计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(3):1−8 [doi: 10.15888/j.cnki.csa.006233]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向动画自动生成的中文短信关系抽取①

李笑妃

(北京工业大学 信息学部, 北京 100124)

摘　要: 手机短信 3D 动画自动生成系统是根据发送方短信的内容, 经过信息抽取、语义分析等一系列步骤, 最终生成一段与短信内容匹配的三维动画并发送给接收方. 信息抽取处于手机 3D 动画自动生成系统首要和关键的位置, 其目的是为 3D 动画自动生成系统的后续环节提供可动画的信息. 本文引入路径特征实现中文短信文本的关系抽取. 利用哈尔滨工业大学 LTP-Cloud 平台对短信进行预处理, 从处理结果中提取路径并泛化得到路径特征, 通过一阶归纳学习器组合特征, 得到匹配规则, 然后再通过匹配到的规则对短信进行预测, 从而抽取出短信中的关系类型及对应的关系组合.

关键词: 信息抽取; 路径特征; 关系抽取; 一阶归纳学习器

引用格式: 李笑妃.面向动画自动生成的中文短信关系抽取.计算机系统应用,2018,27(3):1–8. http://www.c-s-a.org.cn/1003-3254/6233.html

Relation Extraction of Chinese Text Message in 3D Animation for Mobile Phone

LI Xiao-Fei

(Information Science Division, Beijing University of Technology, Beijing 100124, China)

Abstract: SMS automatic 3D animation generating system is based on sender message content, by information extraction, semantic analysis and a series of steps, eventually generating a 3D animation which matches the content of the text message. Information extraction in the 3D animation generating system is primary and key, the purpose is to provide animated information for follow-up links to 3D animation automatically generating system. This paper introduces the path feature to realize the extraction of the Chinese text message. It mainly uses the LTP-Cloud platform of Harbin Institute of Technology to preprocess the short message. The path is extracted from the processing result and the path of feature is obtained. Getting rules by combinating the path of features by the first-order inductive learner, and then predicting the relation of messages by matching rules. Finally, Extracting the type of relationship and relation combination in the text message.

Key words: information extraction; path of feature; relation extraction; first-order induction learner

1　引言

在审视了手机短信和 3G 通信技术的发展现状后, 中科院张松懋研究员于 2008 年提出将 3D 动画自动生成技术应用在手机短信上的想法, 即将发送的中文短信经系统处理分析后生成的 3D 动画发送给接收方, 命名为全过程计算机辅助手机 3D 动画自动生成系统[1] (简称为手机 3D 动画自动生成系统). 处理过程大致分为四个阶段, 短信信息抽取, 情节定性规划, 场景定量规划, 网络渲染. 手机 3D 动画自动生成技术将动画自动生成技术应用于中文手机短信领域, 不仅立足于一个崭新的应用角度, 并尝试研究和解决过程中出现的问题, 这在人工智能领域具有一定的研究意义和价值.

信息抽取处于手机 3D 动画自动生成系统首要和关键的位置, 而实体关系抽取作为信息抽取领域的重

① 收稿时间: 2017-06-05; 修改时间: 2017-06-19; 采用时间: 2017-06-26; csa 在线出版时间: 2018-01-25

要研究课题[2], 其主要目的是抽取句子中已标记实体对之间的语义关系, 即在实体识别的基础上确定无结构文本中实体对之间的关系类别, 并形成结构化的数据便于存储和取用, 例如, 输入一个带有标记实体的句子“< e1 > 曹德旺< /e2 > 任< e2 > 福耀玻璃集团< /e2 > 董事长, 是一名优秀的中国民营企业家. ”, 实体关系抽取系统能自动识别实体“曹德旺”和“福耀玻璃集团”的关系是雇佣关系.

关系抽取技术对自然语言处理的许多应用如本体构建、自动文摘、自动问答、知识库构建等具有重要的意义. 传统的关系抽取依赖于定义好的关系类型体系, 如定义的雇佣关系、整体部分关系、位置关系等. 目前的一系列研究也主要是围绕内容自动抽取会议 (ACE) 所设计的任务展开, 所抽取的关系类型一般也同 ACE 定义的一致.

