主体属性”、“主体数量”、“客体属性”、“客体数量”、“谓词修饰成分”}。事实属性可由单词间的依存关系得到，值是文中的词或带有词性标注的词。

定义2(事实框架) 定义事实框架为由多个事实组成的框架，用于表示某一文本片段中一组实体和它们之间关系的基本语义单元，这里的文本片段以句子为单位。框架概念来源于M. Minsky在研究理解情景、故事时的心理学模型[19]，而后与谓词逻辑融合衍生出语义模型。

多元事实抽取的核心是应用 Stanford Chinese Parser 进行句子依存关系分析。Stanford Chinese Parser 是基于统计学原理的自然语言分析器，它给出了 45 种命名的语法关系（占据了中文命名关系的 91.29%）和一个缺省关系 dep[20]。它接受分词后的文本作为输入源，输出由一系列二元依存关系 rel(head,dep)组成的依存关系结果。本文从依存关系分析结构中总结出 33 条事实抽取规则，主要规则如表 1 所列，涉及到主句的主动语态、被动语态、从句、“把”字句等不同句子类型的处理。

表 1　部分框架元素抽取规则

句子类型	框架元素抽取规则	抽取结果说明
主句	root→root-↓	主动词
	neg∧(neg-↑==root-↓)→(neg-↓+root-↓)	主动词否定形式
	nsubj∧(nsubj-↑==root-↓)→nsubj-↓	主动语态主语
	nsubjpass∧(nsubjpass-↑==root-↓)→nsubjpass-↓	被动语态主语
	dobj∧(dobj-↑==root-↓)→dobj-↓	主动词宾语
	dobj∧(dobj-↑≠root-↓)→dobj-↑	其他动词
	dobj∧(dobj-↑≠root-↓)→dobj-↓	其他动词宾语
	conj∧(conj-↑|conj-↓≠root-↓)→(conj-↑+conj-↓)	并列主语或宾语
	dvpm→dvpm-↑	主动词修饰成分
	nummod→nummod-↓	实体数量属性
	assmod→assmod-↓	实体一般属性
	lobj→lobj-↓	空间界标
	lobj→lobj-↑	空间方位词
	pobj→pobj-↑	介词
	pobj→pobj-↓	主要表示地点、工具和与事等
关系从句	rcomd∧obj(rcomd-↓==obj-↑)→rcomd-↓	从句动词
	rcomd∧obj(rcomd-↓==obj-↑)→rcomd-↑	从句主语
	rcomd∧obj(rcomd-↓==obj-↑)→obj-↓	从句宾语
把字句	(nsubj1 before ba)∧(nsubj2 after ba)∧dobj∧(nsubj2-↑==dobj-↑)→nsubj1-↓	把字句主语
	(nsubj1 before ba)∧(nsubj2 after ba)∧dobj∧(nsubj2-↑==dobj-↑)→nsubj2-↑	把字句动词
	(nsubj1 before ba)∧(nsubj2 after ba)∧dobj∧(nsubj2-↑==dobj-↑)→nsubj2-↓	把字句宾语
	(nsubj1 before ba)∧(nsubj2 after ba)∧dobj∧(nsubj2-↑==dobj-↑)→dobj-↓	把字句与事

注：→表示规则蕴含式，当满足前件时，后件成立。在蕴含式中，↑表示依存关系 relation(head,dependent)的 head 词项，↓表示 relation 的 dependent 词项；∧表示合取；+表示两字符串拼接；==表示两字符串相等，≠表示两字符串不相等。如规则“dobj∧(dobj-↑==root-↓)→dobj-↓”表示存在依存关系 dobj，且 dobj 关系中的动词是主动词，那么 dobj 的 dependent 词就是主动词的宾语。

