文章编号: 1003-0077(2008)05-0045-06

中文语音合成中的文本正则化研究

贾玉祥[1],黄德智[2],刘 武[2],俞士汶[1]

(1. 北京大学 计算语言学研究所,北京 100871; 2. 法国电信北京研发中心,北京 100190)

摘 要: 中文文本正则化是把非汉字字符串转化为汉字串以确定其读音的过程。该工作的难点:一是正则化对象——非汉字串形式复杂多样,难于归纳;二是非汉字串有歧义,需要消歧处理。文章引入非标准词的概念对非汉字串进行有效归类,提出非标准词的识别、消歧及标准词生成的三层正则化模型。在非标准词的消歧中引入机器学习的方法,避免了复杂规则的书写。实验表明,此方法取得了很好的效果,并具有良好的推广性,开放测试的正确率达到 98.64%。

关键词: 计算机应用;中文信息处理;文本正则化;语音合成;最大熵模型

中图分类号: TP391 **文献标识码**: A

Text Normalization in Chinese Text-to-Speech System

JIA Yu-xiang[1], HUANG De-zhi[2], LIU Wu[2], YU Shi-wen[1]

(1. Institute of Computational Linguistics, Peking University, Beijing 100871, China;
2. France Telecom R&D Beijing, Beijing 100190, China)

Abstract: Chinese text normalization is the process of transforming non-Chinese character strings into their corresponding Chinese character strings to determine their pronunciations. The difficulties of this work mainly lie in two aspects: too many non-Chinese character strings of various formats and their high degree of ambiguities. This paper develops an effective taxonomy of non-Chinese character strings with the concept of Non-Standard Words (NSWs). And then a three-layer normalization model is proposed, including NSWs detection, NSWs disambiguation and standard words generation. In the NSWs disambiguation stage, a machine learning method is employed to overcome shortcomings of rule-based method. Experiment results show that this approach achieves a high performance and adapts well to new domains. The accuracy of open test is 98.64%.

Key words: computer application; Chinese information processing; text normalization; text-to-speech; maximum entropy model

1 引言

真实文本中含有大量非标准词(Non-Standard Words,NSWs),这些词在词典中查不到,它们的读音也不能通过正常的拼音规则得到[1]。在中文文本中,非标准词是指包含非汉字字符(如阿拉伯数字、英文字符、各种符号等)的词,其中的非汉字字符需要转换成对应的汉字,这个转换过程称为文本正则化。文本正则化是语音合成的关键环节,也是语音识别的必要步骤。由于非标准词往往是用户关注的焦点,如日期、价格、电话号码、机构名等,因此文本正则化直接影响语音服务的质量。

一个非标准词在不同的上下文中可能对应不同的标准词(汉字词)。如"11"可以读作"十一",在电话号码中读"幺幺",而在"2 米 11"中读"一一"。因此非标准词的消歧问题是文本正则化的难点[2],有些非标准词有相当高的歧义。再加上非标准词类型

收稿日期: 2008-01-08 **定稿日期**: 2008-03-31

基金项目: 国家 973 课题资助项目(2004CB318102)

作者简介: 贾玉祥(1981—),男,博士生,主要研究方向为计算语言学;黄德智(1976—)男,博士,主要研究方向为语音合成;刘武(1978—),男,研发工程师,主要研究方向为语音合成。

多种多样,因此文本正则化也是语音合成的难点。在著名的 Nuance Vocalizer 语音合成引擎中,"20 % 以上的核心应用代码是用来处理文本正则化的,随着新的非标准词的不断出现,代码还会继续增加。"[3]

文本正则化的典型方法是基于规则的,如 LDC (Linguistic Data Consortium)的 Text Conditioning Tools。这种方法的缺点是明显的:规则难于书写、维护,推广性差。作为一个歧义消解问题,机器学习方法被大量采用并显示出了优势。如,决策树和决策列表用于英语和印度语的文本正则化[4],支持向量机用于波斯语非标准词的分类[5],Winnow 用于泰国语的文本分析[6]等。