1998 年, 美国国防高级研究计划委员会 (Defense Advanced Research Project Agency, DARPA) 资助的最后一届消息理解会议 (Message Understanding Conference, MUC) 首次引入了实体关系抽取任务. 1999 年, 美国国家标准技术研究院 (National Institute of Standards and Technology, NIST) 组织了自动内容抽取 (Automatic Content Extraction, ACE) 评测, 其中的一项重要评测任务就是实体关系识别[3]. 与 MUC 相比, ACE 的实体关系语料的语种数量和数据规模都有了大幅度的增加. ACE 2008 的关系抽取任务共定义了 Agent-Artifact、General-Affiliation、Metonymy、Organization-Affiliation、Part-Whole、Person-Social、Physical 7 个大类的实体关系, 细分为 User-Owner-Inventor-Manufacturer、Citizen-Resident-Religion-Ethnicity、Organization-Location 等 18 个子类的实体关系[4]. SemEval (Semantic Evaluation) 是继 MUC、ACE 后信息抽取领域又一重要评测会议, 该会议吸引了大量的院校和研究机构参与测评. SemEval-2007 的评测任务 4 定义了 7 种普通名词或名词短语之间的实体关系, 但其提供的英文语料库规模较小. 随后, SemEval-2010 的评测任务 8 对其进行了丰富和完善, 将实体关系类型扩充到 9 种, 分别是: Component-Whole、Instrument-Agency、Member-Collection、Cause-Effect、Entity-Destination、Content-Container、Message-Topic、Product-Producer 和 Entity-Origin. 考虑到句子实例中实体对的先后顺序问题, 引入“Other”类对不属于前述关系类型的实例进行描述, 共生成 19 种实体关系. SemEval-2010 评测引发了普通名词或名词短语间实体关系抽取研究的新高潮[5].

本文在句法语义分析的基础上对中文短信文本进行关系抽取, 针对于手机 3D 动画系统对动画的表现情况将关系分为 4 种, 包括: 颜色关系、形态关系、描述关系、位置关系, 如短信“我想吃红苹果”, 经过本文处理得到“苹果”和“红”属于颜色关系; 短信“雨下的真大啊”经处理后得到“雨”和“大”属于形态关系, 形态关系即表示物体的大小、长短等的描述; 短信“我的心情很好; ”经本文处理得到“心情”和“好”这样的描述关系. 由于前三种关系可以同属于描述类型, 所以前三种关系用同一语料库进行训练, 得到同一规则集, 只是在用规则集进行关系抽取的过程中细分为了三种关系. 短信“我书包在床上”, 经本文处理后得到“书包”和“床上”属于位置关系. 位置关系单独标注, 单独训练.

2 相关研究

2.1 实体关系抽取技术

在传统的语义关系抽取中, 实体与实体之间的关系是预先定义好的. 在关系抽取中先后出现了基于规则的方法, 其中有基于 ontology 实现信息抽取中的关系抽取[6], 取得比较不错的效果. 随着机器学习的发展, 人们将关系抽取看成一个分类问题, 首先标出句子中的实体, 然后通过一个分类器判断实体对之间的关系. 目前, 有监督学习方法是最基本的实体关系抽取方法, 其主要思想是在已标注的训练数据的基础上训练模型, 然后对测试数据的关系类型进行识别. 有监督学习方法包括基于特征的方法、基于核函数的方法[7]和基于规则的方法.

基于特征向量的方法是一种简单、有效的实体关系抽取方法, 其主要思想是从关系句子实例的上下文中提取有用信息 (包括词法信息、语法信息) 作为特征, 构造特征向量, 通过计算特征向量的相似度来训练实体关系抽取模型. 该方法的关键在于寻找类间有区分度的特征, 形成多维加权特征向量, 然后采用合适的分类器进行分类. 文献[8]在词法特征、实体原始特征的基础上, 融入依存句法关系、核心谓词、语义角色标注等特征, 实验结果表明该方法能有效提高实体关系抽取的性能.

基于核函数的实体关系抽取方法不需要构造特征

向量, 而是把结构树作为处理对象, 通过计算它们之间的相似度来进行实体关系抽取. 在基于核函数的中文实体关系抽取研究方面, 刘克彬[9]利用卷积核函数中的字符串序列核进行实体关系抽取, 并借用《知网》中的词汇语义相似度计算方法计算中文特征词串的相似度, 实验结果表明其 F 值达到了 84%, 这也说明语义信息能提高中文语义关系抽取系统的性能.

基于规则的方法需要对待处理语料通过人工或机器学习的方法总结归纳出相应的规则或模板[10], 然后采用规则或模板匹配的方法进行实体关系抽取. 近年来, 实体关系抽取研究者构建了多个基于规则的实体关系抽取系统[11,12].