在实际应用中我们发现，若将带有词性标注的分词文本作为输入，那么将会得到错误的二元依存关系结果。如句子“春天，小鹿在门前的花坛里，栽了一丛玫瑰。”，经 Stanford Chinese Parser 句法分析器分析后得到的依存关系如下：[tmod(栽-10,春天-1),nsubj(栽-10,小鹿-3),prep(栽-10,在-4),assmod(花坛-7,门前-5),assm(门前-5,的-6),lobj(里-8,花坛-7),plmod(在-4,里-8),root(ROOT-0,栽-10),asp(栽-10,了-11),nummod(丛-13,一-12),clf(玫瑰-14,丛-13),dobj(栽-10,玫瑰-14)]。

但是若该句子带有词性标注，即形如“春天/t,小鹿/ANI 在/p 门前/s 的/udel 花坛/n 里/f,栽/v 了/ule 一/m 丛/q 玫瑰/PLA 。”，则其二元依存关系结果为[nsubj(门前/s-5,春天/t-1),nn(在/p-4,小鹿/ANI-3),nsubj(门前/s-5,在/p-4),root(ROOT-0,门前/s-5),nummod(花坛/n-7,的/udel-6),clf(里/f-8,花坛/n-7),dobj(门前/s-5,里/f-8),advmod(了/ule-11,栽/v-10),conj(门前/s-5,了/ule-11),nummod(丛/q-13,一/m-12),clf(玫瑰/PLA-14,丛/q-13),dobj(了/ule-11,玫瑰/PLA-14)]。两者对比后发现显然后者正确，在前一句中 tmod 表示主动词和时间词间的依存关系，“栽”和“春天”构成该关系；后一句的“门前/s”和“春天/t”不具有主谓关系 nsubj。

另外，需要强调的是，单时间词“早晨”、“傍晚”等的语义角色可以明确地通过依存关系“tmod”获得，但是多时间词像“过了三年”，“春天到了”等难以从依存关系中直接获得，且我们关注的是时间点或时间段本身，在“春天到了”中并不需要关注动词“到了”，所以本文将时间的抽取独立于依存关系分析步骤，利用正则表达式抽取时间短语。同时，我们观察到若当前句子为双宾语、兼语类等特殊句型，则难以从 Stanford Chinese Parser 的依存关系结果中总结出这些句型的事实抽取规则，且与已存在规则不发生冲突，因此我们使用正则式对特殊句型进行事实抽取。例如对于兼语类句法模式“NP1+V1+NP2+V2(+NP3)”的例句“他/rr 看见/v 一/m 只/q 小獾/n 认真/a 地/ude2 学做/v 木工/n 。”，我们使用表达式“(.*)/n(.*?)/v(.*?)/n(.*?)/v((.*?)/n)?”匹配句子类型，并根据 java.util.regex 包的 Matcher 对象的方法 public String group(int i)，其中 $i \geqslant 1$，抽取其相应的框架元素，得到〈主体:他,谓词:看见,客体:小獾〉,〈主体:小獾,谓词:学做,客体:木工〉。故本文提出的抽取事实的步骤如下：

输入：带有词性标注和命名实体识别标注的儿童故事文本

输出：嵌套框架结构表示的多元事实集合

1)Dom 解析 xml 文件得到文本句子集 S。

2)$\forall s \in S$，若 s 不含子句，则进行 3)操作，否则进行 9)操作。

3)使用正则表达式抽取时间实体。

4)判断当前句子是否包含动词、形容词等候选谓词成分，若是则进行 5)操作，否则进行 8)操作。

5)判断当前句子是否为特殊句型，若是进行 6)操作，否则进行 7)操作。

6)使用正则表达式，抽取特殊句型的事实元素，跳至算法结束。

7)去掉当前句子的词性标注，并送入 Stanford Chinese Parser 句法分析器进行依存关系分析，从依存关系中抽取得到主体、谓词、客体、时间和地点，以及主体和客体的数量属性和一般属性等。事实抽取过程结束，跳至算法结束。

