语料资源缺乏的连续语音识别方法的研究

伊·达瓦 [1,2] 匂坂 芳典 [1,2,3] 中村 哲 [1,3]

摘 要 由于少数民族语言有其本身的特点, 不能简单地套用现有的连续语音识别的方法. 本文以蒙古语为例, 研讨了声学和语言模型的建立, 并在日本国际电气通信基础技术研究所的连续语音识别器上实现了蒙古语的语音识别系统. 本文侧重于语言模型的建立, 基于蒙古语黏着性语言特点, 提出用相似词聚类方法建立多类 N-gram 模型. 实验结果显示, 应用我们提出的语言模型, 识别精度比用传统的词的 N-gram 识别法提高了 5.5 %.

关键词 蒙古语, 黏着语言, 相似词分类, 连续语语音识别, 多类语言模型

DOI 10.3724/SP.J.1004.2010.00550

Investigation of ASR Systems for Resource-deficient Languages

I·Dawa[1,3] SAGISAKA Yoshinori[1,2,3] NAKAMURA Satoshi[1,2]

Abstract Because the minority languages in China have their special characteristics, it is not suitable to directly adopt the traditional automatic speech recognition (ASR) methods which are used for some major languages, such as Chinese, English, Japanese, etc. In this paper, we take Mongolian (a resource-deficient language) as an example and build the acoustic and language models for applying the ATRASR system. In this paper, we specially focus on the language modeling aspect by considering the special characteristics of the Mongolian. We trained a multi-class N-gram language model based on similar word clustering. By applying the proposed language model, the system could improve the performance by 5.5 % compared with the conventional word N-gram.

Key words Mongolian language, agglutinative language, similar word clustering, continuous speech recognition, multi-class N-gram model

中国有 56 个民族, 使用文字的语言可达二十多种. 近年来语音通讯技术的开发研究虽然在汉语、英语等语料资源丰富的语言上发展很快, 但是由于缺乏专业人才及完备的语料资源, 目前少数民族语言连续语音识别的开发研究工作尚未起步. 另外, 由于少数民族语言有其本身的特点, 不能简单地套用现有的连续语音识别的方法. 近年来不少使用人口较少的语言如蒙古语、维吾尔语等, 利用通用的语音识别软件 HTK 或 Juilius[1−2], 采取和传统的大语言 (使用人口和地区较多的语言) 同样的方式尝试了语音识别. 因为经济及技术等条件, 这些语言准备的语音–文本数据一般不像大语言那样规模庞大 (大语言一般用 70～500 小时的语音数据来训练声学模型, 7～10 年的报刊杂志等文本数据来训练语言模型). 对于大部分的少数民族语言而言, 虽然目前获得第一手语料较容易, 但语料的手工标注–注释等方面花钱费时困难大. 另外, 他们的声学和语言结构与大语言的差别很大, 即使是仿照大语言的语音识别方式来实现识别器, 最终识别精度远低于大语言的效果[3−4]. 所以对于少数民族语言, 尤其是那些语言资源尚未齐备语言的语音通讯系统的开发研究, 有必要从口语声学特征以及书面语言的实际构造上深入探讨, 找出适合于语言本身的技术途径. 对于语音资源缺乏的语种建模方面有过先行的研究. 如 Schultz[5], Lee[6] 等研讨了借助于齐备的大语言语音数据, 通过统计方法推测小语言的语音参量 (一般用音素单位) 再训练声学模型的方法. 这种方法的出发点在于原语言 (如英语、汉语等) 的发音单位要覆盖并且较接近于目标语言 (如蒙古语、维吾尔语等) 的发音环境. 这种方法较适合于欧美语种, 而不太适合于亚洲语言发音体系. 比如说, 对于蒙古语口语而言, 研究声学模型至少选用 9 个元音, 如: /a, e, i, o, u, ö, ü, ë, æ/, 再加上不同方言的双元音和辅音, 至少设置 40～45 个声学单元. 如果要用日本语或者英语语音数据所设置的语音单元 (通常日本语设置为 37, 英文设置为 32 个单元), 远远覆盖不了目标语 (蒙古语) 所需要的语音环境, 从而难以获得较高的识别精度[7]. 另一方面, 在建立统计语言模型时, 对于文本资源缺少的语言, 目前最常用的方法是利用词类 (Class) N-gram 模型[8]. 但是对于黏着性构造的语言, 由于以下两种原因会使语言信息在词类 N-gram 模型中有可能被丢失: 1) 在

