第 38 卷 第 1 期 计 算 机 工 程 2012 年 1 月
Vol.38 No.1 Computer Engineering January 2012

·人工智能及识别技术· 文章编号：1000—3428(2012)01—0185—03 文献标识码：A 中图分类号：TP18

全过程动画自动生成中的中文文本处理

孙容容，刘椿年

(北京工业大学计算机学院，北京 100124)

摘 要：研究全过程动画自动生成系统中的自然语言处理模块，设计一种面向手机中文短信的信息抽取系统。根据中文语言处理的特殊性，抽取短信中可动画化的信息，并进行否定判断和否定内容识别。实验结果表明，该系统的召回率和准确率较高，可满足动画自动生成系统对信息抽取强度的要求。

关键字：动画自动生成；自然语言理解；信息抽取；模板；否定判断

Chinese Text Processing in Full Cycle Animation Auto-generation

SUN Rong-rong, LIU Chun-nian

(College of Computer, Beijing University of Technology, Beijing 100124, China)

【Abstract】Based on the application of natural language processing in the full cycle computer aided animation auto-generation technology, this paper design a system of information extraction in Chinese mobile phone text messages. The target is to extract animation information from mobile phone message. The system contains negative judgements and negative content recognition . Experimental results show that the recall ratio and precision of this system can meet the Information Extraction(IE) intensity requirements of the animation auto-generation system.

【Key words】animation auto-generation; Natural Language Understanding(NLU); Information Extraction(IE); template; negative judgement

DOI: 10.3969/j.issn.1000-3428.2012.01.059

1 概述

随着人工智能、计算机图形学与硬件技术的快速发展，计算机动画技术已达到一定的水平，但在追求如何将人工智能技术用于计算机动画，提高动画制作的自动化程度和智能性上却提出了更高的挑战。

文献[1]提出了全过程计算机辅助动画自动生成技术，将电影艺术、人工智能和图形学技术引入动画生成全过程。该技术以受限自然语言的形式输入一个适当的故事，由此开始直到动画最终生成，每一步都是在计算机辅助下完成。该技术的第 1 版动画系统《天鹅》已在 1995 年实现，用该系统制作的动画片《三兄弟》也曾在中央电视台播放，初步验证了动画自动生成技术的可行性。

动画自动生成项目组将该技术应用在 2 个系统上：基于语义理解的古建动画辅助生成系统和面向手机中文短信的动画自动生成系统。自然语言处理模块是动画自动生成的起点，也是整个过程中的一个重要的模块。本文从自然语言理解和信息抽取的角度出发，介绍自然语言处理技术在全过程动画自动生成中的应用。

2 相关知识

2.1 受限自然语言理解

自然语言理解(Natural Language Understanding, NLU)的研究开始于 20 世纪 60 年代初，是人工智能的分支学科。自然语言理解是自然语言处理技术中最困难的一项，它研究计算机模拟人的语言交流过程，使计算机能够理解和运用人类的语言，实现人机之间的语言通信，例如查询资料、摘录文献等。

由于用计算机系统来完整地表达和理解中文自然语言是不现实的，因此文献[1]根据儿童童话故事的语言特点总结出一组语法规则，称之为受限自然语言，它是中文自然语言的一个子集，书写格式受限于自然语言单句理解子系统的处理范围。受限自然语言理解主要应用于《天鹅》系统中。

《天鹅》系统通过自己的自然语言单句理解系统对受限自然语言描述的童话故事进行识别处理，经过分词、语法分析、语义分析等步骤将故事转换为第 1 层中间语言 GF1。此外，还需要对故事文本进行常识检查，包括上下文无关常识检查和上下文相关常识检查[1]。上下文无关常识检查是指对故事情节的合理性进行检查，例如“皇后和王子结婚了”是不符合常识的；上下文相关常识检查是指对故事情节进行关联常识检查，例如“公主死了，公主拍球”是不符合常识的[1]。

2.2 受限文本的信息抽取

自然语言处理技术通常用于自由文本的信息抽取(Information Extraction, IE)，不是从文件集中选取一个与用户需求相关的子集，而是从文本中直接抽取与用户需求相关的事实或信息。信息抽取系统中的关键组成部分是一系列的抽取规则，其作用是确定需要抽取的信息。

