汉语自动分词的研究现状与困难

张春霞[1,2], 郝天永[1]
([1]中国科学院计算技术研究所, 北京 100080; [2]中国科学院研究生院, 北京 100039)

摘 要: 汉语自动分词是信息提取、信息检索、机器翻译、文本分类、自动文摘、语音识别、文本语音转换、自然语言理解等中文信息处理领域的基础研究课题。尽管已被研究了二十多年, 分词仍然是中文信息处理的瓶颈问题。基于对汉语自动分词研究的现状分析, 构建了自动分词的形式化模型, 论述了影响分词的诸多因素, 分析了分词中存在的两个最大困难及其解决方法。最后指出了目前分词研究中尤其是在分词评测方面存在的问题以及未来的研究工作。

关键词: 汉语自动分词; 形式化模型; 未登录词; 分词评测
文章编号: 1004-731X (2005) 01-0138-06 **中图分类号**: TP391 **文献标识码**: A

The State of the Art and Difficulties in Automatic Chinese Word Segmentation

ZHANG Chun-xia[1,2], *HAO Tian-yong*[1]
([1]Institute of Computing Technology, Chinese of Academy Sciences, Beijing, 100080, China;
[2]Graduate School of the Chinese of Academy Science, Beijing 100039, China)

Abstract: Automatic Chinese word segmentation is a basic research issue on Chinese information processing tasks such as information extraction, information retrieval, machine translation, text classification, automatic text summarization, speech recognition, text-to-speech, natural language understanding, and so on. Though it has been investigated for more than twenty years, it is still a bottleneck for Chinese information processing. We give a detailed analysis of the state of the art in automatic Chinese word segmentation, build a formal model of word segmentation, discuss factors affecting word segmentation and the two great difficulties in word segmentation and their resolutions, and finally, point out the existing problems, especially those on the word segmentation evaluation, as well as the research problems to be resolved.

Keywords: automatic Chinese word segmentation; formal model; unknown words; word segmentation evaluation

引 言

汉语是一种词根语, 具有如下特点: (1)汉语缺乏形态变化, 没有性、数、格的变化标志, 词本身不能显示与其他词的语法关系, 它们的形式也不受其他词的约束; (2)词序严格, 词序不同, 意义也随之不同(如“打假”和“假打”意义截然不同); (3)虚词是主要的语法手段(如“老师和学生”和“老师的学生”意义截然不同); (4)汉语书写系统采用词标的形式, 词与词之间没有明显的形态界限。因此汉语的这些特征决定了针对其他语言处理的方法并不能完全适用于汉语信息处理。汉语信息处理又称中文信息处理, 是指“用计算机对汉语的音、形、义等信息进行处理, 包括对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作与加工”[1]。汉语自动分词已成为众多中文信息处理任务的一项基础研究课题。例如, 机器翻译、信息检索、信息提取、文本分类、自动文摘、语音识别、文本语音转换、自然语言理解等[2-5, 6-7]。

汉语自动分词已经被研究了二十多年, 但是目前仍然是制约汉语信息处理发展的一个瓶颈[13]。它主要存在语言学和计算机科学等两方面的困难[9]。语言学方面的困难有:

(1)词的定义不统一。调查表明, 对于母语为汉语的被试者, 对中文文本中词语的认同率只有 70%[10]。虽然国家标准《信息处理用现代汉语分词规范》[6]给出了词和分词单位的非形式定义, 但是语言学界对词还没有给出一个为大家广泛接受的、严格且统一的非形式定义。词的形式定义或者抽象定义问题也没有完全解决[9, 11]。

(2)汉语的分词还没有形成一个公认的分词标准。这是人和计算机共同面临的困难。同一文本可能被不同的人划分为几种不同的分词结果[1-2,9,11]。

(3)词的具体判定问题还没有完全解决。尽管《信息处理用现代汉语分词规范》[6]提出了分词单位和一套比较系统的分词规则, 但是由于真实文本的复杂性和多样性, 实践与理论之间的重大差异, 仍然没有能够在词层解决问题。问题的实质在于分词规范和分词词表的构造应该和汉语真实语料库结合起来考虑。同时, 除了定性信息外, 还必须引入定量信息[9,11]。

计算机方面的困难有: (1)没有合理的自然语言形式模型; (2)如何有效地利用和表示分词所需的语法知识和语义知识; (3)如何对语义进行理解和形式化。

收稿日期: 2003-12-23 **修回日期**: 2004-03-02
基金项目: 自然科学基金(#60073017 和#60273019)和科技部重大基础项目基金(#2001CCA03000 和#2002DEA30036)的资助。
作者简介: **张春霞**(1974–), 女, 山西人, 博士生, 研究方向为知识获取和文本挖掘; **郝天永**(1981–), 男, 河南人, 硕士生, 研究方向为知识获取和知识表示。

本文基于对汉语自动分词研究的现状分析，构建了自动分词的形式化模型，论述了影响自动分词的因素，分析了自动分词中存在的两个最大困难及其解决方法。最后指出了目前自动分词研究中尤其是在分词评测方面存在的问题以及未来的研究工作。

