利用遗传算法实现词类标记集的优化①

孙宏林[1,3] 陆勤[2] 俞士汶[1]

(1. 北京大学计算语言学研究所 北京 100871; 2. 香港理工大学电子计算学系 香港 红勘
3. 北京语言文化大学语言信息处理研究所 北京 100083)

摘要: 过去词类标记集的选择主要基于专家的经验知识, 缺乏自动或半自动的方法来辅助这一过程。本文提出了一种利用遗传算法来搜索优化的标记集的新方法。这种方法可以在一个候选标记集集合中自动搜索一个最优或较优的标记集, 并可根据应用的需求调整参数以适应特定任务的需求。实验表明: 遗传算法为标记集的选择提供了一种系统的有效的辅助手段。

关键词: 词性标注; 词类; 标记集; 遗传算法
中图分类号: TP391.1

Using Genetic Algorithms for Optimizing Part-of-Speech Tagset

SUN Hong-lin[1,3] LU Qin[2] YU Shi-wen[1]
(1. Institute of Computational Linguistics Peking University Beijing 100871;
2. Department of Computing, Hong Kong Polytechnic University;
3. Center for Language Information Processing, Beijing Language & Culture University Beijing 100083)
E-mail: sunhl@blcu.edu.cn; csluqin@comp.polyu.edu.hk; yusw@pku.edu.cn

Abstract: POS tagset selection in the past was mainly done by experts using human knowledge manually, since there is no automatic or semi-automatic way to assist the selection process. This paper proposes a novel method to search for an optimal POS tagset using genetic algorithms (GA). The experiment shows that GA provides an efficient optimization of POS tagset and allows for the adjustment of parameters according to user requirement. It provides a systematic way to help people in making an intelligent choice on the selection of a tagset.

Keywords: POS tagging; word class; POS tagset; genetic algorithm

---

① 收稿日期: 2000－05－23
基金项目: 973 项目(G1998030507－4); 国家自然科学基金项目(69973005); 香港理工大学研究基金
作者孙宏林, 男, 1966 年生, 在职博士研究生, 副研究员, 主要研究方向为计算语言学. 陆勤, 女, 1960 年生, 博士, 副教授, 主要研究方向为中文计算系统、信息检索和数据库系统. 俞士汶, 男, 1938 年生, 教授, 博士生导师, 主要研究方向为计算语言学.

一、引言

任何自然语言的词汇都是一个很大的集合，一部中等规模的词典一般也有几万个词项。这使得直接用词来描述语言的结构和建立语言模型都十分困难，甚至难以实现。因而有必要对词汇进行抽象，即对词汇进行分类。词汇分类的依据是词在句法、语义等属性上所具有的共性。至于到底分为多少类，并无一定之规，因为根据抽象程度的不同可以有各种不同的结果。在过去十多年中，词性自动标注技术取得了很大的进展，统计方法的运用使得构造一个高性能的自动词性标注系统变得相对容易，现在世界主要语言几乎都有不少这样的标注系统可供使用。但是，所有的标注系统所使用的词类体系都是依靠人的知觉或经验产生的[4]。由于对词类体系的合理性或合适性缺乏客观的评价，我们在选择或定义一个词类标记集时就缺乏科学依据。同时，由于不同的应用系统对词性知识有不同的需求，对于特定任务到底选择什么样的词类体系也是一个问题。这些问题都属于词类标记集的评价和选择问题。我们曾讨论过词类标记集的评价问题[7]，本文将讨论词类标记集的选择问题。

选择词类标记集的一个直接的简单方法是：对几个候选的词类标记集，用同一标注系统分别进行训练和标注实验，根据实验的结果从中选择一个最合适的标记集。但这种方法的缺点是：候选标记集只能是有限的几个，而可能的标记集几乎是无限的。而且对每一个标记集都需要经过语料标注、训练、标注评测的过程，其中都需要人的介入，工作量很大。

本文针对这一问题，提出了一种基于遗传算法的词类标记集优选方法，该方法可以在一个候选标记集集合中自动搜索一个最优或较优的标记集。该方法的基本思想是：首先定义一个包含一个最大标记集和最小标记集的词类层级体系，然后选择一个语料库，用最大标记集进行标注，然后对介于最大标记集和最小标记集之间的可能的标记集用一个评价函数进行度量，利用遗传算法从中选择最优或较优的标记集。标记集的评价函数综合考虑两个因素：(1)标记集的歧义度，用语料库中每个词例(token)可能的标记数的均值来表示；(2)标记集的信息量，用标记集的熵(entropy)来表示。这一方法的目的是帮助人们确定哪些标记应该包括在标记集中，哪些标记应该排除在外。