8)对于不含谓词成分的当前句子使用正则表达式抽取经 ICTCLAS 分词工具识别标注的地点信息和命名实体识别角色实体，跳至算法结束。

9)将 s 按逗号分割为子句集 s'，对 $\forall s' \in S'$，跳至 2)。

3　本体设计及本体扩充

SOSDL 的特色在于具有定量时空描述能力和基于动词的事件类型系统，适合于故事语义指导的动画生成。SOSDL 主要的概念和关系如图 2 所示。

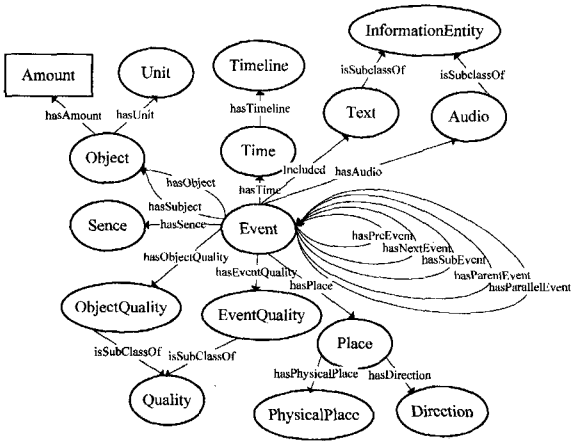

图 2　SOSDL 事件本体模型

3.1　本体模型

定义 3（概念集合）　概念集合 C 表示 SOSDL 本体由抽象实体、事件、信息实体、对象和特征 5 个方面的非空有限概念集合组成。$C=C_{Abstract}\cup C_{Event}\cup C_{InformationEntity}\cup C_{Object}\cup C_{Quality}$，且 $C_i\cap C_j=\Phi(i\neq j\wedge C_i,C_j\in C)$。$C_{Abstract}$ 是抽象实体的概念集合，其子概念有数量、物理属性、时间域和空间域，我们引入艾伦的 13 种时间关系作为定量时间关系[12]，空间域上我们定义空间实体间的 8 种 RCC-8 拓扑关系[22]，即基于锥形域[23]的方位关系，“很近、近、中间、远、很远”5 种定性距离关系以及由数值和距离单位构成的定量距离关系；C_{Event} 是事件类型的概念集合，本文根据词法模型（Lexical Grammar Model）[24]建立事件类型，该模型指出任何语言中的动词都可以分成 10 种类型；$C_{InformationEntity}$ 是多媒体资源的概念集合，如音乐、文本、图片等；C_{Object} 是事件发出者、接受者的概念集合，如人、动物、植物、物品等；$C_{Quality}$ 是对象及事件特征的概念集合，其中对象特征包含对象的颜色、大小、尺寸等固有属性，事件特征包含情感特征、动作频率等事件修饰信息。

定义 4（关系集合）　关系集合 $R=R_h\cup R_a$ 表示 SOSDL 本体中概念与概念之间关系的集合，其中 $R\subseteq C_i\times C_j(C_i,C_j\in C)$。$R_h$ 为概念间的层次关系，包括概念间的继承关系和整体与部分关系；R_a 为概念间的关联关系，表示概念间存在的语义关系。下面着重介绍 3 类关联属性：

1）事件关联属性：事件属性 hasSubject、hasObject、hasPlace、hasTime 分别将事件与主体、客体、时间和地点等事件论元关联起来；属性 hasScene 的值域为时间场景（白天、晚上）和季节场景（春、夏、秋、冬）；属性 hasEventQuality 的值域主要是程度和情感等事件修饰成分；属性 Included 指出事件所属的信息实体，本文主要是故事文本；属性 hasAudio 的值域是对话的音频文件或事件中出现的某些声效等多媒体数据。

2）地点属性：对于含处所词和方位词的空间表达式如“空中”，属性 hasPlaceObject 和 hasDirection 分别将 Place 与 PhysicalPlace 和 Direction 关联起来，方便后续的可视化处理模块根据方位词定性地计算出实体在场景中出现的位置。

3）参与者属性：属性 hasAmount 定量地描述了当前事件参与者的数量关系；hasObjectQuality 定性地刻画了参与者的颜色、大小等属性特征，细化了故事文本可视化表达的粒度。