中文文本正则化主要还是基于规则的[7~10]。文献[10]采用基于外部规则的方法,维护了 400 多条规则,用到了分词、词性标注等信息。文献[11,12]把非标准词的处理和分词、词性标注、命名实体识别等纳入到一个统一的框架。而对于语音合成的应用来说,文本正则化更宜于作为一个独立的模块。建立在分词、标注基础上的正则化模型往往会受到分词、标注错误的影响。如,"中国/ns 共产党人/n 70 年/t 前/f 留/v 在/p 赣南/ns 红土地/n 上/f 的/u 壮举/n",其中"70 年/t"应为"70/m 年/q",正则化的结果是"七十 年",而不是"七零年"(例句中的分词、标注格式参见规范[13])。

本文通过考察大规模语料库中非汉字串的分布情况,制定了一个全面的非标准词分类体系。提出了文本正则化的三层模型:非标准词识别、消歧和标准词生成。在非标准词的消歧中引入最大熵模型。此方法直接处理真实文本,无须进行分词、标注。

2 非标准词分类体系

非汉字串形式复杂多样,为了更好地进行处理,引入非标准词的概念。非标准词是符合一定构成模式的非汉字串或非汉字字符和汉字字符的混合串。通过非标准词的识别、消歧,标准词的生成而达到对所有非汉字串的正则化。

非标准词的分类体系是文本正则化的基础。作者考察了 2000 年《人民日报》语料中非汉字串的出现情况,制定了非标准词的分类体系,并统计了非标准词的分布情况。一共定义非标准词 48 类,表 1 给出了非标准词的分类简表。语料中一共出现非标准词276 525 个,其中 95 %是阿拉伯数字表达式,包括纯数字串、数字串和".,-,/,:"的组合、数字串加后接成分、范围(如 100-200 人:100 到 200 人,无歧义)等。后接成分包括特殊符号如"%"、汉语数词(万、亿、多等)、量词(个、年、平方米等)、名词(人、港币等)、字母单位(如 m^2)等。阿拉伯数字的其他用法还有:'99(九九年)、1 000 元(一千元)等。单个符号(-,/,:,.等)也作为一个非标准词。其他类型还有 URL、Email、英文字母串等。在实际处理中,所有符号均考虑了半角和全角编码。

表 1 非标准词按构成形式分类

阿拉伯数字表达式	纯数字串	110, …
	数字 + .	1.29, 2000.9.10, 162.105.81.14, …
	数字 + -	1998-2002, 2000-9-10, 4-3-2-1, …
	数字 + /	1/3, 2000/9/10, …
	数字 + :	10:15, 10:15:20, …
	数字 + 后接成分	80.5%, 100 万, 50 多, 75 年, 500 人, 10cm, …
	范围	100-200 人, …
	其他	'99, 1 000 元, …
符号		-, /, :, ., ×, >, =, <, …
其他		URL, Email, Alphabets, …

按照非标准词是否有歧义可以划分为:基本非标准词(无歧义)和歧义非标准词。基本非标准词是主体,如表 2 所示,其分布约占到整个非标准词的 84 %,其中仅数字串加后缀的情形就占整个非标准词的 55 %,这里后缀主要包括汉语量词(如:个、平方米等),个别数词(如:多)、名词(如:人、港币)等。

表 2 基本非标准词

类 别	举 例	百分比
数字 + 后缀	35 人	55 %
整数 + 量级	100 万	8 %
百分数	10 %, 12.5 %	6 %
日期	2007 年 10 月	4 %
小数 + 量级	1.5 亿	3 %
范围	10-15 厘米	2 %
年数	50 年历史	2 %
其他	'99, Win32	4 %

歧义非标准词占总体的 16 %,其中有些类型使用简单的启发式规则就可以消歧,有些类型则需要

长距离上下文信息甚至全局信息。根据情况不同分别采用基于规则或基于机器学习的方法处理。2000 年《人民日报》全年语料的统计显示,机器学习方法处理的非标准词占总体的 14 %,规则方法处理的仅占 2 %。表 3 给出了歧义非标准词的一些例子,可见一些类别的歧义性很高。

