基于类图语义框架的中文需求分析方法

利锦标, 李 童, 刘
(清华大学软件学院, 北京 100084)

摘 要: 需求文本的分析和建模是需求工程中一个重要环节, 其获取建模过程的自动化也渐渐成为了需求工程中一项重要研究内容. 本文针对中文自然语言处理和需求分析中的难点, 提出了基于面向类图语义框架的中文需求类图半自动建模方法. 该建模方法的流程包括: 文本分词与词性标注, 基于语义框架的类图模型提取, 基于问卷的模型改进和手动模型编辑. 该方法具有较高精确度, 能够显著地提高中文需求建模的效率. 本文设计实现了半自动建模系统来完整地支持整个建模方法流程. 最后, 本文通过对 3 个实际需求文本的样例进行实验, 检验了基于语义框架抽取类图元素的效果.

关键词: 语义框架; 需求分析建模; 类图; 自然语言处理; 面向对象分析
中图分类号: TP311 文献标识码: A 文章编号: 0372-2112 (2011) 3A-094-05

Chinese Requirements Analysis Based on Class Diagram Semantics

LI Jin-biao, LI Tong, LIU Lin
(*School of Software, Tsinghua University, Beijing 100084, China*)

Abstract: Requirement analysis and modeling is an important part of requirements engineering, and how to automate this course has gradually become an important research issue. Facing the difficulties in Chinese natural language processing and requirements analysis, this paper proposes a semi-automated method based on the semantic analysis. The modeling method includes the following steps: word segmentation, class diagram elicitation based on the semantic framework, model improvement with intelligent questionnaire and manual revision. The method can achieve higher accuracy, and it can obviously improve the efficiency of Chinese requirement modeling. This paper designs and implements a modeling tool to provide full support to the proposed process. Finally, we use three practical requirements specifications as show cases to evaluate the result of model elicitation with semantic framework.

Key words: semantic framework; requirement analysis and modeling; class diagram; NLP(Natural Language Processing); OOA(Object-Oriented Analysis)

1 引言

需求分析与建模是需求工程中的重要环节, 是问题描述和目标系统之间的桥梁. 尽管需求分析在传统上往往由专业人士来完成, 但随着自然语言处理(NLP)技术的产生和渐趋成熟, 需求的自动获取和建模也成为了可能. 这意味着可以结合 NLP 技术和面向对象分析(OOA)概念对需求文本进行自动处理, 形成初步的面向对象模型, 最后通过人工的调整完成模型编辑, 这样可以显著提高需求工程的效率.

尽管基于英文 NLP 的需求自动建模方法已经较为完善, 但由于中英文语法形式等方面有很大不同, 这些方法并不能直接用于中文需求工程中. 刘群[1]提到, 汉语在语法形式上有如下一些特点: 词之间没有空格分开, 即形式上不存在"词"这个单位; 汉语词形态上变化很弱, 即同一个词在句子中充当不同语法功能时, 形式是完全相同等. 此外, 汉语中存在大量歧义和多义的现象, 同一个句子在不同的语境或者场景下, 可以理解成不同的词串、词组串等, 并有着不同意义. 而需求的自动建模要求必须消除自然语言中的歧义, 得出需求工程所需要的面向对象模型. 而要消除需求文本中的歧义, 则需要大量的相关领域的知识和语义材料, 以及形成一个合理的推理模式. 这些都是工作量极大且十分困难的工作, 仍然需要更多的人进行长期、系统的研究.

