# 基于知识图的语义分析

张 蕾[1],李学良[2],刘小冬[2]

(1. 西北工业大学 计算机科学与工程系;2. 西北工业大学 数学与信息科学系,陕西 西安 710072)

**摘要**:探讨了一种适合于汉语理解的、面向语义的、新的语言分析方法。该方法是以知识图这种新的知识表示方法为语义模型,模拟人的语言感知过程,先对整个句子进行语义片段的划分,再对每个片段进行分析。

**关 键 词**:自然语言处理;知识图;语言分析;语义块

**中图分类号**:TP391 **文献标识码**:A **文章编号**:1000-274X(2002)02-0153-04

语言分析作为自然语言处理领域中一个独立的分支,长期以来一直受到世界各国研究人员的高度重视。在过去的几十年里,由于受到以美国著名语言学家乔姆斯基(Chomsky)为代表的转换生成语法理论和当时整个科学界唯物主义思潮的巨大影响,基于规则的句法分析方法和系统一直是分析研究的主流。的确,基于规则的方法和系统在诸如自然语言接口、天气预报等应用领域取得了一定的成功。

近几年来,随着计算机科学技术的飞速发展和 Internet 上信息的与日俱增,人们对语言信息处理的要求不断提高,不仅要求智能化的信息处理系统能够自动检索信息,而且要求能对信息和知识进行分类、提炼、过滤、翻译等基于理解的高层次加工。而这些高层次的加工必须是,也只能是建立在语义的基础上。传统的知识表示方法不能确切地表达深层语义,因此,知识图方法(属于概念结构范畴)的诞生[1],对推动信息时代向知识处理为主的高级阶段转变和发展产生积极的作用。

汉语的理解存在着特殊的困难[2,3],不能生搬硬套传统的分析方法。因此,迫切需要研究出面向语义的、适合于汉语理解的分析方法。

## 1 基于知识图的语义分析

传统分析技术经历了一个漫长的发展过程。不同的分析方法反映了不同的分析深度。早期的模式匹配技术,主要依靠关键词匹配技术来识别输入句子的意义。系统将当前输入的句子与规定好的模式进行近似匹配,并没有真正意义上的语法分析。先句法后语义技术,强调在语言分析过程中存在一个相对独立的句法分析阶段,输出的结果是输入句子的一棵(或多棵)句法分析树。再经过语义分析阶段的处理,获得句子的深层语义表示。实践证明,这类系统的设计可以不依赖于某个特定的应用领域,因而具有更好的可扩展性和可移植性。无语法分析技术,抛弃传统的句法分析模式、不产生句法分析树这样的中间结果。以 Schank 为代表的美国耶鲁学派创导的一体化概念分析方法就是这种观点的典型代表。他们设计的 MARGIE,SAM,PAM 都采用了这种分析策略。这种分析策略的主要依据是心理学方面的合理性,直接从输入句子生成其意义表示。对那些有语法错误或信息不全的句子,可以根据语义信息获得解释,这是传统的句法分析难以对付的问题。然而,对于比较复杂的句子、没有句法指导,语义分析往往难以奏效。

无论单独依靠句法信息还是单独依靠语义信息,它们各自所依据的信息都是不完全的、在分析的过程中应该相互得到对方的帮助。另一方面,理解最终还是要基于语义,因此,语义研究应该是基础。

基于知识图的语义分析是以知识图为语义解释

**收稿日期**:2001-04-01

**基金项目**:国家自然科学基金资助项目(NSFC19971069)

**作者简介**:张 蕾(1964-),女,陕西西安人,西北大学副教授,西北工业大学博士生,从事知识表示、自然语言理解方面的研究。

模型,模拟人的语言感知过程。人在理解说话的意思时,不是听完了整个句子,分析出主语、谓语、宾语以后才去理解句子的含义;而是一个词一个词地去理解,再由词组成一个一个的片段,每个片段的含义理解了,整个句子的含义也就随之理解了。具体方案如下:①用句法词图和语义词图[5~6]表示词的多层面信息结构;②句法词图辅助语义分析的进行;③明显不同于现有分析方法的是,先从说话路径和语义块的感知入手,然后再由句法分析辅助、引导分析过程。与传统分析方法相比,基于知识图的语义分析的特点是:

1) 模拟人的认知过程。人在理解的时候不是先分析整个句子的主、谓语结构才理解句子的意思,而是一个词、一个词地听,一个语义块、一个语义块地去理解。

2) 不产生句法分析树这样的中间结果。句法分析不是独立的一个过程,而是隐含在整个分析中。

3) 基于语义的理解。传统的方法多是以句法为主,用语义信息消解句法分析产生的歧义。而结构分析是面向语义的分析,句法信息辅助、引导正确语义的产生。

4) 语言表示模型优于复杂特征集,表达形式更简洁、直观。具有独立于语种的、广泛的语言表达范围和深刻的语义刻画能力,传统方法所能表达的知识图都能表示。

5) 基于图的合一运算。新一代语言学模型的共同特点都是基于复杂特征集和合一运算的,知识图在这一点上与概念图类似,也有相应的运算。

6) 结构分析对汉语尤为适用。这是因为,第一,汉语拥有比较明确的语义块区分标志,如汉语有"把、对、被、向、就……而言、与……相比"之类的语义块指示符;西语有比较完善的短语指示符标记,如"the, a, for, with, ……"等,汉语没有那么完备。第二,汉语语义块的封闭性优于西语,像众所周知的"I saw a girl with a telescope near the bank"之类典型和普遍的语义块构成模糊,汉语是不存在的。

7) 便于机器实现。因为图的存储结构易于机器实现,具有结构稳定、易搜索的特点。

## 2 说话路径和语义块

### 2.1 说话路径和语义块

一个句子的含义由句子图表示,而句子图通过句子中的单词图"粘"合而成的。说话者按照一定的语序来说话,与此顺序相对应的是构成句子图的一个一个子图的顺序,每个子图代表句子中的一个语义块的含义。这些子图的不同排列顺序就隐含了不同的说话路径,即语序。

考虑下面的4句话:

1) The volcano, that lies in Alaska, 130 kilometers from Anchorage, erupted in 1992.

2) The volcano, that erupted in 1992, lies in Alaska, 130 kilometers from Anchorage.

3) 130 kilometers from Anchorage, Alaska, lies the volcano, that erupted in 1992.

4) In Alaska, 130 kilometers from Anchorage, lies the volcano, that erupted in 1992.

这4句话表达是同一个意思,虽然说话时采用的语序不同,句子的结构不同,但含义相同。按照知识图"结构即含义"的观点,应该具有相同的句子图(如图1所示)。我们从中划分出一些语义块。这四句话中都有的语义块是:"130 kilometers from Anchorage";"erupted in1992"和"the volcano, that", "in Alaska"在第3句话中没有,但是,"in"在这里也可以使用。图1给出了简化以后的句子图,每个语义块用一个框架表示,并用标号①到⑥标记。上面的4句话可以用语义块(从①到⑥)的不同序列来表示。

即

- ①→②→④→⑤→⑥→③
- ①→②→③→④→⑤→⑥
- ⑥→⑤→④→①→②→③
- ⑤→⑥→④→①→②→③

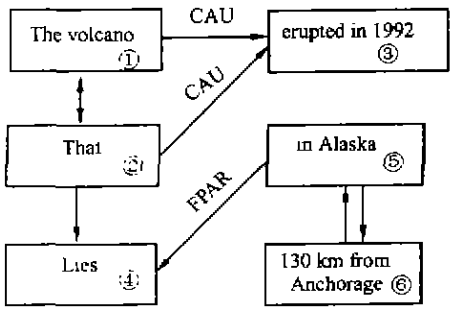


图1 语义块示意图

Fig. 1 Semantic chunks

注意到,在这些连贯的语义块之间出现了"跳跃"——这些语义块的框架在句子图中是不相邻的。句1中存在跳跃"⑥→③";句2中存在跳跃"③→④";句3中存在跳跃"④→①";句4中存在跳跃"⑥→④"和"④→①"。