本文第二节简单介绍遗传算法的基本思想，第三节给出问题的表示，第四节详细讨论标记集评价函数，第五节给出实验结果，最后是全文总结。

二、遗传算法

遗传算法是根据自然进化中“适者生存”的原则提出的一种概率型优化方法[5]。这种方法适合于具有很大搜索空间的优化问题，下面简要介绍遗传算法的主要思想，详细情况可参考文献[3, 6]。

在遗传算法中，有一个包含个体 x_i 的群体 $P=\langle x_1, \cdots, x_n\rangle$，个体 x_i 代表问题的一个解，群体就是问题的一些解的集合。某一评价函数 $F(x_i)$ 被用来对这些候选解进行评价，目标是优化该评价函数(搜索该函数的最大值或最小值)以解决给定的问题。这些候选解通常用位串(bit string)的形式表示。把解表示为位串的过程称为编码，编码后的每个位串就表示一个个体，即问题的一个解。评价函数用以评价群体中每个个体的适应度(fitness)。在算法的每次迭代(借用生物学术语称作一代)中，评价函数按照优化标准对每个个体 x_i 进行度量，计算其适应度 f_i，适应度最高的个体被选择允许再生，以产生新的一代。下面是算法的形式化描述：

```
{
gen=0;        //代号初始化
initialize();    //初始化第一个群体
evaluate(P(gen));    //P(gen)是  第 gen 代的群体
do  {
gen=gen+1;
generate new  population P(gen) from P(gen-1) through select, crossover, and
        mutation; //通过选择、交叉和变异从 P(gen-1)生成新的群体 P(gen)
evaluate(P(gen));    //  评价 P(gen)
}  while(gen<=maxgen)
}
```

遗传算法中的再生过程主要包括三个遗传算子:(1)选择;(2)交叉;(3)变异。在选择过程中,适应度高的个体被直接复制到下一代群体中。适应度越高的串,产生后代的概率就越高。在交叉过程中,两个串的部分位(称为基因)进行交换从而产生一个新串作为下一代的个体。变异用来随机地改变位串中的某些位。交叉和变异的使用都有一定的概率,分别称为交叉概率和变异概率。

三、问题的表示

词汇的分类可以用图 1 所示的树型结构来表示。树中的一个结点对应于一个词类标记,树中若干个结点构成一个标记集。在遗传算法中,我们用一个位串来表示一个标记集。分类树中除根结点以外的所有结点可以用一个位来编码,每一位有两种可能的值:1 表示该结点在分类树中存在,0 表示该结点在分类树中不存在。如果最大结点数为 n,即最大标记集 M 中有 n 个标记,那么,n 个标记构成的所有可能的标记集数是 2^n-1 个(M 的所有可能子集的数量,不考虑空集),当然由于树结构的约束,实际的数目并没有这么多,但可能标记集的数目仍然是很大的。我们的目标就是在一个如此大的空间中搜索一个最优或较优的标记集,即找到适应度最高的位串。

由于分类树是树型结构,因此位串中各个位的值并不是相互独立的。如果一个结点的值为 0,那么它的下位结点必为 0。所以,在样本产生过程中违反以下限制的位串将被删除:

如果任一子结点非空,则父结点不能为空。

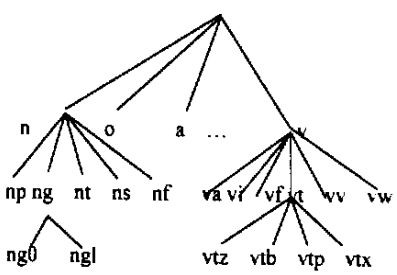

图 1　词分类树

例如,附录中的标记集有 107 个标记,分为三个层级(根结点为空),如图 1 所示。在第一层上有 15 个结点,第二、三层上分别有 41 和 51 个结点。整个标记集可以用图 2 所示的位串来表示,该位串的长度为 107。