机器学习中规则归纳即“规则学习”是从训练数据中学习出一组能用于对未见实例进行判别的规则. 与神经网络、支持向量机这样的“黑箱模型”相比, 规则学习具有更好的可解释性, 能使用户更直观地对判别过程有所了解. 另外, 数理逻辑具有极强的表达能力, 绝大多数人类知识都能通过数理逻辑进行简洁的刻画和表达. 如: “爸爸的爸爸是爷爷”这样的知识不易用函数式描述, 而用一阶逻辑可以方便的写成“爷爷 (X, Y)← 爸爸 (X, Z)∧爸爸 (Z, Y)”. FOIL (First-Order Inductive Learner)[12]是著名的规则学习算法, 首次由 Quinlan 在 1993 年提出, 该算法分为正例和负例提取规则, FOIL 算法采用信息增益来提取最好的一个属性值生成规则, 而且一次只生成一条规则, 再生成规则之后, 将被规则覆盖的训练集删除, 继续从剩余的训练集中寻找最好的属性值. 因为它是把命题规则学习过程通过变量替换等操作直接转化为一阶规则学习的, 因此比一般的归纳逻辑程序设计技术更高效. 文献[13]结合了 Apriori 算法和 FOIL 算法实现文本分类, 准确率达到了 99%.

2.2 句法、语义分析

句法分析[14]将句子由一个线性序列转化为一棵结构化的依存分析树, 通过依存弧上的关系标记反映句子中词汇之间的句法关系. 与短语结构相比, 句法结构具有形式简洁、易于标注、便于应用等优点, 逐渐受到学术界和工业界的重视. 语义分析默认要建立在句法分析的基础上, 中文的句法是从西方引进来的, 而中文严重缺乏形态的变化, 词类与句法成分没有严格的对应关系, 导致中文句法分析的精度始终上不去. 目前 LTP-Cloud 已经联合北京城市学院标注了 1 万句中文语义依存分析树[15], 且已经有初步的实验结果. 如句子“男孩跑步, 女孩跳舞”得到的句法分析与语义分析分别如图 1 和图 2 所示, 所以为了提高关系抽取的准确率, 本文采用句法分析与语义分析相结合的方式进行训练与测试.

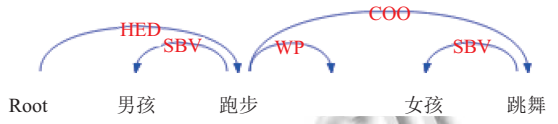


图 1 句法分析示例

图 2 语义分析示例

2.3 同义词词林

《同义词词林》是一部汉语分类词典, 其中每一条词语都用一个编码来表示其语义类别. 本文所用的《同义词词林》为《同义词词林 (扩展版)》, 是哈尔滨工业大学信息检索研究室在《同义词词林》的基础上研制的. 最终的词表包含 77 429 条词语, 其中一词多义的词语为 8860 个, 共分为 12 个大类, 94 个中类, 1428 个小类, 小类下再以同义原则划分词群, 最细的级别为原子词群, 这样词典中的词语之间就体现了良好的层次关系. 不同级别的分类结果可以为自然语言处理提供不同颗粒度的语义类别信息, 《同义词词林》语义信息能显著提高中文关系抽取的性能, 文献[16]就是根据《同义词词林》完成了实体关系抽取, 最高 F 值达到 81.8%.

3 本文的方法

3.1 基本流程

LTP-Cloud 是由哈尔滨工业大学社会计算与信息检索研究中心研发的云端自然语言处理服务平台. 后端依托于历时 10 年形成的语言技术平台, 语言云为用户提供了包括分词、词性标注、依存句法分析、命名实体识别、语义角色标注、语义依存分析在内的丰富高效的自然语言处理服务[17]. 本文在哈尔滨工业大学 LTP-Cloud 平台的基础上, 对语料进行初步处理, 获取含有句法语义分析的 XML 文档, 对 XML 文档进行特

征路径的提取, 然后经过一阶归纳学习器进行训练, 得到匹配规则. 最后通过规则进行预测, 得到关系抽取结果, 并对实验结果评估. 具体过程如图 3 所示. 下面章节将对主要过程进行详细介绍.

