汉语句型自动分析和分布统计模型的研究

罗振声　郑碧霞

清华大学中国语言文学系

摘　要

汉语句型的自动分析与分布统计是一项重要的基础性研究课题。本文就以结构特征为标准的句型系统，提出了以谓语为中心的句型成分分析和句型匹配相结合的汉语句型分析算法与策略，给出了实验模型系统的结构、初步的实验数据与分析。

关键字：句型分析，分布统计，句型成分，句型匹配

一、引　言

汉语句型自动分析和分布统计是继我国字频统计和词频统计之后的又一亟待解决的重要基础性研究课题。基于语料库大规模真实语料的汉语句型综合频度统计工作的最后完成，将为现代汉语、基础汉语教学、自然语言处理以及机器翻译等领域的研究工作提供强有力的统计数据和科学依据。本文就以结构为标准的句型系统讨论了句型系统的建立，提出了以谓语为中心的句型成分分析和句型匹配相结合的汉语句型自动分析算法与策略。并给出了实验模型的具体算法、体系结构以及基于大约二十五万字真实语料的初步实验结果与分析。

二、句型系统的确定

本文以北京语言学院句型研究小组赵淑华教授的《现代汉语基本句型》为主要参考，采用结构特征为标准、建立了实验模型的汉语句型系统。其方法可归纳为以下四个步骤：

1. 确定句型成分。在六大句子成分中，根据是否影响句子的基本结构，确定主语、谓语、宾语、状语和补语为句型成分，定语为非句型成分。

2. 区别质成分和一般成分。主语和谓语为质成分，宾语、补语和状语为一般成分。

3. 根据质成分确定句型系统中高层次句型。具体情况如下：

根据句子是否具有主语和谓语，把句子分成主谓句和非主谓句；根据谓语成分的特征，把主谓句又分为名词谓语句、形容词谓语句、动词谓语句和主谓谓语句；把非主谓句分为无主句和独词句等。如图1所示。

4. 根据一般成分和关键字确定下位具体句型。如，根据状语的有或无以及状语的位置，可把动词谓语句分为“主语＋状语＋谓语”句、“状语＋主语＋谓语”句、“主语＋谓语＋宾语”句等句型；根据关键字“是”的有或无，可把动词谓语句分为“是”字句和非“是”字句等等。

此外，我们还设立了“杂类”句型，即本文标准句型系统中没有与之相对应的被测试句型，均称为杂类句型。我们的句型系统共有209个句型。

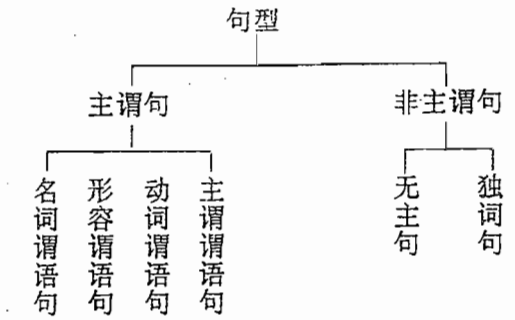

图1　句型系统中的高层句型

三、句型分析的策略

句型分析的策略可用图2表示如下：

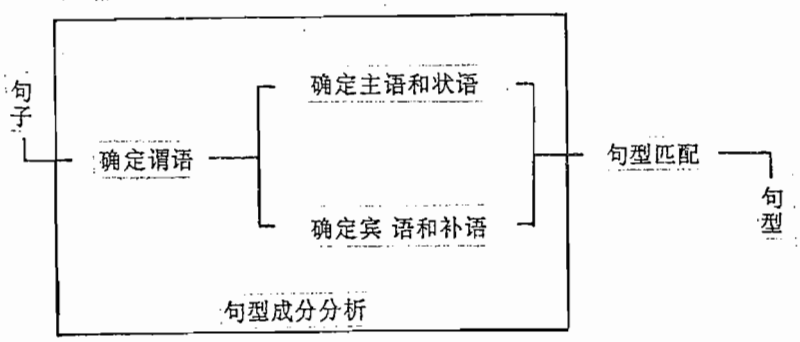

图2　句型分析的策略

句型成分分析确定句子的句型成分，从而确定句子的基本结构模式；句型匹配将被测试句子的结构模式与句型库中的标准结构模式比较，以确定句型类属。

本系统采用以谓语为中心的句型成分分析法。首先确定句子的谓语，然后根据谓语把句子分成两部分，对谓语前面的部分分析主语和状语，对谓语后面的部分分析宾语和补语。若句子中出现多个谓语(假设n个)，且这n个谓语组成连谓式，则这n个谓语把句子分成n+1个部分。对第一个部分，分析主语和状语，对第n个部分，分析宾语和补语，对其余n−1个部分分析状语、宾语和补语。若句中出现兼语式或小句宾语，则采用递归的策略进行分析。

