第 37 卷 第 21 期
Vol.37 No.21

计 算 机 工 程
Computer Engineering

2011 年 11 月
November 2011

• 软件技术与数据库 •　　文章编号：1000—3428(2011)21—0049—03　　文献标识码：A　　中图分类号：TP393

面向中文短信的信息抽取方法

吴中彪，刘椿年

(北京工业大学计算机学院，北京 100124)

摘　要：在手机 3D 动画自动生成系统中，研究面向中文短信的信息抽取方法。设计一种基于上下文无关文法的模板定义方式，以及对应的模板知识库与模板解析器。在模板解析器处理数据的过程中，通过最左规约算法保证中文短信的信息抽取效率。实验结果表明，该方法在扩展抽取内容范围的同时，能提高信息抽取的准确性。

关键词：手机 3D 动画自动生成系统；模板知识库；模板解析器；信息抽取

Information Extraction Method for Chinese Text Messages

WU Zhong-biao, LIU Chun-nian

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

【Abstract】In the application domain of mobile phone 3D animation automatic generation system, resesrches the information extraction method for Chinese text messages. It proposes a method to do the information extraction on Chinese text messages. A domain template definition method based on the limited context-free grammar is defined. After that designs and implements a template base with the corresponding template parser. The template parser uses the left-first deduction algorithm to ensure the efficiency. Experimental results show that this method can expand the extracted range and improve the accuracy of information extraction.

【Key words】mobile phone 3D animation automatic generation system; template knowledge base; template resolver; information extraction

DOI: 10.3969/j.issn.1000-3428.2011.21.017

1　概述

在信息抽取技术的发展过程中，消息理解会议(MUC)[1-5]、自动内容抽取会议(ACE)、多语言实体任务会议(MET)[2]等起到重要的推动作用。其中，MUC 会议的主旨是建立信息抽取系统的评价指标体系；ACE 的主要研究内容是从新闻语料库中自动抽取实体、关系、事件等内容。

国内有关中文信息抽取的研究与国外相比起步较晚。对信息抽取系统构建的尝试较早的是北京大学会议新闻抽取系统[6]和上海交通大学多语种投资信息抽取系统[7]。近年来，中文信息抽取技术发展较快。北京邮电大学的面向手机短信的命名实体识别系统[8]通过研究手机短信及其命名实体的语言及构词特点，构建相应的专家知识库，然后混合运用专家知识和条件随机场模型进行手机的命名实体识别。文献[9]的信息抽取系统从全信息角度综合分析语法、语义、语用信息构建中文的信息抽取系统。

文献[10]提出一种全过程计算机辅助动画自动生成技术。面向手机短信 3D 动画自动生成的信息抽取相对于一般意义上的信息抽取而言，在系统目标、性能的设计和考虑上有很多差别。本文重点阐述基于中文短信的信息抽取模块的实现细节和技术路线，探讨的信息抽取的相关方法全部在全过程计算机辅助手机 3D 动画自动生成系统(以下称手机 3D 动画自动生成系统)应用范畴内。

2　面向中文短信的信息抽取总体设计

2.1　中文短信及其信息抽取的特点分析

短信息作为人们交流的信息载体，在格式上一般是没有任何限制的，这为分析短信的规律带来了一定困难。中科院心理研究所曾经对人们短信交流的主题和内容进行过调查和统计，并将人们短信交流的主题分为了 2 类：工具类短信和表达类短信[11]。工具类短信又分为协调类、询问信息、礼貌回应等子类别；表达类短信分为闲谈聊天，关心问候等子类别。

受此启发，本文采用从不同角度对短信进行分析。从客观上来看，每条短信都包含不同角度的信息，当确定到一个特定的角度上时，发现短信中出现的某特定角度的信息是有规律可循的，可以发现特定的信息出现频率非常高。

例如，对于短信“我对神许愿：愿你永远快乐。神说不行，只能 4 天。我说春天夏天秋天冬天。神愣了：2 天。我笑：黑天白天；神惊：一天！我笑：生命中的每一天!”这属于祝福类的短信。在祝福类的短信中，有关祝福“快乐、健康、幸福、好运”的内容出现频率是非常高的。

本文对搜集到的具有一定数据量的短信进行了高频信息元素的分类和总结，将总结出的信息元素放到模板中。将短信文本中出现的内容进行分类和总结，在一定程度上提高抽取信息的确定性，另外还通过多角度的信息增强抽取能力。