1 汉语自动分词的研究现状及分析

根据是否利用机器可读词典和统计信息，可将汉语自动分词方法分为三大类：基于词典的方法、基于统计的方法和混合方法[12]。

1.1 基于词典的分词方法

基于词典的分词方法的三个要素为分词词典、文本扫描顺序和匹配原则。文本的扫描顺序有正向扫描、逆向扫描和双向扫描。正向扫描是指从待切分语句的开头开始扫描，而逆向扫描是指从待切分语句的末尾开始扫描。双向扫描是正向扫描和逆向扫描的组合。匹配原则主要有最大匹配、最小匹配、逐词匹配和最佳匹配。

最大匹配法的基本思想：(1)取待切分汉语句的 m 个字符作为匹配字段，其中 m 为机器可读词典中最长词条的汉字个数；(2)查找机器可读词典并进行匹配。若能匹配，则将这个匹配字段作为一个词切分出来；若不能匹配，则将这个匹配字段的最后一个字去掉，剩下的字符串作为新的匹配字段，进行再次匹配。重复以上过程，直到切分出所有词为止。最小匹配法的基本思想是使待切分语句分词后得到的词最少。逐词匹配法是指把词典中的词按由长到短的顺序在待切分语句中进行搜索和匹配，直到把所有的词都切分出来为止。最佳匹配法的基本思想是词典中的词条按照词频的大小顺序排列，以求缩短分词词典的检索时间，从而降低分词的时间复杂度。

赵曾贻[13]提出了一种改进的最大匹配分词算法，分词词典支持词首字 Hash 查找和标准的不限词条长度的二分查找。李振星[14]提出了全二分最大匹配快速分词算法，采用首字 Hash 和完全二分查找，分词词典存放于内存中，不用进行 I/O 操作。这种分词方法查找单字词无需匹配，查找多字词的平均匹配次数为 1.56。杨建林[15]提出了一种基于词链的自动分词方法。Sproat 等人[4]提出了一种随机有限状态的分词算法。吴胜远[16]根据多级内码理论，提出了一种并行分词方法。孙茂松[17]通过实验考察了三种典型的分词词典的数据结构：整词二分、TRIE 索引树和逐字二分。实验表明对最大匹配分词法和全切分分词法而言，基于逐字二分的分词词典机制的处理速度较基于整词二分的处理速度分别提高了 15.3 倍和 16.6 倍。吴胜远[18]提出了一种单扫描的汉语分词方法，词典采用首字索引结构，分词的时间复杂度为 2.89，但是文中没有给出分词的正确率。

基于词典的分词方法的优点是易于实现[11,15,19]。其缺点是：(1)匹配速度慢；(2)存在交集型和组合型歧义切分问题；(3)词本身没有一个标准的定义，没有统一标准的词集；(4)不同词典产生的歧义也不同。

对于基于词典的分词方法，影响其精度的因素有[20]：(1)机器词典中词目的选择和词条的数量；(2)机器可读词典与待切分文本中词汇的匹配关系；(3)切分歧义；(4)未登录词；(5)分词方法。词典对分词精度造成的影响远远大于分词方法本身产生的歧义切分错误和未登录词问题。影响其速度的因素有[20]：机器可读词典的组织结构、匹配的原则和扫描的顺序。

1.2 基于统计的分词方法

基于统计的分词方法所应用的主要的统计量或统计模型有：互信息、N 元文法模型、神经网络模型、隐 Markov 模型和最大熵模型等。这些统计模型主要是利用词与词的联合出现概率作为分词的信息。

基于统计的分词方法的优点是：(1)不受待处理文本的领域限制；(2)不需要一个机器可读词典。缺点是：(1)需要大量的训练文本，用以建立模型的参数；(2)该方法的计算量都非常大；(3)分词精度与训练文本的选择有关。

(a) 互信息

互信息是一种度量不同字符串之间相关性的统计量。对于字符串 x 和 y，其互信息的计算公式如下：

$$\mathrm{MI}(x,y)=\log_2\frac{p(x,y)}{p(x)p(y)}$$

其中 $p(x,y)$为字符串 x 和 y 共现的概率，$p(x)$和 $p(y)$分别为字符串 x 和 y 出现的概率。

互信息 MI(x, y)反映了字符串对之间结合关系的紧密程度[21]：

(1) 互信息 MI$(x, y)>>0$，则 x, y 之间具有可信的结合关系，并且 MI(x, y)越大，结合程度越强；

(2) MI$(x, y)\approx 0$，则 x, y 之间的结合关系不明确；

(3) MI$(x, y)<<0$，则 x, y 之间基本没有结合关系，并且 MI(x, y)越小，结合程度越弱。

(b) N 元文法模型

高军[19]提出了一种无监督的动态分词方法，采用了一种可变长的 N 元文法模型。以词信息理论中极限熵的概念为基础，运用汉字字符串间最大似然度为匹配原则。文中没有给出实验结果与语料的选择和规模的关系。

(c) 神经网络模型

尹锋[22]提出了一种神经网络的分词方法。韩客松[8]提出了一种无词典的分词模型系统。其思想是：(1)采用人工神经网络的方法解决汉语歧义字段切分存在的问题；(2)分析网络结构和两类学习算法(BP 和概率网 PNN)对歧义词切分的影响。

该方法需要进一步研究下面的问题：(1)分析大样本中隐含层节点数、网络层次、样本数量、和学习次数等因素对分词精度的影响；(2)当语料庞大时，对分词精度的测评；(3)