3.2　本体扩展

SOSDL 本体由本体编辑器 Protégé 建模得到。事件类型识别围绕事件类型触发词，采用《同义词词林（扩展版）》进行触发词扩展，构建事件类型-触发词表。将事件框架映射为本体模型的实例，其过程包含两部分：（1）定义事件元素静态类型信息，根据框架的属性值生成为 SOSDL 对应的本体概念的实例；（2）生成事件间的动态信息，包括事件与事件元素间的关联关系、定性排序关系（前一个，后一个，并发）、事件间的定量时间关系（根据事件的开始时间和持续时间判断）和事件和子事件的组合关系，后两个关系主要是在 Protégé 手工编辑得到。选择小学课文《小鹿的玫瑰花》的第一个句子“春天，小鹿在门前的花坛里，栽了一丛玫瑰。”进行应用分析。使用前面的事实框架元素抽取算法得到事件“栽”及该事件的事件元素。通过预定义的事件类型触发词表判断出“栽”为抽象动词，将“栽”分解为系统预定义的元动词“拿起”和“放下”，属性 hasSubEvent 实现“拿起”和“放下”和“栽”的事件组合关系；每个事件实例即动作都有相应的执行时间段，且这些时间间隔满足艾伦的 meet 和 during 时间关系，从而得到事件间的定量时间关系，而且所有的时间间隔通过属性 ontimeline 关联到全局时间轴上；地理实体“花坛”和“门”具有方向关系“前”，而空间表达式“花坛里”给出了对象“玫瑰花”最终在花坛中的摆放位置；对象“玫瑰花”则有数量属性。例句的语义关系如图 3 所示。

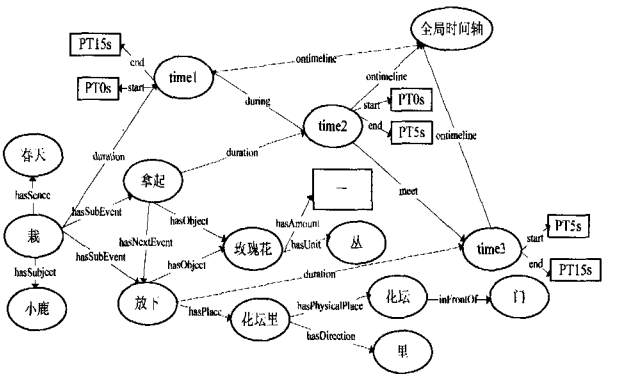

图 3　事件语义关系实例

4　评估和讨论

4.1　实验数据集和事实抽取测评标准

本文从人教版、苏教版和北师版的 1－4 年级小学语文课本中选择 154 篇叙事课文作为实验语料。小学语文教材中 80%左右属于儿童文学，取材包含寓言、儿童小故事、童话故事等多种儿童文学文体，具有代表性。本文使用查全率 R、查准率 P、$F1$ 值和完整率 C 作为评估指标。设原文中人工判断应抽取的事实总数为 N_{total}，系统抽取的事实总数为 N_{found}。对于系统抽取的事实，对照原文句子人工地判断事实的正确性和完整性：（1）正确且完整记为 $N_{t\&c}$，（2）正确不完整记为 $N_{t\&inc}$，（3）错误，其中“正确不完整”定义为正确抽取出句子的主谓或动宾关系，但其他参数缺失；若被抽取句子有一个事实出错则视为错误。那么 $R=(N_{t\&c}+N_{t\&inc})/N_{total}$，$P=(N_{t\&c}+N_{t\&inc})/N_{found}$，$F=2RP/(R+P)$，$C=N_{t\&c}/N_{found}$。

4.2　事实抽取的实验结果与讨论

本文共设置 4 个实验，System 为本文提出的事实抽取方法，Baseline1 为语料文件不经角色实体、动词变化形式、时间

实体识别处理的事实抽取方法；Baseline2 为语料文件经过完整的预处理，但事实抽取规则只含主谓宾及时空信息的事实抽取方法；方法 Liuyaohua 为本文实现文献[11]的中文事件元素抽取规则。