手机 3D 动画自动生成系统应用范畴之内的信息抽取具有它自身的一些特点：(1)对抽取出的信息整体关联程度要求较高；(2)抽取的信息元素具有一定的侧重性。

2.2　模板知识库的设计

鉴于以上分析，本文设计了基于模板知识库的信息抽取系统框架。依据短信所表达的内容将短信分为不同的主题，

基金项目：国家自然科学基金资助项目(60496322)

作者简介：吴中彪(1985－)，男，硕士研究生，主研方向：信息抽取；刘椿年，教授、博士生导师

收稿日期：2011-05-17　**E-mail：**wzbclock@126.com

在每个主题下的短信都对应一个模板的集合，抽取信息时先依据一定的方法将短信归到某一个主题范围内，然后用该主题下的所有模板对短信依次进行内容的抽取。模板所表达的内容是短信某一方面的属性。

比如说人物模板，它将短信中经常出现的人称、人物关系等以规则的形式表达出来，然后利用模板解析器进行模板和短信的匹配，匹配的模板就是该短信对应的人物相关信息。

模板的定义采用了扩展的巴克斯范式(EBNF)的上下文无关文法，存储格式为可扩展标记语言(eXtensible Markup Language, XML)文件。

下面以模板库中“人物”模板的部分内容为例说明模板定义所采用的文法和存储形式。模板文法：

人物=人物称谓|人称代词|人物名称；

人物称谓=[人称代词]，人物关系；

人称代词=第一人称代词|第二人称代词|第三人称代词；

第一人称代词=第一人称单数|第一人称复数；

第一人称单数=“我”|“自己”；

第一人称复数=“我们”|“咱”|“咱们”|“我”(“们”)?“俩”|“我们仨”|“咱”(“们”)?“俩”|“咱”(“们”)?“仨”；

第二人称代词=第二人称单数|第二人称复数；

第二人称单数=“你”|“您”|“君”；

第二人称复数=“你们”|“你”(“们”)?“俩”|“你”(“们”)?“仨”|“您二位”|“几位”|“大家”|“大伙”|“各位”|“诸位”|“同志们”|“同学们”|“孩子们”；

第三人称代词=第三人称单数男性|第三人称单数女性|第三人称复数男性|第三人称复数女性|中性单数|中性复数；

第三人称单数男性=“他”；

第三人称单数女性=“她”；

第三人称复数男性=“他们”|“他”(“们”)?“俩”|“他”(“们”)?“仨”；

第三人称复数女性=“她们”|“她”(“们”)?“俩”|“她”(“们”)?“仨”；

中性单数=“它”；

中性复数=“它们”|“别人”；

…

其对应的模板 XML 部分如下：

```
<root name="人物" flag="0" expr="人物称谓|人称代词|人物名称 "value="">
 <child name="人物称谓" flag="0" expr="人称代词?人物关系" value="">
  <child name="人称代词" flag="0" expr="第一人称代词|第二人称代词|第三人称代词" value="">
   <child name="第一人称代词" flag="0" expr="第一人称单数|第一人称复数" value="">
    <child name="第一人称单数" flag="0" expr="" value=""/>
    <child name="第一人称复数" flag="0" expr="" value=""/>
   </child>
   <child name="第二人称代词" flag="0" expr="第二人称单数|第二人称复数" value="">
    <child name="第二人称单数" flag="0" expr="" value=""/>
    <child name="第二人称复数" flag="0" expr="" value=""/>
   </child>
   <child name="第三人称代词" flag="0" expr="第三人称单数男性|第三人称单数女性|第三人称复数男性|第三人称复数女性|中性单数|中性复数" value="">
    <child name="第三人称单数男性" flag="0" expr="" value=""/>
    <child name="第三人称单数女性" flag="0" expr="" value=""/>
    <child name="第三人称复数男性" flag="0" expr="" value=""/>
    <child name="第三人称复数女性" flag="0" expr="" value=""/>
     <child name="中性单数" flag="0" expr="" value=""/>
     <child name="中性复数" flag="0" expr="" value=""/>
    </child>
  </child>
```

在 XML 文件中，叶子节点下边的终结符并没有存储到模板对应的 XML 文件中，而是存储到了数据库中，这样做的目的是为了提高效率。因为 XML 文件的存储格式本身比其他的文件格式更耗费内存，而终结符相对于非终结符来说数据量大很多，将终结符存入数据库能够降低内存消耗，提高系统运行效率。