分词的速度与语料大小的关系。

基于神经网络方法的优点是：(1)神经网络具有自学习、自组织、并行、非线性处理方式等特点，从而使该方法具备知识表达简洁、学习功能强、开放性好、知识库易于维护和更新的优势；(2)分词速度快；(3)精确度较高。缺点是：(1)容易陷入局部极小值点；(2)学习算法收敛速度慢；(3)网络层数及隐含节点选取无确定原则；(4)新加入样本对已学完样本有一定影响。

(d) 隐 Markov 模型

李家福[11]提出了一种基于 Markov 模型的分词系统，主要利用了汉语词长的分布规律。文中没有给出实验语料的来源和选取准则，以及使用统计方法的分词结果(包括分词的有效性和准确程度)能够完全满足文本分类需要的原因。

Masao Utiyama[23]等人提出了一个与领域无关的、不需要训练数据的文本切分统计模型，以找到文本的最大概率的切分。这种方法的优点是：(1)不需要训练数据；(2)能够利用切分的特征信息，例如切分的平均长度。但是这种方法把文本切分为较少的部分，不能将文本切分为很多部分。

基于隐 Markov 模型的分词方法的优点是[11]：降低了未登录词和专有名词的影响，只要有足够的训练文本就易于创建和使用。

(e) EM 模型

李家福[24]提出了一种基于 EM 算法的分词方法，EM 算法是在极大似然原则下的一种建模方法，存在模型和数据的过度拟合问题。

(f) 关联词统计语言模型

金凌[25]提出了一种距离加权的关联词统计语言模型，通过距离加权函数来引入距离信息。其平滑方法采用了基于图灵估计的退化算法。该方法应用到了一个中文整句拼音输入法系统中。实验结果表明该模型比 N-gram 统计语言模型的性能有了一定的提高，但是汉字的识别率有所降低。

该方法的优点是：在关联模型中只保存词对之间的关系。随着 M 的增大，只需增加一部分在语料中距离不大于 N−1 且没有出现过的词对之间的关系，模型的规模大小只是略有增加。M 可远大于 N，从而使模型具有更好的性能。由于 N-gram 模型保存的是各个 N 元组的相关信息，随着 N 的增大，N-gram 模型的规模将呈指数级别增长。该方法的缺陷是：(1)关联词模型预测每一个词出现的概率依赖于前面的词；(2)训练语料的大小限制了模型对这些词之间关系的准确表达。

(g) 其他方法

黄萱菁[26]提出了一个基于机器学习的无需人工编制词典的切词系统。该系统利用 x^2-统计量和广义似然比计算双字、三字和四字的候选词的相关度，将切词问题转化为一个有向图中求解最大加权路径问题。该方法以句中相邻字之间的间隙作为图中的顶点，候选词的对数相关度作为边的权重。

金翔宇[20]提出了一种中文文档的非受限无词典抽词方法。该方法通过自增长算法获取中文文档中的汉字结合模式，并用支持度、置信度等概念来筛选词条。实验结果表明：使用无词典抽词模型的平均分类精度为 94.3，而使用基于词典抽词模型的平均分类精度为 83.6。该方法对不同字串的包含关系进行了置信度分析，并且利用出现的频率来测量它们之间的关系。

1.3 混合的分词方法

王锡江[9]提出了一种基于邻接知识的汉语自动分词系统。该系统利用了最大匹配法、结合标志分词法、辅以词法和语义的邻接知识，并采用记忆学习和告知学习两种学习方式。文中没有给出如何获得词法和语义的邻接知识、以及解决冲突的方法。

赵铁军[27]提出了一种提高汉语自动分词精度的多步处理策略，包括：(1)消除伪歧义；(2)部分确定性切分；(3)数词串处理；(4)重叠词处理；(5)基于统计的未登录词识别；(6)使用词性信息消除切分歧义的一体化处理。伪歧义是指包含交集歧义但实际文本中只能有或几乎只有一种切分可能的字段。确定性切分是指成语、惯用语、叹词、语气词等的切分。

Palmer[2]提出了一种可训练的基于规则的分词算法，其核心部分是基于转换的学习机制。需要进一步测试该算法能否应用到实际的系统中。因为需要 80%的数据用于训练，20%的数据用于测试。

Kok[28]首先利用文本的统计信息给出可能性最大的词语边界，然后分析存在于句子中的语法和语义关系。文中没有给出该方法的实验结果。

2 汉语自动分词的形式化模型

本文在王明会等人[29]的工作基础上，进一步全面提出了汉语自动分词的形式化模型。并在此基础上，分析汉语自动分词所面临的问题和可能的解决方法。通过构建汉语自动分词的形式化模型，使得人们从根本上了解汉语自动分词的本质，揭示出汉语自动分词的困难和面临的问题，从而有效地解决汉语自动分词存在的问题。

设汉语文本中的符号集为有限集 $T=\{S_1, S_2, S_3, \ldots, S_m\}$，其中 $m\in\mathbf{N}$，$\mathbf{N}$为自然数集合，S_i可能为汉字、外文字母、阿拉伯数字、标点符号或者空格。设全体汉字的集合为有限集 $C=\{C_1, C_2, C_3, \ldots, C_n\}$，其中 $n\in\mathbf{N}$，C_i $(i=1, 2, \ldots, n)$为一个汉字。