信息抽取的研究对象主要分为 3 种[2]：结构化文本，自由文本，半结构化文本。

基于语义理解的古建动画辅助生成系统中信息抽取是以受限的中文文本为研究对象的。系统针对不同的用户设计并实现了 2 个不同的版本：

(1)针对没有古建知识的用户，采用完全受限的输入，界面实现采用的是选择式的下拉列表框的方式，该版系统适用于会展等公共场所的展览。

(2)针对具有一定古建知识的用户，要求输入的是局部受

基金项目：国家自然科学基金资助项目(60496322)

作者简介：孙容容(1985－)，女，硕士研究生，主研方向：文本信息抽取，人工智能；刘椿年，教授、博士生导师

收稿日期：2011-07-11 **E-mail：**sunrong199@126.com

限的中文文本，系统通过一些规则从中抽取建筑类型所有的参数信息，不关心其他的输入，这就需要用户的输入必须包含建筑的所有参数。

本文将主要通过面向手机中文短信信息抽取系统介绍自由文本信息抽取的全过程。

2.3 信息抽取研究现状

从自然语言文本中获取结构化信息被看作是信息抽取技术[3]的初始研究，它最早开始于 20 世纪 60 年代中期，以美国纽约大学的 Linguistic String 项目和耶鲁大学 Roger Schank 及其同事开展的有关故事理解的研究 2 个长期的、研究性的自然语言处理项目为代表。

20 世纪 80 年代以来，美国政府一直支持消息理解会议(Message Understanding Conference, MUC)对信息抽取技术进行评测。MUC 系列会议使信息抽取发展成为自然语言处理领域一个重要分支，并一直推动该领域向前发展。1987 年开始到 1998 年，MUC 会议共举行了 7 届，MUC 的特点在于对信息抽取系统的评测[4]。除了 MUC 外，Tipster 文本项目、自动内容抽取会议(Automatic Content Extraction, ACE)、多语言实体任务会议(Multilingual Entity Task, MET)也推动了信息抽取的发展。

过去几年信息抽取研究成果丰硕。英语和日语姓名识别已达到了人类专家的水平。但是中文信息抽取研究起步的比较晚，目前基本集中在命名实体识别和自然语言处理上，实现完整的中文信息抽取系统还处于探索阶段。

3 面向手机中文短信的信息抽取系统

信息抽取的处理步骤包括：句法分析，词性标注，命名实体识别和抽取规则。具体说就是把文本分割成多个句子，对一个句子的成分进行标记，然后将分析好的语法结构和事先定制的规则进行匹配，获得所需的内容。规则可以由人工编制，也可从人工标注的语料库中自动学习获得。在面向手机中文短信的信息抽取系统中，规则是由人工编制的。系统的研究对象是自由的短信文本，任务是抽取可动画化的信息。

中文自然语言处理在算法、理论、系统实现等方面都取得了显著进步。尤其统计语言模型引入该领域后，中文分词、词性标注的准确度和处理速度都有了很大提升。面向手机中文短信的信息抽取系统借助了哈尔滨工业大学信息检索研究室的自然语言处理系统，系统中的命名实体模块是在自然语言处理系统命名实体模块后设计的，Topic 和模板的规则涉及到了分词、词性标注和命名实体模块。否定分析器是在语义角色标注模块后设计并实现的。

3.1 系统框架

信息抽取系统的研究成果丰硕，其中也有很多手机短信的信息抽取的研究，例如手机短信服务中的信息抽取[5]，它是将短信中的信息进行结构化处理，直接将用户关心的焦点内容发送到手机上，使得用户不需多次翻屏，只需一两次翻屏就可得到所需要的信息；中文短信的过滤[6]，研究特定领域的短信的倾向性进行识别。还有面向手机短信中的命名实体识别的研究[7]。而本系统是以动画化为目的，抽取出可用动画表现的信息。具体流程如下：

(1)文本预处理。

(2)命名实体识别。识别短信文本中的专有名称，如人名、饭店名、购物场所等。

(3)Topic 分析。匹配短信文本中存在的 Topic 信息，划分 Topic 所属的句子范围并在句子范围内进行专有模板的匹配。输出为 Topic 及其专有模板信息。Topic 匹配失败，对整条短信进行共享模板的匹配，输出为所有匹配成功的共享模板的信息。共享模板匹配失败，输出为模板的中间信息。