表 3　歧义非标准词

类　别	读　法	举　例
纯数字串	数码	2 米 11
	整数	110
	音变	110
	英语	p2p
数字-数字	年-年	1998-1999
	电话	010-12345678
	数码-数码	波音 737-200
	整数-整数	200-300
	比分	比分 2-3
	减	100-1 = 99
数字/数字	分数	1/3
	不发音	T65/66
	日期	2001/01
数字:数字	时间	上午 10:15
	比值	比分 10 ∶15
年	某年	1999 年
	多年	1000 年
年-年	某年	1998-1999 年
	多年	30-50 年
符号“-”	到	北京-上海
	不发音	时代-华纳
符号“/”	每	元/天
	不发音	张三/李四

3　正则化过程

根据上述分类体系,本文提出文本正则化的三层模型(如图 1 所示)。有限自动机(Finite State Automata,FSA)从真实文本中识别非标准词,并给出非标准词的具体类别标记;基本非标准词直接进入标准词生成模块;歧义非标准词进入消歧模块。消歧模块根据歧义非标准词的不同类型采用简单规则(Simple Rules, SR)方法或最大熵模型(Maximum Entropy Model,MEM),得到非标准词的具体读法(子类标记)。前两个阶段是分析阶段,分析结果——非标准词及其类别标记,作为标准词生成阶段的输入、输出为正则化的文本。标准词生成是一一映射,没有歧义,使用规则加映射表的方法进行处理。

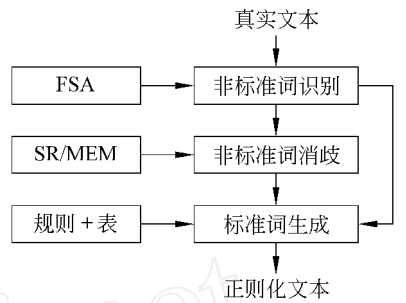

图 1　文本正则化流程图

3.1　非标准词的识别

FSA 从真实文本中识别非标准词,并给出非标准词的类标记。当存在非标准词嵌套的情况时,采用最长匹配策略,即最长串为非标准词,而不是它的子串。因为串越长,串内所含信息越多,歧义越小;需要处理的串的数量也越少。

为了识别“数字 + 后缀”的情况,FSA 中使用一个后缀(见基本非标准词)列表,由汉语量词、个别数词、名词等构成,如“人/天/时/元/角/分/厘米/公斤……”。对于单字后缀的情况,可能会出现少量切分歧义,如“1999 人才工程”,其中“1999 人”会被识别出是“数字 + 后缀”的非标准词,因而“1999”被认为是整数。实际上,“1999”指的是“1999 年”,应为数码读法。

为了解决这个问题,引入一个辅助列表,列表中的词是以后缀打头的词,如“人才/天津/时代/元月/角逐/分别/分区……”。这样根据最长匹配策略,“1999 人才”认为是一个单位,在后续处理中会给“1999”一个“数码读法”标记。后缀列表和辅助列表从《现代汉语语法信息词典》[14]及《人民日报》语料中抽取。后缀列表包括 388 个词,辅助列表中的词主要是以单字后缀开头的词,且在实际考察的语料中出现,经筛选后含有 42 个词:时代、人才、天津、余额、元月、角逐、分别、分区、区域、项目、位居、条款、金秋、班机、笔记本、拨号、步枪、部队、部长、车间、车

组、道路、幅面、副总、集团、列车、双座、台阶、行动、站点、支线、支持、环境、环保、代表、股票、开始、开通、问题、包机、包揽、手机、手写。后缀列表和辅助列表中的词可以根据实际情况增加或删除。

3.2 非标准词的消歧

歧义非标准词在不同上下文中有不同的读音,需要进行消歧处理,这是正则化过程中的难点。本文采用最大熵模型进行消歧,对于简单的或不适合机器学习的情形,采用简单规则方法加以处理。

3.2.1 简单规则

FSA仅仅利用词形信息进行非标准词的识别和初始分类。要进行非标准词的消歧,则需要利用更多的信息,包括非标准词内部的信息以及外部上下文信息,这些信息可以通过简单规则加以利用。