定义 1 汉字段

汉语言文本中出现的由若干个汉字构成的字符串称为汉字段。

定义 2 待切分汉语句

待切分汉语句 $S=S_1S_2S_3\ldots S_p$ 定义为一个含有 $p(p\in\mathbf{N})$个字符且任意两个字符之间没有空格的字符串，其中 $S_i\in T$ $(i=1, 2, \ldots, p)$。字符可能为汉字、外文字母、阿拉伯数字、

或标点符号。

定义 3 汉语自动分词系统

汉语自动分词系统 ACWSS(Automatic Chinese Word Segmentation System)定义为一个六元组(St, Lk, Si, A, Ewsr, Gwsr)，

(1) 源文本 St (Source Texts)，

(2) 语言学知识 Lk (Linguistic Knowledge)，

(3) 统计信息 Si (Statistic Information)，

(4) 分词算法 A (Algorithm)，

(5) 系统分词结果 Ewsr (Experimental Word Segmentation Result)，

(6) 目标分词结果 Gwsr (Goal Word Segmentation Result)。

对于汉语自动分词系统 ACWSS，基于词典匹配的汉语分词系统利用了语言学知识机器可读词典 Mrd(Machine Readable Dictionary)。基于统计方法的汉语分词系统则利用了统计信息 Si。在基于混合方法的分词系统中，上述两种信息被同时利用。

定义 4 未登录词

对于待切分汉语句 S，设 S'为 S 的词语切分结果，$S'=W_1/W_2/W_3/\ldots/W_m/$，其中“/”为切分标记，如果 $W_i \notin Mrd$ ($i=1, 2, \ldots, m$)，则称 W_i 为未登录词。

一个词是否为未登录词是相对于一个特定的汉语分词系统而言的。对于一个分词系统，它可能为未登录词，而在另外一个分词系统中却不一定是未登录词。

定义 5 正确词语切分句和错误词语切分句

设 S 为一个待切分汉语句，若 S'是 S 的一个已切分语句，定义 $S'=W_1'/W_2'/W_3'/\ldots/W_m'/$为 S 的正确词语切分句当且仅当 S'的 W_i'之间满足语法、语义规则以及上下文语境限制，否则称为错误词语切分句。

一个已切分汉语句 S 的正确性判断是一个很复杂的问题。一个待切分语句对不同的人或系统而言，可能会产生不同的分词结果。并且与上下文语境、分词面向的应用领域紧密相关。因此，关键问题在于我们能否给出一个待切分汉语句 S 的统一的分词结果。

定义 6 切分歧义词串

假设一个待切分语句 S 在某个分词算法的作用下，产生一组不同的词语切分语句 $S_1', S_2', \ldots S_n'$，定义集合

$$\bigcup_{i,j=1}^{n} (S_i' \cap S_j')$$

为待切分语句 S 的歧义切分词语集，并称歧义切分词语集在 S 中所对应的极大连续串为待切分汉语句 S 的切分歧义词串。一个句子可能包含多个切分歧义词串。

例如，待切分语句 $S=S_1S_2\ldots S_{10}$ 在某个分词算法的作用下，产生不同的词语切分句 S_1', S_2', S_3'：

$S_1'=S_1S_2/S_3S_4/S_5S_6S_7/S_8S_9S_{10}$

$S_2'=S_1/S_2S_3S_4/S_5S_6S_7/S_8S_9/S_{10}$

$S_3'=S_1S_2/S_3S_4/S_5S_6S_7/S_8/S_9S_{10}$

那么 S 的切分歧义词串为 $S_1S_2S_3S_4$ 和 $S_8S_9S_{10}$。

切分歧义词串的情况相当复杂，有两种典型类型：即组合型切分歧义词串和交集型切分歧义词串。

定义 7 组合型歧义切分词串

假设一个待切分语句 S 包含歧义切分词串 S_1S_2，如果 S_1S_2 被切分成：$S_1/S_2/$ 和 $S_1S_2/$，则称 S_1S_2 为组合型歧义切分词串。其中 S_1, S_2, S_1S_2 可能为机器可读词典中的词语、或未登录词、或不构成词语。

定义 8 交集型歧义切分词串

假设一个待切分语句 S 包含歧义切分词串为 $S_1S_2S_3$，如果 $S_1S_2S_3$ 被切分成：$S_1S_2/S_3/$和 $S_1/S_2S_3/$，则称 $S_1S_2S_3$ 为交集型歧义切分词串。其中 S_1, S_1S_2, S_2S_3, S_3 可能为机器可读词典中的词语、或未登录词、或不构成词语。

其他切分歧义串为这两种类型的组合。例如，待切分语句 $S=S_1S_2\ldots S_{10}$ 在某个分词算法的作用下，产生不同的词语切分句 S_1', S_2'：

$S_1'=S_1S_2S_3/S_4S_5S_6/S_7S_8S_9S_{10}$

$S_2'=S_1S_2/S_3S_4/S_5S_6/S_7S_8S_9S_{10}$

那么歧义切分词串 $S_1S_2\ S_3S_4\ S_5S_6$ 为组合型和交集型切分歧义的组合词串。

不同的分词算法会得到不同的歧义切分语句。对一个分词算法而言，如果一个待切分语句存在多个分词结果，必然存在歧义切分词串。可见，歧义切分语句的存在构成了汉语自动分词的困难之一。