2 相关工作

目前在英文需求文本的处理上已经存在若干需求分析和自动建模的方法, 并且基于这些方法实现了相应的自动建模工具. Mich[2]在 1996 年开发的基于一个 NLP 系统的自动生成对象图的工具 NL-OOPS. Harmain 和

收稿日期: 2010-10-29; 修回日期: 2011-01-20
基金项目: 国家自然科学基金面上项目(No. 60873064); 国家自然科学基金"可信软件"重大专项重点项目(No. 90818026); 国家 973 重点基础研究发展规划(No. 2009CB320706); 核高基国家科技重大专项(No. 2009ZX01045-001-001-02)

Gaizauskas[3]在 2000 年开发的 CM-Builder 系统, 实现了从自然语言文本中直接抽取初始类图的功能. Börstler[4]在 2001 年构建了一个名为 RECORD 的系统, 该系统能够通过用例描述中特定的关键字, 抽取对象图及自动生成简单类图. Overmyer 和 Rambow[5] 在 2001 年开发了名为 LIDA 的系统, 其主要功能为通过自然语言需求描述生成对象图以及类图, 此外, LIDA 还提供了一系列模型转换器, 能与多个 CASE 工具兼容. 总体来说, 基于英语 NLP 的需求自动建模工具已经发展的较为成熟, 并在实际应用中有一定成效.

国内目前还没有较为成熟的中文需求自动建模工具, 纪磊[6]在 2008 年提出了 CREAT“知文”半自动建模系统. 该系统是一个领域独立的基于 NLP 和 OOA 的 CASE 工具. 其主要功能是对自然语言进行中文分词, 然后通过简单的句型进行匹配, 抽取并输出参考类以及对应候选元素(关联、属性、操作等). 需求分析师对系统的输出进行判断选择, 通过人工修改, 得到 UML (Unified Modeling Language) 类图. 但由于 CREAT 系统采用的句式较为简单, 只能识别出参考类, 而不能直接识别类, 这也限定了其使用对象是需求分析师. 而在中文文本信息抽取方面, 国内已经存在较成熟的研究成果. 林亚平[7]等提出了基于最大熵模型, 利用上下文特征等信息改进隐马尔可夫模型的方法, 获得良好的抽取效果; 周顺先[8]等提出了基于二阶隐马尔可夫模型的文本抽取方法, 提高了对错误信息的识别能力. 这些方法在文本信息抽取的精确度和召回率方面都能体现出良好的性能.

3　基于语义框架的中文需求类图建模方法

为了能够准确高效地对中文需求文本进行分析, 并生成对应的类图模型, 本文提出了基于语义框架的中文需求类图建模方法. 如图 1 所示, 本方法包含 4 个主要的流程: 文本分词与词性标注, 基于面向类图的语义框架生成初步模型, 基于问卷的模型改进和手动模型编辑. 本文使用 Flex + J2EE 设计并实现了该建模系统, 对建模方法中的各个步骤进行了支持, 实现了中文需求的半自动获取和建模. 下面本文将针对建模方法中的每一个步骤进行详细地介绍.

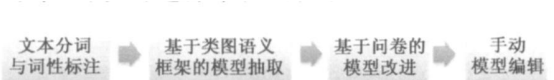

图1　基于语义框架的中文需求类图建模方法流程

3.1　文本分词与词性标注

获取需求文本后, 首先需要对文本内容进行基本的NLP处理——分词和词性标注. 本文使用 ICTCLAS[9] 分词系统对需求文本进行分词操作.

词性的识别和提取效果将会直接影响到语义模型的分析, 为了得到更加详细的词性标注结果, 我们采用了中科院计算所的二级词性标注集. 此外, 根据需求文本中语言的特性, 我们对 ICTCLAS 现有标注集进行了的定制. 具体的修改包括以下两个方面:

(1) 修改部分副词、代词等词性标注为数词, 如“任意 mm”、“多 mm”、“每个 ml”、“每 ml”等, 其中第一个“m”为二级标注中数词标记符, 第二个字符中“m”表示“more”、“l”表示“less”;

(2) 添加标记符 g, 增加自定义词性, 用于表示“至少”、“至多”一类形容数词的词, 如表 1 所示.

表1　自定义词性 g

词性	词语
gm	至少　多于　大于　最少　不少于
gl	至多　少于　小于　最多　不多于
…	…

上述两种对词库的修改是针对类图中的关联多样性而设定的, 其目的是便于匹配类图多样性的语义框架. 例如, 如句子“每个教师至少担任两门课程”, 在经过分词和词性标注处理后会得到: “每个/ml 教师/n 至少/gm 担任/v 两/m 门/q 课程/n”.

