基于语法和语义分析的同音词识别模型

李素建 刘群 白硕

中国科学院计算技术研究所软件研究室 北京 100080

摘要 我们从汉语输入中的同义词识别问题出发、实用为主，提出了一个关于语法分析触发语义分析和计算的框架。语义分析是在语法分析基础上进行的,语法属性关系激发语义分析。语义分析实际上就是语法分析的进一步细化，解决语法分析不能解决的歧义。我们建立了一个进行语法和语义分析的拼音—汉字转换系统。语法分析使用了 LR 语法分析器，语义分析的核心是具有类属层次结构的语义知识库,.使语义分析根据主位语义类进行属性计算。

关键词 语法分析，语义分析，自然语言处理，拼音输入法

1. 引言

各种汉字输入方法中，拼音输入逐渐成为主流趋势。汉语有相对简单的音节结构，只有 400 个左右的无调音节是人们常用的输入，而汉字数量却极为庞大，仅国标一、二级汉字就由 6763 个。因此汉语音节集到汉字集是一对多的映射。汉语句子中普遍存在着词，利用平面的约束可去掉大量的候选词。即使使用大词库，同音词多选问题仍不能得到很好的解决，从而大大降低了输入速度。据估计，拼音汉字以词为单位输入，一次转换成汉字，其正确率平均在 95%左右。则其余的 5%是对于同音词而言的。因此，要继续提高准确率，只用利用更高层次的语言学知识，所以我们提出通过借鉴语法和语义分析方法来处理同音词多选问题。

语法和语义分析方法包括 ATN 网络[3]，配价语法[5]，依存语法[6]，HNC 理论[7]等，这些方法对于自然语言处理都起了一定的促进作用。我们从汉语输入中的同义词识别问题出发，实用为主，由 Wittgenstein[8]的“语义即用法”理论，提出了一个关于语法分析触发语义分析和计算的框架。

2. 语法和语义分析的作用

语法分析就是应用语法知识，将输入句子中单词之间的线性次序，变化成象句法树那样的某种数据结构。语法分析是 NLP 中一个重要的步骤，它实际上是对词之间增添了一个限定条件，使以词为单位构造语句时减少了随意性，因此语法分析可以看作是语句合法性检查的第一步。例 1 说明了语法分析如何解决部分歧义的.

例 1，输入拼音：　　tongzhi.women(‘.’ 表示假设词已经分好)

语法分析前输出的可能候选词：通知（同志）.我们

语法分析后输出的可能候选词：通知.我们

例 1 通过动词加上名词构成动词短语，把“tongzhi”的词性限制到了动词，从而排除了名词词性的“同志”。

语义分析是识别一句话所表达的实际含义，赋予由语法分析所建立的数据结构所含的“意义”，在句法结构及任务的领域内的物体间进行映射变换。由此看出，语义分析是以语法分析为

基础的，对于语法分析所不能解决的问题进一步处理。结合同音词消歧问题，语法分析已经把问题限制到同种词性的同音词上，进一步的语义分析就是根据实际的意义和更细致的上下文进一步进行限制。

例 2，输入拼音：　　zai.beijing

语法分析后输出的可能候选词：在.北京（背景）

语义分析后输出的可能候选词：在.北京

语法分析之后即使“beijing”的词性限制到名词上，仍没有彻底解决歧义问题。在语义知识库，限制作为抽象事物的名词“背景”，使它的介词属性值不能为‘在’，而专有名词“北京”符合语义，从而去掉多选的可能性。

通过以上两个事例，我们了解了语法分析和语义分析的具体工作。语法分析是通过分析语句中不同成分之间的搭配，对多选起了限定作用；而语义分析的基础是本体论，通过词义和具体事物的关系，对词汇的选择进一步限制，从而使同音词多选的几率再次降低。这里，只有语法成功的基础上我们才进行语义分析，语义分析实际上就是语法分析的进一步细化，解决语法分析不能解决的歧义。