我们可以根据待切分语句的歧义切分句中切分符号“/”的数目来判断其包含的歧义切分串的类型。

定义 9 极小歧义切分串

假设一个待切分语句 S 对应的歧义切分句 S_1'和 S_2'，它们共有 k 个切分符号的位置相同，不妨设为：

$S_1'=S_1\ldots S_{i1}/S_{i1+1}\ldots S_{i2}/S_{i2+1}\ldots S_{ij}/S_{ij+1}\ldots S_{ik}/S_{ik+1}\ldots S_m$

$S_2'=S_1\ldots S_{i1}/S_{i1+1}\ldots S_{i2}/S_{i2+1}\ldots S_{ij}/S_{ij+1}\ldots S_{ik}/S_{ik+1}\ldots S_m$

则称连续两个相同位置的切分符号之间的字符串为 S 的极小歧义切分串，即 $S_1\ldots S_{i1}$，$S_{i1+1}\ldots S_{i2}$，…，$S_{ij+1}\ldots S_{ik}$，$S_{ik+1}\ldots S_m$ 都为 S 的极小歧义切分串。

定理 1 假设一个待切分语句 S 对应的歧义切分句为 S_1'和 S_2'并且 $S_1' \neq S_2'$，定义 SSN(S')表示已切分语句 S'包含的切分符号“/”的数目，如果 T 为 S 的极小歧义切分串，则当 $SSN(T_{S1'}) = SSN(T_{S2'})$ 时，T 含有交集型歧义切分串，其中 $SSN(T_{S1'})$和 $SSN(T_{S2'})$分别表示 T 在 S_1'和 S_2'中包含的切分符号的数目。

3 汉语自动分词的困难及其解决方法

确定分词单位的非语法因素有语义因素和语音因素[30]。冯志伟提出了将形式词作为分词的单位，其中形式词定义为

语言中能够自由运用的最小单位。同时，他提出了视读原则、多元化原则、领域针对原则等确定切词单位的非语言学的原则[30]。但是这些原则都是描述性的，主观的因素很大，而且其实用性还有待验证。汉语自动分词的两个最大困难是未登录词的识别和切分歧义的消除[27]。

3.1 切分歧义的消除

解决歧义的方法可以分为两类：基于规则的方法和基于统计的方法。基于规则的方法主要有：李家福[11]提出了基于规则的消除切分歧义的方法，并根据句法、语义规则和语法、语义解析进行分词判断。这些规则仅涉及若干毗邻词之间的线性关系，没有反应句子中各成分之间的层次关系，可靠性不强，难以建立完整、有效、无矛盾的体系。刘开瑛提出了一种基于词性的方法[31]。

基于统计的方法主要有：基于互信息和 T-Test 的方法[32,33]、基于隐 Markov 模型的方法[34]、基于 SVM 和 k-NN 结合的汉语交集型歧义切分方法、基于 EM 的方法[35]等。

谭琼[32]提出了采用双向扫描法识别歧义字段，然后利用互信息和 T-测试解决歧义。但是需要进一步解决下面的问题：(1)对于存在多处切分歧义的交集型字段，如何计算每种切分的可能性；(2)当每一句子都有多种切分方法时，该方法的计算量较大。

孙茂松[33]提出了利用句内相邻字之间的互信息和 T 测试这两个统计量来解决汉语自动分词中交集型歧义切分字段的方法。其优点是：(1)不需要人工标注语料，直接从生语料库出发，通过字的统计信息模拟词频，进而设计交集型歧义切分字段的算法；(2)字的统计信息获取过程是全自动的，因而避免了人工标注语料可能引起的各种问题，保证了数据的准确性、一致性、方法的简明性和移植性。存在的不足是：(1)连续字对的出现与语料的大小不一定有直接的关系；(2)必须保证训练语料要足够的大；(3)二元语法模型的固有缺陷导致了那些不常见用法的错误。例如：化工厂，分为：化工/厂 而不是：化/工厂。孙茂松还提出了一种消解中文三字长交集型分词歧义的算法[34]。该算法利用了词的概率消息、词性 Bigram 和常用字分合法。该方法可以称为不考虑上下文制约关系的零阶 Markov 模型。主要优点是回避了训练代价比较高昂的词性消息。

李蓉[35]提出了将基于 SVM 和 K-NN 结合的分类方法用于解决交集型歧义切分。将交集型歧义切分的切分过程形式化为一个分类过程；然后从歧义字段中挑选出一些高频伪歧义字段，人工将其正确切分并代入 SVM 训练；最后对于待识别歧义字段通过使用 SVM 和 k-NN 相结合的分类算法即可得到切分结果。

王伟[36]提出了一种基于 EM 非监督训练的自组织分词歧义解决方案。分三步进行：(1)得到每个句子的所有分词结果，构成训练集；(2)基于已构建的训练集和初始语言模型，利用 ME 算法估计出一个新的语言模型；(3)最终的语言模型通过多次迭代而得到。在 E-step 中，基于给定的模型参数，计算每个隐变量对于样本的概率。其中隐变量是样本的扩充。在 M-setp 中，根据上步计算出来的概率分布和扩充，重新计算模型参数。通过交替使用这两个步骤，EM 算法逐步改进模型的参数，是参数和训练样本的似然概率逐渐增大，最后终止于一个极大点。该方法的优点是具有非监督方法的自动化的特性。

