

基于语义句模的语义理解方法研究

马玉慧, 谭 凯, 尚晓晶

(渤海大学 教育技术研究所, 辽宁 锦州 121000)

摘 要:针对目前情境信息丰富问题语义理解的困难,提出了一种基于语义句模的语义理解方法。该方法借鉴汉语言学中句模的研究成果,以及问题的分类,构建语义句模,将语义信息蕴含在句模中,利用文本中能够穷举的信息,为千变万化的情境信息赋予相应的语义。实验构建了基于自然语言处理工具 Gate 的语义句模 jape 规则库,以收集的 102 道四则运算应用题为例进行语义理解,实验结果为完全理解的题目为 82.4%,部分理解为 17.6%,完全不理解的 0%。得出的实验结论是,该方法能够较好地实现情境信息丰富问题的语义理解。

关键词:语义理解;句模;知识表示;自动求解

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2012)10-0117-04

Research on Method of Semantic Comprehension Based on Semantic Sentence Template

MA Yu-hui, TAN Kai, SHANG Xiao-jing

(Institute of Technological Education, Bohai University, Jinzhou 121000, China)

Abstract: According to the difficulty of semantic comprehension for the problems with rich contextual information, present a method of semantic comprehension based on semantic sentence template. This method is to construct the semantic sentence templates based on the research of sentence template in the field of Chinese linguistic and the classifications of problems. The semantic is contained in the semantic sentence templates. The semantics of the contextual information can be recognized by the exhaustive information. Experiment constructs jape rule library of semantic sentence template based on natural language processing tool Gate, 102 word problems are collected as examples for semantic comprehension, and the results demonstrate that the full comprehension topic accounts for 82.4%, part comprehension 17.6%, without not comprehension. It is concluded that this method is effective for semantic comprehension of the problems with rich contextual information.

Key words: semantic comprehension; sentence template; knowledge representation; automated solving

0 引言

目前在计算机自动求解的研究领域中,大多数学者主要关注的是基于自然语言的代数问题的自动求解、几何作图,以及几何定理的自动证明等,并已取得了不少研究成果^[1~4]。相对而言,应用题等丰富情境信息问题的自动求解研究发展却很缓慢^[5]。主要原因可归结为以下两点:

(1)相对于几何命题等文本中词汇量少、语义范围窄的特点而言,丰富的情境信息,增加了语义理解的难度;

(2)理解应用题等问题,不仅仅包括通常的浅层语义分析,以及不同符号之间的转换,更关键的是识别

题目中各个数据对象显式或隐式的数量关系等深层次的语义信息,为自动求解提供基础。

随着自然语言理解技术的发展,应用题的自动求解研究也取得了一定的进展。早期的 ARITHPRO 系统,通过构建词语的分类词典,以及相应激活问题框架的产生式规则,将用户输入的谓词逻辑形式的文本转换成相应的问题框架,实现了最简单的一步加减算术应用题的语义理解^[6]。该系统模拟人问题求解的思路——先将文本转换成认知表征模型,再基于表征模型进行问题解决,为后来应用题自动求解的研究奠定了很好的基础。近期比较成熟的应用题自动求解系统主要有 LIM-G 系统以及整数一步和部分两步应用题自动解答系统。LIM-G 系统(2007 年)通过构建几何应用题本体,抽取出应用题文本中的几何概念来确定问题类型,实现了一步几何应用题的语义理解^[7]。我国有研究者(2008 年)则根据不同题型的结构特征,利用知网,构建了应用题关键词串,将输入的应用题文本

收稿日期:2012-02-17;修回日期:2012-05-20

基金项目:辽宁省教育科研计划项目(w2010042)

作者简介:马玉慧(1974-),女,博士,研究方向为人工智能及其教育应用。

与关键词串进行匹配,实现了整数一步、部分二步应用题的语义理解^[8]。这两个系统采用的语义理解思路都是:通过识别某些词语或词语组合的语义完成整个题目的语义理解。这种方法的局限在于以某些词语的语义代替整个题目的语义,使得题目的求解只能是直接与特定题型匹配,而不能实现类似人解题时的推理过程,最终导致能够自动求解的题目受题型限制和解题步数限制。此外,由于不同系统针对的问题类型不同,因而对什么是应用题的语义理解,或者说如何界定系统是否理解了应用题有不同的解释。正因为没能在语义层面上达成共识,所以导致针对一种题型提出的语义理解方法不适用于其他类型。