以"数字:数字"类型的非标准词"x:y"为例。内部信息的使用:当其中一个数字含有小数点时,表示比值;当$0 \leq x \leq 24$且$0 \leq y \leq 59$时,既可表示时间也可表示比值,否则只能表示比值。上下文信息的使用:利用上下文特征词,例如"时间"特征词有"年/月/日/周/当天/上午/下午/夜/播/航线/起飞/到达/返回"等,"比值"特征词有"比分/局/盘/决赛/首节/半场/选手/对手/名将/赢/胜/负/败/淘汰/不敌/领先/落后/超出/比/投入/供需/率/战/打"等,根据上下文中出现特征词的不同进行分类。

简单规则方法,主要用于处理那些容易消歧,或语料稀少,或语料存在类间严重不均衡的非标准词。

3.2.2 最大熵模型

最大熵模型是一种有效的分类模型。熵是度量随机性的指标,熵越大说明随机性越大。在只掌握未知分布的部分信息时,可以使用最大熵模型对未知分布进行估计。最大熵模型认为:在符合已知信息的条件下,未知情况最合理的分布是熵最大的那个分布,即最随机的那个分布。该分布满足下面的形式:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left\{ \sum_i \lambda_i f_i(x,y) \right\} \quad (1)$$

因此又称最大熵模型为指数模型或对数线性模型。其中,x是已知条件或历史,y是估计结果或类标记。

$$Z(x) = \sum_y \lambda_i f_i(x,y) \quad (2)$$

上式是归一化因子。模型含有一系列特征函数$f_i(x,y)$,例如:

$$f_i(x,y) = \begin{cases} 1 & \text{if } y \text{ 取数码读法}, NSW = 110 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

表示特征i(当前非标准词)取"110"的时候,读音取数码读法。

λ_i是模型要训练的参数,表示特征对预测结果的贡献。参数估计方法有GIS(Generalized Iterative Scaling),IIS(Improved Iterative Scaling),L-BFGS,BLMVM[15]等。数据平滑方法包括高斯平滑、指数平滑、不等式平滑[16]等。本文算法使用BLMVM进行参数估计,使用不等式方法进行数据平滑。

最大熵模型的优点是,使用者可以把注意力集中在特征的选取上,任何有用的特征都可以拿进来,不要求特征之间的条件独立性假设,特征的使用也非常灵活。

3.2.3 特征模版

为每一类非标准词建立一个分类器。这些分类器有公共特征模版,也有私有特征模版。公共特征模版为所有分类器共享,选取特定窗口内的n元字特征。如下所示(窗口大小为4):

$$Uni\text{-}gram: C_n (n = -4, -3, -2, -1, 0, 1, 2, 3, 4)$$

$$Bi\text{-}gram: C_n C_{n+1} (n = -4, -3, -2, -1, 0, 1, 2, 3)$$

$$Tri\text{-}gram: C_n C_{n+1} C_{n+2} (n = -4, -3, -2, -1, 0, 1, 2)$$

$$4\text{-}gram: C_n C_{n+1} C_{n+2} C_{n+3} (n = -4, -3, -2, -1, 0, 1)$$

其中,C_n可以是一个汉字、数字串、字母串或符号。如果上下文中的C_n是数字串,则用串"num"代替,因为在这里具体数字并不重要;也可以缓解数据稀疏现象。

$$C_n = \begin{cases} \text{"num"} & \text{if } n \neq 0 \wedge C_n\_is\_digits \\ C_n & \text{otherwise} \end{cases} \quad (4)$$

私有特征是一些启发性信息,不同类别的非标准词,私有特征会有不同。比如,对"纯数字串"类,私有特征包括:非标准词中数字的位数、是否以零开头、前驱是否是字母、后继是否是字母等。而对"年"这个类,私有特征则是"年"前面的数字的范围。

3.3 标准词的生成

一个非标准词的类别经过识别、消歧阶段后确定下来。标准词生成模块根据类别把非标准词中的非汉字符号转化为汉字,这是一个确定的转化过程,没有歧义,由转换规则加映射表实现。映射表中存储符号到汉字的映射,如"%"对应"百分之","v"对应"伏","Microsoft"对应"微软"等。转换规则包括整数的转换、连续字符读音顺序的调整等。例如,"10-15%"转换成"百分之十到十五","百分之"要放到数字的前面读出。

