基于德语语料库词性标注和统计方法的研究

姚天昉　　林　莉

（上海交通大学计算机科学与工程系，上海　200030）

Lexical Tagging The Research for the Method of the German Corpus-based and Statistics

Yao Tianfang　　Lin Li

(The Dept. of CS & E, Shanghai Jiao-Tong University, Shanghai 200030)

目前，以语料库为基础的机器翻译系统的研究是自动翻译研究的一个发展方向。本文论述了基于德语语料库的研究工作，介绍了一种德语语料词性标注方法以及基于词性标注的统计方法。初步实验证明了上述方法对德语语料标注和标注后的语料进行单词、词类、短语结构和句子的统计是正确和有效的。

本德语语料库的标记集由124个标记符号组成，每个标记由 2－8 个字符组成。其中，124个标记又可分为以下两大类：(1)词类标记：共计111个，由 2－7个字符组成。(2)　标点符号标记：共计13个 ，　由 2－8个字符组成。所定义的机器内部表示格式为词类标记、子类标记和属性标记。其中词类标记、子类标记用两个字节表示，属性标记也用两个字节表示。

语料标注工作是在已开发的“德汉题录机器翻译系统”的基础上进行的。该系统词法分析器是处理短语的，而本系统需处理句子。因此上述系统的词法分析器所能提供的语料信息是远远不够的。为了解决这一问题，采用了上述系统的分析机制与电子词典相结合的方法进行处理。以上述系统词法分析结果为面，而以电子词典提供的句法属性为纵深，构成一个立体知识库。根据不同的处理要求，利用这些句法属性信息，给语料库中的语料标注上标记。采用这一设计方法，语料标注的基本过程是：(1)词法分析器短语级词类分析，查询动词词典分析动词；(2)查询句法属性附加词典，补充词类的句法属性；(3)对应标记集，分析所得词类句法属性，将语料赋上标记。

由于对词性标注后得到的熟语料进行统计的内容涉及面大，在设计本系统过程中采用的基本思想是：从简单到复杂，从基础到上层，即从单词-->词类-->短语-->句子的步骤进行统计。系统统计的内容有：(1)单词出现频率统计；(2)词类出现频率和分布情况统计；(3)短语结构出现频率统计；(4)句型出现频率统计等。

根据上面介绍的方法，笔者已实现了德语语料库词性标注及统计管理系统。经测试，本系统运行是正确和有效的。系统的主要功能是：(1)对语料库中的生语料和熟语料进行管理。(2)完成生语料(包括短语、句子)的词类标记；(3)可查询单词和语料库中的所在句子(短语)，以及所在句子(短语)的类型；(4)　可统计某一语料文件中词类出现频率，以及该词类与其他词类的位置关系；　(5)　可统计某一语料文件中某一短语结构的出现频率，以及某一句型的出现频率；(6)可统计整个语料库中词类出现频率，以及该词类与其他词类的位置关系；(7)可统计整个语料库中某一短语结构的出现频率，以及某一句型的出现频率。