收稿日期 2009-02-06 录用日期 2009-05-04
Manuscript received February 6, 2009; accepted May 4, 2009
日本独立行政法人情报通信研究机构多语言高新技术语音–文本处理研究项目资助
Supported by Multi-lingual Advanced Speech and Text (MASTAR) Research Project of National Institute of Information and Communications Technology (NICT), Japan
1. 日本独立行政法人信息通信技术研究所 京都 日本 619-0288 2. 日本早稻田大学国际信息通信研究科 东京 日本 169-8552 3. 日本国际电气通信基础技术研究所 京都 日本 619-0288
1. National Institute of Information and Communications Technology (NICT), Kyoto 619-0288, Japan 2. Global Information and Telecommunication Institute (GITI), Waseda University, Tokyo 169-855, Japan 3. Advanced Telecommunications Research Institute International (ATR), Kyoto 619-0288, Japan

蒙文中, 虽然一个字符串和字符串之间像英文那样用空格分开, 但是一个字符串不一定是个孤立词, 通常由词根 + 词尾 + 介词连接而成. 实际上可能是一个短语句. 对于这种结构的语言, 像英文中 /in the room/ 介词 + 名词的规则建立 Class N-gram 语言模型当然与实际发音不符合, 导致误识; 2) 在这类语言中有许多种词 (名词、形容词、动词等), 由于词尾的变化而后续词的类型不同而词义变化. 如果用通常前后词的位置关系建立 Class N-gram 模型就会丢失词连接的有用信息而达不到预期的效果. 因此, 本文中我们侧重于这类语言的特征以及语料的现状研讨了连续语音识别器的建立过程. 我们基于 ATR (Advanced Telecommunications Research Institute International) 的语音识别系统展开工作. 对于声学模型的建立, 提出借助于少量母语语料种子 (Seed) 语音样本引导切分语音单元建立语音模型的方法. 对于统计语言模型, 提出基于词性标注的标准词典的相似词法标注文本词性建立多类 (Multi-class) N-gram 语言模型的方法. 本文以蒙古语喀尔哈发音 (蒙古国使用标准发音) 讨论连续语音识别技术, 在整个系统中侧重于语言模型的建立. 基于蒙古语的特点, 提出用相似词聚类方法建立多类 N-gram 模型. 本研究是在日本 NICT (National Institute of Information and Communications Technology) 的 MASTAR (Multi-lingual advanced speech and text research) 课题研究资助下为实现蒙古语对多语言口语翻译 (Speech-to-speech translation) 系统而进行的研究课题.

本文第 1 节简单地介绍蒙古语说话人人口分布以及使用语言文字 – 发音特征的大体情况. 第 2 节将重点讨论黏着语言通过相似词法自动分类词的类型以及实现多类 N-gram 语言模型的具体过程. 第 3 节介绍训练 – 测试语料库. 最后在第 4 和第 5 节中给出本次实验的评测比较结果以及在第 6 节简述结论及今后的工作.

1　蒙古语的基本特点

蒙古语族民众居住在亚洲到欧洲之间的许多国家和地区. 他们使用的口语和文字与中文和西方语言文字比较有许多独特性. 目前蒙古民众还没有使用通用的文字系统. 虽然各地区和国家把各自使用的文字系统称为蒙古语或者蒙文, 但是由于不存在通用的文字语言, 所以所谓的蒙古语实际上是多方言 – 多种文字为前提的抽象语言. 各地使用的文字 – 口语之间有较大的差距, 但书面语言的语法顺序 (SOV (Subject object verb) 结构) 基本相同. 在中国的蒙古族使用的语言被列为少数民族语言.

