基于概念层次网络的小学应用题句类分析和知识提取①

HNC-Based Approach of Sentence Category Analysis and Knowledge Extraction for the Primary Problem

潘 黎 冯 速 (北京师范大学 信息科学与技术学院 北京 100875)

摘 要： 提出一套适用于小学应用题解题系统的句类分析及知识提取方法,以概念层次网络(HNC)理论为基础，从概念间的联系出发，重点研究概念间的数量关系，对自然语言描述的应用题进行句类分析，由此提取应用题中的知识，特别是数量上的知识，并用形式化的符号进行描述。在机器自动知识提取方向上所做的探索，可以为小学应用题自动解题系统提供一种思路。

关键词： HNC 自然语言理解 自动解题 句类分析 知识提取

1 引言

智能化是计算机应用的必然趋势。目前已经有很多成功的专家系统和解题系统能够比较合理地解决系统相应的问题并能够达到一定的正确性。但是，通过对得到一定应用的多个系统的调查和研究发现，这些系统并不能把由自然语言描述的题目作为直接输入。例如，中学平面几何解题系统[1]由操作者阅读以自然语言描述的题目，用形式化的计算机符号描述题目所述图形的坐标、长宽、直径、交点等参数，进而用已形式化的定理及公理库进行推导解题。另例，主观题自动批改系统[2]对关键词的语义相似度进行计算，未提取语义知识及处理语句上下文的关系。目前的解题系统较多集中于推理过程的研究[3,4]，自动推理与教育软件智能平台的基本构成及功能主要是学科知识库以及自动推理系统[5]，这些系统都是对已经形式化存储的知识进行处理、展现，并不能完成从原始题目到最终解答的运算。针对以上现状，本文探讨如何将自然语言描述的题目以字符串的形式直接作为解题系统的输入，然后让计算机从中提取有效信息，并最终解决问题。

本文首先对 HNC 与自然语言理解做简单介绍,然后详细讨论小学应用题中涉及的句类分析和知识提取方法，给出分析算法示例，并使用两个具体例子说明分析算法的运用。

2 HNC与自然语言理解

以自然语言作为原题输入的解题系统研究的难点之一是自然语言理解。汉语是意合型语言，具有如下特点：1、有明确的语义块标记；2、“字义基元化，词义组合化”特征明显；3、以词义组合的方式表达其词性；4、可以出现没有动词的句子；5、句子经常可以调换语义块的次序而意义不变[6]。这些特性给汉语机器理解带来了特殊的难点。

然而,黄曾阳提出的 HNC(Hierarchical Network of Concepts，概念层次网络)理论[7]的出现为汉语机器理解的研究带来了新的契机，使汉语的自然语言理解向前迈进了一大步。HNC 理论建立了一套较完备的关于自然语言理解处理的理论体系，该理论在深入挖掘汉语特点的基础上,以意义表达和语言理解为主线，建立了一种模拟大脑语言感知过程的自然语言表述模式和计算机理解处理模式。

本文以 HNC 作为理论基础，利用 HNC 的概念层

① 基金项目:国家自然科学基金(60273015)
收稿时间:2009-03-18

次网络和概念间的相互联想，可以比较有效地提取概念间的关系。

3　句类知识与知识提取

由于汉语的上述主要特点，将 HNC 理论直接用于汉语整体的理解和信息提取，其复杂性和语言概念的广泛性都使寻找自动解题方法变得较为困难。而小学生应用题有如下特点：语法规范、文字叙述简明、上下文关系清晰、逻辑关系严谨简单，且仅含有加减乘除四则运算。因此，本文从小学应用题入手，探索自然语言理解与知识提取的途径，寻找一种句类分类、分析方法，以期得以较为有效地提取知识。