从表 2 可以看到，System 在 R、P、F 和 C 上均高于其他方法；Liuyaohua 在查全率 R 上明显低于其他方法，该方法信息损失率大，不适合故事文本理解的事实信息抽取；从 System 和 Baseline1 的对比结果可以看到，Baseline1 找到的事实数目多于 System，经分析发现未经时间短语识别、动词变化形式识别处理的语料，特别是动词的多种变化形式会使得输入句子成分变得复杂；Stanford Parser 因缺乏对这类句型的上下文训练，导致分析出多个冗余的谓词信息，因此短语边界的识别可以预先为句法分析器扫清障碍，从而提高事实抽取的查全率、查准率和完整率；从 System 和 Baseline2 的对比结果可以看到，本文提出的规则对于查全率、完整率有较大影响，能有效地提高事实抽取的准确性。

表 2　System、Baseline1、Baseline2 和 Liuyaohua 4 种方法的评估结果

Method	N_{found}	$N_{t\&c}$	$N_{t\&inc}$	N_{total}	R	P	F	C
System	4078	3012	340	4344	0.7716	0.8220	0.7960	0.7386
Baseline1	4114	2700	356	4344	0.7035	0.7428	0.7226	0.6563
Baseline2	3922	2484	538	4344	0.6957	0.7705	0.7312	0.6333
Liuyaohua	3388	2149	431	4344	0.5939	0.7615	0.6673	0.6342

从图 4 可以看到，儿童故事文本主要由一元、二元、三元和四元事实构成。对抽取事实结构进一步分析发现，System 抽取出来的事实能比较全面地反映文本的语义信息；但 System 的启发式框架元素抽取规则没有将 Stanford Chinese Parser 的缺省依存关系 dep 包含进来，因为 dep 关系表示的语法关系不明确，可以是动宾，也可以是主谓或者其他任何关系，跳过 dep 关系就导致了部分事实信息的缺失；特殊句型的规则考虑也不完全，即使是兼语类型的句子，当含有大量数量、地点等修饰成分时，其抽取结果也不理想。另外，分词和词性标注出错（如“背/v 上/f”、“花/v 裤子/n”）以及后面句法分析结果的累计错误，还有汉语的副词、语气词等虚词，都影响到本算法的效果。

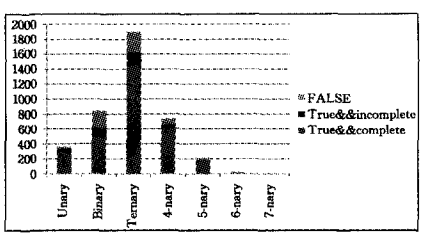

图 4　本系统 System 抽取的多元组分布及其准确性

4.3　本体模型验证

为检验 SOSDL 设计的合理性，本文采用与文献[14]类似的本体评估方法。将模型描述事件的能力分为 7 个维度：事件的动作、实体、时间、空间、事件间的关系、媒体和动画，与现有模型的描述能力进行了比较。表 3 给出了各模型的描述能力比较结果，SOSDL 与本体 E、Event-Model-F 和 NOEM 相比，其描述能力相当，但包含丰富的针对动画自动生成的故事语义理解任务需要的类和关系。

表 3　模型描述能力比较

模型	动作	实体	时间	空间	事件关系	媒体	动画
E	√	√	√	√	√	√	×
Event-Model-F	√	√	√	√	√	√	×
NOEM	√	√	√	√	√	√	×
SOSDL	√	√	√	√	√	√	√

为检验 SOSDL 在实际应用中描述故事事件的能力，我们组织实验室若干名研究生一起对约 200 篇故事文本进行了人工分析，超过 90% 的故事可以很好地用 SOSDL 模型表示故事涉及到的事件和事件的语义要素，以及语义要素本身的属性。