1.1　人口分布

据语言学者的划分, 蒙古语族分类为: 蒙古语、满族语、锡伯语、達斡尔语等多民族语族类. 而且, 蒙古语本身划分为: 喀尔哈 (Halha) 方言 (蒙古国), 内蒙古方言 (Inner Mongolia) (中国内蒙古自治区), 卫拉特方言 (Oirat) (中国新疆、俄罗斯的卡尔梅克联邦国) 以及布里亚特方言 (Buirat) (俄罗斯布里亚特联邦国) 等. 据最近统计, 蒙古语族人口约有 850 万人口, 其分布见图 1[9].

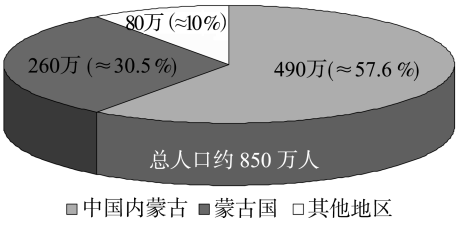

图 1　话者人数占地区人口比例

Fig. 1　Proportion of the local population with speakers

1.2　语言 – 文字

蒙古族民众目前仍使用如图 2 所示的 3 种形式的电子化文本. 比如图 2 中的传统蒙文 (Traditional Mongolian (TM)), 托忒文 (Todo) 及新蒙文 (New Mongolian (NM), 或者称 Cyrillic 文字).

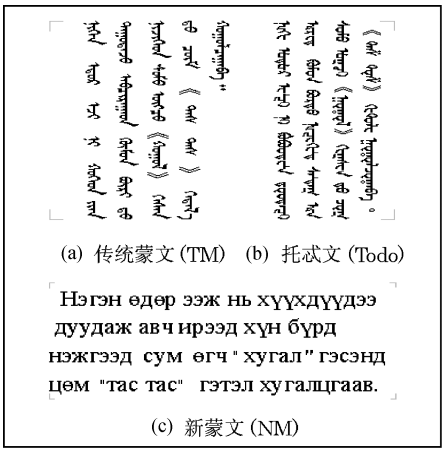

图 2　电子化文本种类

Fig. 2　Digital versions by Mongolian

其中, TM 现在主要使用于中国内蒙古地区, Todo 现使用在中国新疆和俄罗斯的卡尔梅克等地区. 相比较 TM, Todo 文字有易于读写、便于机器处理等特点[10]. 大约 70 年前蒙古国由于 TM 文字的结构复杂、不易学习使用等原因停止使用 TM, 而改用了基于俄罗斯字母的斯拉夫文, 即新蒙文[11−12]. 至于 NM, 目前除了蒙古国之外还有俄罗斯国的卡尔梅克、布里亚特等地区作为常用媒体文字使用. 蒙文中, 不论是纵向竖写的文本, 还是横向写体 NM 文, 每个输入量 (严格说不是一个词) 之间有空格区分. 在 TM 中, 一个输入量由词根 (固定量) 和词尾 (变量 → 引起词意变化) 组成, 而新蒙文中, 这种组合甚至可以达到词根 + 词尾 + 复数形 + 介词等多个变量的组合 (见图 3).

TM (短语): /n /p 把孩子们

Cyrillic (短语): хүүхдүүдээ = хүүх/н+ дүү/рl+ дээ/р

图 3 TM 和 NM 文本中短语表现形式

Fig. 3 Phrase form in Mongolian

1.3 发音特征

图 4 为蒙古语口语中 9 个元音的发音位置图[13], 而图 5 和图 6 分别给出蒙古语喀尔哈方言发音和卫拉特方言发音 7 个元音的共振峰频率 F_1-F_2 分布图[14]. 为便于比较, 在图 7 中给出了日本语 5 个元音 (/a, e, i, u, o/) 的共振峰频率 F_1-F_2 分布图[15].