3.1 句类知识

HNC 把自然语言理解定位在概念联想脉络表示式的激活与运作。它将自然语言抽象为概念，并用形式化的符号描述概念，而具体的词语则被描述成结合了一系列概念的一组符号。

从自然语言提取知识必然要分析、处理句子，而要分析句子结构，需要对句子进行分类。不同的句类有各自的特征，称为句类知识[8]。HNC 理论提出了 57 种基本句类及常见混合句类，其目标是要解决广义上的自然语言理解。考虑到本文研究对象的特征，小学生应用题系统句类知识的基本内容包含两个方面：(1)语义块构成知识；(2)语义块之间的概念关联知识，其中，概念间的关系主要为数量关系。因此有必要对 HNC 的句类进行精简，并针对出现频率较高的句类进行一定的扩展和修改，以便于数量概念的提取和数量关系的构成。以下是小学生应用题系统涉及的句类知识：

(1) 基本动作句 BAS(basic action sentence)句类表示式是：BAS=JK1+E+JK2

例如：妈妈(JK1)　吃了(E)　一些橘子和 1 个香蕉。(JK2)

学校食堂(JK1)　这个月　用(E)　煤 200 千克。(JK2)

基本动作句是最基本也是最重要的句类，广义对象语义块 JK1 和 JK2 近似于传统语言学中的主语和宾语。JK1 多为人物和组织团体，JK2 中一般含有数量块和名词短语。特征语义块 E 的核心动词是一般的动作。

(2) 存在句 ES(existence sentence)

ES31 = JK1 + E(“有”)+ JK2

ES32 = JK2 + E(“有”)+ 数量块

ES33 = JK1 + JK2 + E(“有”)+ 数量块

例如：图书室(JK1)　有(E)　科技书 240 本。(JK2)

科技书和故事书(JK2)　共　有(E)　多少本?(数量块)

某厂(JK1)　男工(JK2)　有(E)　300 人。(数量块)

ESij 表示 ES 句共有 i 种句式，此句式是第 j 种。在 ES31 中 JK1 一般是存在的广义空间，包括物理空间、社会空间和抽象空间，也包括某些时间。社会空间包括一些社会组织。JK2 表示该空间内存在的事物，多带有数量语义块修饰。ES32 将 ES31 中的 JK2(即空间内存在的事物)提到“有”的前面，而省略了 ES31 中的 JK1(广义空间)。ES33 其实是 ES31 的变型，即将空间内存在的事物提到了特征语义块之前。E(“有”)表示特征语义块以动词“有”为核心动词，语义块通常由核心部分和说明部分构成，特征语义块核心部分的前后都可以有说明部分，HNC 分别称前后的说明部分为特征语义块的上装和下装。

(3) 简明状态句 SS(status sentence)

SS = JK1 + 数量块

例如：每支铅笔(JK1)　0.25 元。(数量块)

李江(JK1)　今年 12 岁，他的爸爸(JK1)　今年 40 岁。(数量块)

简明状态句是最重要的无特征语义块的句类，数量块可以像特征语义块一样有上装和下装。当 JK1 中含有数量块时，两者位置可交换。例如：一个苹果 0.25 元 0.25 元一个苹果。

(4) 交换句 EXS(exchange sentence)

EXS31=JK1+E(“隐含变化方向的动词”)+ JK2

EXS32=JK1+E+hv(“来、进、走、出、入”)+JK2

EXS33=JK1+E(“显含变化方向的动词”)+数量块

例如：中心小学五年级共有图书 102 本，又卖(E)了 12 本。(数量块)

学校里原来有白粉笔 40 盒，学校(JK1)又买(E)来(hv)30 盒红粉笔。(数量块)