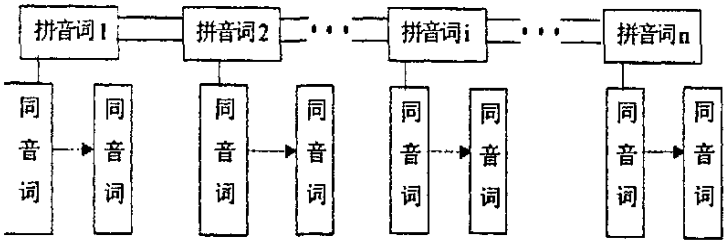

图 3.1 同音词序列链

3. 同音词识别的系统设计和语法分析

上面已经说明了语法分析和语义分析在解决同音词歧义中的作用，因此我们要建立需要语法和语义分析的拼音—汉字转换系统。首先要对输入拼音串进行分词，由于篇幅有限我们不讨

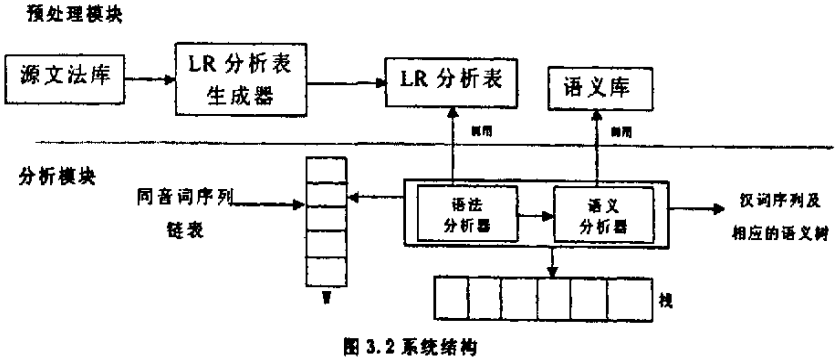

图 3.2 系统结构

论分词问题，而假设输入拼音串已经正确分词，所以只需把以词汇为最小单位的汉语拼音串转换成合法的汉字序列。语法、语义分析器的输入为图 3.1 所示的同音词序列链表结构，每一个拼音词链接着一个或多个同音汉字词。

同音词链表结构进入系统经过分析，理想的输出是每一个拼音词节点最终只对应着一个汉字词。整个系统结构如图 3.2 所示。

系统中，语法分析使用了 LR 语法分析器，同时语法分析激发语义分析进行属性计算。由图 3.2 可知，整个系统的设计分为两大模块，一是预处理模块，源文法库经 LR 分析表生成器生成供分析器使用的 LR 分析表，它是进行语法分析的核心。同时建立一个语义知识库，这是语义分析的基础；另一模块是分析模块，它对输入的同音词链表序列进行句法分析和语义检查，从而输出合法的汉词序列和相应的语义树。

对于拼音输入系统，进行分析时的效率是一个必须要考虑的问题，所以语法分析采用了广义的 LR 分析算法[2]。该算法的核心数据结构是 LR 分析表，每一个表项可以保留多个动作。事先根据语法构造分析表，可以使分析时的计算基本上变成了查表和按照表上的动作机械执行的过程，因此 LR 算法可以达到差不多同输入字符串的长度呈线性关系的速度[2]。这里，由于输入为同音词序列，因此效率与输入拼音词的数目和每个拼音词对应的同音词的词性数目有关。LR 分析算法是一种比较成熟的语法分析算法[1]，对于输入的每个拼音词，可能存在着多种词性，LR 分析的任务就是选取合法的词性。每个词的某个词性在出现错误时就选取下一个词性或进行回溯操作。

例如：下面的一段语法规则：

1) *VP* → *VP* *NP*
2) *VP* → *VP* *Adj*
3) *VP* → *Verb*
4) *NP* → *VP* *'de'* *NP*
5) *NP* → *SHU* *NP*
6) *NP* → *Noun*
7) *SHU* → *SHS* *Unt*
8) *SHS* → *Numb*

由这些规则，拼音串“yi.zhi.yao.si.lieren.de.gou”经过分析得到如图 3.3 所示的语法树，确定了每个词所对应的属性分别是：数词、量词、动词、形容词、名词、‘的’、名词。