在完成分词和词性标注后, 我们根据得到的结果将需求文本抽取为词、句、段、结构这四个列表. 其中结构列表包含需求文本中每个句子对应的词性标注模型, 这样就完成了对非结构化的文本的结构化处理.

3.2　面向类图的语义框架

基于 NLP 领域对于浅层语义处理的相关研究成果, 本文针对需求文本和类图模型的特点, 提出了面向类图的语义框架, 用于自动提取初步结构化文本中类图模型信息.

3.2.1　语义框架结构

面向类图的语义框架是包含词语层、句式层和语义层的三层架构, 如图 2 所示. 其中, 词语层包含文本经过分词和词性标注后的句子信息; 句式层分析句子成分的相关信息; 语义层主要描述句子所包含的语义信息. 对于特定的语句, 从不同视角进行分析会得到不同的语义信息, 本文提出的语义框架的语义层模型是面

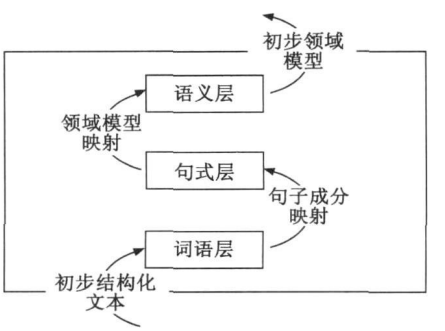

图2　面向类图的语义框架结构

向类图的语义模型. 利用该语义框架, 通过对初步结构化的文本进行两次映射即可获取初步的类图模型.

3.2.2 语义框架模型

基于三层的语义框架结构, 针对需求文本中常出现的句式, 本文设计了 6 个具体的面向类图的语义框架模型. 其中包括 4 个普通模型(复杂模型还包括子模型) 和 2 个特殊模型. 下面将对这些模型进行详细地介绍.

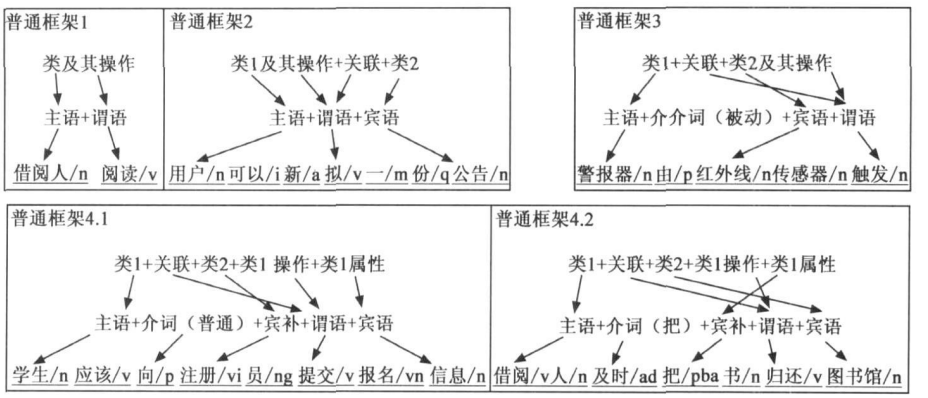

图3 普通语义框架模型

根据句子中的特征词, 中文句子一般可以分为把字句, 被字句, 比字句, 对字句, 是字句和给字句. 而在对需求文档进行深入分析后, 我们发现在实际需求文档中, 把字句和被字句常被用到. 而其他的句式则几乎不会用到, 即使偶尔用到也往往都是表达与类图模型无关的信息. 因此, 本文中所设计的语义模型会针对“把”和“被”等特殊介词设计特定的语义模型.

针对上述特点, 按照句式的复杂程度, 本文设计了 4 种普通语义框架模型, 如图 3 所示.