(4)模板分析。根据划定的 Topic 的句子范围进行模板的匹配。

(5)否定分析。判断短信中是否存在否定信息。

(6)后续处理。将输出的数据转化为统一格式。如时间：2009 年 12 月 24 日 22 点 10 分，将转化为诸如 24/12/2009/10/22 这样的格式。

面向手机短信的信息抽取系统的系统框架如图 1 所示。

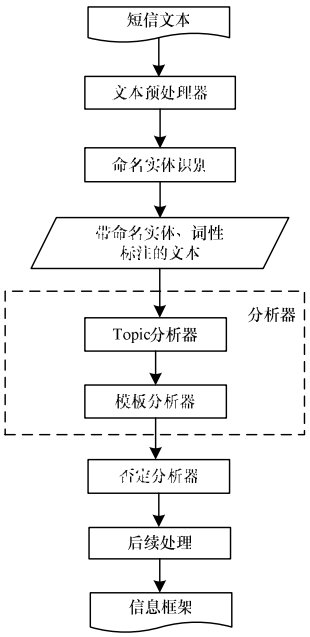

图 1 面向手机短信的信息抽取系统框架

3.2 系统知识库

信息抽取系统中的关键组成部分是一系列的抽取规则，其作用是确定需要抽取的信息。系统分为 3 个信息层次：Topic 层，模板层，信息层。

系统以短信表达的主旨内容的不同划分为不同的 Topic，每个 Topic 下对应一个模板的集合，称为 Topic 专有模板。模板所表达的是短信中存在的可动画化的信息，它将短信中经常出现的如时间、人物等信息以规则的形式表达出来。

(1)Topic

规则采用了正则表达式的形式。下面以约会为例说明 Topic 的定义文法和存储形式：

Topic 文法如下：

约会=约会|[到去在][^，：]{1,10}碰面|一起[^，：]{1,10}看电影|一起[^，：]{0,10}散步|[^，：]{2,4}我们见面|出去走走|一起共度

其中，表达式等号右侧的信息称为原子，属于 Topic 信息层。由于 Topic 的结构单一，而且原子的数据量很大，因此笔者将 Topic 的种类及原子信息存储在 SQL Server 数据库中，从而提高系统运行效率。

(2)模板

规则采用了 EBNF 的上下文无关文法与正则表达式相结合的格式。以部分地点模板为例说明模板的定义文法和存储形式，如下所示：

模板文法(#nn 表示名词词性，#nb 表示数词)：

地点=学校地点|社交场所

学校地点=图书馆|教室|讲台|操场
社交场所=饭店|娱乐场所
娱乐场所=舞厅|酒吧|歌厅
饭店=餐馆|餐厅
餐馆=饭店|酒店|食堂|餐馆
餐厅=餐厅数字，修饰餐厅
餐厅数字=(阿拉伯数字|个位数字|十位)
修饰餐厅=(#nb)?餐厅#nn|(#nb)?食堂#nn

规则的终结状态表达式右侧信息称为原子，所有左侧信息称为非原子信息。

在模板中，原子信息采用正则表达式的格式，而所有非原子层即非终结状态采用 EBNF 格式。系统中模板分为原子信息和非原子信息两部分存储，前者数据量很大，存储在数据库中，结构性很强的非原子层以 XML 形式存储。

系统中 Topic、模板的总结及分类都是以动画化为标准的，是通过阅读动画自动生成项目组成员搜集的 1 845 条短信文本(网络搜集和个人收到的短信)总结的，并在测试中不断填充和修改，其中 1 500 条用来总结 Topic 和模板，345 条用于系统测试。

(3)Topic 专有模板

与 Topic 紧密相关的模板。一个模板可被多个 Topic 所用。

(4)共享模板

Topic 所共有的不依附于 Topic 存在的模板。系统部分 Topic 与模板信息如表 1 所示。

表 1　部分 Topic 与模板信息

Topic 类别	Topic 名称	Topic 专有模板	共享模板
运动	打篮球、足球、跑步	天气、服饰、生活品、交通	时间、地点、情绪、人物
节日	新年、节日、中秋、元旦、纪念日、结婚、圣诞、七夕、典礼	天气、语气、人物角色、动作、花草、娱乐	
社会	上下班、社会通用、购物、旅游、出行	天气、生活品、交通、服饰、食物	
情绪	关怀、思念、喜、悲、歉意、生气、谢意	语气、动作、动物、花草	
学校	学校日常、毕业、假期、开学、考试	天气、生活品、动作、交通、服饰	
聚会	相聚、约会、唱歌、吃饭、跳舞	天气、交通、服饰、花草、娱乐	