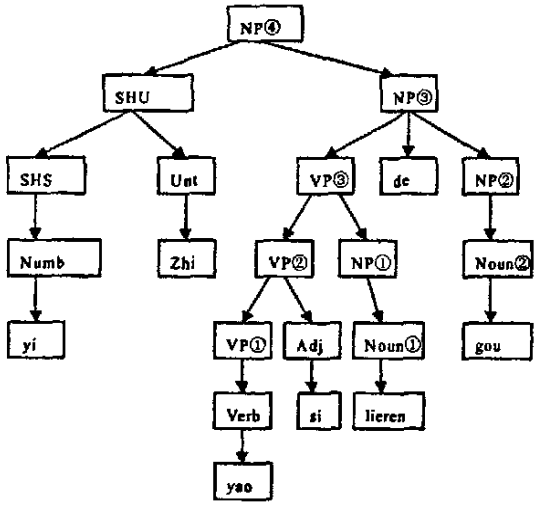

图 3.3: 拼音串"yi.zhi.yao.si.lieren.de.gou"的语法树

4. 语义分析方法

LR 分析采用自上而下的分析算法，从词出发逐层向上归纳为短语。在向上归纳的过程中遇到歧义就进行相关的语义检查，调用语义分析子程序从语义角度对词重新分析。

语义分析的核心是具有类属层次结构的语义知识库，这里借鉴了《同义词词林》对汉语的语义分类，分为 12 大类，94 中类，1426 小类。它是一种网络层次结构。下面是相关的一些定义：

W 表示汉语所有词的集合，m 表示所有词的数目；每个词不论在语法和使用上都有一些特点，把这些特点作为属性，A 就表示这些属性的集合，每个属性在计算时都得到了赋值；SCi 表示一个语义类，它可以有一个或多个上位语义类 SC_{m1}、SC_{m2} 等；在每个语义类中都有一些具体的汉语词汇 W_{i1}、W_{i2},...相对应，如果该语义类存在着反义语义类，则为 SC_{ia}。SC_i 包括 q_i 个属性 $SC_i:\langle a_{ij}\rangle(1\leqslant j\leqslant q_i)$，它们的值可以预先定义或者通过继承得到。语义知识库就是所有语义类 SC_i 通过一些关系构成的语义网络。这些关系包括：同义关系、反义关系、逻辑多继承关系、整体部分关系和非单调推理关系。

(1) $W=\{w_1,w_2,\cdots,w_m\},\ m>0$

(2) $A=\{a_1,a_2,\cdots,a_n\},\ n>0$

(3) $SC_j:SC_{m_1}:SC_{m_2}:\Lambda$

$\{\ w_{i1},w_{i2},\Lambda,w_{ip_i};$

$!\ SC_{ia};$

$SC_i:\langle a_{i1}\rangle;$

$SC_i:\langle a_{i2}\rangle;$

$\Lambda;$

$SC_i:\langle a_{iq_i}\rangle;$

$\}$

$w_j\in W(0<j\leq p_i),\ a_j\in A(0<j\leq q_i)$

语义分析是在语法分析基础上进行的。对于 $R_{SA}=x\cdot y,\ x\in SA,\ y\in SA$，SA 表示语法属性，该式是句法分析的逻辑表示形式。它得到的结果确定了 x，y 的主位和从位关系，并且返回语法属性关系 R_{SA}。对于语义分析，我们定义了公式 $R_{sc}=X\cdot Y,\ X,\ Y\in SC\cup SA$，其中 X 与 x，Y 与 y 在句中的位置分别相对应。这样两个语义类 X,Y 也默认所得到的主位和从位关系。语法属性关系激发语义分析。使语义分析根据主位语义类进行属性计算。根据实际的需要，我们规定了 15 种激发语义分析计算的语法属性关系： a.量名关系(UNr)；b.名偏关系(NPr)；c.名名关系(NNr)；d.动名关系(VNr)；e.名方关系(NOr)；f.介宾关系(POr)；g.数量关系(VPr)；h.动偏关系(VPr)；i.主谓关系(SPr)； j.动补关系(VCr)；k.动量关系(VUr)；l.动宾关系(VOr)；m.连动关系(VVr)；n.形偏关系(APr)；o.形补关系(ACr)。语义类正是由这些关系触发进行属性计算和语义检查的。