### 2.2 语义块指示符

所谓语义块就是一个句子中有固定含义的一部分(子句)。在一个句子中,特别是汉语中,词的顺序是灵活多变的,但是,一个语义块的内部结构是相对稳定的。例如,"那个学生今天上午读完了这本小说",也可以说成"今天上午那个学生把这本小说读完了"。这两句话中,词的顺序是完全不同,但是其中词组,如"今天上午"、"那个学生"、"这本小说"和"读完了"的结构是固定不变的。这些词组就是一个个语义块。

**定义 1** 语义块是语句的下一级语法和语义构成单位。

换句话说,句子是由语义块构成的。它可以是一个词、一个短语,甚至可以包含另外一个句子,或由另一个句子退化而来。

在句子中如何来界定一个语义块呢?

**定义 2** 语义块指示符是一个语义块开始或结束的语言标志。

寻找一个句子中的语义块,就是要找到句子中语义块的指示符。我们归纳出下列语义块指示符:

1) 指示符 0:逗号和/或句号对;

2) 指示符 1:助动词;

3) 指示符 2:指代词;

4) 指示符 3:语义上不可能联系的"跳跃";

5) 指示符 4:介词。

对于汉语而言,则有更多的信息可以作为语义块指示符。例如,虚词"的"、"地"、"得"都是很重要的语义块指示符。从语法上讲,"的"可以作为一个名词性词组的标记,与其前面的名词词组、动词词组、数量词词组、介词词组、形容词词组等组成一个相当于名词成分的"的"字结构。所以,"的"应该是一个语义块指示符。"地"和"得"的用法也有相应的情况。详细情况参见文献[7]。此外,一些特殊字,如"把"、"是"、"有"、"被"等有明确的语法和语义含义,也应该看作是语义块指示符。综上所述,汉语的语义块指示符除了与英语相同的 5 类以外,至少还应该增加以下几类:

6) 指示符 5:"的"、"地"、"得"。

7) 指示符 6:"是"、"把"、"被"、"有"。

## 3 分析策略

由于在句法分析的过程中会遇到各种各样的歧义问题,所以歧义处理方法的设计就成为分析系统效率的关键问题。传统的歧义处理方法有两种:①回溯(backtracking);②并行(paralleling)。这两种情况都需要大量的簿记,以便保存多种可能性的踪迹。对于并行算法来说,由于它必须保存所有的局部分析,所以它对于存储空间的要求相对于回溯处理来说大得多。反之,回溯算法只要求比存储一棵单独的句法树略多一点的空间。如果对一个规模较大的语法来分析一个长句子,有可能产生数以千计的局部分析,这对于并行算法的句法分析器来说是一个沉重的负担。

### 3.1 控制策略

语法分析控制策略可以分成两大类:①自底向上,它是由数据驱动的;②自顶向下,它是由预期驱动的。汉语的分析中如果采用单纯的自底向上和自顶向下,都可能产生组合爆炸的情况。所以,结构分析以词图驱动的自底向上分析为主,在分析过程中适当地引用"预期信息"进行制导的控制策略。

### 3.2 双向扫描

结构分析要先划分大的语义块,再将大语义块划分成小语义块,因此一遍扫描完成这个工作很困难,应适当安排扫描次数。

### 3.3 采用确定性算法

在采用 ATN 这类非确定算法时,因回溯过多,对它的效率人们并不十分满意。此外,ATN 并不能方便地提供足够大的观察空间,只允许有节制地间接访问已处理过的某些节点(通过寄存器组),不允许向前看若干节点,而确定性算法提供了这种机制。因此,对汉语的分析算法宜采用确定性算法。

算法由划分语义块和生成句子圈两个部分组成。划分语义块的算法指述如下:① 当前结点与表的当前状态是否能够引发某种合并? ② 如果是,则实行这种合并,且将合并的结果存入当前结点中;③ 否则,便暂不处理,将指针指向下一个结点,等待合适的时机。通过移动指针,将新的结点作为当前结点,重新开始新的一轮合并。

## 4 结束语

对于本文提出的分析方法正在应用于一个中文信息自动抽取系统中,这将有助于系统对语义信息的处理,克服传统方法中基于简单模式匹配的弊端,使系统真正能够达到"信息理解"的这个高级知识处理阶段。

## 参考文献:


[1] 张 蕾,李学良,刘小冬.自然语言处理中的逻辑词[J].小型微型计算机系统,2000,21(2):520-523.