4　实验及分析

4.1　语料

本文在 2000 年《人民日报》全年语料上考察非标准词。“纯数字串”“年”等六类非标准词采用最大熵模型进行消歧。模型的训练语料从《人民日报》中抽取,平均为每个分类器随机抽取含有非标准词的句子约 1 000 个。封闭测试语料是从《人民日报》中随机抽取含非标准词的句子 6 986 个,含有 13 468 个非标准词。开放测试语料从网上抓取,题材与《人民日报》差别比较大,包含体育、数码产品、军事、BBS 等内容,共 6 007 个含非标准词的句子,含有 12 219 个非标准词。封闭测试语料中基本非标准词的比例为 83.59 %,歧义非标准词的比例为 16.41 %,与第 2 节中全年语料中的分布基本一致;而开放测试语料中基本非标准词的比例为74.77 %,歧义非标准词的比例为 25.23 %。

从非标准词的具体类型来看,封闭测试语料中包括本文定义的 48 种非标准词类型中的 40 种;开放测试语料覆盖其中的 39 种,没有超出本文定义的类别。单个类型非标准词的数量,封闭测试语料最多为 7 474 例(数字 + 后缀),占总体的 55.49 %,最少 1 例;开放测试语料最多为 6 187 例(数字 + 后缀),占总体的 50.63 %,最少 2 例。表 4 给出了测试语料中最大熵处理的各类型非标准词的数量,表 5 列出了其他类型非标准词的数量,其中最后一列“其他”是所有表中未详细列出的各类型非标准词数量之和,分别占封闭测试集和开放测试集中非标准词总数的 7.63 %和 12.24 %。

表 4　最大熵处理的各类型非标准词的数量

	纯数字串	年	年-年	数字-数字	符号“/ ”	符号“- ”
close	275	1 136	53	21	34	90
open	615	667	27	164	121	338

表 5　其他类型非标准词的数量

	数字 + 后缀	整数 + 量级	百分数	日期	小数 + 量级	范围	年数	数字:数字	数字/ 数字	其他
close	7 474	1 018	846	547	369	201	208	103	65	1 028
open	6 187	743	677	503	209	159	83	197	34	1 495

语料的标记方法是:采用一对中括号“[”]”把非标准词括起来,后面跟上非标准词的类型。如果是歧义非标准词,类型后面加上小类,用“/ ”隔开。例如,“拨打 [110] digits/ dd”表示“110”是非标准词,类别是“纯数字串”,按数码方式读。

首先制定语料库的标记规范。在规范的指导下,语料库的标记工作由两人共同完成,标记的一致性用 Kappa 统计量(K)来度量。拿最可能出现不一致的“纯数字串”类来说,随机抽取 250 个样本,其中只有两个不一致,即“桑塔纳 2000 型轿车”中“2000”的读法,一个标记者标记为整数读法,另一个标记者标记为数码读法。计算得到,P(A) = 0.992,P(E) = 0.531,K = (P(A) - P(E))/ (1 - P(E)) = 0.983,具有很高的一致性。

4.2　实验结果

本文非标准词的定义,确保所有的非汉字串都能够通过某一种非标准词识别出来,并得到处理。标准词生成是一个确定的过程,非标准词的识别和消歧决定了正则化的结果。因此,使用非标准词分类(识别 + 消歧)正确率来度量系统的性能。

$$\text{正确率} = \frac{\text{正确分类的非标准词个数}}{\text{非标准词总数}}$$

基准方案(Baseline)是 FSA 识别出非标准词后,对歧义非标准词,赋予它最常见的子类标记。采用消歧模块后的结果和基准方案的结果(基准值)做对比(见表 6),可以看出封闭测试正确率提高 2.96 %,开放测试正确率提高 5.34 %。基本非标准词确保了系统有一个较高的基准值。

表 6　文本正则化的正确率

	close	open
FSA	96.99 %	93.30 %
FSA + SR/ MEM	99.95 %	98.64 %

具体地,非标准词识别的正确率封闭测试为 100 %,开放测试的 12 219 个非标准词中只有一个错误。简单规则消歧的正确率,如“数字:数字”类,封闭测试为 103/103 = 1,开放测试为 180/197 = 0.913 7。不同特征模版下的最大熵消歧结果如表 7 所示。其中,粗体数据为最好结果,符号“-”类型没有使用私有特征,以后会加入地名列表,来确定类似“北京-上海”中“-”的读音。可见,公共特征加私有特征的方案融合了统计属性和先验知识,取得了较好的效果。另外,每个分类器可以选择使各自效果最好的特征模版。