利用词频信息构造算法解决切分歧义十分有效[33]，关键在于词频信息的获取。但仍然面临下面的问题：(1)需要相当规模、已经人工标注的语料作为训练样本；(2) 词频对领域有一定的敏感性；(3)人工分词和机器校正都不同程度地依赖于人的语感，而语感因不同的人和不同的时间而不同。

3.2 未登录词的识别

产生未登录词的原因主要有：(1)机器可读词典中词目的选择和词目的数量；(2)机器可读词典与待处理文本中的词汇的匹配关系，包括机器可读词典对待处理文本中词汇的覆盖率。覆盖率指待处理文本的词汇在机器可读词典中所占的比例。如果待处理文本中含有 m 个不重复出现的词，其中有 n 个词在机器可读词典中出现，则机器可读词典对待处理文本词汇的覆盖率为 n/m。

目前主要有基于分解与动态规划策略的汉语未登录词识别和基于语料学习的未登录词检测等方法。

吕雅娟[37]利用词频、上下文信息和未登录词候选词表，采用如下分解处理策略：(1)预处理；(2)二字、三字候选字段的处理；(3)多字候选字段的处理。未登录词识别的过程如下：给出所有可能的未登录词估计，采用动态规划策略选出最有可能的识别结果。该方法使用评价函数对未登录词进行估计，评价函数包含自身可信度和上下文认可度(利用启发式规则)。这种方法的优点是：(1)统计中国人名、中国地名、外国译名，得到每个字作为姓氏、名字、地名首、中、尾、译名首、中、尾的频率等级；(2)评价函数包含了自身可信度和上下文可信度。但是还需要解决下面的问题：(1)以字为单位作为评估函数的考察单元，对于姓氏可以，但是对于地名就有问题；(2)可信度的给出依赖于语料的选择。

Keh-Jiann Chen[38]提出了一种基于语料学习的未登录词检测方法。其方法能够提取那些用来区分单音节词和多音节词词素的语法规则。该方法的优点是：(1)进行自动规则学习；(2)自动评测每个规则的性能；(3)通过动态选择规则集能够平衡召回率和正确率。但是，这些优点是依赖于语料库的。该文也没有给出如何提取规则和如何运用规则来检测未登录词。同时，也需要考虑由多音节构成的未登录词的识别方法。Chiang[39]也提出了一种处理分词和未登录词识别的统计模型。

4 存在的问题及未来的工作

目前汉语自动分词研究主要存在下面的几个问题：

(a) 面向领域语料库的分词方法

目前的自动分词方法主要是针对新闻报纸等的语料，对于各种专业领域语料是否适用，是否需要寻找另外的方法来对专业术语进行切分，这些问题还有待进一步研究。

(b) 汉语语料的建设

严格地讲，在世界范围内，还没有一个真正经得起各方面推敲并形成一定影响的大型汉语分词语料库[40]，分词质量远不能达到人们期望或者想象的水准[33]。目前主要的汉语语料库有中央研究院语料库、香港城市大学语料库、宾州大学语料库和北京大学语料库等，而领域专业文本的语料却很少。

(c) 汉语自动分词的评测

由于分词的应用目的的不同，进行分词的语料也不一定相同。我们把不同分词系统的评测条件分为强条件和弱条件。

定义 10 分词系统评测强条件

对于两个分词系统 $ACWSS_1=(St_1, Lk_1, Si_1, A_1, Ewsr_1, Gwsr_1)$和 $ACWSS_2= (St_2, Lk_2, Si_2, A_2, Ewsr_2, Gwsr_2)$进行评测的强条件是：

(1) $St_1 =St_2$；

(2) $Gwsr_1=Gwsr_2$。

定义 11 分词系统评测弱条件

对于两个分词系统 $ACWSS_1=(St_1, Lk_1, Si_1, A_1, Ewsr_1, Gwsr_1)$和 $ACWSS_2 =(St_2, Lk_2, Si_2, A_2, Ewsr_2, Gwsr_2)$进行评测的弱条件是：

(1) $Gwsr_1$ 和 $Gwsr_2$ 遵循相同的分词规范；

(2) St_1 和 St_2 都为领域专业文本或者非领域专业文本。

定义 12 分词系统的定量评测

分词系统定量评测的三个指标是分词速度、分词精度、未登录词的识别率以及 F-measure，其中 $F=(1+\beta)PR/(\beta P+R)$[3]。

目前不同分词系统的性能比较主要是在源文本和目标分词结果都不相同的意义上进行的，即 $St_1\neq St_2$ 且 $Gwsr_1\neq Gwsr_2$。由于没有统一的评测语料和评测结果，不同自动分词方法的实验结果比较会有很大差别。还有结果评估的人为性和主观性，大大影响了汉语自动分词的研究进展。

(d) 汉语自动分词的应用

汉语自动分词仅仅是中文信息处理任务的手段，并不是最终目标。因此我们更应关注自动分词系统在实际应用中的效果。比如汉语分析与理解、机器翻译、中文文献自动标引、中文信息检索、汉字识别、汉语语音识别与合成、中文简繁体自动转换及文本处理、中文文稿自动校对等。

5 结论

随着自然语言处理任务的不断出现和发展，汉语自动分词面临着新的机遇和挑战。本文对汉语自动分词研究现状进行了综述分析；论述了影响自动分词效果的因素；讨论了自动分词的两个最大困难消除切分歧义和识别未登录词，及其解决方法；最后指出了汉语自动分词研究中存在的问题以及下一步的研究工作。

参考文献：

[1] 《汉语信息处理词汇 01 部分：基本术语(GB12200.1-90)》[M]. 北京：中国标准出版社, 1991.