由于总结模板的短信内容的局限性和有限性，系统现有模板的覆盖率较低，但已经基本满足了系统测试的目标，并且系统具有良好的可扩展性，当系统需要扩展时，只需添加新的模板 xml，并将所需信息添加到数据库中，不需要关心程序的修改。

4　实验结果与分析

4.1　否定识别实验

否定性词语对于确定文本中的事件发生与否和是非评价起重要作用，对于信息抽取也具有决定性的影响。

信息抽取系统通常采用有无否定副词作为否定判断的标准，并在一些特定领域，例如医学领域[8]取得了很好的成果。这些领域主要是对结构化的专业术语的否定判断，对自由文本尤其是短小、结构简单的短信，还需要判断否定词否定的内容，否定内容是指被最终确定下来的实际被否定项。

由于否定的识别非常复杂，而哈尔滨工业大学的自然语言处理系统中没有否定处理的模块，因此系统设计了自己的否定的识别方法：系统否定模板识别和系统设计的否定分析器，而对间接否定、隐含否定、无标记否定[5]等不做研究。

系统否定模板识别是基于规则的，是在短信分词后进行的。否定分析器是在分析语法、语义基础上实现对否定和否定内容的判断。系统通过语义分析出否定词，再以否定词为中心找出语法上与否定词相关的否定内容。学者钱敏汝提出对否定载体“不”应从语义、语法 2 个角度在各语言层面上的否定内容进行了探讨，即做语义的否定分析。

否定识别的实验结果是以 102 条含有否定词的短信进行测试的，系统 2 种否定识别方法的实验结果如表 2 所示。

表 2　否定识别结果　(%)

识别方法	召回率		准确率	
	否定词	否定内容	否定词	否定内容
否定模板	79.4	65.7	82.8	89.6
否定分析器	85.3	75.5	93.8	97.4

由表 2 可以看出，否定模板方法的召回率和准确率较低，并且否定内容的匹配也是不完整的，而动画生成更注重的是完整的否定内容。例如：

短信：正在睡觉呢，你到学校怎么也不打个电话回家啊？

根据规则“否定词”+“动词”，否定模板只能识别出“不”+“打”的否定信息，而否定分析器能识别出“否定词”+“否定内容”：“不”+“打电话”，识别出了否定的具体信息。

否定分析器针相对于否定模板的识别，在召回率和准确率上有了很大的提升，否定和否定内容的识别在一定程度上已经达到了系统的需求。

4.2　信息抽取实验

短信召回率和短信准确率是指一条短信中信息点的召回率和准确率，而召回率和准确率是对所有短信的召回率和准确率求平均值。召回率和准确率计算公式如下：

$$\text{短信召回率}=\frac{\text{短信中正确匹配的信息点}}{\text{短信中的信息点}}$$

$$\text{短信准确率}=\frac{\text{短信中正确匹配的信息点}}{\text{短信中匹配成功的信息点}}$$

$$\text{总召回率}=\frac{\sum\text{短信召回率}}{\text{短信的数量}}$$

$$\text{总准确率}=\frac{\sum\text{短信准确率}}{\text{短信的数量}}$$

例如：

(1)短信：这是我的新号码，你买了新手机啊？

短信中共有 1 个 Topic 信息点和 5 个模板信息点：购物 Topic 信息“买”，语气词“疑问语气”，抽象物“号码”，人物“我”、“你”，生活用品“手机”。但是生活用品“手机”没有匹配成功，所以这条短信模板的召回率为 80%。

(2)短信：我还有几十分钟就到学校了，累啊。

短信中共有 4 个模板信息点：时间“几十分钟”，地点“学校”，人物“我”，动作“累”。其中时间匹配为“十分”，匹配的信息不完整，所以这条短信模板的准确率为 75%。

在测试的 345 条短信中，匹配出 Topic 的短信有 148 条，占 42.9%，匹配出共享模板的短信有 148 条，占 54.8%，只匹配出中间信息的有 8 条，占 2.3%，其中，匹配出 2 个以上 Topic 的短信占 6.09%。

(下转第 191 页)