汉语语音识别用电子词典的自动建立方法研究

张树武　　黄泰翼　　徐波　　马斌
(中国科学院自动化研究所国家模式识别实验室)

摘要:
本文从统计的角度，分析和研究汉语语音识别系统中自然语词的构词原则，语词选择、语词属性收集方法及电子词典的自动建立方法。
关键词: 电子词典, 语音识别, 自然语言处理.

An Automatic Building Method of Electronic Dictionary Used for Chinese Speech Recognition

Zhang Shuwu , Huang Taiyi , Xu Bo,　Ma Bin
Institute of　Automation ,　Chinese Academy of Sciences
(Box 2728, Beijing　100080,　China)

ABSTRACT:
The paper , based on statistical　theory,　analysed　and　researched　a　series of　questions　about dictionary construction of Chinese speech recognition, including natural word defination, word selection, word attribute collection, etc. Of　them, some new algorithms or methods　have been offered firstly.　The method is not only available for automatic building of　Chinese speech recognition (CSP) using electronic dictionary , but also is useful of　constructing the other　information processing　using　electronic dictionary .
KEYWORD:　electronic dictionary (ED) , speech recognition, natural language processing.

一、　引　　言

在汉语大词汇量语音识别系统中，词典是系统的一个重要组成部分，电子词典的构造也是影响系统性能好坏的一个决定性因素。但是，电子词典的建立是非常复杂的，它包括构词法、词典条目的选择、属性分类与收集、词频及属性分布频度的统计等诸多问题。从语言处理的研究角度出发，人们已经提出了一些分词及构词原则、词条选择及规模、词条信息收集等有关电子词典建立的建议和方法。其中一些已应用在机器翻译、全文检索等语言处理应用系统中，取得了较好的效果。然而，要想建立一部通用的电子词典仍存在很多困难。这主要是由以下几个因素造成的：一、汉语分词及构词方法的模糊性：迄今为止，人们对汉语分词及构词方法众说纷纭。同时,由于汉语词汇构成的无限性，很难建立一个标准的汉语基本词汇集，所以电子词典的完备性及覆盖程度受到局限，另外，在一些语法和语义属性划分上也莫衷一是，使得电子词典的建立没有一个统一的依据及评判标准；二、应用场合的局限；对于不同的应用，对词典规模及信息条目的要求也各有差异；三、手工工作量较大：无论是词典条目的收集，或是词条的分类属性等都需要进行大量的人工收集与整理工作。由于在建立电

子词典上存在这样一些困难，因而在一定程度上也影响了一些具体应用系统的发展。

对于汉语语音识别来讲，在经历了汉语全音节识别的研究和系统研制之后，目前研究的热点已集中到了大词汇,连续语音识别的研究和系统研制上。这就需要一部好的词典的支持。针对具体的语音识别任务，对电子词典的要求综合起来可以归纳为如下几个方面：一、要求能基本覆盖现代汉语的常用词汇；二、构词原则应考虑汉语发音的自然韵律；三、一定的新词预测和收容机制；四、适当的分类属性及统计频度信息，借助一定的统计规律或规则，能有效地指导识别中声学判别和语言处理过程中的搜索和决策。目前，国内在识别用电子词典的建立上，由于受条件的限制，所采用的方法或是借用现成的词典或是手工整理一些较小规模的实验用词典。其主要弊端表现在于：一、缺乏对语音识别具体问题的考虑；在构词及词类划分上，照搬一些笼统的分词规范及分类原则，很少考虑语音识别自身的特殊性。二、信息不全；由于词典和汉语语料库是脱离的，所以词典中不能全面地反映识别中需要依据的一些统计特性，因而不能有效地指导识别过程中的搜索和决策；三、手工工作量大；由于缺乏词典的自动建立方法，所以难以做到使词典能够包容更全面的语言现象、拥有更多的统计属性以及更合理的语词类属信息。这样就造成了识别用电子词典的建立缺乏系统性和科学性。

本文提出一种基于语料统计的汉语语音识别用电子词典的自动建立方法。旨在以大规模真实文本语料为基础，借助一些原始的统计数据收集词条、统计频度、标引类属。研究和实现电子词典的自动建立技术。

二、语音识别用电子词典构词原则及语词自动收集方法

1、 语音识别用电子词典条目构成及其信息在识别中的作用

正如通常的词典，识别用电子词典条目也是由以下几部分构成的：

词条	词频	类别属性	统计信息	规则信息

虽然其表面形式与通常的词典别无二致，但是其中每个子项在语音识别中都有其新的内涵和作用。词条定义应考虑汉语的发音自然分段特征，其定义的好坏直接影响着声学建模及识别的性能。词频、统计信息及规则信息是语言模型建立的主要依据。它指导着汉语同音词的辨识及音节词到文本的自动转换。词典中词条、词频及一些统计信息是从大量真实语料中处理和统计得到的。类别属性采用语词类属自动标引算法获得。规则信息则是与一定的词法和句法规则知识库相连接。