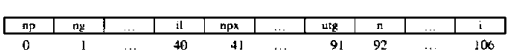

图 2　标记集的表示

例如,在图 2 中,第 0 位是“np”(专有名词),如果第 0 位的值为 0 则表示标记集中没有此标记,

否则表示有此标记。

四、评价函数

对群体中的每一个个体,要计算其适应度值作为从前一代“进化”到下一代的过程中选择的标准。在标记集优选问题中,需要考虑两个参数:(1)标记集对词性标注系统准确率的影响;(2)标记集为下一步处理所提供的信息的丰富程度。词性标注追求的目标是:词性标记包含的信息越丰富越好,同时,自动标注的准确率越高越好。显然,这两个目标是有矛盾的:一方面,词类分得越粗,自动标注的准确率就越高,但词类标记提供的信息量就越少;另一方面,词类分得越细,给高一级处理提供的信息也就越丰富,但自动标注的准确率就越低。因此,我们需要在词类标记的信息量与自动标注的准确率之间找到一个最佳的平衡点。以下将分别讨论这两个参数,并在此基础上提出一个综合这两个参数的对词类标记集的评价函数。

4.1 标记集对标注准确率的影响

为了度量标记集对标注程序准确率的影响,我们可以使用直接的方法,即用不同的标记集分别进行语料的标注、标注系统的训练和标注评测。如前所述,这种方法的缺陷是:考察的标记集数量有限,而且费时费力。所以我们应该寻找间接的方法。我们曾经通过实验证明,词类标记集的动态歧义指数(语料库中每个词例可能的标记数)与标注系统的准确率成正比例关系[7]。这一点是比较容易理解的,因为词性标注系统的主要工作就是进行词类的消歧,语料中每个词例的平均歧义指数越高,那么标注系统面临的消歧困难也就越大,因而标注准确率也就越低。标记集 TS 的动态歧义指数(dynamic ambiguity index,DAI)可以通过下面的公式从标注好的语料库中很容易统计得到:

$$DAI(TS)=\frac{\sum_{i=1}^{N} AMB(W_i)\times f(W_i)}{\sum_{i=1}^{N} f(W_i)} \tag{1}$$

这里,N 是语料库中词型(word type)的数目,$f(W_i)$ 是词 W_i 在语料库中的出现次数,$AMB(W_i)$ 表示 W_i 可能的标记数。

在计算一个标记集的 DAI 之前,必须有用该标记集标注好的语料库。在遗传算法的进化过程中,需要对大量的标记集进行评价,不可能为每一个标记集准备一个标注版本。但是,如果语料库事先用最大标记集进行了标注,就可以利用最大标记集到其子集的多对一(有些标记是一对一)的映射关系自动得到用新标记集标注的语料库版本。

4.2 标记集的信息量

词的分类可粗可细。分类越粗,得到的类就越少,同时给后一步处理提供的信息也就越少。以名词的分类为例,我们分别考虑以下三种情形:

(1)对名词不作进一步分类;

(2)把名词分成两类:普通名词和专有名词;

(3)在(2)的基础上,进一步把专有名词分成四个子类:人名、地名、机构名和其他专名。

对于命名实体任务[2]来说,很显然以上第三种情形可以提供更多的信息。

但要想准确地估计一个标记集所包含的信息量却并非易事,一个标记所包含的信息在下一步的处理中到底起什么作用,跟整个系统的理论体系有关,很难给出一致的度量。我们这里

所说的信息量跟标记的具体属性含义无关，它只是一个一般性的定量的度量，而非定性的描述。

标记集所包含的标记数目可以作为一个度量标准，但它是静态的，不能反映每个标记对系统的实际贡献，因为不同的标记在文本中出现的概率不同。因此必须考虑标记的实际出现概率。从这个角度来说，熵(entropy)是对标记集信息量的一个合适的度量。

从信息论的观点看，文本中的词性标记序列可以看成一个随机过程，标记集是一个随机变量，标记序列中的特定标记就是这个随机变量的值。熵是随机变量平均不确定性的度量，不确定性越高，熵值就越大。标记的可能的取值越少，我们在猜测序列中标记的出现时得到的信息就越少；相反，标记可能的取值越多，那么我们得到的信息就越多。标记集 TS 的熵可以通过下面的公式来计算：

$$H(TS) = -\sum_{t \in TS} P(t)\log P(t) \tag{2}$$

这里，t 是标记集 TS 中的一个标记，$P(t)$是 t 的出现概率，对数以 2 为底。

4.3 评价函数

在进化过程中，对每一个标记集都需要给出一个适应度的度量以指导再生过程。优化的过程就是搜索最大或最小评价函数值的过程。

正如在第一节中所讨论的，两个相关的参数(准确率和信息量)成反比关系。事实上，我们通常要根据特定任务的需求在这二者之间找到一个平衡点。不同的任务对词性标注有不同的要求，在某些情况下，准确率是优先考虑的因素，而在另一些情况下标记的信息量则更重要。所以，应该给这两个参数赋以不同的权值，而且权值可以根据应用的要求进行调整。