树上有 6 只鸟，又 飞　来 5 只。
　　E　hv 数量块

花金鱼 减少 120 条。
JK1　　E　数量块

因为小学生应用题很大程度是处理数量关系，而交换句的核心动词又包含数量变化的方向，所以这里作为一个单独句类。隐含数量变化方向的词包括：买、卖、借、贷等。显含数量变化方向的动词包括：增加、减少、节约、增产、提前等。EXS31 和 EXS32 中 JK1 常省略，JK2 中包含数量块和名词短语，名词短语可省略。当一个句子中同时出现隐含变化方向的动词和后缀时，将该句归为 EXS31 类型，比如例句："学校又买来 30 盒红粉笔"。hv 是动词后修饰成分的缩写，它包含两大类：一类是对 E 块的事态修饰，例如："着、了、过"等；另一类是对 E 块的效应和空间特性的说明，例如："到、来、出、去"等，EXS32 中出现的通常是第二类。

EXS33 是一个强调数量变化的句类，它忽略引起数量变化的原因，只考虑数量变化的结果，即增加或减少。EXS33 将 EXS31 和 EXS32 的 JK2 中的名词短语提到核心动词前作为 JK1，表明变化的量。

(5) 参照比较句 ACS(according comparison sentence)

ACS31 = JK1 + "比" + JK2 + UC + 数量块

ACS32 = JK1 + "比" + JK2 + UC + E + 数量块

ACS33 = JK1 + "比" + JK2 + E("显含变化方向的动词")+数量块

例如：故事书的本数 比 科技书 多 123 本。
　　JK1　　JK2　UC 数量块

小华种了 8 棵树，小明 比 小华 多 种 3 棵。
　　JK1　　JK2　U CE 数量块

二月份 比 一月份 增产 54 吨。
JK1　　JK2　　E　数量块

参照比较句 ACS31 和 ACS32 中，比较表现一般具有形容词特性，而不具有动态特性，不能形成特征语义块，所以用 UC 表示，两个例句中的"多"都是 UC。UC 前也可以用前说明成分(QE)来修饰。ACS33 与 EXS33 类似，只是强调数量变化方向。

参照比较句 ACS 前，一般会出现一个基本动作句 BAS = JK1 + E + JK2(也可以是存在句或者交换句，它们的句子结构类似，只是特征语义块不一样，在这里以 BAS 为代表)。ACS31 中的 JK1 或者 JK2 一般为 BAS 中的 JK2，其比较表现 UC 后不接动词。ACS32 中的 JK1 或者 JK2 一般为 BAS 中的 JK1，其比较表现 UC 后接特征语义块 E，且 E 的核心动词与 BAS 中 E 的核心动词一致。ACS33 中的 JK1、JK2 常为时间块 TK 和前说明成分 QE。

(6) 相互比较句 ICS(inter-comparison sentence)

ICS = JK1 + "和、与、跟、同" + JK2 + "同样、相同、一样、差不多" + UC

例如：橘子 和 香蕉的个数 同样 多。
　　JK1　　JK2　　UC

王军 和 刘明 一样 高。
JK1　　JK2　　UC

ICS 中 JK2 多为单一语义块("刘明")或者构成简单的复杂语义块("香蕉的个数")，JK1 可以是单个的语义块，也可以是另一个句类。比如："花金鱼减少 120 条就和红金鱼同样多了"，这个 ICS 中 JK1 就是一个交换句 EXS33。

(7) "是"状态句 JS(judgment sentence)

JS21=JK1+E("是")+数量块

JS22=JK1 + E("是")+ JK2 + "的" +数块+ "倍"

例如：地球表面的海洋面积 是 3.61 亿平方千米。
　　JK1　　数量块

运来酸奶的瓶数 是 鲜牛奶瓶数 的 1.5 倍。
JK1　　JK2　　数块

"是"状态句是预期知识最贫乏的句类，它一般用来揭示某个概念的数量特性或者概念之间的数量关系，其中 JK1 可以是各种形式的句蜕，例句 2 中 JK1 即为交换句的句蜕。JS21 的句类结构与简明状态句 SS 一致。