结束语　针对儿童故事文本的特点，基于机器学习的方法学习过程中参数设置复杂的缺点，本文使用句法依存分析和正则表达式结合的方法，研究从儿童故事文本中抽取多元事实，由此得到结构化的文本浅层语义表示结构，并生成 SOSDL 本体的实例。多元事实抽取的效果直接影响到语义模型生成的有效性。本方法在准确率 P、完整率 C、召回率 R 和 F 值都取得了不错的效果。但方法也易受噪声和不合语法规范的文本的影响。对于输入句法规范的文本，本方法的抽取结果的准确性可高达 97%，满足故事文本的语义理解及故事可视化需要，但对于句法成分复杂且不符合语法规范的文本，其准确率则大概在 65% 左右，信息缺失量大。本文还存在以下不足：(1) 故事语料预处理各个模块未考虑完全，例如人物对话内容识别只考虑显示的人物对话，故事角色实体、时间实体等命名实体的识别规则还不完备；(2)语义框架元素的抽取算法还要进一步提高，以满足故事可视化需要。这些不足是我们未来的改进方向。在后期的工作中，将进一步研究事件本体的查询和推理，得到故事的隐含信息，用于动画创作中的导演分镜头剧本生成和幼儿舞蹈创编，从而辅助幼儿的叙事教育和游戏教育。

参 考 文 献

[1] Fan J, Ferrucci D, Gondek D, et al. PRISMATIC: Inducing Knowledge from a Large Scale Lexicalized Relation Resource [C]// Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading. Los Angeles, CA, 2010: 122-127

[2] Etzioni O, Banko M, Cafarella M J. Machine reading[C]// In AAAI Conference on Artificial Intelligence. London, 2006: 1517-1520

[3] Barker K, Agashe B, Chaw S Y, et al. learning by reading: A prototype system, performance baseline and lessons learned[C]// AAAI2007. Vancouver, Canada, 2007: 280-286

[4] Banko M, Cafarella M J, Soderland S, et al. Open information extraction from the web[C]// In International Joint Conference on Artificial Intelligence. Hyderabad, India, 2007: 2670-2676

[5] 赵军，刘康，周光有，等. 开放式文本信息抽取[J]. 中文信息学报，2011，25(6)：98-110

（下转第 264 页）

本文方法相对于均值融合、中值融合来说，都有非常大的改观，这和直观的效果是一致的，证明了本文方法的有效性。

结束语 本文提出了一种在畸变环境中对图像序列进行融合的方法。该方法首先利用基于B样条的非刚性配准技术，对图像序列中每幅图像的局部畸变进行校正，之后，通过计算每幅校正后图像的质量指数，充分利用图像的互补信息进行融合。通过在真实数据集上的对比实验证明，该融合算法能有效地完成畸变环境中的图像序列校正。

在今后的研究中，将进一步研究提升图像融合质量的方法，提升对畸变的校正效果、消除环境的影响。同时进一步寻找在畸变环境中对图像质量进行判定的更客观的方法，对实验结果进行定性分析。

参考文献

[1] 刘纯胜，卢晓芬，洪汉玉，等. 基于特征点配准的气动光学图像校正方法研究[J]. 系统工程与电子技术，2006，28(10)：1468-1472

[2] Rueckert D, Sonoda L, Hayes C, et al. Nonrigid registration using free-form deformations: application to breat mr images [J]. Medical Imaging, 1999, 18(18): 712-721

[3] 李雄飞，张存利，李鸿鹏，等. 医学图像配准技术进展[J]. 计算机科学，2010，37(7)：27-33

[4] 王伟，苏志勋. 基于移动最小二乘法的医学图像配准[J]. 计算机科学，2010，37(9)：270-272

[5] Shimizu M, Yoshimura S, Tanaka M, et al. Super-resolution from image sequence under influence of hot-air optical turbulence [C]// Computer Vision and Pattern Recognition, 2008: 1-8

[6] Wang Zhou, Sheikh H R, Bovik A C. A universal image quality index [J]. IEEE Signal Processing Letters, 2002, 9(3): 81-84

[7] Zhu X, Milanfar P. Removing atmospheric turbulence via space-invariant deconvolution [J]. IEEE Transactions Pattern Analysis and Machine Intelligence, 2012, 35(1): 157-170