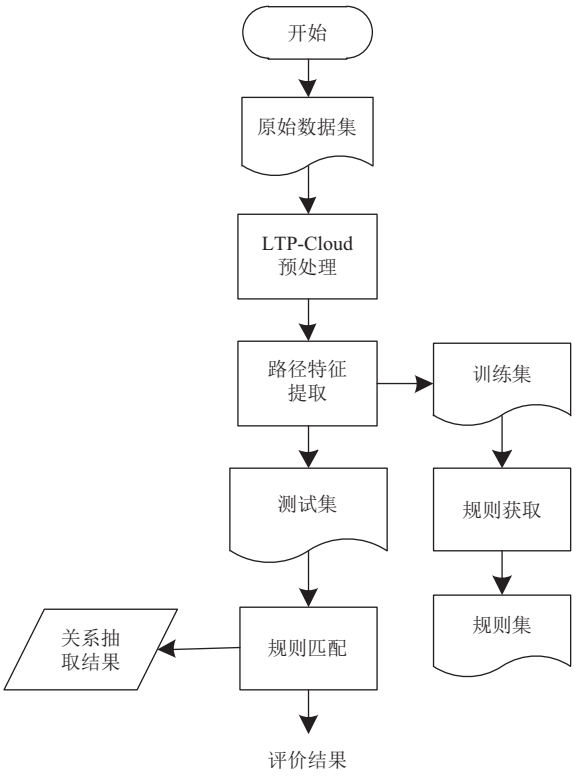


图 3 基于句法语义分析的关系抽取过程

3.2 路径特征形式化表示

短信中的实体本身以及实体之间有多方面的属性, 每一个属性刻画的信息可以将关系组合的具体化, 所以关系抽取问题可以转化成路径特征组合问题, 从短信文本中抽取出关于实体的路径特征, 然后使用一阶归纳学习器的思想来组合这些路径特征.

比如短信“黄色的苹果”, 经过 LTP-Cloud 处理后得到如图 4 所示结果.

```
<sent id="0" cont="黄色的苹果">
    <word id="0" cont="黄色" pos="n" ne="O" parent="2" relate="ATT" semparent="2" semrelate="Feat">
        <sem id="0" parent="2" relate="Desc" />
    </word>
    <word id="1" cont="的" pos="u" ne="O" parent="0" relate="RAD" semparent="0" semrelate="mAux">
        <sem id="0" parent="0" relate="mAux" />
    </word>
    <word id="2" cont="苹果" pos="n" ne="O" parent="-1" relate="HED" semparent="-1" semrelate="Root">
        <sem id="0" parent="-1" relate="Root" />
    </word>
</sent>
```

图 4 LTP-Cloud 处理结果示意

带标记的路径提取结果为:

Path1:{"id1":"0", "cont1":"黄色", "id2":"2", "cont2":"苹果", "pos1":"n", "pos2": "n", "relate":"ATT", "semrelate":"Feat"}

Path2:{"id1":"1", "cont1":"的", "id2":"0", "cont2":"黄色", "pos1":"u", "pos2": "n", "relate":"RAD", "semrelate":"mAux"}

Path3:{ "id1":"2", "cont1":"苹果", "id2":"–1", "cont2":" ", "pos1":"n", "pos2":" ", "relate":"HED", "semrelate":"Root"}

Path1-Path3 表示短信各个分词实体之间的关系以及实体本身的性质, path1 表示“黄色”词性是“n”, “苹果”词性是“n”; “黄色”与“苹果”之间的句法关系是“ATT”, 语义关系是“Feat”; “id1”和“id2”分别表示实体在 XML 结果中的位置, 是一种唯一性标识. 如果把实体“黄色”、“苹果”等变量替换成对应的词性, 则得到带标记的路径 path1-path3 泛化后的结果 F1-F3 即为路径特征.

F1: (n, n, n, n, ATT, Feat)

F2: (u, n, u, n, RAD, mAux)

F3: (n, –1, n,, HED, Root)

同样对于短信“我看见有红色的苹果”得到带标记的路径为:

Path1:{"id1":"0", "cont1":"我", "id2":"1", "cont2":"看见", "pos1": "r", "pos2": "v", "relate":"SBV", "semrelate":"Aft"}

Path2:{"id1":"1", "cont1":"看见", "id2":"–1", "cont2":" ", "pos1":"v", "pos2":" ", "relate":"HED", "semrelate":"Root"}

Path3:{"id1":"2", "cont1":"有", "id2":"1", "cont2":"看见", "pos1": "v", "pos2": "v", "relate":"VOB", "semrelate":"dCont"}

Path4:{"id1":"3", "cont1":"红色", "id2":"5", "cont2":"苹果", "pos1": "n", "pos2": "n", "relate":"ATT", "semrelate":"Feat"}

Path5:{"id1":"4", "cont1":"的", "id2":"3", "cont2":"红色", "pos1": "u", "pos2": "n", "relate":"RAD", "semrelate":"mAux"}

Path6:{"id1":"5", "cont1":"苹果", "id2":"2", "cont2":"有", "pos1": "n", "pos2": "v", "relate":"VOB", "semrelate":"Belg"}

泛化后的路径特征为:

F1: (r, v, r, v, SBV, Aft)

F2: (v, –1, v,, HED, Root)

F3: (v, v, v, v, VOB, dCont)

F4: (n, n, n, n, ATT, Feat)

F5: (u, n, u, n, RAD, mAux)

F6: (n, v, n, v, VOB, Belg)

可以看到第一条短信的 F1 与第二条短信的 F4 是一样的, 并且 F1 与 F4 所对应的带标记的路径 path1 与 path4 就是表示颜色关系的实体对的组合. 所以 (n, n, n, n, ATT, Feat) 可以作为一条匹配规则.