2、 构词原则

词条构词原则是词典建立的关键。在识别中，词条的定义注重的是语词发音的自然韵律分割，而不是语法意义上的功能定义。所以应在通用的信息处理用现代汉语分词规范的基础上，结合语音识别自身的特点来制定构词原则。我们参照北航和北大提出的关于分词和文本切分的基本规范[1][2]，对其进行了一些必要的修改形成了"语音识别用汉语构词原则"。

3、 语词自动预切分

我们自行设计了一个“汉语语词自动预切分系统”，其目的是为了从大量语料素材中自动建立一个真实的粗加工词典。在此基础上，借助大量的统计数据，自动收集生词，同时对切分过程中因歧义而造成的统计偏差问题利用均衡算法进行自动平滑，对词典进行二次再加工。预切分方法采用双向扫描最大匹配法，结合“语音识别用汉语构词原则”，制定一些构词规则，在切分中进行辅助判断，因而它是一种基于规则的自动分词方法。切分用基础词典的建立也是影响切分质量的一个重要因素。我们采用了滚雪球的方式，即初始化以一个常用词典为基础，对二万字的语料进行预切分，然后依据构词原则，对已切分的语料人工检查调整，一个专门的新词回收机制从人工调整的语料中收集新词，并将其加入分词词典中，重复该过程对更大量的语料进行预切分，选择一定比率的语料进行人工调整使正确率基本恒定。切分系统测试结果见表3.1

切分集(字)	2万	5万	10万	50万	100万	500万	1000万	2000万
测试集(字)	2万	2万	2万	2万	2万	2万	2万	2万
正确率	94.4%	96.9%	97.3%	98.4%	99.1%	98.4%	99.2%	98.9%

表 2.1 汉语语词自动预切分系统测试一览表

4、 语词收集及词频、词对同现频度统计

在对大量语料进行语词切分的基础上，我们可以动态地从切分语料中收集语词。同时，借助这些已加工过的语料，可以实时地得到一些相关的统计信息。其中，词频和词对同现频度是最基础和最容易利用的信息。而且，它们的计算复杂度和空间占用相对较小，因而较容易实现。我们在对约二千万字的语料进行自动预切分后，收集到的统计信息见表3.2

语料规模	实际词次	词条数目	词对数目
1,897万字	10,135,827	59,040	2,429,930

表 2.2 统计信息

5、 生词收集及统计偏差平滑算法

在“汉语语词自动预切分系统”中，由于受分词词典词条数目限制和施加的有限数量构词规则的制约，在语料切分加工中,难免会遇到一些自然词被误切为单字词或较小单位的词。同时，还会有相当数量的由于切分歧义带来的词对同现统计偏差问题有待消除。根据概率原理，引用统计频度信息弥补或局部修正这些问题是一种有效的方法。这里我们介绍一组利用

词频和词对同现频度信息收集生词、平滑词对同现统计偏差问题的方法。

a). 生词收集策略：

设：词Wi,Wj有同现频度F；f(Wi),f(Wj)为Wi,Wj对应之词频；

最大定义词长为ML；

若：L(Wi)+L(Wj) ≤ ML ；L(Wi) 为Wi之词长；

且：$F/f(Wi)\geq\delta_L$ 或 $F/f(Wj)\geq\delta_L$ （δ_L 为与词长有关的门限值）

则：$WiWj$ 组成新词 Wij；修改相应的频度信息，加Wij到词典中。

词对	哥 们	倒 序	蹒 跚	港澳 台	良 莠	这 是
新词	哥们	倒序	蹒跚	港澳台	良莠	这是

表 2.3 生词收集示例

b). 统计偏差平滑算法：

平滑切分歧义而造成的统计偏差的原则是：依据大数定理 当corpus $\rightarrow\infty$ 时，

$$p(w_i)\approx f(w_i)/\sum_j f(w_j)$$

设语料足够大，则每个词的词频可以反映该词在语料中的分布情况。即使有少量的切分错误，但与整个语料数量相比，可以忽略不计。因而，可以用词频的分配比率平滑切分中的歧义问题，达到局部修正词对同现统计偏差的目的。

设：具有同现频度f_0的词对 $w_{i0}w_{j0}$ 存在m个歧义切分 $w_{i1}w_{j1},\ldots,w_{im}w_{jm}$，对应同现频度分别为 $f_1,\ldots,f_m$；

则修正词对同现频度 $f_l'=\lambda_1 f_l+\lambda_2 f_l\times(f(w_{il})\times f(w_{jl})/\sum_{n=0}^{m}(f(w_{in})\times f(w_{jn})))$