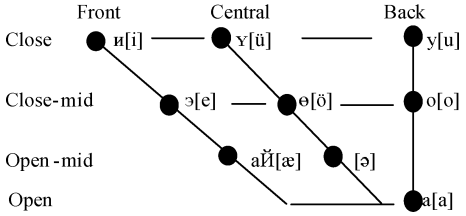

图 43 蒙古语基础元音发音位置图

Fig. 4 Articulatory pattern of Mongolian

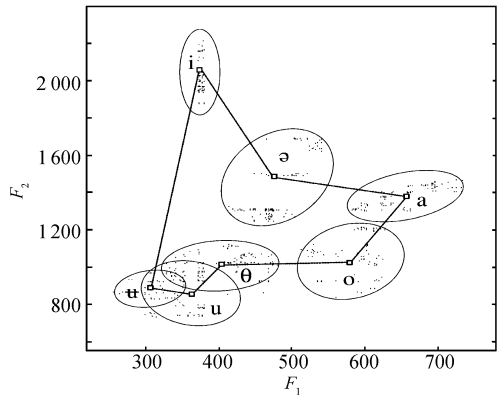

图 5 喀尔哈方言发音 F_1-F_2 分布图 ($\partial=$ ë)

Fig. 5 F_1-F_2 distribution of seven vowels by Halha dialect

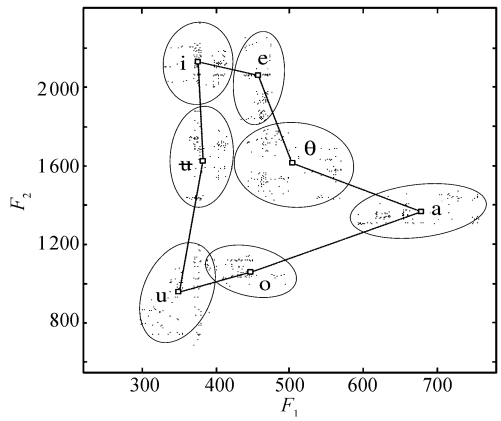

图 6 卫拉特方言发音 F_1-F_2 分布图

Fig. 6 F_1-F_2 distribution of seven vowels by Oirat dialect

通过比较可以发现, 各图中元音 /a, i/ 的发音分布基本相似. 而蒙古语喀尔哈方言发音的/$\partial=$ ë/, 卫拉特方言发音的 /o/ 音分别接近于日本语的 /e, o/ 音. 卫拉特方言发音的 /ʉ/ 音相似于日本语的 /u/ 音. 尤其是图 5 中的元音 /ʉ, θ, u/ 的分布重叠在一起, 这与图 6 和图 7 中的分布有较大的差别. 基于这样的比较, 我们可以看出, 不同语言系统的声学特征空间不能相互覆盖. 所以在声学模型的建立中必须考虑语言的特定声学空间特点. 由于声学模型的建立不是本文的重点, 我们在系统实现部分将作简单介绍.

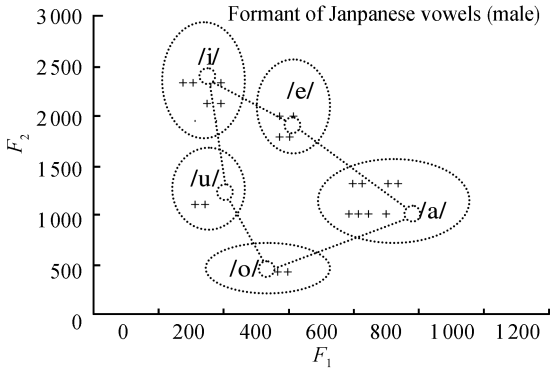

图 7 日本语发音 F_1-F_2 分布图

Fig. 7 F_1-F_2 distribution of five vowels by Japanese

1.4 黏着语言 (Agglutinative language)

蒙文分类为阿勒泰语系语族, 属于黏着性语言, 其特点是: 1) 基本词素 (Morpheme) 按词中形态连接而构词; 2) 词根 (Root/stem) 固定不变; 3) 词根添加词尾 (Suffix/affix) 构成新词; 4) 词尾可以按语言学规则有限变化, 从而改变词义; 5) 由于词尾的不同而后续词的词性连接关系发生变化 — 词义变化[16]. 基于这种特点, 在构造蒙古语识别系统时, 其语言模型的建立也需要相应的变化, 这是本文的侧重点. 我们将在下一部分作详细介绍.