[2] HOEDE C, LI X, LIU X, *et al*. Knowledge graph analysis of some particular problems in the semantics of Chinese[J]. Memorandum,2000,516(1):1-10.

[3] 刘小冬,李学良,张 蕾.量词在知识图中的分类与表示[J].小型微型计算机系统,2000,21(5):153-157.

[4] HOEDE C, LI X. Word graphs: the first set[A]. EULUND P W, ELLIS G, MANN G. Proceedings of the 4th International Conference on Conceptual Structures[C]. Australia: Addison-Wesley, 1996. 81-93.

[5] HEODE C,LIU X. Word graphs: the second set[A]. MUGNIER M L, CHEIN M. Proceedings of the 6th International Conference on Conceptual Structure[C]. Australia: Addison-Wesley, 1998. 375-389.

[6] HEODE C,ZHANG L. Word graphs: the third set[A]. HARRY S. D, GERD S. Proceedings of 9th International Conference on Conceptual Structures[C]. Stanford:Springer,2001. 15-27.

[7] 詹卫东.面向中文信息处理的现代汉语短语结构规则研究[M].北京:清华大学出版社;南宁:广西科学技术出版社,2000.


(编 辑 曹大刚)

# Semantic analysis based on knowledge graphs

ZHANG Lei[1], LI Xue-liang[2], LIU Xiao-dong[2]

(1. Department of Computer Sciences and Engineering, Northwestern Polytechnical University, Xi'an 710072, China; 2. Department of Mathematics and Information Sciences, Northwestern Polytechnical University, Xi'an 710072, China)

**Abstract**: A new method of semantic analysis for natural language processing is proposed. The method is an oriented semantic understanding and suitable for Chinese natural language understanding. With this method based on knowledge graphs, each sentence is chunked into semantic segments, then each semantic chunk is analyzed.

**Key words**: natural language processing; knowledge graph; semantic parsing; semantic chunks

(上接第150页)

## 参考文献:


[1] WILLIAMS C L. Importance of dietary fiber in childhood [J]. J Am Diet Assoc,1995,95:1 140-1 146.

[2] SHINNICK F L, MATHEWS R, INK S. Serum cholesterol reduction by oats and other fiber sources[J]. Cereal Foods World,1991,36:815-821.

[3] 郑建仙,高孔荣.论膳食纤维[J].食品与发酵工业,1994,(4):71-74.

[4] 闻芝梅.现代营养学.第7版[M].陈君石译.北京:人民卫生出版社,1988.

[5] 张小玲.果胶的咔唑硫酸分光光度测定法研究[J].甘肃农业大学学报,1999,(3):75-78.

[6] 韩会新.食品卫生理化检验文集(2)[M].北京:北京大学出版社,1990.

[7] 李雄彪.植物细胞壁[M].北京:北京大学出版社,1993.


(编 辑 姚 远)

# A study on the extraction of soluble dietary fiber from apple pomace

SHI Hong-bing[1], SONG Ji-rong[1], HUANG Jie[1], DENG Hong[2]

(1. Department of Chemical Engineering, Shaanxi Key Laboratary of Physico-Inorganic Chemistry, Northwest University, Xi'an 710069, China; 2. Department of Food & Engineering, Shaanxi Normal University, Xi'an 710062, China)

**Abstract**: Apple pomace was used as raw material to extract soluble dietary fiber(SDF). The effect of solution concentration, Liquid-to-solid ratio stuff ratio, reaction temperature and time on the yield and quality of the product were studied. The optimum condition was obtained.

**Key words**: apple pomace; dietary fiber; technological condition