我们把标记集优选的过程看成一个适应度值最大化的过程。DAI 和适应度值成反比例关系，熵和适应度值成正比例关系，所以，标记集 TS 的适应度可以定义如下：

$$Fitness(TS) = \frac{\lambda_1 \times \left(\frac{1}{D_T} - \frac{1}{D_{\max}}\right)}{\frac{1}{D_{\min}} - \frac{1}{D_{\max}}} + \frac{\lambda_2 \times (H_T - H_{\min})}{(H_{\max} - H_{\min})} \tag{3}$$

这里，D_T 和 H_T 分别表示 TS 的 DAI 和熵，$D_{\max}$和 $D_{\min}$分别是最大标记集和最小标记集的 DAI，$H_{\max}$和 $H_{\min}$分别是最大标记集和最小标记集的熵值。这里，$D_{\max} > D_{\min}$，$H_{\max} > H_{\min}$。λ_1 和 λ_2 是两个参数的权值，$\lambda_1+\lambda_2=1$。当优先考虑标注系统的准确率时，可以给 λ_1 赋一个大于 0.5 的值，如果优先考虑标记集的信息量，则可以给 λ_2 赋一个大于 0.5 的值。当 $\lambda_1=\lambda_2=0.5$时，两个参数的优先级相等。实际的取值可以在实验中调试。显然，λ_1 越大，结果的标记集就越小，λ_2 越大，结果的标记集就越大。

五、实验结果

5.1 实现

实验中所用的标记集是一个汉语语料库所用的标记集[8]。该分类系统是一个层级体系，分类树中共有 107 个结点，我们把这 107 个结点构成的标记集定义为最大标记集。其第一级包含 15 个标记，我们把它定义为最小标记集，这 15 个标记包含在进化过程中生成的任一标记集中，因此共有 92 个变量。我们用长度为 92 的位串来表示这 92 个结点，变异率为 1/92，遵循这样的原则：位串中的每一位有 1/L 的变异概率(L 为位串中位的数量)[11]。交叉率为0.8，交叉采用了简单的单点交叉。

为了评价每一个标记集，我们使用了以上语料库中的一个子集作为测试语料，其中包含157篇选自《人民日报》的文章，共有203,499个词例。这些语料用以上的最大标记集进行了标注，并经过细致的人工校对。因为最大标记集和生成的标记集之间有多对一的对应关系，该语料库可以自动地转化为用其他标记集标注的版本，所以在程序运行过程中不需要人的介入。

在评价函数中，选择 $\lambda_1=\lambda_2=0.5$，给标注的准确率和标记集的信息量以相同的权值。表1给出了几个标记集的适应度，我们可以发现，在运用公式(3)中，D_{max}和 D_{min}的值分别是2.42和1.49，H_{max}和 H_{min}的值分别是4.54和2.83。表1中的Tagset 1是最大标记集，它包含107个标记，在所有的候选标记集中DAI值和H值最大。对于最大标记集，因为 $D_T=D_{max}$且 $H_T=H_{max}$，所以其适应度值等于 λ_2。Tagset2是最小标记集，它在所有候选标记集中DAI值和H值最小。因为 $D_T=D_{min}$且 $H_T=H_{min}$，所以其适应度值等于 λ_1。

表1　部分标记集的适应度

Tagset	DAI	H	Fitness(%)	标记集说明
1	2.42	4.54	50	最大标记集，107个标记
2	1.49	2.83	50	最小标记集，15个标记
3	1.74	3.87	61.83	所有的第二集结点都不为空，所有的第三季结点皆为空
4	1.62	3.64	63.43	vt^*和 vw^*为空，其余结点皆不为空