2 基于类语言 (Class N-gram) 语言模型

训练 N-gram 语言模型时, 对容量为 V 的训练集要产生 V^N 个 N-gram 参量, N 增大参量就要急剧增大, 所以一般采取词类 N-gram 语言模型的方法. 对于词串 $W=w_1, w_2, \cdots, w_Q$, 假设语料中词与词类标记的对应情况都是独立的, 即对于当前词类与前 $N-1$ 个词类出现的漂移概率 (即语言模型) 由下式给出:

$$P(W)=\prod_{i=1}^{Q}P(c_i|c_{i-1},\cdots c_{i-N-1})P(w_i|c_i) \quad (1)$$

其中, c_i 为当前单词 w_i 所属的类, Q 是词数. 这样, 对于词类为 C 的词, 要推测的参量应该从词 N-gram 推测时的 V^N 降到 C^N. 因此, 由于使用的类

标记数较少, 在规模不大的训练集上也能够可靠地推测 N-gram 频度. 可是, 由于这种方法使得词与词间的连接性不能够充分地表现, 即使是用足够大的训练集, 词间连接性的推测一般也不如孤立词时的精度[17].

2.1　词类及多类 N-gram 模型

对于类 N-gram, 词类分类是用来反映属于词类的词间的连接关系的. 而在常用的词类分类法中, 把当前词的前后词视为同等的连接性而不区分. 如果对前述的黏着性语言也采用这种方法, 由于词尾的变化而引起当前词、前后词的词类性不同而词义变化会丢失大量的有用的词连接关系信息. 这是中文和黏着语言的根本性差异. 所以对于本文讨论的 NM 文本建立类 N-gram 语言模型时, 有必要分开考虑一个词前方的连接关系和后方的连接关系. 因此, 我们提出词多类分类 (Multi-class) 和词多类 N-gram 的方法. 也就是说, 借助于我们提出的相似词推类词性方法 (见 2.2 节介绍) 获得标注语料, 然后用式 (1) 推测词多类 N-gram 语言模型. 当 $N = 2$ 时, 式 (1) 应改写为式 (2):

$$P(W) = \prod_{i=1}^{Q} P(c_i^t | c_{i-1}^{f^1}) P(W | c_i^t) \tag{2}$$

这里, c^t 表示后续词类 (To class), c^f 为先行词类. 这样, 词的分类有以下步骤进行:

步骤 1. 给每个词分配一个特定类.

步骤 2. 对一个类或者对任意一个词 x 分配一个向量用来反映词间的相关性. 即

$$v^t(x) = [c_1(p^t(w_1|x)), \cdots, c_N(p^t(w_N|x))] \tag{3}$$

$$v^f(x) = [c_1(p^f(w_1|x)), \cdots, c_N(p^f(w_N|x))] \tag{4}$$

这里, $c_i(p^{t(f)}(w_i|x))$ 表示在第 i 类 x_i 中, 从词 x 到词 i 的经平滑后的先行或者后行词的 N-gram 值.

步骤 3. 通过下式 (5) Merge cost 把合并损失最小的两个类合并为一个类:

$$\text{merge cost} = U_{\text{new}} - U_{\text{old}} \tag{5}$$

其中

$$\begin{aligned} U_{\text{new}} &= \sum_w p(w) D(v(c_{\text{new}}(w)), v(w)) \\ U_{\text{old}} &= \sum_w p(w) D(v(c_{\text{old}}(w)), v(w)) \end{aligned} \tag{6}$$

步骤 4. 重复以上循环步骤 2 至步骤 3, 直到计算全类数 C 为止.

2.2　相似词分类方法

为了获得训练集中输入量 (某单元) 词性的大体分类 (Parts of speech, POS), 本文采用了基于编辑距离 (Edit distance, ED) 法的相似词分类方法[18]. 图 8 显示本文提案的实现相似词分类的示意图.