3.2 句类激活

在系统中应用句类知识，其第一步是句类激活。在本文中把特定句类的触发称为句类激活，而句类激活的第一步是 lv 激活及关键字激活。

lv 激活即寻找句子中的 l 类概念和 v 类概念，形成一个 lv 序列。l 类概念包括所有语义块切分符和语义块组合符，v 类概念包括所有能构成特征语义块的核心成分和附属成分(即特征语义块的上下装)。例如，在句子"妈妈把苹果吃了。"中，"把"是 l 类概念，

属于语义块切分符，“吃”和“了”是 v 类概念，“吃”是特征语义块的核心成分，而“了”是附属成分。lv 激活方法如下：对句子中每个词语的语义进行判断，如果语义中包含特定类型的概念，则激活相应的类型。如果某个词语激活了一个类型，则这个位置称之为激活点。对于上例，在“把”、“吃”和“了”处共有三个激活点，分别激活了语义块切分符、特征语义块核心、特征语义块下装三个类型。这个过程可以形式化为如下激活规则：

<激活类型>→<概念类别><HNC 符号>

其中，“概念类别”和“HNC 符号”对某个“概念激活类型”所要求的语义信息进行刻画，这两部分信息来源于词语知识库。

关键字激活是激活上文所述的 7 种基本句类中的核心词，即“是”，“有”，“比”等，通过核心词匹配，当核心词命中时，激活相应句类，提取知识。

当一个句子不含 v 概念或关键字时，则用简明状态句的句类结构检测该句，如符合，则激活简明状态句，否则该句不处理。

3.3 知识的提取

当句类激活完成后，即形成一个已知句类的句块序列，此时可进行知识信息的提取。对于不同的句类需要使用不同的知识信息提取算法。下面给出参照比较句的知识提取算法：

(1) 判断 JK1 是否为空，如果不空，分别分析 JK1 和 JK2 结构，按最大原则补充完整，如果 JK1 和 JK2 不是简单语义块，则提取名词短语作为 JK1 和 JK2，一般如果 JK1 和 JK2 均为一个词，我们认为他们是对仗的，不需要处理。转(3)。

(2) 如果为空，分析 JK2 的结构类型，在前句中找到和 JK2 类型相同的内容，把 JK1 补充上。如果 JK2 为特殊块，例如组织团体块，则找前句的组织团体块，否则计算 JK2 与前句中 JK1 和 JK2 的相似度，取相似度最大的 JK 充当该句 JK1。(相似度计算方法参阅语义距离计算[6])。

(3) 在上文中找到 JK1、JK2 哪个是已知量。如果 JK1 和 JK2 都找不到，设置标志位，保存 JK1 和 JK2，跳过 4，进行后面句子的处理。后面句子中一旦出现 JK1 或 JK2，立即转(4)。

(4) 以已知量的句型为依据，复制句型结构，用比较者更换其内容。

3.4 知识的提取

本节以两个例句对第 3.3 节所述的知识提取方法进行说明。

例句 1：体育室有篮球 24 个，比排球少 6 个。

① 第一句句类激活，得到句类表示式：ES=JK1(体育室)+“有”+JK2[名词短语(篮球)+数量块(24 个)]；

② 第二句句类激活得知此句为参照比较句 ACS31，“比”前为空，即 JK1 为空，找到 JK2(排球)；

③ 计算“排球”与“体育室”和“篮球”的词语相似度，得到 ACS31 中 JK1 为篮球，即 ACS31 句补充为：篮球比排球少 6 个；

④ JK1(篮球)在上句出现过，判断 JK1 是已知量；

⑤ 复制句型：体育室有篮球 24 个。更换：体育室有排球 24 个；

⑥ JK1 是已知量，JK2 是未知量，u 转变方向：少(-)->(+)得到：体育室有排球(24+6)个。

例句 2：一架飞机每小时比一艘轮船要快 1780 千米，一架飞机 2 小时飞行 3600 千米。一艘轮船每小时行多少千米?