四、以谓语为中心的句型成分分析算法的实现

句型成分分析算法是本系统的核心，其主要算法及特点是：

1. 自左向右四次扫描

第一次扫描，处理介词结构。

第二次扫描，处理粘着性较强、优先级较高的短语。

第三次扫描，分析并确定句子的谓语。

第四次扫描，分析并确定其它句型成分。

2. 词类驱动

上述四次扫描分别访问四个规则库，规则库中每条规则包含六个部分：

①驱动词类，在分析过程中，若被分析词的词类与规则的驱动词类一致时，唤醒该规则。

②左模式，描述被分析词的上文环境。
③右模式，描述被分析词的下文环境。
④操作，当上、下文环境同时满足时，执行该操作，并标记该词的语法与功能（见⑤和⑥）。
⑤语法标注。
⑥功能标注。

3. 采用改进的确定性算法

本文句型分析算法采用了 Marcus 确定性算法的主要思想，并对 Marcus 确定性算法作了以下两个方面的修改：

①用三个指针，即当前结点指针 p—cur、当前结点左侧指针 p—left 和当前结点右侧指针 p—right，取代 Marcus 确定性算法中的 3 单元缓冲器 3bf 和激活结点栈 ans。

②分析窗口大小等于句子长度。

五、实验系统及其体系结构

1. 设备与环境

Sun4/65 机器，UNIX 操作系统，C 语言编程。

2. 实验系统体系结构

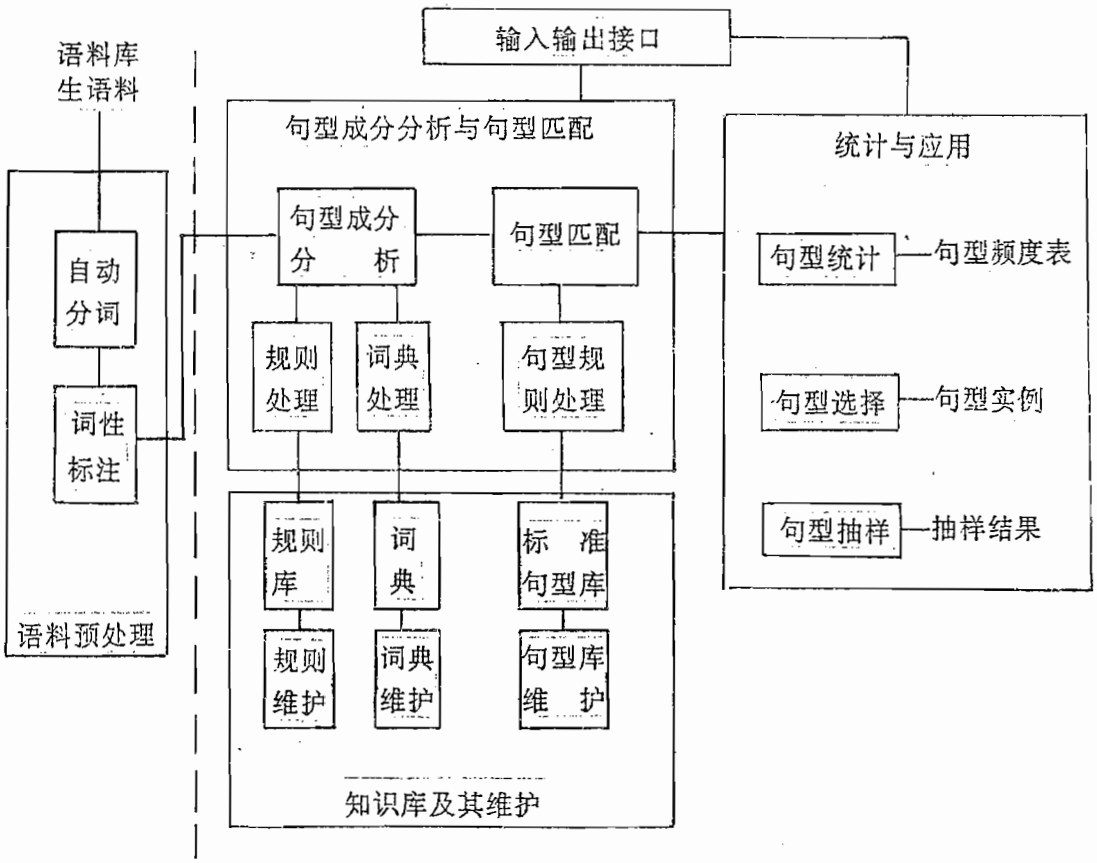

图 3　实验模型体系结构

3. 系统五大模块及其主要功能

1）预处理模块，其主要功能是对语料库中被抽样的大规模真实语料进行自动分词、自动词

性标注以及相应的人工纠错处理。(其中,自动分词采用北航的自动分词软件,自动词性标注采用清华大学黄昌宁、白拴虎等研制的自动词性标注软件。)