[8] Zhu X, Milanfar P. Image reconstruction from videos distorted by atmospheric turbulence [C]// SPIE Electronic Imaging, Conference on Visual Information Processing and Communication, 2010

[9] Oreifej O, Shu G, Pace T, et al. A two-stage reconstruction approach for seeing through water [C] // Computer Vision and Pattern Recognition, 2011: 1153-1160

[10] Szeliski R, Coughlan J. Spline-based image registration [J]. International Journal of Computer Vision, 1997, 22(3): 199-218

[11] Myronenko A, Song X. Intensity-based image registration by minimizing residual complexity [J]. IEEE Trans. on Medical Imaging, 2010, 29(11): 1882-1891

[12] Aubailly M, Vorontsov M A, Carhat G W, et al. Automated video enhancement from a stream of atmospherically-distorted images: the lucky-region fusion approach [C] // Proceedings of SPIE, 2009

[13] Aubailly M, Vorontsov M A, Carhat G W, et al. Image enhancement by local information fusion with pre-processing and composed metric [C]// Proceedings of SPIE, 2008

[14] Tian Y, Narasimhan S. Seeing through water: Image restoration using model-based tracking [C] // International Conference on Computer Vision, 2009: 2303-2310

(上接第225页)

[6] Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction[C] // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 1535-1545

[7] Akbik A, Bross J. Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns [C] // Proceedings of The WWW 2009 workshop on semantic search. Madrid, Spain, 2009: 6-15

[8] Akbik A, Löser A. KRAKEN: N-ary Facts in Open Information Extraction[C] // Knowledge Extraction Workshop at NAACL-HLT 2012. Montréal, Canada, 2012: 52-56

[9] 吴平博，陈群秀，马亮. 基于事件框架的事件相关文档的智能检索研究[J]. 中文信息报，2003，17(6)：25-30

[10] Wang W, Zhao D Y, Wang D. Chinese news event 5 w 1 h elements extraction using semantic role labeling[C]// Proceedings of the third International Symposium on Information Processing. China, Qingdao, 2010: 484-489

[11] 刘耀华. 基于句法分析的中文事件抽取方法研究[D]. 上海：上海大学，2009

[12] Westermann U, Jain R. Towards a Common Event Model for Multimedia Applications[J]. IEEE MultiMedia, 2007, 14(1): 19-29

[13] Scherp A, Franz T, Saathoff C, et al. F—a model of events based on the foundational ontology dolce+ DnS ultralight [C]// Proceedings of the fifth international conference on Knowledge capture. Redondo Beach, California, 2009: 137-144

[14] 王伟，赵东岩. 中文新闻事件本体建模与自动扩充[J]. 计算机工程与科学，2012，34(4)：171-175

[15] http://nlp.stanford.edu/software/lex-parser.shtml

[16] HtmlParser. http://htmlparser.sourceforge.net/

[17] ICTCLAS. http://www.ictcals.org/index.html

[18] 鲁川. 知识工程语言学[M]. 北京：清华大学出版社，2010

[19] Minsky M. A Framework for Representing Knowledge [EB/OL]. http://dspace.mit.edu/bitstream/handle/1721.1/6089/AIM-306.pdf? sequence=2

[20] Chang P C, Tseng H, Jurafsky D, et al. Discriminative reordering with Chinese grammatical relations features[C]// Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation. Boulder, Colorado, 2009: 51-59

[21] Allen J F. Maintaining Knowledge About Temporal Intervals [J]. Communications of the ACM, 1983, 26: 832-843

[22] Randell D A, Cui Z, Cohn A G. A Spatial Logic Based on Regions and Connection[C] // Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning. Morgan Kaufmann, San Mateo, 1992: 165-176

[23] Frank A. Qualitative Spatial Reasoning: Cardinal Directions as an Example[J]. In International Journal of Geographic Information Systems, 1996, 10(3): 269-290

[24] Faber P, Mairal R. Constructing a Lexicon of English Verbs [M]. Berlin/New York: Mouton de Gruyter, 1999