语义类属性值通过继承得到计算，可以看作是一种局部的类属关系计算。如果从短语和语句角度上建立起语义类的属性计算，则称为是全局的类属关系计算。根据语法分析的二分性和语法分析激活语义分析，语义分析就是通过语义类属性计算检查语义类的组合关系。

我们把所有语义类划归到三类顶层语义类：事物类、动作类和特征类。根据触发语义分析的语法关系，首先我们必须为语义类设计属性描述。对于事物类所设计的可能属性有：量词修饰、修饰前缀、包含对象、并列搭配、方位搭配、介词搭配。为动作类所设计的属性有：修饰属性、动作主体、动量属性、动补属性、动作对象、并列搭配。为特征类所设计的属性有：修饰属性、形补属性。这些属性的设置正是为了具体到某种语义类具有特定搭配时进行检查和限制。我们举个具体的例子看如何语义属性的计算解决同音词歧义的。例如

拼音串	汉字串
Yi.zhi.hua	一枝花
Yi.zhi.ji	一只鸡
Yi.zhi.qian	一支枪

表 1: 数量名短语事例

这里由三个从拼音到汉字的例句，都是数词+量词+名词的结构，并且中间的拼音都为“zhi”，因此同音词消歧比较困难。但是，我们注意到这么一个规律：和植物类名词搭配的“zhi”转换为汉字“枝”，和动物类名词搭配的“zhi”转换为汉字“只”，和工具类名词搭配的“zhi”

转换为汉字“支”。因此在植物类、动物类、工具类中分别定义它们的量词修饰属性为：

〈植物类〉：〈量词修饰〉＝‘枝’；

〈动物类〉：〈量词修饰〉＝‘只’；

〈工具类〉：〈量词修饰〉＝‘支’；

在语法分析成功后，量名关系(UNr)激活名词所在语义类的属性计算，通过继承关系得到该词的量词修饰属性的值，可以有效地消除同音词歧义。

5. 结束语

我们所采用的语法分析基础是短语结构语法，语义分析基础是基于类属分析的语义类属性计算。通过语法分析规约时激活语义分析，从上下文环境中解决同音词歧义问题，在很大程度上降低了同音词的多选问题。LR 分析效率高，因为对发生变化的规则集只需重新编译一次得到分析表，分析算法就可以一直使用该分析表，从而分析表的预先处理保持了分析过程的速度。同时基于类属关系的语义分析从实用角度使语法分析进一步细致化。这里语义知识库是一个焦点问题，它的组织和形式化还是一个有待继续深化的课题。

参考文献：

[1] 编译原理和技术，陈意云，马万里 中国科学技术大学出版社

[2] 白硕，计算语言学教程讲义

[3] 汤建华，利用句法、语义循环递归网络实现汉语拼音-汉字转换，《中文信息学报》，Vol.5，No.3，1989

[4] 基于二元语义关系的句法和语义分析 万建成，姚文琳 Comm. COLIPS(新加坡)，Vol.8，No.1，1998

[5] Shen Yang, Zheng Dingou, Studies on Valent Grammar in Modern Chinese, Peking University Press, 1995

[6] 冯志伟，从属关系语法的某些形式特性，1998 中文信息处理国际会议论文集，清华出版社

[7] Wang Hou-feng, The ProcessingTactics based on HNC for Delimitation of Chinese sentences, Comm. COLIPS(Singapore), Vol.9, No.8, 1999

[8] Andrew Lilico, Wittgenstein & the Augustinian Picture of Language, http://ourworld.compuserve.com/homepages/Andrew_Lilico/augustin.htm

基于语法和语义分析的同音词识别模型

作者： 李素建， 刘群， 白硕
作者单位： 中国科学院计算技术研究所软件研究室,北京,100080