表 7 采用不同特征模版的正确率 (%)

类 别	公共特征		私有特征		公共 + 私有特征	
	close	open	close	open	close	open
纯数字串	97.82	76.75	91.27	86.99	**99.27**	**88.78**
年	96.30	95.35	**99.82**	**98.20**	99.74	**98.20**
年-年	67.92	66.67	**100**	**100**	**100**	**100**
数字-数字	**100**	76.22	42.86	23.17	**100**	79.27
符号“/”	**100**	79.34	91.18	90.08	**100**	**94.21**
符号“-”	**100**	**93.79**	-	-	**100**	**93.79**

图 2 给出了“纯数字串”类的消歧正确率与样本个数的关系,可见样本个数比较少时,同样可以取得很好的效果(200 个样本时,开放测试正确率接近 80 %)。主要是因为私有特征受样本个数的影响较小,并且特征数量虽少(“纯数字串”27 个,“年”5 个,“年-年”6 个),对消歧的贡献却很大(见表 7)。

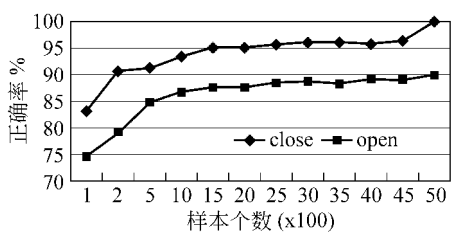

图 2 最大熵模型消歧正确率与训练样本个数的关系

4.3 错误分析

非标准词识别错误。“皇家社会 2-3 瓦伦西亚”中的“2-3 瓦”作为一个表示“范围”的非标准词识别出来。解决办法可以是引入地名或球队名列表(包括瓦伦西亚)等知识。这样“2-3”和“瓦伦西亚”分别作为两个完整的单元,“2-3”被识别出是一个“数字-数字”的类型,并且表示“比分”。

非标准词消歧错误。“数字-数字”类,开放测试的正确率只有 79.27 %。原因是,开放测试集中用“数字-数字”的形式表示比分,如“皇家社会 2-3 瓦伦西亚”,而在训练集中,比分只用“数字:数字”的形式来表示。解决方法是,采集“数字:数字”表示比分的样本添加到“数字-数字”的训练集中。这是样本不均衡现象的一个极端的例子,一个子类的训练样本个数为零。

5 总结

本文对中文文本正则化问题进行了深入的考察,提出了一个清晰的解决方案,引入了机器学习的方法,取得了很好的效果和推广性。从实现上,本文方法避免了复杂规则的书写,不需要分词和标注处理。

作为下一步的工作,第一,改进消歧模块。引入更多知识,如姓氏列表,用于提高“陈 ×”等的处理;地名列表,用于改进“北京-上海”等的处理。对比其他的机器学习方法,如支持向量机等。第二,改进样本选择方法。本文采用随机采样的方法,力求体现样本的自然分布。下一步将重点考察难度更高的样本,研究如何改进效果较差类型的处理,考虑使用主动学习的采样方法。第三,探索文本正则化的实际应用。语音合成的应用有领域之分,如体育、财经、政治、银行等,如何使文本正则化模块易于按领域定制和优化,提高文本正则化模块的鲁棒性,是下一步探索的重点。

参考文献:

[1] Richard Sproat, Alan Black, Stanley Chen, et al.