5.2　实验结果

实验共生成了100代，表2给出了其中部分结果。在表2中，第一栏是代号，第二栏是群体

表2　部分实验结果

代	代最小适应度(%)	代最大适应度(%)	代平均适应度(%)	总的最大适应度(%)	标记集的位串表示(hex)
1	48.67	56.24	52.74	56.24	acfbc04b5f2881000574574
2	48.92	67.91	55.83	67.91	e5e3fa974ad7880005b90dc
6	53.44	70.67	67.52	70.67	e1e3ea974ad7000005790dc
16	53.44	70.67	70.21	70.67	e1e3ea974ad7000005790dc
23	54.40	71.22	70.28	71.22	b1e3ead75af4000001a9014
50	54.66	71.22	70.95	71.22	b1e3ead75af4000001a9014
100	54.66	71.22	70.89	71.22	b1e3ead75af4000001a9014

中最小的适应度，第三栏是群体中最大的适应度，第四栏是群体中所有标记集适应度的均值，第五栏是已获得的最佳适应度，最后一栏是具有最佳适应度的标记集的位串表示(用16进制表示)。标记按照附录中的序号编码，第0号标记对应于位串中的第0位，第1号标记对应于位串中的第1位，其余依此类推。附录中序号为92—106的15个标记是分类树中的一级标记，如前所述，我们规定这些标记包含在生成的任一标记集中，所以不在位串中编码，因此位串的长度为92。

图3显示了在前50代中适应度的变化情况，横轴上的11个点分别对应于代号1，5，10，15，…，50。图中，横轴表示代号，纵轴表示适应度。上面的一条曲线表示最佳适应度的变化情况，下面一条曲线表示平均适应度的变化情况。

从图3中我们可以发现，在进化过程中所生成的群体的最大适应度和平均适应度都随着代号的增加而单调增加，而且在第23代基本达到收敛。

具有最高适应度的位串是 b1e3ead75af4000001a9014，在第 23 代就得到。与该位串对应的标记集所包括的标记在附录中的序号后加了星号。该标记集包含 51 个标记，在第一、二、三级上分别有 15、25 和 11 个标记。下面以名词为例对得到的标记集进行简单的讨论。

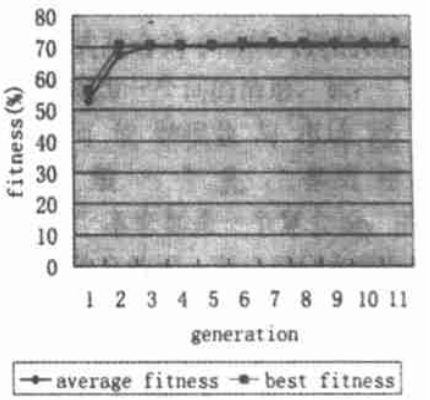

图 3　进化过程中适应度的变化情况

在最大标记集中，在第一级上的名词在第二级上分为两类：普通名词和专有名词。专有名词在第三级上进一步分为五类，普通名词在第三级上又分为两类：普通名词和离合名词。其中离合名词是指一些动宾式复合词中间插入别的词语之后，原来在复合词内部只作一个构词成分的语素不得不独立成为一个词的情形，如：

如何 使 种植业 与 市场 接 上 轨

……向　老师们　敬　一　个　礼　，然后　他　深深　鞠　了　一　躬

“接轨”、“敬礼”和“鞠躬”本来都是一个复合词，由于中间插入了别的成分，使得“轨”、“礼”和“躬”都变成了独立的词。但这些词和前面的动词显然有很强的依赖关系，有的离了前面的动词之后根本就不可能存在，如“躬”。这些词和一般的名词显然不同，如果把这些词单独加上标记，对下一步的句法和语义分析显然是有益的，但这样会增加更多的兼类现象，如上例中的“礼”在上面的语境中它是一个离合名词，但在别的情况下它也可以是一般的名词。是不是需要这一类要根据任务的要求综合考虑准确率和信息量两个因素来作决定。在上面的实验中，我们设 $\lambda_1=\lambda_2=0.5$，即给予两个参数相同的权重时离合名词就没有被选中，在另外的改变参数权重的实验中我们发现，当 λ_2 大于0.7时，离合名词就总是被选中。而当 λ_2 小于或等于 0.7 时，离合名词总是不被选中。