除上述四个基本句式外, 本文还针对类图模型元素的特殊性, 设计了匹配多样性和匹配“的”字的语义模型, 如图 4 示. 特殊语义框架模型的匹配基于普通框架模型的已识别类, 旨在发掘更为细致的类图模型信息.

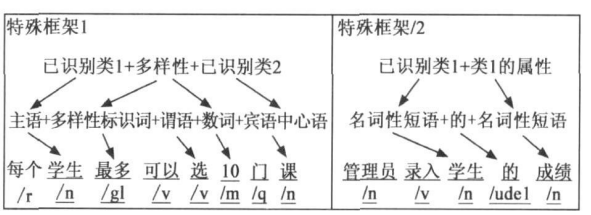

图4 特殊语义框架模型

3.2.3 模型的映射匹配策略

如图2 所示, 在利用语义框架进行语义模型分析匹配的过程中, 主要涉及到两层映射关系. 保证这两层关系的准确映射是实现语义模型分析匹配的关键. 其中, 中文自然语言固有的复杂性对词语层与句式层映射有很大影响. 本节中将主要介绍词语层与句式层映射中的问题, 以及本文针对这些问题的具体解决方案.

(1) 领域词汇识别的问题

在对于中文自然语言进行分词处理时, 一个完整的领域词汇往往会被切分为若干个不同的部分. 例如, 对“档案管理员”进行基本分词处理得到的结果是“档案/n 管理员/n”. 这个问题会导致在语义框架模型的映射中, 无法准确地将词语映射为适当的句子成分. 解决这个问题的方法是, 在对文本进行分词前将其领域专业词汇添加至分词系统中. 本文针对需求文本的特点, 提出以出现频率为主要特征, 结合领域词汇的组成规律来提取领域词汇, 所有满足组词规律并达到出现频率阀值的词语都将被抽取为领域词汇.

(2) 复杂句式匹配问题

在句式层进行句子成分分析时, 只关注句子的主干, 然而实际需求文档中的句子结构往往比较复杂. 例如“学生应该向注册员及时提交编辑好的报名信息”中需要被分析和映射的内容只是加粗的文本, 而如“应该”、“及时”等修饰语都是可有可无的. 如果使用严格的句式匹配方法, 往往会遗漏掉重要的信息, 因此本文中采用了模糊匹配的策略. 模糊匹配是指根据需求文本句式的特点, 在基本句式的适当位置添加若干任意词汇的匹配项. 以语义模板 1 中的句式层为例, 用 X 表示任意词汇, 经过对大量需求文本的人工分析总结, 本文在系统实现中将该句式的匹配改进为“主语 + X{0- 2} + 谓语+ X{0- 3} + 宾语”, 其中 X 可以表示情态动词、状语、定语等.

3.2.4 模型匹配流程

本文提出的基于类图语义框架的模型匹配抽取的具体流程如图 5 所示.

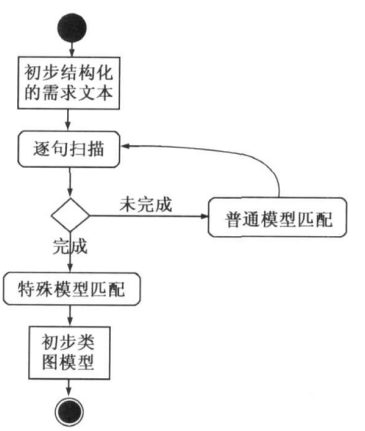

图5 基于类图语义框架模型的自动建模过程

在上述的模型提取流程中, 普通模型匹配是按照“先繁后简”的顺序进行判定的. 本文所设计的 4 个普通语义框架模型中, 复杂的模型会包含简单的模型. 因此, 为了尽可能抽取所有的语义信息, 在进行模型匹配时会首先尽可能匹配复杂的模型, 如不成功再尝试匹配简单的模型.