本研究首先以认知心理学中问题解决的相关研究成果为理论基础,以应用题的语义理解为研究对象,对应用题的语义理解进行了界定。然后利用汉语言学关于语义平面中句模的研究成果,提出了基于语义句模的语义理解方法。该方法可以使得自动求解系统能够在此语义理解的基础上进一步实现不同类型的、多步应用题的自动求解。

1 应用题的语义

语义理解,也就是所谓的“知道了一个词或一个句子的意思”,包含了两层含义:一是能够实现不同符号之间的变换;二是能够进行推理^[9]。由此可知,应用题的语义理解是实现由自然语言表述的应用题到某种符号表示的转换,应用题的自动求解则要基于这些转换的符号,运用数学知识进行推理求解。

认知心理学对这些转换的符号进行了深入的研究。Kintsch 和 Greeno 从认知心理学的研究视角提出了小学算术一步加减应用题的问题表征模型——问题框架。问题框架包含四个槽:分别为对象槽、数量槽、说明槽和角色槽。对象槽的槽值是指该句子中所描述的数据对象。数量槽是数据集合的具体数据信息,若此数据未知,则槽值为“一些”或者“多少”。说明槽的槽值用于区别其他数据对象的信息,例如该数据对象的所属、地点、时间等信息。角色槽用于说明该数据对象在某类型题中的数量关系^[10]。虽然 Kintsch 和 Greeno 的研究仅针对一步加减应用题,但他们提出的问题框架则帮助人们理清了什么是应用题的理解:即由自然语言到问题框架的转换。

由此可知,应用题的语义理解即为由自然语言到问题框架的转换,也就是填充问题框架中各个槽槽值的过程。但是毕竟 Kintsch 和 Greeno 的问题框架是针对一步加减应用题的,无法承载多步应用题的语义信息。为此,有研究者以他们的研究为基础,从计算机自动求解的视角,提出了能够表示多步应用题语义信息

的问题框架构建方法^[11]。

本研究以多步四则运算应用题的语义理解为例(四则运算应用题界定为能够分解为多个一步算术应用题的应用题),说明如何将自然语言文本转换为相应的问题框架。

具体的多步四则运算应用题的问题框架如图 1 所示。

槽	值
框架名	
对象	<名词>、<动词>
数量	<数值>、一些、多少、几
数量单位	<名词>
所属	<名词>
角色	起始量、结果量、移入、移出、大有作为要、小集、比较多、比较少、相等、子集、总集、单位量、份数、总量、1 倍数、倍数、几倍数
关系框架	<框架名>

图 1 多步四则运算应用题的问题框架

2 语义句模

2.1 汉语言学中句模的研究

汉语言的研究者提出,句模是根据句子语义平面的特征分类处理的句子类别^[12]。句模研究首先是将句子的核心部分——动词进行了分类,然后再依据动核结构的必需成分分析出语义角色,并按照各个语义角色在句中可能出现的先后顺序进行排列,构成句模。

句模的研究对实现语义理解有着积极的实用价值。例如:如果知道“告诉”这个词造句时必须支配三个从属成分:一个是施事,即发出告诉动作的人;一个是当事,即告诉的对象;一个是受事,即告诉的具体内容,而且知道其排列顺序为“施事——动核——当事——受事”,只要识别出动词,计算机就较容易分析出词语间的语义关系。

2.2 构建应用题的语义句模

应用题表述的是现实世界中对象的数量以及对象间的数量关系,因此除了动词起着主要的核心作用外,数词、量词以及表示数学信息的词语(例如:“倍”,“一共”等)也起着重要的作用。参照句模的构成方式,以上述词为核心,按照词语在应用题中的语义成分:所属成分、对象成分以及它们在句子中可能出现的位置,构建应用题的语义句模,再将应用题的句子与句模进行匹配,即可实现应用题情境信息的识别,并可在此基础上进一步实现由自然语言到问题框架的转换。