这里： $(l=0\ldots m \quad and \quad \lambda_1+\lambda_2=1)$

三、 词典语词分类信息及语词类属自动标引算法

1、 词典语词的语法分类

关于语词的语法分类问题，无论是词法层次上的词性划分或是语义环境下的义项分类，目前尚不统一。对于词性的分类，清华大学计算机系和山西大学计算机系曾提出了108 个细类的词性划分方法[3] [4]，北京大学计算语言学研究所也提出了 31 类的词性划分原则[2]。其分类原则各有特色，但从语音识别的角度来考虑语词的分类，也都存在一些问题。在遵循语词划分的通用原则（即完备性、确定性、交叉性最小和分布性准则）的前提下，还必须注重如下一些问题：

a). 同音、近音词的分布特征：要求同音、近音词尽可能的有不同的类属关系；

b). 分类粒度：在语音识别过程中，统计语言模型的建立很大程度上依赖于词类的数目，从特征提取和统计的角度来考虑，分类过细，容易造成语词的兼类增多和统计数据的

稀疏；而分类过粗又容易造成语词特征的重叠过多，使语词分类特征模糊,难以区分。对于 Bi-Class模型，词类数在1000以内都是可实现的。而对于Tri-Class模型的建立，类型数应限制在100-200范围内。否则难以实现。

综合词法和语义的语词分类原则，我们希望建立一个适合汉语语音识别的Tri-Class 语言模型。由此，在借鉴国内一些较好的词性分类方法的基础上，结合语音识别的特定问题对他们进行改进，我们得到了一个 121类的语词分类体系，并在一定的语料基础上，对其进行了标注和统计。

2 、词典语词类属自动标引算法

在已经人工对现代汉语进行词法、语义分类的基础上，如何对词典中词条标注类属并得到其相应的统计信息是摆在我们面前的一个困难问题。以往采用的有些是手工标注的方法。这种方法人工工作量非常大而且不能得到相关的真实语言环境的使用频度信息；另外一种方法便是滚雪球的方法。它是一种半自动的标注方法。这种方法在一定程度上提高了标注的自动性，但仍然需要相当多的人工工作和时间耗费，很难扩展到巨量语料的实际处理中。自动聚类的方法也已经有许多人提出和实验，然而，它有一个缺陷就是聚类的结果很难赋予一定的词法或语义意义上的内涵。这样很难再做更进一步的句法基础上的语言分析和处理。这里,结合实际的语法分类体系和自动聚类方法，我们提出一种语词类属自动标引算法。其主要思想是，以一个小规模的有语法意义的类别标注的熟语料为初始训练集，结合巨量语料基础上的词频及词对同现频度信息,对词典中词条的类属关系进行模式评判并对其频度作概率估算。

为了发现一个未知的映射 G：v→G(v)，我们可以用最大可能准则：

$$\max_{G}\sum_{n}\log P(w_n|G(w_{n-1})) = \max_{G}\sum_{w}N(vw)\log P(w|G(v))$$

这里概率P(w|G(v))是未知的，但对于每一个固定的映射 G，P(w|G(v)))的最大可能估值可以为：

$$\tilde{p}(w|g)\approx N(gw)/N(g)$$

这里N(gw),N(g)分别对应二元和一元的频度值这样我们就可以采用如下的算法对汉语语词作自动类属标引。

1. 初始化：小规模（ 10% 左右）已标注类属及频度的词库，词频及词对同现频度,
   词类数tm;
2. 对未知类属的词条V，选择映射 G： $V\to g_i$ (i=0...tm-1)
   对每个与V有后继关系的词w_t，计算 $N(vw_t)\log\tilde{p}(w_t|g_i)$
3. 评价映射g_i,确定最优等价类 $G(v)=\arg\max_{g}\sum_{w}N(vw)\log\tilde{p}(w|g)$；
   赋词V 类别标志G(v)；
4. 循环执行2，3 直至所有词条均已标引类属标志。

目前，这方面的工作正在进行中。

四、 结　　论

电子词典是语言信息处理的一个基础工具。基于大量真实语料统计基础上的词典自动构造方法是建立该工具的一条有效途径。它的最大特点在于其真实完备性和便捷性。基于语料统计的方法能够真实反映和全面覆盖汉语的各种语言现象，同时，自动建立技术借助计算机自动完成词典条目的收集与整理，基本摆脱了人工约束而造成的时间耗费。或许，采用该方法在实际的词典构造过程中会产生局部数据的不精确。但是，相对于足够充足的统计语料和一些平滑算法，这些问题是可以忽略和克服的。实际应用表明，本文所提出的电子词典自动建立方法对汉语语音识别中声学判别和语言模型的建造起到了非常积极的作用。同时，对于别的语言信息处理用电子词典的构造，它也不失为一种有效的方法。

参考文献

[1] 刘　源等　<<信息处理用现代汉语分词规范及自动分词方法>>
清华大学出版社，1994,6.

[2] 俞士汶等　"现代汉语文本切分和词性标注规范（草案）"
北京大学计算语言学研究所技术报告，1994，12.

[3] 白拴虎　"现代汉语词性自动标注方法研究" 清华大学硕士论文，1992.

[4] 赵　军　"现代汉语词性标注算法研究" 山西大学硕士论文，1992.

[5] H. Ney 、U.Essen " On smoothing Techniques For Bigram-Based Natural Language Modeliing", *In Proc. of ICASSP-91, pages 825-828, Toronto,Sept. 1991.*

[6] E.Charniak " Statistical Language Learning " *MIT Press,1993.*

[7] K.W.Church and R.L.Mercer, "Introduction to the Special Issure On Computational Lingustics Using Large Corpora", *Computational Lingustics Vol.19 No.1,pp1-24.*