3.3 智能问卷处理完善建模元素

由于中文自然语言的多样性和复杂性, 现有 NLP 技术并不能完全处理语义, 因此需求文本在经过面向类图的语义框架模型的自动分析和抽取后, 很难得到完整准确的类图. 为了完善类图中建模元素, 需要在建模过程中添加人工的处理操作, 本文提出并实现了基于智能问卷的模型求精方法来解决这个问题. 智能问卷是基于 NLP 和类图模型而构造的, 问题主要针对文本中无法自动识别的语法元素提出, 问题的题干和可能的选项都是自动生成的. 通过智能问卷的作答, 问卷能够获取建模元素碎片并将其整合到初始模型当中. 因此, 用户不需要需求工程方面的专业知识也能顺利完成类图的建模. 智能问卷主要处理代词指代和模型细化两类问题.

3.3.1 代词指代处理

汉语中代词的指代具有不确定性, 很难单纯从语法结构上确定代词所指代的名词. 代词在需求文本中的出现频率比较高, 且其指代对象往往对应类图中的类元素, 因此代词处理是否准确会影响到类图信息的完整.

本文利用问卷来处理代词指代的问题. 智能问卷系统会针对存在人称代词的句子进行提问, 并通过分析该句子及其前后一个句子, 将可能的名词或名词性短语抽取出来作为问题的选项. 在用户选定正确的指代关系后, 系统将对其所在句子重新进行语义模型的分析匹配, 获取该句中所包含的模型碎片.

3.3.2 模型细化处理

通过上述自动建模过程, 可以从需求文本中抽取大部分类图信息, 但由于句式的复杂和文档本身可能存在语义模糊等原因, 抽取结果会遗漏小部分信息. 因此, 本文设计通过智能问卷实现进一步细化模型.

智能问卷系统会主要从以下两个方面自动地分析现有类图: 类的属性和操作是否很少或者为空, 是否存在没有任何关联的孤立类. 系统根据分析评估的情况, 将自动生成问题. 问题是针对删除冗余类、完善所需类的信息等方面自动生成. 问卷系统通过分析类词语所在的句子, 抽取其中对应的语法元素(名词、动词) 等作为类的候选属性或者操作, 交由用户选择. 如此便可以通过问卷系统, 用户可以很方便的完善类图信息, 提高建模效率.

经过上述三个步骤处理后可以得到一个较为完整的 UML 类图. 接下来, 该类图模型的信息将导入系统的模型编辑模块, 用户可以图形化地浏览类图模型, 进行进一步的编辑和修改.

4 实验及分析

为了测试本文提出的基于语义框架模型的类图建模方法的正确性, 本文对 3 段中文需求文本(文本规模约 1000 字) 分别进行了建模实验, 并自动建模结果与人工参与的问卷完善后的结果进行比较, 其实验结果如表 2.

从表 2 中可以看到, 利用语义框架模型抽取类图元素的召回率是比较令人满意的: 类图四项元素的平均召回率分别 75.8%. 这说明当前语义框架模型能够覆盖需求文档中的绝大部分句式结构. 较之于召回率, 抽取建模元素的准确率并不是很高: 四项元素平均准确率约 71.9%. 这表明当前语义框架模型三层之间的映射关系还不够准确, 需要进一步提高精度. 而随着需求文本规模的扩大, 复杂句式出现的频率也逐渐增加, 当前语义框架匹配失败率也在提高, 说明当前语义框架的语义规则还需要进一步完善.

表 2　自动建模实验结果

模型元素 / 需求文档	类		关联		属性		操作	
	准确率	召回率	准确率	召回率	准确率	召回率	准确率	召回率
1	61.5%	72.7%	46.7%	63.6%	65.9%	73%	66.7%	81.8%
2	81%	70.8%	63%	81%	83%	76.5%	74.1%	90.9%
3	76.5%	72.2%	78.3%	81.8%	75.7%	68.3%	90.5%	76.2%
平均值	73.00%	71.90%	62.67%	75.47%	74.87%	72.60%	77.10%	82.97%

5 结论和后续工作

本文针对中文自然语言需求文本自动建模中存在的语义分析难度大、建模元素识别率低等问题, 提出了基于面向类图语义框架的中文需求类图半自动建模方法.