2.2.1 语义句模的基本构成要素

应用题语义句模(以下简称语义句模)是指由问题框架中的所属成分、对象成分、动词、数量、数量单位以及应用题中表示数学信息的关键词(以下简称关键

词),按照在句中可能的位置组合而成的句子模板。构建语义句模的原则是将能够穷尽的词语尽量一一列出,避免计算机自动分析。

遵循这样的原则,将语义句模的基本元素界定为:<动词>、<数量>和<数量单位>、<所属>、<对象>和关键词。

(1)动词。在日常语言中,动词是句子的核心,制约着句子中其他词语的使用及其出现的位置。在应用题中,动词的这些功能依然没有变化,因此动词仍然是构成句模的核心。

(2)数量和数量单位。数量表示了数据对象的数值,是进行数学运算的基本单位。因此,数量是构成句模的基本元素之一。数量单位即量词,是表示事物或动作的数量单位的词。

(3)所属和对象。在日常语言中,动核、施事、受事、成事、共事等语义角色构成了句模的基本元素。但在应用题领域,一个命题集合内词语间的语义关系不再是施事、受事、成事等,而是问题框架中的所属和对象。

(4)关键词。在应用题句子中,有些词语是用于表达特定的数学关系的,例如“还剩下多少只”中的“还剩下”。这些词语都具有固定的数学含义。本研究将这些词语称为关键词。获得关键词的方法是进行词频统计。本研究收集了 1083 道四则运算应用题,将这些应用题加载到 ROST 系统中,进行分词和词频统计,规定出现次数大于 8 的即为关键词,共得到的关键词 65 个。

2.2.2 应用题的语义句模

本研究依据应用题句子中是否包含数据,以及“多少”、“一些”,将句子分为情境句和非情境句。例如“同学们参加跳绳比赛,男同学有 8 人,女同学的人数同男同学一样多,一共有多少人参加跳绳比赛”,该题中的“同学们参加跳绳比赛”为情境句。其他的为非情境句。对于非情境句,若不包含关键词,则称为赋值句,否则为关系句。依据所包含的关键词及其所对应的题型,将关系句进一步分为子集句、总量句、比较句、单位量句、等分句、起始量句、转移句、结果量句。本研究共收集了 1083 道小学四则运算应用题,以这些应用题的句子为单位,依据句子的分类,按照构建语义句模的方法,共构建了 79 个 10 类应用题语义句模:赋值句模、子集句模、总量句模、比较句模、单位量句模、等分句模、起始量句模、转移句模、结果量句模、情境句句模。这些句模的分类与命题集合的分类一一对应。在句模中,“|”表示或者,“[]”表示可选,“其他”表示该部分文本不是解题必须的信息。其中部分子集句模及其例句如表 1 所示。

表 1 部分子集句模及其例句

序号	句模	例句
1	其中~[动词]~数量~数量单位~对象	其中有 4 个足球是新买的
2	其中~对象~[动词]~数量~数量单位	其中苹果有 9 个
3	其中~[动词]~数量~数量单位~对象	其中 8 头大牛
4	还~动词~数量~数量单位~对象	还种了 4 行果树
5	其中~所属~和~所属~一共 共~ [动词]~数量~数量单位	其中小华和小红共做 12 朵

3 基于语义句模的语义理解方法

如上所述,从计算机理解的角度,四则运算应用题的语义理解是将应用题文本转换为其所对应的问题框架,也就是对象、数量、数量单位、所属、角色、关系框架六个槽的填充。计算机能很容易识别出数据,通过分词和词性标注能准确地识别数量单位(即量词),因此语义理解主要是指所属、对象、角色和关系框架槽的填充。

本研究提出了基于语义句模的语义理解方法,其核心思想是将每个经过预处理(分词、词性标注等)的句子与语义句模相匹配,由此填充对象槽、数量槽、数量单位槽,以及所属槽。再依据所匹配的句模类型、关键词、上下文,以及常识知识库填充角色槽和关系框架槽。