3.3 规则获取

3.3.1 规则学习算法

类似于一阶归纳学习器 FOIL, 使用从一般到特殊的策略来组合路径特征, 与 FOIL 不同的是, 在学习规则的时候, 不以单个实体作为规则中的基本单位, 而是以路径特征为基本单位. 规则获取算法流程如下.

算法. 规则获取 (Acquire Rules)

Input: Training Set D=P∪N, P: positive dataset, N: negative dataset
Output: Mapping rules set R for D

```
1. Rule R←Φ
2. While |P|>min_message do
3.      Selected path feature set Sf←Φ
4.      P′←P N′←N
7.          for message a∈P′ do
5.      while |N′|>0 and r.length<Maxrule.length do
6.          Candidate path feature S^P←Φ S^N←Φ
8.            and f^a to S^p
9.          end for
10.           for message b∈N′ do
11.             and f^b to S^p
14.             Computer FoilGain of f
12.         end for
13.         for path feature f∈S^p do
15.         end for
16.         find feature f_opt from S^p with maximum FoilGain
17.         add f_opt to Sf
19.     end while
18.         remove from P′、N′ all example not satisfied f_opt
20.     get rule r from Sf and add r to R
21.     remove all the message that satisfied r from P
22. end while
```

其中第 3–20 行描述了如何通过组合路径特征来学习匹配规则. 首先目标特征路径集合 S_f 初始化为空集, 正负训练数据集 P 和 N 分别初始化为 P′和 N′; 再通过最大信息增益值获取当前最优路径特征, 并把选择的特征 f_{opt} 添加到特征集合 S_f 中, 循环该过程直到 N′为空, 即选择的路径特征组合没有匹配到 N′中的短信; 在内层循环中第 5–19 行, 当 N′为空时结束, 得到一条规则, 然后删除所有的 P′中的匹配短信, 当 N′不为空时加特征进行路径特征组合, 直到 N′为空为止.

FoilGain 即为信息增益, 可以度量当前路径特征集合 S_f 添加路径特征后所增加的信息量. 假设 S_f 是当前选择的路径特征集合, |P|和|N|分别表示数据集中满足 S_f 的正例与反例的个数, 如果添加一个新的路径特征 f, 路径特征集合变成 S_f', 使得 S_f'的正例个数和反例个数变成|P′|和|N′|则添加路径特征 f 后获得的信息增益是:

$$\text{FoilGain(f)} = |P'| * (\log\frac{|P'|}{|P'|+|N'|} - \log\frac{|P|}{|P|+|N|})$$

信息增益值最大的被选择加入到路径特征集合 S_f 中, 路径特征组成的集合则构成了一条关系抽取规则.

4 实验结果与分析

4.1 实验结果评价指标

根据手机 3D 动画自动生成系统的表现能力将关系抽取分为颜色关系、位置关系、形态关系和描述关系四种, 由于本文将关系抽取过程看作是分类的过程, 所以这里的评价方式也采用常规的准确率 *P*、召回率 *R* 和 *F* 值. 准确率使针对预测结果而言的, 它表示的是预测为正的样本中有多少是真正的正样本. 公式表达如下:

$$P = \frac{TruePositives}{\#PredictedPositives} = \frac{TruePositives}{TruePositives + FalsePositives}$$

召回率是针对原来的样本而言的, 它表示的是样本中的正例有多少被预测正确. 公式表达如下:

$$R = \frac{TruePositives}{\#ActualPositives} = \frac{TruePositives}{TruePositives + FalseNegatives}$$

由于 *R* 和 *P* 指标有时候会出现矛盾的情况, 这样就需要综合考虑他们, 最常见的方法就是 *F* 值, 通过计算 *F* 值来评价结果, 常见的 *F* 计算方法如下:

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

4.2 实验设计

本文用同样的设计方案对有无借助语义分析结果两种情况做对比实验, 如下文所示.

4.2.1 训练实验设计

本文的关系抽取包括颜色关系、形态关系、描述

关系、位置关系四部分, 考虑到符合前三者关系的短信中路径特征相同, 所以将颜色关系、形态关系和描述关系结合在一起进行规则学习, 而位置关系则单独处理.