$$Ed(A, B) = 1 - 2 \times \frac{d(A, B)}{I + J} \tag{7}$$

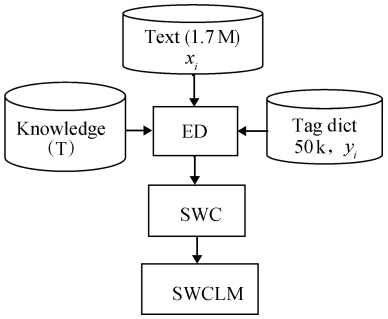

图 8　相似词分类概要图

Fig. 8　Block diagram of similar word clustering

首先, 把表 1 中所示训练文本综合为一个大的文本 (词量 1.7 M); 再用标点符号 (如: [/, . ? ; : ! /] 等) 信息切分文本成若干个短文 (本实验中切分短文数为 170 k); 然后对文本的每一个输入量 x_i 在第一个音节相同的条件下, 根据词尾知识库 T 提供的规则信息, 用式 (7) 计算出参考量 y_j 间的相似度 Ed, 最后从中选取相似度高的词并添加词性码标注存放于 SWC (Similar word clustering) 中 (如图 9 所示). 剩余的未能够标注的词抽出后用人工标注. 本实验使用的词性信息标注词典共有 5 万个常用词, 含 75 类标注码. 图 8 中, SWCLM (Similar word clustering language model) 是用这种相似词训练的聚类 N-gram 语言模型 (Similar word class N-gram model). 图 10 给出了相似词实现的一个实例. 从这个实验结果可以知道, 对于输入量 ($A = /bolj/$) (动词), 当设定一个临界值 ($\alpha < 0.6$) 时, 可以选出候选词 B_1, B_2 及 B_3 分类为一类[19].

表 1　NM 文本语料库

Table 1　NM text database

	会话	NN 出版	下载	总词
训练集	160 k	920 k	620 k	1.7 k
标注词典		50 k		15 M
测度集	2.5 k			

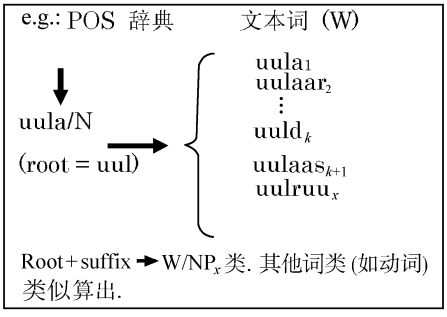

图 10　相似词分类举例

Fig.10　An example of similar word tagging

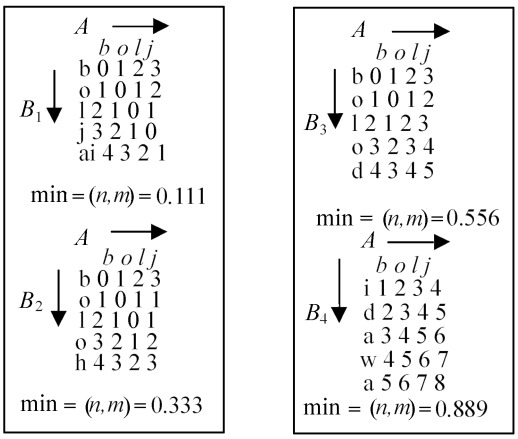

图 10 输入量 /*bolj*/ 的 ED 算法实例

Fig. 10 An example of string /*bolj*/

3 数据准备

为建立蒙古语的识别系统, 我们收集建立系统所需要的语音和文本语料. 语料库的规模、内容等实验数据如表 1～3 所示. 有关语料方面的细节阅文献 [20−21]. 其中测试集中含有会话语句 500 个, 教材用标准短语 2 000 个.

表 2 蒙古语喀尔哈发音口语语料

Table 2 Speech corpus of Mongolian Halha

Halha	发话人数	发话句/小时
训练集	77 (40 男, 37 女)	3 000/10.5
测试集	6 (3 男, 3 女)	500/0.7

表 3 ATR BTEC3 (旅游会话文本语料)

Table 3 ATR BTEC3 (Basic travel expressions corpus)

	BTEC3	蒙文 (NM)	中文	日本语
句子数	133 454	133 454	133 454	

蒙古语使用人口较多的地区是内蒙古地区 (见图 1), 本地区使用 TM 文字. 考虑到 TM 文本中习惯书写用语较多出现, 常常口语和文字语不对应. 这在口语识别时会引起未登录词 (Out of vocdbulary, OOV) 结果, 使得识别精度大幅度下降. 为此我们在实验中选用蒙古国现用的 NM 文字文本 (Text) 以及喀尔哈发音口语语料.