本文系统地设计了 6 种面向类图的语义框架模型, 用于自动地抽取需求文本中的类图元素, 该框架表示了从半结构化的需求文档到类图模型的三层结构和两层映射关系. 针对中文自然语言的特殊性, 本文在自动

建模的基础上设计并添加了智能问卷模块, 用户只需要简单作答问卷即可丰富和完善模型. 本文设计实现的建模系统, 完整地支持了上述各步骤, 可以帮助不具备需求领域知识的用户, 高效准确地生成类图模型. 最后本文对 3 篇需求文本样例进行了实验验证了建模方法的正确性和实用性. 结果显示, 本方法在模型抽取中能够识别出绝大部分主要的类图元素, 平均召回率约 76%. 模型抽取的平均准确率约 72%, 且随着文本规模的扩大, 准确率也有所提高. 但复杂句式匹配的失败率较高, 说明当前的语义规则还不够完善.

未来工作围绕以下几点展开: 首先, 将进一步改进和扩充现有的类图语义框架模型, 从而更好地完对文本分析和匹配. 其次, 引入数据挖掘和机器学习方法, 结合现有类图语义框架匹配规则, 实现模型提取, 提高系统自动化程度. 最后, 通过实际应用建模系统后得到的用户反馈, 提高系统易用性和实用价值.

参考文献

[1] 刘群. 汉语词法分析和句法分析技术综述[R]. 北京: 中国科学院计算技术研究所, 2002.

[2] P L Mich. NL-OOPS: from natural language to object oriented requirements using the natural language processing system LOLITA[J]. Natural Language Engineering, 1996, 2(2): 161-187.

[3] H M Harmain, R Gaizauskas. CM-Builder: An automated NL-based CASE tool[J]. Automated Software Engineering, 2003, 2(2): 157-181.

[4] J Börstler. User-centered requirements engineering in RECORD - an overview[A]. Proceedings NWPER'96[C]. Aalborg, Denmark: Programming Environment Research, 1996. 149-156.

[5] Scott P Overmyer, P Benoit Lavoie, P Owen Rambow. Conceptual modeling through linguistic analysis using LIDA[A]. Proceedings of the 23rd International Conference on Software Engineering[C]. NW Washington: IEEE Computer Society, 2001. 401-410.

[6] 纪磊, 刘 ."知文" ——基于自然语言的需求分析和建模方法[J]. 计算机科学(专刊), 2008, 35(11): 48-52.
Lei Ji, Lin Liu. NLP based requirements modeling method[J]. Computer Science(special issue), 2008, 35(11): 48-52. (in Chinese)

[7] 林亚平, 刘云中, 等. 基于最大熵的隐马尔可夫模型文本信息抽取[J]. 电子学报, 2005, 32(02): 236-240.
Lin Ya-ping, Liu Yun-zhong, et al. Using Hidden Markov Model for text information extraction based on maximum entropy [J]. Acta Electronica Sinica, 2005, 32(02): 236-240. (in Chinese)

[8] 周顺先, 林亚平, 等. 基于二阶隐马尔可夫模型的文本信息抽取[J]. 电子学报, 2007, 35(11): 2226-2231.
Zhou Shun-xian, Lin Ya-ping, et al. Text information extraction based on the second-order hidden Markov model[J]. Acta Electronica Sinica, 2007, 35(11): 2226-2231. (in Chinese)

[9] Hua-ping Zhang, Hong-kui Yu, et al. HHMM-based Chinese lexical analyzer ICTCLAS[A]. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing[C]. Morristown: Association for Computational Linguistics, 2003. 184-187.

作者简介

利锦标 男, 1988 年 2 月出生于广东省河源市. 现为清华大学软件学院研究生, 主要研究方向为需求工程.
E-mail: linjy10212@gmail.com

李 童 男, 1986 年 7 月出生于北京市, 现为清华大学软件学院研究生. 主要研究方向是需求工程、安全需求分析.
E-mail: litong08@mails.tsinghua.edu.cn