3.1 语义句模的知识表示

本研究采用产生式规则对语义句模进行表示。以“草地上一共有山羊和绵羊 100 只”为例,其所对应的句模为“[所属]~一共~有~对象~和~对象~数量~数量单位”,该句模所对应的产生式规则的形式化描述如下所示:

Rule 总量句

IF

Tokens1 不等于“一共”

Token 等于“一共”

Token 等于“动词”

Tokens2 不等于“和”

Token 等于“和”

Tokens3 的数据类型不是数值的字符串

Token 的数据类型为数值

Token4 等于“数量单位”

Then

将 Tokens1 标识为“所属”

将 Tokens2 标识为“对象 1”

将 Tokens3 标识为“对象 2”

将 Token4 标识为“数量单位”

若某个句子与该句模匹配成功,即能够激活规则的前件,那么其对象、数量、数量单位、所属就得到了识别。将所有的句模转换为相应的产生式规则,即可生

成语义句模规则库。

3.2 对象槽和所属槽的填充

问题框架中的所属槽和对象槽的填充是指从若干的情境信息中识别出表示所描述的数据对象的信息,以及表示该数据对象的所属信息,并将其填充到问题框架的所属槽和对象槽中。所属和对象的识别通过与语义句模的匹配实现。例如上例中,与句模匹配成功后,将 Tokens1 标识为“所属”,即“一共”之前的字符串“草地上”为所属槽的槽值;将动词与字符“和”之间的 Tokens2 标识为“对象 1”,字符“和”与数据之间的 Tokens3 标识为“对象 2”,将 Tokens2 与 Tokens3 用“&&”符号连接后填充对象槽。

3.3 角色槽和关系框架槽的填充

依据角色槽值识别的过程不同,将其分为两类:第一类是能够通过分析该句子的信息即可识别的角色,这类角色包括比较多、比较少、相等、倍数、部分的总集、起始量、结果量、份数、子集;第二类是仅依靠单个句内的信息不能确定其角色,还需要考虑上下文情境,要依靠多个句子的信息才能识别。这类角色包括子集、大集、小集、移入、移出、单位量、份数、总量、1 倍数、几倍数和部分的起始量、结果量、总集。下面分别对这两类角色的识别进行阐述。

(1) 通过句模和关键词识别角色。

应用题中的部分数学关系通过一个命题集合内的信息即可识别。识别的方法是判断该命题集合匹配的句模类型和出现的关键词。例如:“苹果比梨多 3 个”匹配的句模类型是比较句模,该命题集合中的关键词是“多”,因此其角色槽值为“比较多”;“绵羊的数量是山羊的 5 倍”表示的是“倍数”关系。

(2) 通过上下文识别角色。

有些角色不能仅依靠单个的命题集合中的信息进行识别,对于这种情况需要进一步分析上下文情境来确定。例如下列情况:

①一个句子承担多个角色。例如“小明有 10 本故事书,比小红少 2 本,是小亮故事书数的 2 倍,他们一共有多少本故事书”,题目中“小明有 10 本故事书”相对于“比小红少 2 本”的角色是“小集”,相对于“是小亮故事书数的 2 倍”的角色是“几倍数”,相对于“他们一共有多少本故事书”的角色是“子集”。

②需要常识知识识别部总关系和转移关系。例如“合唱队有 65 名同学,有 20 名男同学,男同学比女同学少多少名”,该题目中男同学和女同学都是合唱队的组成部分。只有知道了“合唱队”、“男同学”、“女同学”之间的上下位概念关系,才能够成功解题。

对于第一种情况,首先要判断是否存在一个问题框架承担多个角色的情况。判断的方法是若一个问题

框架的所属槽或对象槽槽值是两个或两个以上其他问题框架的所属槽或对象槽的一部分(例如“小明”是“小明 && 小红”的一部分),则该框架具有多个角色。解决的方法是对该问题框架进行复制,若有 3 个这样的问题框架,则复制 2 个,然后逐个进行角色和关系框架槽的填充。对于第二种情况,需要构建常识知识库进行解决。

4 结束语

进行语义分析前,应先对应用题文本进行预处理。预处理的主要功能是对用户提交的待解应用题文本进行处理,为语义分析中的所属和对象识别、角色和关系框架识别提供符合其相应输入要求的文本。预处理主要包括:分词、词性标注、动词替换为“动词”,量词替换为“量词”,删除助词等。