4 语言模型的性能评估

基于上面的讨论, 在建立整个语音识别系统中, 我们针对蒙古语的特点, 建立语言模型和声学模型. 语言模型的建立在第 2 节中作了详细的讨论. 在测试语言模型在识别中的作用前, 我们先用词分类实验及语言模型的信息熵和困惑度来评价其优劣.

4.1 词分类实验

首先我们评价本文提出的基于词分类实验的语言模型. 在表 1 (长度为 1.7 M) 文本中, 通过上述 ED 法推测相似词并自动标注时, 在设定的临界值 $\alpha < 0.6$ 时得到了 437 个词类. 没有能够自动标注的词有 23 000 个, 即正确标注率为 98.64 %. 剩余的未能标注的词大部分都是地名或者人名等固定名词类, 本实验中实施了人工添加词性的方法.

4.2 2-gram 语言模型的评估

目前评估语言模型性能的常用尺度是困惑度 (Perplexity). 其算法由式 (8) 给出:

$$\begin{aligned} \text{Entropy} &= \frac{1}{L}\sum_{i}\log_2(p(w_i)) \\ \text{Perplexity} &= 2^{\text{Entropy}} \end{aligned} \tag{8}$$

其中 $p(w_i)$ 为单词 w_i 在长度为 L 的文本中出现的概率. 本次实验中, 我们评测了基于两种类型的数据集的 2-gram 语言模型的 Perplexity 和熵 (Entropy). 首先, 利用表 1 所示本文的综合数据 (词量 1.7 M), 实验集量为 6 000 词. 图 11 中给出了本文提出的多类方法产生的 Class 2-gram 语言模型和常用词的 2-Gram 模型时的 Perplexity 比较结果. 从图 11 可知, 在分类数计数到 1 200 附近时类 2-gram 的 Perplexity 取最低值 24.6. 这个值低于词单元 2-gram 的值 32.2.

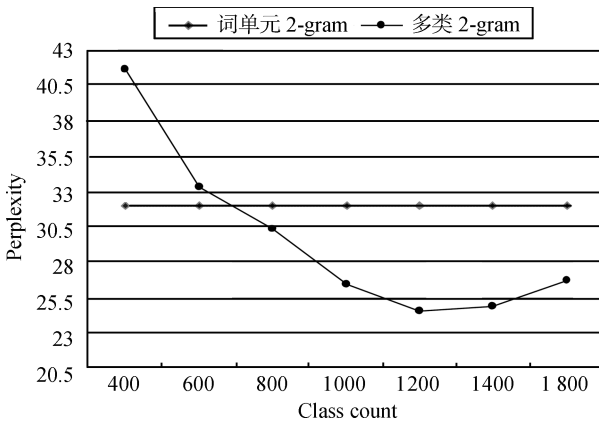

图 11 多类 N-gram 的 Perplexity 比较

Fig. 11 Evaluation of multi-class N-gram in perplexity

其次, 利用表 3 所示的语料来比较提案方法获得的多类 2-gram 与其他语言 (ATR BTEC3 中中文和日本语) 数据获得的多类 2-gram 模型在含信息量上的差异. 各语言文本信息量可以利用式 (8) 算出. 从实测结果 (表 4) 可知, 即使是同量同意义的语料, 每文所需要的平均熵为中文时最大, 为蒙文时最小. 这意味着, 对于 ATR BTEC3 语料中的蒙文, 要推测所需要的信息量小于日本语和中文的信息量. 这可能是黏着性语言独特的语言构造特点.

表 4　多语言并行文本信息量比较

Table 4　Comparisons of entropy in parallel corpus

	蒙文	日本语	中文
测试集	6 k	6.2 k	5.8 k
平均熵	157.6	165.8	294.7

5　语音识别实验

有了整个识别系统, 可以评价本文提出的语言模型在识别性能上的作用. 为此首先建立声学模型, 我们在这里作一个简单的介绍.