使用 Java 语言实现了本文中的规则获取算法考, 虑到手机 3D 动画自动生成系统处理的文本短小精悍, 包罗万象, 所以语料库主要来自三个方面:

(1) 手机 3D 动画自动生成系统历来的测试短信, 经处理去重随机抽取 1000 条文本.

(2) 北京邮电大学处理后的 10 万条短信中提取 8000 条.

(3) 1998 年 1 月份《人民日报》随机提取 4000 条句子.

其中表示颜色关系、位置关系和描述关系的短文本有 8546 条, 表示位置关系的短文本有 1697 条. 使用 LTP-Cloud 对短文本进行预处理, 从中提取出路径特征, 用规则学习算法进行学习. 考虑到算法复杂度以及文本的特点, 需要对路径特征组合的最大长度做出限制, 多次试验最终把最大长度设置为 8, 即规则包含的路径特征个数最大为 8.

4.2.2 测试实验设计

同样使用 Java 语言设计实现测试系统, 该测试系统即为关系抽取系统, 该系统通过匹配规则集可以抽出短信中包含的关系以及关系组合. 系统主要分两个部分, 第一部分是颜色关系、形态关系、描述关系的抽取, 本文把这三种关系统称为描述型关系, 第二部分是位置关系的抽取. 测试预料主要来自两方面, 一方面是手机 3D 动画自动生成系统中除去训练集的部分短信 300 条, 另一方面是北京邮电大学 10 万条短信中抽取的 550 条, 总共 850 条短文本.

描述型关系抽取过程如图 5 所示, 在颜色关系与形态关系的抽取过程中, 需结合《同义词词林 (扩展版)》获取表示颜色和形态的类别, 同时得到该类别下的所有词群. 如果带标记的路径中所包含的实体能够在词群中找到所对应的原子, 则表示短信中含有颜色关系或者位置关系, 然后结合带标记的路径推导出相应的关系组合; 否则可判定为描述关系, 同样结合带标记的路径抽取出描述关系的组合. 与描述型关系抽取过程类似, 位置关系的抽取首先是进行规则匹配, 得到带标记的路径, 然后再根据带标记的路径分析结果, 找到关系组合.

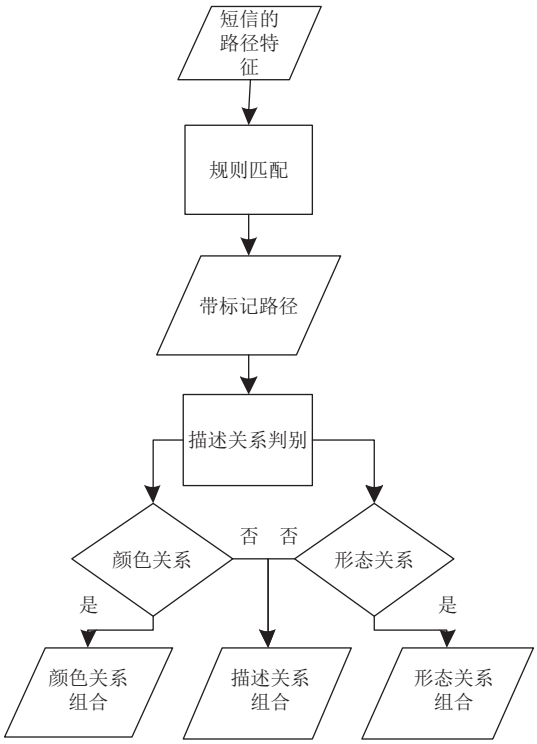


图 5 描述性关系抽取过程

4.3 实验结果

本文根据不同的路径特征进行对比实验, 分析借助语义分析后的关系抽取效果. 通过训练实验得到借助语义分析的描述型关系的规则集条数为 126 条, 未借助语义分析的规则集条数为 103 条, 位置关系的规则学习也得到两个数据 24 条与 32 条, 表 1 为得到的描述型关系与位置关系规则集示例.