本研究收集了 102 道小学四则运算应用题,句子个数共 363 个。实验中,采用中科院的 ICTCLAS 实现分词和词性标注,自然语言处理工具 GATE^[13]实现语义句模的匹配。为验证该方法的有效性,进行预处理后重新进行了人工校验。在此基础上得出的实验结论是:问题框架中对对象成分识别的正确率为 88.6%,召回率为 87%,所属成分识别的正确率为 91.2%,召回率为 84.8%。若将语义理解的程度分为三个层次:完全理解、部分理解和不理解。其中完全理解是指问题框架中的各个槽都得到了正确的填充;部分理解是指部分槽填充的不正确;完全不理解是指这些槽都未能填充或者填充的都不正确。则实验样本中完全理解的题目为 82.4%,部分理解为 17.6%,完全不理解的 0%。

本研究提出了一种基于语义句模的语义理解方法。该方法借鉴汉语言学中句模的研究成果,以动词、数据、量词,以及关键词为核心,依据应用题表征模型,构建了应用题的语义句模。将文本与句模相匹配,利用可以穷举的信息去识别千变万化的情境信息,实现了应用题这种情境信息丰富问题的自动语义理解,并取得了较好的实验结果。该方法不仅适用于应用题的语义理解,只要能够找出问题情境中的所有语义,也适用于其他领域情境信息丰富问题的语义理解。

参考文献:

- [1] Kook H J. Geometry Tutoring Supported by an Intelligent Drawing Interface and Automatic Problem Solving[J]. International Journal of Computational Intelligence Systems, 2010, 3(1): 21-27.
- [2] Wong Wing-Kwong, Huang Chunwei, Yin Shengkai, et al. A

(下转第 124 页)

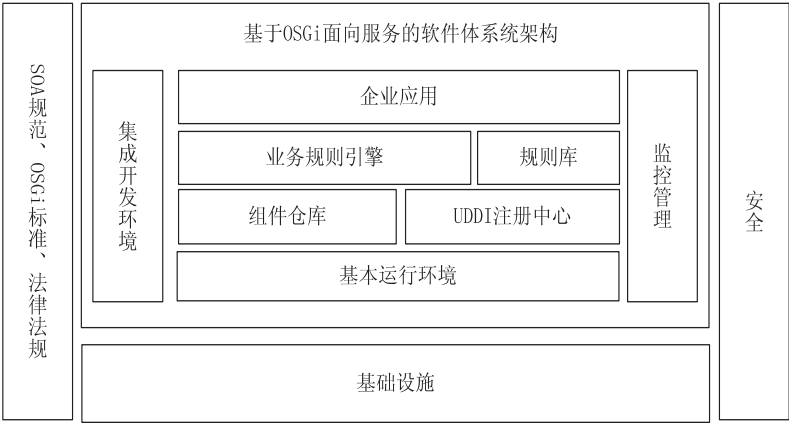


图 4 统一服务架构平台

服务架构平台。通过使用统一服务架构平台,可将软件产品或系统设计成可插拔、可动态改变行为、稳定、高效、规范、可扩展的模块,以达到最大程度的复用,提高开发、测试效率。目前,西安未来国际信息股份有限公司已经逐渐形成了统一的基础架构、开发方式、部署方式,同时也使公司的组件仓库丰富起来。今后统一服务架构平台还可以通过云平台进行统一管理,统一部署入口,使云计算真正落地,通过 SOA 架构提供源源不断的服务供应。

参考文献:

[1] Hu J, Khalil I, Han S, et al. Seamless integration of dependability and security conception SOA: A feedback control system based framework and taxonomy [J]. Journal of Network and Computer Applications, 2011, 34(4): 1150-1159.

[2] Zmuda D, Psiuk M, Zielinski K. Dynamic monitoring framework for the SOA execution environment [J]. Procedia Com-

puter Science, 2010(1): 125-133.

[3] 汪林林, 张 春, 刘 歆, 等. SOA 全生命周期建模方法综述 [J]. 计算机应用研究, 2011, 28(1): 37-41.