在上面的结果标记集中我们发现，原来的专有名词分为五个小类，在结果标记集中保留了其中的四个。这主要是因为专名跟其他类兼类的现象比较少，因此尽管专名分得这样细，但引起的词性歧义并不多。不过，这并不说明专名的归类就很容易了。事实上，专名的细分类并不是一种语法分类，而是语义分类。专名的归类更多的要依靠词性标注以外的技术来解决。

显然，所得到的标记集并不能直接应用，因为它在某些方面缺乏系统性。比如，在词缀中，它选择了名词前缀(kh)、动词后缀(kv)和可能中缀(kp)，但没有选择名词后缀(kn)。从词缀的系统性考虑，这显然是不合理的，所以有必要对得到的标记集进行适当的调整。在实际的标记集选择过程中，往往要经过多次实验，选择不同的 λ_1 和 λ_2，看在不同权值下标记集的变化情况，据此决定具体标记的取舍。

六、结论

本文提出了一种利用遗传算法优选词类标记集的方法。该方法的目的是帮助人们根据任务的需求选择一个最优或较优的标记集。该方法自动生成可能的标记集，并对每一个标记集给出一个评价函数值，在一个很大的标记集候选集中搜索评价函数值最大的标记集。该评价函数考虑了两个参数：标记集的歧义度和信息量。参数的权值可以进行调整。

本文给出了一个应用遗传算法进行标记集优选的实用方法。首先，在一个树型的分类体

系中定义一个最大标记集和最小标记集。然后选择一个语料库,用最大标记集进行标注,并经过人工校对。在进化过程中,当一个新的标记集生成时,该语料库自动转化为用新标记集标注的版本。我们的实验表明:该方法对词类标记集的优选提供了一种有价值的辅助手段。

参 考 文 献

[1] Bäck T. Optimal mutation rates in genetic search. In: Proceedings of the 5th International Conference on Genetic Algorithms (ICGA'93). Morgan Kaufmann, 1993, 2—9

[2] Chinchor N. MUC—7 Named Entity task definition. In: Proceedings of the Seventh Message Understanding Conference(MUC—7). 1998, http://www.muc.saic.com/proceedings/muc_7_toc.html

[3] Goldberg D E. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, 1989

[4] Halteren H. Syntactic Wordclass Tagging. Kluwer Academic Publishers, 1999

[5] Holland J H. Adaptation in Natural and Artificial Systems. University of Michigan Press, 1975 (Second edition: MIT Press, 1992)

[6] Mitchell M. An Introduction to Genetic Algorithms. MIT Press, 1996

[7] Sun H, Yu S, Lu Q. Evaluations on Part-of-speech Tagset. In: Proceedings of the 5th Natural Language Processing Pacific Rim Symposium(NLPRS(99). Tsinghua University Press, 1999, 25—31

[8] 孙宏林等."现代汉语研究语料库系统"概述. 见:罗振声、袁毓林. 计算机时代的汉语和汉字研究. 北京:清华大学出版社, 1996

附录 实验中使用的最大标记集及获得的优化标记集

序号	标记	说 明	序号	标 记	说 明	序号	标记	说 明
0*	np	专有名词	36*	kp	可能中缀	72*	mam	中助数词
1	ng	普通名词	37	in	名词性成语	73	mab	后助数词
2*	nt	时间词	38*	iv	动词性成语	74*	qns	个体量词
3*	ns	处所词	39	id	副词性成语	75	qnu	集合量词
4	nf	方位词	40*	il	连接语	76*	qnk	种类量词
5	ag	一般形容词	41*	npx	汉人姓氏	77	qng	量词"个"
6	az	状态词	42*	npm	人名	78	qnm	度量词
7*	ab	区别词	43*	npu	机构名	79*	qnc	不定量词
8*	va	助动词	44	nps	地名	80	qnt	临时量词
9*	vi	系动词	45*	npr	其他专名	81	qv0	动量词
10*	vf	形式动词	46	ng0	普通名词	82	qvt	临时动量词
11	vv	"来/去"+ VP	47	ng1	离合名词	83	usd	助词"的"
12	vt	动词用作体词	48	ag0	普通形容词	84	usz	助词"之"
13	vw	动词用作谓词	49	agz	形容词作状语	85	usy	助词"似的"
14*	mg	一般数词	50	agb	形容词带宾语	86	usi	助词"地"
15*	ma	助数词	51	ags	形容词作主宾语	87*	usf	助词"得"
16*	qn	名量词	52	agx	形容词作 NP 中心语	88	uss	助词"所"
17*	qv	动量词	53	vtz	动词作主语	89*	utl	助词"了"
18*	qt	时间量词	54	vtb	动词作宾语	90	utz	助词"着"
19	ra	代词作定语	55	vtp	动词作定语	91	utg	助词"过"
20*	rs	代词作主宾语	56	vtx	动词作 NP 中心语	92*	n	名词
21	rp	代词作谓语	57	vw0	动词不带宾语	93*	a	形容词
22*	rd	代词作状语	58	vwn	动词带 NP 宾语	94*	v	动词