[2] Robert Dale, Herman Moisl, Harold Somers. Handbook of Natural Language Processing [M]. New York: Marcel Dekker, Inc, 2000.

[3] David D. Palmer. A Trainable Rule-based Algorithm for Word Segmentation [A]. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics [C]. 1997.321-328.

[4] Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. A Stochastic Finite-State Word-Segmentation Algorithm for Chinese [J]. Computing Linguist, 1996, 22(3): 377-404.

[5] 尹锋. 汉语自动分词研究的现状与新思维[J]. 现代图书情报技术, 1998, 4: 22-26.

[6] 刘源，谭强等. 信息处理用现代汉语分词规范及自动分词方法[M]. 北京：清华大学出版社. 1994.

[7] 许嘉璐. 现状和设想：试论中文信息处理与现代汉语研究[J]. 中文信息学报, 2001, 15(2): 1-8.

[8] 韩客松，王永成，陈桂林. 汉语语言的无词典分词模型系统[J]. 计算机应用研究, 1999, 16(10): 8-9.

[9] 王锡江，王启祥，陈家俊. 基于邻接知识的汉语自动分词系统[J]. 计算机研究与发展, 1992, 29(11): 54-58.

[10] 黄昌宁，高剑峰，李沐. 对自动分词的反思[C]. 全国第七届语言学联合学术会议. 2003.26-38.

[11] 李家福，张亚非. 一种基于概率模型的分词系统[J]. 系统仿真学报, 2002, 14(5): 544-550.

[12] 付国宏，王晓龙. 汉语词语边界自动划分的模型与算法[J]. 计算机研究与发展, 1999, 36(9): 1143-1147.

[13] 赵曾贻，陈天娥，朱兰. 一种基于语词的分词方法[J]. 苏州大学学报 [J], 2002, 18(3): 44-48.

[14] 李振星，徐泽平，唐卫清，唐荣锡. 全二分最大匹配快速分词算法[J]. 计算机工程与应用, 2002, 38(11): 106-109.

[15] 杨建林，张国梁. 基于词链的自动分词方法[J].情报理论与实践, 2000, 23(2): 84-87.

[16] 吴胜远. 并行分词方法的研究[J]. 计算机研究与发展, 1997, 34(7): 542-545.

[17] 孙茂松，左正平，黄昌宁. 汉语自动分词词典机制的实验研究[J]. 中文信息学报, 2000, 14(1):1-6.

[18] 吴胜远. 一种汉语分词方法[J]. 计算机研究与发展, 1996, 33(4): 306-311.

[19] 高军. 无监督的动态分词方法[J]. 北京邮电大学学报, 1997, 20(4):66-69.

[20] 金翔宇，孙正兴，张福炎. 一种中文文档的非受限无词典抽词方法[J]. 中文信息学报, 2001, 15(6): 33-39.

[21] Yubin Dai, Teck Ee Loh, Christopher Khoo. A New Statistical Formula for Chinese Text Segmentation Incorporating Contextual Information [C]. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999.82-89.

[22] 尹锋. 基于神经网络的汉语自动分词系统的设计与分析[J]. 情报学报, 1998, 17(1): 41-50.

[23] Masao Utiyama, & Hitoshi Isahara. A Statistical Model for Domain-Independent Text Segmentation[C]. The 39th Annual Meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter of the Association for Computational Linguistics. 2001. 491-498.

（下转第 147 页）

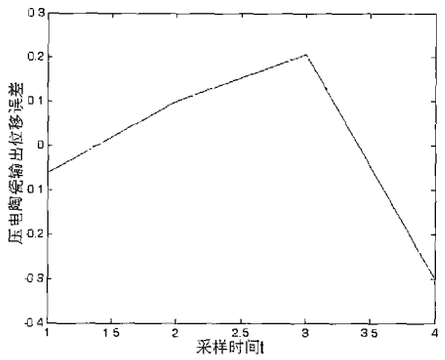

图 10 模型预测误差

4 结论

从多维空间全新的角度， 提出了压电陶瓷迟滞回环特性三维空间模型。三维空间模型的输入信号是电压信号和采样时间，输出信号是位移。在三维空间中，压电陶瓷状态唯一确定，输入和输出为单值映射。采用了运算量小，学习速度快的 RBF 神经网络逼近压电陶瓷的迟滞特性。在周期性输入电压值（形成主环）和任意输入电压值的两种情况下，进行了模型逼近和模型预测，结果表明在三维空间中对迟滞回环特性进行建模是有效的，并具有较高的模型精度。该方法适合于压电陶瓷短时间动作的场合。对长时间连续工作的情况，须对时间轴进行处理，有待进一步研究。

参考文献：

[1] Mayergoyz I D, Friedman G, and Salling C. Comparison of the classical and generalized preisach hysteresis models with experiments[A]. Magnetics Conference, Digests of NTERMAG '89. International[C]. 1989, 28-31.HC2 -HC2.

[2] Jiles D C, Atherton D L. Theory of ferromagnetic hyseteresis[J]. J. Magn. MagnMater, 1986, 61:48-60.