[4] 李伟平, 褚伟杰, 杨维强, 等. 一种支持 SOA 系统开发的服务注册库 [J]. 小型微型计算机系统, 2011, 32(1): 34-38.

[5] Zhang Jingjun, Wang Lei, Li Hui, et al. Research Java Web framework based on OSGi [J]. Procedia Engineering, 2011, 15: 2374-2378.

[6] de Vergara J E L, Villagra V A, Fadon C, et al. An autonomic approach to offer services in OSGi-based home gateways [J]. Computer Communications, 2008, 31(13): 3049-3058.

[7] Chen Min-Xiou, Tzeng Tze-Chin. Integrating service discovery technologies in OSGi platform [J]. Computer Standards & Interfaces, 2011, 33(3): 271-279.

[8] Vilas A F, Redondo R P D, Arias J J P, et al. Context-aware personalization services for a residential gateway based on the OSGi platform [J]. Expert Systems with Applications, 2010, 37(9): 6538-6546.

[9] 孙力军, 陈德人, 施敏华. 基于 OSGi 的自适应可进化软件框架 [J]. 江南大学学报(自然科学版), 2007, 6(2): 140-143.

[10] 冯志宇, 黄林鹏. 基于 OSGi 的两层服务模型 [J]. 计算机应用研究, 2009, 26(7): 2590-2592.

[11] 张 仕, 黄林鹏. 基于 OSGi 的服务动态演化 [J]. 软件学报, 2008, 19(5): 1201-1211.

[12] 陈志伟, 贾可荣, 张志祥, 等. 一种基于 OSGi 的动态演化方法 [J]. 计算机工程与科学, 2008, 30(11): 102-122.

(上接第 120 页)

computer-assisted environment for understanding geometry theorem proving problems and making conjectures [J]. International Journal of Intelligent Information and Database Systems, 2009, 3(3): 231-245.

[3] 王海林. SWRL 推理规则在平面几何证明中的应用 [J]. 计算机技术与发展, 2010, 20(9): 218-221.

[4] 常建鹏, 赵 克, 亿珍珍, 等. 初中几何专家系统中的知识获取及实现 [J]. 计算机技术与发展, 2008, 18(7): 156-159.

[5] Iwane N, Takeuchi A, Otsuki S. A network intelligent educational system for arithmetic word problems [J]. Systems and Computers in Japan, 1997, 28(10): 30-39.

[6] Dellarosa D. A computer simulation of children's arithmetic word problem solving [J]. Behavior Research Methods, Instruments & Computers, 1986, 18(2): 147-154.

[7] Wong Wing-Kwong, Hsu Sheng-Cheng, Wu Shi-Hung. LIM-

G: Learner-initiating instruction model based on cognitive knowledge for geometry word problem comprehension [J]. Computers & Education, 2007, 48(4): 582-601.

[8] 程 志. 小学算术应用题自动解答系统-以整数一、二步和分数基本应用题为例 [D]. 北京: 北京师范大学, 2008.

[9] 俞士汶. 计算语言学概论 [M]. 北京: 商务印书馆, 2007.

[10] Kintsch W, Greeno J G. Understanding and solving word arithmetic problems [J]. Psychology Review, 1985, 92: 109-129.

[11] Ma Yuhui, Tan Kai, Shao Limin, et al. Constructing the representation model of arithmetic word problems for intelligent tutoring system, 2011 [C]. Singapore: Computer Science & Education, 2011.

[12] 朱晓亚. 现代汉语句模研究 [M]. 北京: 北京大学出版社, 2000.

[13] GATE [EB/OL]. 2012. <http://gate.ac.uk/>.

基于语义句模的语义理解方法研究

作者：

[马玉慧](#)，[谭凯](#)，[尚晓晶](#)

作者单位：

[渤海大学 教育技术研究所, 辽宁 锦州 121000](#)

刊名：

[计算机技术与发展](#)

英文刊名：

[Computer Technology and Development](#)

年，卷(期)：

2012(10)

本文链接：http://d.g.wanfangdata.com.cn/Periodical_wjfz201210032.aspx