23	pg	一般介词	59	vwv	动词带 VP 宾语	95 *	m	数词
24 *	pa	介词“把”	60	vwa	动词带形容词宾语	96 *	q	量词
25 *	pe	介词“被”	61	vws	动词带小句宾语	97 *	r	代词
26	pz	介词“在”	62	vwd	动词带双宾语	98 *	p	介词
27 *	db	否定前副词	63	vwj	动词带兼语	99 *	d	副词
28	dd	程度副词	64	vwc	动词作补语	100 *	c	连词
29 *	dr	其他副词	65	mgx	基数词	101 *	u	助词
30 *	us	结构助词	66	mgw	位数词	102 *	y	语气词
31 *	ut	时态助词	67	mgg	概数词	103 *	o	拟声词
32	ur	其他助词	68	mgm	数量词	104 *	e	叹词
33 *	kh	前缀	69	mgh	数词“半”	105 *	k	词缀
34	kn	名词后缀	70	mgo	数词“零”	106 *	i	成语
35 *	kv	动词后缀	71 *	maf	前助数词			

[消息]

中科院自动化所模式识别国家重点实验室正式成为国际语音翻译研究协会核心成员

语音翻译(Speech-to-speech Translation)是近几年来国际上发展迅速的热点研究领域，为了推动语音翻译技术研究的快速发展，由美国CMU(Carnegie Mellon University)、日本ATR、德国Karlsruhe大学等单位联合发起，于1991年正式成立了国际语音翻译研究协会(Consortium for Speech Translation Advanced Research，简称C-STAR)。到目前为止C-STAR已经历了三个发展阶段，今年10月正式转为第三阶段C-STARⅢ。C-STAR发展阶段的提升，标志着国际上语音翻译技术的不断进展。

C-STAR是目前国际上语音翻译研究领域最具权威性的学术机构，1993年和1999年C-STAR曾两次组织进行了语音翻译系统的国际性联合实验，翻译语种主要是英语、日语和德语，在国际上产生了很大影响。协会的会员分为核心成员(Partner Member)和一般联系成员(Affiliate Member)两种。核心成员之间具有更密切的合作关系，并通过国际互联网或国际长途电话进行双向多语种之间的语音翻译联合实验。由一般联系会员转为核心成员手段十分严格。协会不仅要对申请单位的技术力量和研究水平、软硬件设施等进行严格的考察，而且要对其经费状况和进行国际间合作交流的可行性等进行必要的考察，通过后方允许答辩，答辩通过后才能成为正式的核心成员。截止到C-STARⅡ，该组织已拥有十二个国家的20个著名大学、研究机构和企业作为会员，其中，美国CMU，德国Karlsruhe大学，日本ATR，意大利的科学技术研究所(IRST)、韩国的高级网络服务技术部(ETRI)和法国的自动翻译研究所(GETA-CLIPS)六家单位为核心成员。由于中科院自动化所模式识别国家重点实验室(NLPR)在汉语语音识别技术和口语信息处理领域的成就，1996年9月，该实验室被邀请作为联系会员加入C-STARⅡ。其后，经过对NLPR的综合考察，C-STAR于2000年9月正式接受该实验室加入核心成员的申请。在2000年10月召开的C-STAR核心成员会议上，通过正式答辩，全票一致同意NLPR正式成为C-STARⅢ的第七个核心成员，这标志着我国的口语自动翻译技术研究已进入国际最先进行列。

C-STAR Ⅲ的目标是推动口语自动翻译技术的实用化，研究开发基于电话语音(包括移动电话语音)的多领域、多语言(中、日、英、德、法、意、韩)的双向语音自动翻译系统，实现任意时间、任何地点的多语言电话语音自由通讯。自动化所模式识别国家重点实验室将承担汉语口语识别、汉语口语理解、口语生成和中文语音合成的研究任务。

(宗成庆)