(下转第 55 页)

[3] 孙基寿. 汉字输入编码优劣评测方法的探讨[J]. 中文信息学报, 2006,20(5): 97-104.
[4] 张玉华,周克兰. 基于规则库的汉字输入法自动评测系统的设计与实现[J]. 中文信息学报, 2004,18(4): 50-54.
[5] GB/T 19246 信息技术,通用键盘汉字输入通用要求[S]. 2003.
[6] 汉语拼音方案.《中华人民共和国国家通用语言文字法》第十八条[S],1996.
[7] 王晓龙. 拼音语句输入系统 INSUN[J],中文信息学报,1993,7(2):45-54.
[8] 谢锦辉,潘小兵. 连续语音识别系统性能评估软件[J]. 计算机应用与软件,1994.
[9] 吕军, 曹效英. 基于语音识别的汉语发音自动评分系统的设计与实现[J]. 计算机工程与设计, 2007,(5).
[10] 王晓龙. 音字流切分及相互转换的理论研究与系统实现[D]. 哈尔滨工业大学博士学位论文,1989.
[11] Cai, L., & Hofmann. Hierarchical document categorization with support vector machines[C]//11 ACM CIKM, 2004.
[12] Dekel, O., Keshet, J., &Singer, Y. Large margin hierarchical classification[J]. ICML '04, 2004, 209-216.
[13] Juho Rousu, Craig Saunders, Sandor Szedmak, John Shawe-Taylor. Kernerl-Based Learning of Hierarchical Multilabel Classification Models [J]. JMLR '06, 2006, 1601-1626.

(上接第 50 页)

Normalization of Non-Standard Words [J]. Computer Speech and Language, 2001, 15(3): 287-333.
[2] Jan van Santen, Richard Sproat, Joseph Olive, et al. Progress in Speech Synthesis [M]. New York: Springer, 1996.
[3] Andrew Breen, Barry Eggleton, Peter Dion, et al. Refocusing on the Text Normalization Process in Text-to-Speech Systems [C]// Proc. ICSLP 2002. 2002: 153-156.
[4] K. Panchapagesan, Partha Pratim Talukdar, N. Sridhar Krishna, et al. Hindi Text Normalization [C]// Proc. KBCS 2004. 2004: 19-22.
[5] M. H. Moattar, M. M. Homayounpour, D. Zabihzadeh. Persian Text Normalization Using Classification Tree and Support Vector Machine [C]// Proc. ICTTA 2006. 2006: 1308-1311.
[6] Virongrong Tesprasit, Paisarn Charoenpornsawat, Virach Sortlertlamvanich. A Context-Sensitive Homograph Disambiguation in Thai Text-to-Speech Synthesis [C]// Proc. HLT-NAACL 2003. 2003: 103-105.
[7] Chilin Shih, Richard Sproat. Issues in Text-to-Speech Conversion for Mandarin [J]. Computational Linguistics and Chinese Language Processing, 1996, 1(1): 37-86.
[8] Min Chu, Peng Hu, Yong Zhao, et al. Microsoft Mulan--a bilingual TTS system [C]// Proc. ICASSP 2003. 2003: 264-267.
[9] 蔡莲红, 魏华武, 周俏峰. 汉语文—语转换中的语言学处理 [J]. 中文信息学报, 1995, 9(1): 31-36.
[10] 陈志刚, 胡国平, 王熙法. 中文语音合成系统中的文本标准化方法 [J]. 中文信息学报, 2003, 17(4): 45-51.
[11] Jianfeng Gao, Mu Li, Changning Huang, et al. Chinese Word Segmentation and Named Entity Recognition, a Pragmatic Approach [J]. Computational Linguistics, 2005, 31(4): 531-574.
[12] Guohong Fu, Min Zhang, Guodong Zhou, et al. A Unified Framework for Text Analysis in Chinese TTS [C]// Proc. ISCSLP 2006. 2006: 200-210.
[13] 俞士汶, 朱学峰, 段慧明. 大规模现代汉语标注语料库的加工规范[J]. 中文信息学报, 2000, 14(6): 58-64.
[14] 俞士汶, 朱学峰, 等. 现代汉语语法信息词典详解(第二版)[M]. 北京:清华大学出版社, 2003.
[15] Steven J. Benson, Jorge J. More. A Limited-Memory Variable-Metric Method for Bound-Constrained Minimization [D]. Technical Report ANL/MCS-P909-0901, Argonne National Laboratory, 2001.
[16] Jun'ichi Kazama, Jun'ichi Tsujii. Evaluation and Extension of Maximum Entropy Models with Inequality Constraints [C]// Proc. EMNLP 2003. 2003: 137-144.