① 第一句句类激活，得到 JK1：一架(数量块)飞机(名词短语)每小时(时间块)，JK2：一艘(数量块)轮船(名词短语)，按照最大原则，补充 JK2：一艘 l 轮船每小时，并将 JK1 和 JK2 分别简化为名词短语飞机和轮船；

② JK1,JK2 在已知知识中均不存在，设置标志位，处理下一句；

③ 第二句句类激活得到句类表示式：基本动作句 BAS=JK1(一架飞机)+时间块(2 小时)+E(飞行)+数量块(3600 千米)；

④ 发现时间块(TK)的时间为复数，建立时间量关系<小时，千米，2，3600>，进一步得到<小时，千米，1，1800>；

⑤ 发现第一句 ACS31 中 JK1 的 TK 为单数时间，转换基本动作句 BAS 得到：基本动作句 BAS’=JK1(一架飞机)+TK(1 小时)+E(飞行)+数量块(1800 千米)；

⑥ BAS’中发现 ACS31 的 JK1(飞机)出现，还原标志位，进入参照比较句 4 阶段；

⑦ 复制句型：一架飞机 1 小时飞行 1800 米。更换：一艘轮船 1 小时飞行 1800 米；

⑧ JK1 是已知量，JK2 是未知量，u 转变方向：快(+)->(-)得到：一艘轮船 1 小时飞行(1800-1780)。

4 结语

研究和验证本文所述的方法时，我们选取了具有代表性的 200 道应用题实例。按照解题所需的运算分为一则、二则、三则以及四则运算型四类，其中一则运算型可细分为加法型、减法型、乘法型和除法型，二则、三则和四则运算型即为一则运算型的组合或叠加。本文探索的方法对其中 189 道应用题可得到满意的结果，但是对于若干题目还无法正确解题。例如：

(1) 正方形池塘每边有 6 棵树，四边一共有多少棵树?

(2) 一列火车身长 90 米，以每秒 160 的速度过一个山洞，用了 5 秒钟。问山洞长多少米?

(3) 一桶油连桶共重 18.5 千克，倒出油的一半后，剩下的油和桶共重 13.5 千克。桶重多少千克?

(4) 小林看一本小说，6 天看了 120 页。照这样计算，又看了 2 天，前后一共看了多少页?

这些题目涉及到概念与概念之间非单纯数量的关系，比如正方形和边之间的关系，山洞本身所具备的属性即什么是山洞，桶和油的集合关系，以及“前后”在句子中指代的对象，等等。处理概念之间更为复杂的关系是本方法今后需要改进的地方。

应用本文所述方法，可将应用题中的数量知识形式化表述，有利于计算机的实现，对开发小学生应用题解题系统会有比较大的帮助。然而，要确立概念间的关系则系统必须以句类知识为依托，对于广泛的、由自然语言描述的问题的解答，本文提供的句类知识还远远不够，而完善知识库是一个庞大的系统工程，需要巨大的人力物力投入。本文旨在探索一种合理的方法，从简入繁，循序渐进。

参考文献

1 杨俊梅,许威,赵克.平面几何智能辅导系统的设计与实现.航空计算技术, 2004,34(4):102－104.

2 肖雪莲.基于 HNC 理论的主观题自动批改算法设计与系统实现[硕士学位论文].上海:华东师范大学, 2006.

3 张景中,李传中.自动推理与教育软件智能平台.广州大学学报(综合版), 2001,15(2):1－6.

4 徐茜.双向推理系统在初等几何自动解题中的实现.计算机应用研究, 2004,21(11):232－234.

5 李涛,张波,李传中.基于前向推理的平面解析几何自动推理系统研究与实现.计算机应用,2006,26(7):1715－1717,1720.

6 晋耀红. HNC(概念层次网络)语言理解技术及其应用.北京:科学出版社.2006.13－14,50－60.

7 黄曾阳.HNC(概念层次网络)理论.北京:清华大学出版社, 1998.1－20.

8 苗传江.HNC(概念层次网络)理论导论.北京:清华大学出版社, 2005.48－67.