2.3 模板解析器的框架设计

模板匹配短信的具体处理思路为：以人物模板为例，模板就是一个产生式，人物模板中的内容为一条短信中有可能出现的各种形式的与人物相关的信息。在利用模板解析器对短信和模板进行匹配时，将短信中的文本作为终结符，然后对模板的产生式进行规约，如果能够规约到开始符号，则认为是匹配成功，对应短信文本中的人物相关的信息已经提取出来；如果不能规约到开始符号，则认为是匹配失败，但是此时也可能匹配出一些信息，经过一定处理后可以作为有用信息。模板解析器的处理流程如图 1 所示。

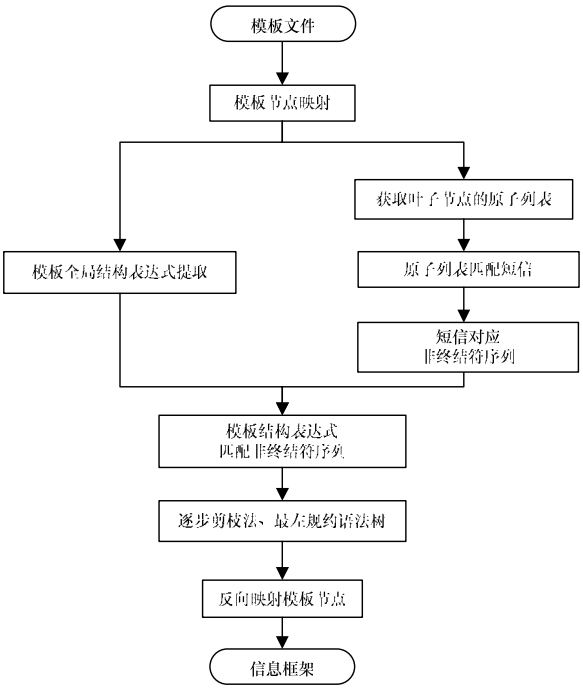

图 1 模板解析器的处理流程

3 模板解析器的核心算法

在模板解析器处理数据的过程中，对产生式的规约和记录规约路径是整个解析器的核心问题，设计到的算法的性能关系到整个信息抽取模块的效率，因此，对产生式进行规约的算法，即模板匹配的最左规约算法是模板解析器的核心算法。

模板匹配的最左规约算法描述如下：

输入 短信的非终结符序列，记为 TL(Token List)；映射后的模板，记为 MT(Mapped Template)；模板全局结构表达

式，记为 GSE(Global Structure Expression)

输出 短信的非终结符序列或者匹配语法树

(1)以 GSE 作为匹配模式与 TL 进行匹配，如果匹配失败，执行步骤(2)；如果匹配成功，则执行步骤(3)。

(2)输出 TL 作为中间结果，算法结束。

(3)将匹配成功 TL 的所有子串存储为列表 MTL(Matched Tokens List)，存储内容包括匹配位置、对应的终结符。

(4)利用 MTL 后序遍历 MT 进行无用叶子节点及非终结符节点的剪除。

(5)在 MT 的叶子节点挂载 MTL 中的终结符：

1)按照 MTL 中的顺序先序遍历 MT 进行终结符的挂载。

2)按照中序遍历 MT 的顺序找第 1 个这样的节点：节点 A，他有若干个孩子，其中存在匹配成功的节点同时也存在有匹配失败的节点，则将节点 A 的第 1 个匹配失败的节点从 A 上剪除。

3)先序遍历 MT 将已经挂载的终结符全部清空。

4)依次重复步骤 1)~步骤 3)，直至没有符合上述条件的节点 A 为止。

(6)后续遍历 MT 再次判定当前匹配结果是否能够使模板匹配短信成功。如果匹配失败，转步骤(7)；如果匹配成功，输出 MT，算法结束。

(7)输出 TL 和 MT 作为中间结果，算法结束。

4　实验结果与分析

首先利用 MUC 和 MET 的评估指标对信息抽取系统进行一个整体的评测，该信息抽取系统含有模板元素、模板关系、场景模板。

系统中的模板知识库现在有模板 22 个，分别为地点、动作、动物、服饰、关怀、花草、交通工具、经典人物角色、情绪、人物、日化、生活用品、时间、食物、数字、思念、天气、天气及抽象物、网络用语、问候内容、娱乐节目、语气词。这些模板基本涵盖了短信所涉及到的各个角度的内容，能够为短信的信息抽取提供很好的支持。