[3] Serpico C, Vison C. Magnetic Hysteresis Modeling via Feed-Forward Neural Networks [J]. IEEE Transactions on Magnetics, 1998,34(3): 623-628.

[4] Cincotti S, Daneri I. Neural network identification of a nonlinear circuit model of hyseteresis[J]. Electronics Letters, 1997, 33(13): 1154 –1156.

[5] Cruz-Hernández J M, Hayward V. On the linear compensation of hyseteresis[A]. in 36th IEEE Conf. Decision Contr.[C]. 1997, 1:1956–1957.

[6] Cirrincione M， Miceli R, Galluzzo. G R and Trapanese M. Preisach function identification by neural networks[J]. Magnetics, IEEE Transactions on, 2002,38(5): 2421 –2423.

[7] Cincotti S, Marchesi M, and Serri A. A neural network model of parametric nonlinear hysteretic inductors [J]. Magnetics, IEEE Transactions on 1998, 34(5): 3040-3043.

[8] 赵 辉，蒲昭邦，刘国栋等. 压电器件在微位移系统中应用[J]. 压电与光电, 2000, 22(4): 244-245.

[9] 张 波，王纪武，陈 恳，等. 压电驱动晶体的特性[J]. 中国机械工程 ，2002, 13(5): 446-449.

（上接第 143 页）

[24] 李家福、张亚非. 基于 EM 算法的汉语自动分词方法[J]. 情报学报, 2002. 21(3): 269-272.

[25] 金凌，吴文虎，郑方，吴根清. 距离加权统计语言模型及其应用[J]. 中文信息学报, 2001, 15(6): 47-52.

[26] 黄萱菁. 基于机器学习的无需人工编制词典的切词系统[J]. 模式识别与人工智能, 1996, 9(4): 297-303.

[27] 赵铁军,吕雅娟,于浩,杨沐昀,刘芳.提高汉语自动分词精度的多步处理策略[J]. 中文信息学报, 2001, 15(1): 13-18.

[28] Kok Wee Gan. Integrating Word Boundary Identification with Sentence Understanding [C]. The 31st Annual Meeting of the Association for Computational Linguistics. 1993. 301-303.

[29] 王明会，钟义信，田中英严. 汉语文本切分的形式化和难点分析[A]. 中国人工智能学会第 9 届全国学术年会论文集：中国人工智能进展[C]. 北京：北京邮电大学出版社. 2001. 987-991.

[30] 冯志伟. 确定切词单位的某些非语法因素. 中文信息学报, 2001, 15(5): 9-14.

[31] 刘开瑛. 中文文本自动分词和标注[M]. 北京：商务印书馆. 2002.

[32] 谭琼，史忠植. 分词中的歧义处理[J]. 计算机工程与应用, 2002, 38(11): 125-127.

[33] 孙茂松等. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义[J]. 计算机研究与发展, 1997, 34(5): 332-339.

[34] 孙茂松，左正平，黄昌宁.消解中文三字长交集型分词歧义的算法[J]. 清华大学学报, 1999, 39(5): 101-103.

[35] 李蓉，刘少辉，叶世伟. 基于 SVM 和 K-NN 结合的汉语交集型歧义切分方法 [J]. 中文信息学报, 2001, 15(6): 13-18.

[36] 王伟，钟义信，孙建，杨力. 一种基于 EM 非监督训练的自组织分词歧义解决方案[J]. 中文信息学报. 2001, 15(2): 38-44.

[37] 吕雅娟，赵铁军，杨沐昀，于浩，李生. 基于分解与动态规划策略的汉语未登录词识别[J].中文信息学报, 2000, 15(1): 28-33.

[38] Keh-Jiann Chen, Ming-Hong Bai. Unknown Word Detection for Chinese by a Corpus-based Learning Method [J]. International Journal of Computational Linguistics & Chinese Language Processing. 1998, 3(1): 27-44.

[39] Tung-Hui Chiang, Jing-Shin Chang, Ming-Yu Lin, and Keh-Yih Su. Statistical Models for Word Segmentation and Unknown Word Resolution [C]. Proceedings of ROCLING-V, ROC Computational Linguistics Conference. 1992.123-146.

[40] 孙茂松. 谈谈汉语分词语料库的一致性问题[J]. 语言文字应用, 1999, 8(2): 88-91.

[41] 揭春雨等. 论汉语自动分词方法[J]. 中文信息学报, 1989, 3(1): 1-9.

[42] 吴立德. 大规模文本处理[M]. 上海：复旦大学出版社, 1997.

[43] 赵军，黄昌宁.汉语基本名词短语结构分析模型[J]. 计算机学报, 1999, 22(2): 141-146.

[44] 张树武，黄泰翼. 汉语统计语言模型的 N 值分析[J]. 中文信息学报, 1998, 12(1): 35-41.

[45] 黄昌宁. 中文信息处理中的分词问题[J]. 语言文字应用, 1997, 6(1): 72-78.

[46] 孙茂松，邹嘉彦. 汉语自动分词研究中的若干理论问题[J]. 语言文字应用, 1995, 4(4):40-46.

[47] 刘开瑛. 现代汉语自动分词评测技术研究[J]. 语言文字应用. 1997, 6(1): 101-106.

[48] 王彩荣，李晓毅，黄玉基. 汉语自动分词系统的评价[J]. 微处理机, 2003. 25(5): 28-30.