表 1 规则集示例

路径特征类型	描述性关系部分规则
借助语义分析	(n, a, n, a, SBV, Exp) (a, n, a, n, ATT, Feat) (n, n, n, n, ATT, Feat) (d, a, d, a, ADV, mDegr);(a, –1, a,, HED, Root)
无语义分析	(n, a, n, a, SBV) (z, n, z, n, ATT) (n, v, n, v, SBV);(v, –1, v,, HED);(n, v, n, v, VOB)
特征路径类型	位置关系部分规则
借助语义分析	(n, n, n, nd, ATT);(n, v, n, v, Loc) (n, v, nl, v, Loc) (n, n, n, nd, ATT);(n, v, n, v, Loc);(n, v, n, v, VOB, Exp) (n, v, n, v, Loc);(n, n, nd, n, mRang);(n, v, n, v, VOB, Exp)
无语义分析	(n, n, n, nd, ATT);(n, v, n, v, SBV) (n, v, n, v, SBV);(n, v, n, v, VOB) (n, v, n, v, SBV);(v, –1, v,, HED);(n, v, n, v, VOB)

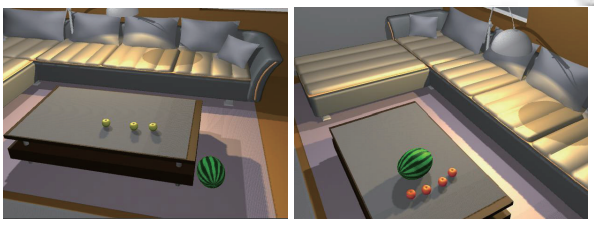
短信“看见桌子上有红色苹果和大西瓜, 心情好呀”, 通过带语义分析的规则匹配, 得到如图 6 所示的 IE 输出结果结果. 其中的 Relation 标签下的文本是本文关系抽取结果的结构化表示形式. 短信包含有四种关系, 其中颜色关系有两个组合一个是“苹果“与”红”, 表示形态关系的标签为 Form, 关系组合为“西瓜”与“大”; “心情”与“好”构成描述关系的组合; 最后一条 Location 表示的是位置关系, 即“苹果; 西瓜”与“桌子上”构成位置关系组合, 表示前者的位置是“桌子上”. 通过这些关系输出可以为手机 3D 动画系统提供可供动画表现的信息, 比如可以刻画水果的颜色与大小, 还能对物体出现在动画中的位置做出规划. 图 7(a) 与图 7(b) 即为手机 3D 动画自动生成系统生成在关系处理前和处理后的动画截图, 由图 7(b) 可以看出苹果是红色的, 并且在桌子; 西瓜也在桌子上. 表现了位置关系和颜色关系, 更能表现短信所要表达的内容. 并对预测结果进行评估得到表 2 的评估结果. 另外, 文献[18]所提出的中文实体关系抽取方法是中文实体关系抽取领域较为经典的方法之一, 本文将关系分成两类描述性关系与位置关系, 同时变成了二分类问题. 将本文的基于语义分析的实验结果与文献[18]的研究结果进行了比较得到图 8 所示对比图.

表 2 实验评估结果 (单位: %)

特征路径类型	颜色关系		形态关系		描述关系		位置关系	
	有语义分析	无语义分析	有语义分析	无语义分析	有语义分析	无语义分析	有语义分析	无语义分析
准确率P	100	100	94.12	83.56	88.73	83.83	84.21	87.5
召回率R	97.72	82.21	82.47	74.39	74.71	65.61	65.30	57.14
F	98.85	90.24	87.91	79.05	81.11	73.61	67.47	69.10

```
<result negType="" negCont="">
  <message value="看见桌子上有红色苹果和大西瓜，心情好呀"/>
  <nemessage value=" 看见桌子上有红色苹果和大西瓜，心情好呀"/>
  <segmessage value="看见#vb 桌子#nn 上#nd 有#vb 红色#nn 苹果#nn
                和#cj 大#aj 西瓜#nn ，#wp 心情#nn 好#aj 呀#ax"/>
  <topic name="" key="">
    <root name="生活用品" flag="" value="家具:桌子" />
    <root name="食物" flag="" value="水果:苹果;西瓜"/>
  </topic>
  <relation>
    <root name="Color" cont="苹果"  value="红" />
    <root name="Form" cont="西瓜"  value="大" />
    <root name="Description" cont="心情"  value="好" />
    <root name="Location" cont="苹果;西瓜"  value="桌子上" />
  </relation>
</result>
```

图 6 短信关系抽取结果示例



(a) 处理前 (b) 处理后

图 7 手机 3D 动画生成系统最终动画截图

4.4 结果分析

分析上述结果可以看出, 本文所述方法在借助语义分析情况下颜色关系和形态关系抽取方面准确率比较高, 原因是在关系抽取过程中结合了《同义词词林(扩展板)》, 从而囊括了颜色与形态的几乎所有情况, 并且表示颜色和形态的实体词词性也比较单一, 主要是名词或者形容词, 所以准确率比较高. 而位置关系抽取效果相对较差, 召回率低, 只有 65%, 造成这种情况的原因一方面是位置关系训练语料库规模比较小; 另一方面是表示短文本的路径特征的选取以及路径特征间的顺序不太合适; 再一方面就是在对语料库的结果标注存在很大的人为因素. 考虑到目前手机 3D 动画自动生成系统的表现能力, 关系抽取主要要求准确率高. 在使用经典关系抽取算法得到的结果中, 可以看出在手机 3D 动画自动生成系统中, 本文的方法取得了比较不错的结果, 可以应用到目前的手机 3D 动画系统中.