对该系统进行测试的短信库含有 965 条短信，均经过人工标注信息元素。信息元素是指短信中能够被模板匹配成相应节点的词语。据统计，短信库共有信息元素 1 250 个，用 SUM(info)表示。实验抽取出信息元素 1 143 个，用 SUM (extraction)表示。1 250 个信息元素中有 43 个未被抽取出来。经过人工筛选，发现抽取出的错误信息元素有 64 个，用 SUM(error)表示。其中，召回率计算如下：

$$\text{召回率}=\frac{\text{SUM(extraction)-SUM(error)}}{\text{SUM(info)}}=86.32\%$$

准确率计算如下：

$$\text{准确率}=\frac{\text{SUM(extraction)-SUM(error)}}{\text{SUM(extraction)}}=94.40\%$$

调和平均值计算如下：

$$\text{调和平均值}=\frac{2(\text{召回率}\times\text{准确率})}{\text{召回率}+\text{准确率}}=90.18\%$$

对短信库测试完成后，还对系统的模板知识库依据出错位置进行相应的修改，从而提高匹配的准确率。由于本系统是手机 3D 动画自动生成系统中的信息抽取模块，动画自动生成系统对本模块要求是尽最大努力的抽取出信息，因此要求召回率要高，从计算结果可以看到，在测试中的召回率达到了 86.32%，已经能够满足动画自动生成系统对信息抽取强度的要求。

5　结束语

本文提出一种面向中文短信的信息抽取方法，该方法具有良好的扩展性。从上述的分析可以看到，模板解析器与模板本身是高度分离的。只要按照模板书写文法规范书写模板，模板解析器就能够对其进行正确的解析，还可以对现有模板进行修改，使其更加符合短信中内容的形式和规律。良好的扩展性使得本文方法能够在不断扩展抽取内容范围的同时，提高抽取信息的准确率。

从应用角度来说，动画自动生成系统要求对每一条短信都能够提取一定的信息生成相应动画，而对生成的动画而言，并不要求动画所表达的内容与短信文本表达的意思完全吻合。因此，在设计信息抽取系统时采用了一些技术专门提高信息抽取的召回率，这从一定程度上降低了信息抽取的准确率。从实验可以看出，系统的召回率达到了一个较理想的状态。除了上述主观因素外，可能还受到了短信库来源比较单一、数量不足的影响。这些因素在一定程度上提高了召回率，从而使系统召回率偏离了真实水平。下一步改进的内容为：(1)由于短信库中短信的数量多和来源广，因此从更广泛的人群中收集短信，更能体现人们平时短信交流时的主要内容；(2)通过总结更多的短信扩展模板知识库，能够从更细的粒度上进行信息抽取。

参考文献

[1] Grishman R, Sundheim B. Message Understanding Conference-6: A Brief History[C]//Proceedings of COLING'96. Copenhagen, Denmark: [s. n.], 1996.

[2] Chinchor N. Overview of MUC-7/MET-2[C]//Proceedings of the 7th Message Understanding Conference. San Diego, USA: [s. n.], 1998.

[3] Chinchor N, Lewis D D, Hirschman L. Evaluating Message Understanding Systems: An Analysis of the 3rd Message Understanding Conference[J]. Computational Linguistics, 1991, 19(3): 409-449.

[4] Appelt D, Bear J, Hobbs J, et al. SRI International FASTUS System MUC-4 Evaluation Results[C]//Proceedings of the 4th Message Understanding Conference. [S. l.]: Morgan Kaufmann, 1992.

[5] Appelt D, Hobbs J, Bear J, et al. SRI Description of the JV-FASTUS System Used for MUC-5[C]//Proceedings of the 5th Message Understanding Conference. Baltimore, USA: [s. n.], 1993.

[6] Choi B D, Kim B. M/G/1 Queuing System with Fixed Feedback Policy[J]. The ANZIAM Journal, 2002, 44(2): 283-297.

[7] 李　芳, 盛焕烨, 张冬茉. 多语种投资信息抽取系统的实现[J]. 上海交通大学学报, 2004, 38(1): 21-25.

[8] 刘海鹏. 面向手机短信的命名实体识别研究[D]. 北京: 北京邮电大学, 2009.

[9] 李　蕾, 周延泉, 王菁华. 基于全信息的中文信息抽取系统及应用[J]. 北京邮电大学学报, 2005, 28(6): 48-51.

[10] 陆汝钤, 张松懋. 从故事到动画片——全过程计算机辅助动画自动生成[J]. 自动化学报, 2002, 28(3): 321-348.

[11] 蒋　丽, 马达飞. 手机短信中的沟通主题[J]. 人类工效学, 2008, 14(3): 35-37.

编辑　陆燕菲