2)输入输出接口模块,包括菜单管理、语料输入、结果输出等功能。

3)句型分析模块,包括句型成分分析和句型匹配,是本系统的核心。

4)统计与应用模块,包括句型统计、句型选择和句型抽样。句型统计是本系统的主要目标,它计算出所测试语料的各种句型的频度统计数据,给出汉语句型的综合频度表或某一专业领域的汉语句型频度表;句型选择是从语料库中选择指定句型的实例,它为计算语言学和现代语言学研究提供一个实用性工具与手段;句型抽样是从所测试语料中按指定百分比随机地抽样出一定数量的句子,用于本系统的抽样检查。

5)知识库与知识维护模块,包括句型成分分析规则库、词典和标准句型库等知识库及其维护软件。

六、实验结果与分析

6.1 句型统计结果

本文对大约七十七万字的真实语料进行了测试。其中二十五万字左右的语料是经过精加工的(所谓精加工是指手工检查并改正自动分词、自动词性标注中的错误),我们对这二十五万字语料进行了分析与统计,得到了实验模型的句型频度表(限于篇幅,本文略)。

6.2 抽样检查结果

为了评价系统对大规模真实语料进行句型分析与统计的正确性与可信度,我们对分析结果采取随机抽样、人工检查的方法,从被统计的语料中按指定的百分比随机地抽样一定数量的句子,并人工逐句进行检查。

我们从二十五万字(约13000个句子)的语料中随机抽取了1039个句子,通过检查,得出其中101个句子分析有错误,句型分析的正确率为90%。这101个句子又可分为以下四种情况:

1)原始语料本身为病句,共6句,占分析有错误句子总数的6.0%。

2)自动分词有错误(人工未及纠正),共4句,占分析有错误句子总数的4.0%。

3)自动词性标注有错误(人工未及纠正),共23句,占分析有错误句子总数的22.7%。

4)本系统分析有错误,共68句,占分析有错误句子总数的67.3%。

七、结束语

本文目前仅完成了汉语句型自动分析与分布统计实验模型的研究。本课题与清华大学ZW大型通用汉语语料库的建设同步进行,将进一步考虑语料的分类与分布体系,提高语言的覆盖率,扩大精加工库存语料的规模,例如,至少达二至三百万字左右,并于近期(视语料精加工进度)完成首次汉语句型综合频度统计工作。

我们还将进一步考虑汉语句型系统中的问题,以及考虑以语义待征为标准的汉语句型系统及其分析与统计工作,以适应语言学研究的不同需求。

主要参考文献

[1]赵淑华,《谈80年代与90年代的句型研究》,《80年代与90年代中国现代汉语语法研究》,北京语言学院出版社,1992。

[2]北京语言学院句型研究小组,《现代汉语基本型》,《世界汉语教学》,1989,3.

[3]朱德熙,《语法讲义》,商务印刷馆,1984.

[4]徐　枢,《宾语和补语》,黑龙江人民出版社,1985.

[5]吴竞存、侯学超,《现代汉语句法分析》,北京大学出版社,1988.
[6]黄昌宁等,《汉语词性自动标注系统技术报告》,清华大学计算机科学与技术系,1992.
[7]孙茂松,《汉语句法分析的一种多扫描确定性算法及其在篇章理解中的应用》,清华大学计算机系硕士论文,1988.
[8]Mary Dee Harris,《Introduction to Natural Language processing》,Reston Publishing Company,1985.

An Approach to the Automatic Analysis and Frequence Statistics of Chinese Sentence Patterns

Zhensheng Luo Bixia Zheng

Dept. of Chinese Language & Literature
Tsinghua Univesity

ABSTRACT

The automatic analysis and frequence statistics of Chinese sentence patterns is an important foundamental problem. Based on a Chinese sentence pattern system of which structure is the standard, puts forward the strategy of Chinese sentence pattern analysis which includes sentence pattern analysis and sentence pattern matching. This paper concludes with the introduction of the overall structure of our system as well as the result of our experiment.

Keywords: Sentence Pattern Analysis, Frequence Statistics, Sentence Pattern Components, Sentence Pattern Matching.