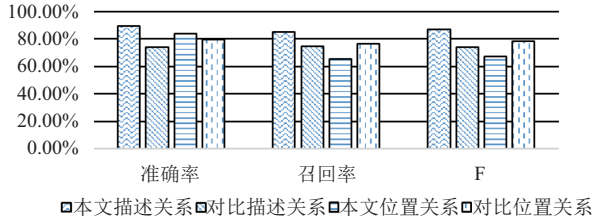


图 8 实验结果对比图

5 总结

本文研究的主要内容是首次在手机 3D 动画信息抽取系统中添加关系抽取. 提出了一种基于规则学习的短文本关系抽取方法. 首先结合手机 3D 动画自动生

成系统, 定义了颜色关系、形态关系、描述关系和位置关系四种类型, 然后在句法、语义分析的基础上, 通过一阶规则学习算法获取关系抽取的规则集, 测试集通过匹配规则集得到关系类型并抽取出对应的关系组合, 最后以结构化的形式将关系输出到信息抽取结果中, 为手机 3D 动画系统提供更多可供动画表现的信息.

本文的研究是在句法分析、语义分析的基础上进行的, 研究对象是中文的短文本, 而目前中文的语义分析效果还不是很理想, 这就降低了关系抽取的准确率. 另外, 人为标注语料库存在很大的局限性和主观性, 限制了语料库的规模, 质量也不高, 进而影响规则的学习. 针对以上不足, 在后续关系抽取的研究过程中, 需要充分利用自然语言处理的最新研究成果, 实现自动化或半自动化标注语料库, 提高关系抽取的准确率.

参考文献

1 吴中彪. 全过程计算机辅助手机 3D 动画自动生成系统的设计与实现[硕士学位论文]. 北京: 北京工业大学, 2011. 11–38.

2 陈宇, 郑德权, 赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取. 软件学报, 2012, 23(10): 2572–2585.

3 http://www.ldc.Upupenn.edu/Projects/ACE/.

4 Chan YS, Roth D. Exploiting background knowledge for relation extraction. Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China. 2010. 152–160.

5 Hendrickx I, Kim SN, Kozareva Z, *et al*. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Boulder, CO, USA. 2009. 94–99.

6 Chen GC, Zhao JY, Cohen T, *et al*. Using ontology fingerprints to disambiguate gene name entities in the biomedical literature. Database, 2015, (2015): bav034.

7 王敏. 基于多代理策略的中文实体关系抽取[硕士学位论文]. 大连: 大连理工大学, 2011. 1–55.

8 郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取. 中文信息学报, 2014, 28(6): 183–189.

9 刘克彬, 李芳, 刘磊, 等. 基于核函数中文关系自动抽取系统的实现. 计算机研究与发展, 2007, 44(8): 1406–1411.

10 Du XZ, Doermann D, Abd-Almageed W. Signature matching using supervised topic models. Proceedings of the 22nd International Conference on Pattern Recognition. Stockholm, Sweden. 2014. 327–332.

11 McDonald DM, Chen H, Su H, *et al*. Extracting gene pathway relations using a hybrid grammar: The Arizona relation parser. Bioinformatics, 2004, 20(18): 3370 –3378. [doi: 10.1093/bioinformatics/bth409]

12 Quinlan JR, Cameron-Jones RM. FOIL: A midterm report. European Conference on Machine Learning: ECML-93. Vienna, Austria. 1993. 1–20.

13 汪雪君. 基于规则的分类方法研究[硕士学位论文]. 漳州: 闽南师范大学, 2013: 1–47.

14 刘挺, 车万翔, 李正华. 语言技术平台. 中文信息学报, 2011, 25(6): 53–62.

15 邵艳秋, 邱立坤, 梁春霞, 等. 中文语义依存关系资源建设及分析技术研究. 第十一届全国计算语言学学术会议. 洛阳, 中国. 2011.

16 刘丹丹, 彭成, 钱龙华, 等. 《同义词词林》在中文实体关系抽取中的作用. 中文信息学报, 2014, 28(2): 91–99.

17 http://www.ltpc loud.com/intro/.

18 徐芬, 王挺, 陈火旺. 基于 SVM 方法的中文实体关系抽取. 第九届全国计算语言学学术会议论文集. 大连, 中国. 2007. 497–502.