5.1　声学模型的建立

由于第一手语料数据的收集切分以及人工注释 – 标注需要投入大量的人力和物力, 因此, 蒙古语的语音处理方面的研究目前还处在初期阶段. ATR 语音识别系统声学模型的建立需要预先切分标注的数据, 所以在利用该系统建立声学模型时需要准备切分标注语料. 由于表 2 所示的口语语料较大 (约 10.5 小时), 没有条件进行人工切分 – 标注. 所以本文研讨的蒙古语语音翻译系统的语音模型部分是经过以下途径实现的:

1) 语音特征分析: 特征量 25 维 Mel 频率倒谱系数 (Mel-frequency cepstrum coefficients, MFCC), ΔMFCC, Δ对数能量. 采样周期为 16 kHz, 帧宽 10 ms;

2) 种子 (Seed) 模型的建立: 从表 2 语料中选择 10 个发话人语音 (5 男, 5 女, 发话时间 = 1.5 小时), 设置 40 个声学建模单元 (见表 5) 并进行人工标注, 再利用 HTK toolkit 对以上 10 个发话人语料进行声学模型的训练, 产生种子 (Seed) 模型.

表 5　音素建模单元的定义

Table 5　Definition of phone set

单元类性	建模单元
元音/双元音	a e i o u ox v ex ai/aa ee ii oo uu ox2 v2
辅音	b p h k g :h l m s s: t d q c j z y r w f
韵尾	B G R S D N L ng S: sil

3) 语音数据切分: 如图 12 所示, 对表 2 语料中剩余的 67 个发话人语音 (WAVE), 通过信号处理抽出语音特征量 (MFCC), 并添加文本 TRS (Transcribution) 文件. 利用 Viterbi alignment 算法对特征量借助于标记符 (Label) 切分. 并且对每个切出的标记符, 按前后两个标记符的组合产生学习用数据. 再利用学习用数据在 ATRASR 上训练新的声学模型. 实现过程是: 学习数据的生成 ⇒ Topology 学习 ⇒ Label 学习 ⇒ 连接学习. 训练出来的语音模型是 3 个因子 (Triphone) HMnet 格式.

4) 重复训练: 把第一次的样本 Seed 模型 (2) 中生成的) 用新的声学模型来替换重复多次训练出最终的声学模型 HMnet (Aconstic model, AM).

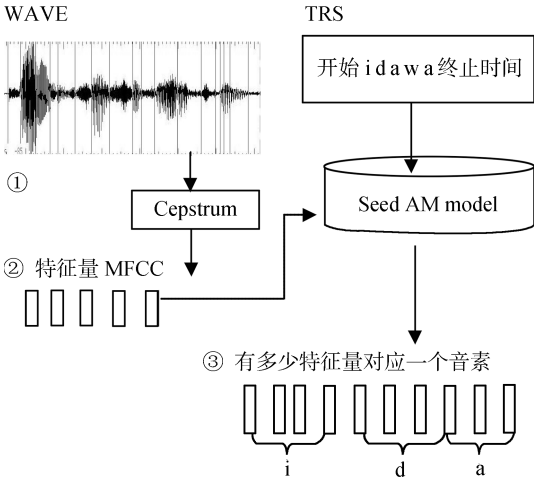

图 12　HMnet 语音模型生成过程原理图

Fig. 12　Creating acoustic model HMnet

本文尝试了 3 种识别实验. 即系统对单音素的识别精度, 系统在不同的语言模型上的单词识别精度及蒙古语的识别精度与其他语的识别精度的比较. 识别器用 ATR 开发的 ATRASR 连续语识别器[22−23].

5.2　单音素 (Phoneme) 识别

图 13 出了单音素识别率 (Phoneme recognition rate, PRR), 其中, LV 为蒙文的长元音, C 为辅音. 从图 13 可以看到, 单音素的平均识别率约为 71 %. 另外, 除了摩擦音 /k, c/ 及标记外来语音标 /f/ 之外, 大部分音素识别率均大于 65 %. 其中元音和长元 /ox, v, ox2/ 的识别率低于其他元音和长元音的识别率, 这可能由于图 5 中这些元音 (ox = θ, v = ü) 的发音特征重迭而造成的.

SV	a	e	i	o	u	ox	v
PRR	81.2	73.6	70.3	74.6	71.7	59.8	58.8
LV	aa	ee	ii	oo	uu	ox2	vv
PRR	65.3	66.4	78.3	67.0	81.2	53.4	67.9
C	n	b	p	h	k	:h	g
PRR	79.0	77.0	67.5	78.2	57.2	66.4	68.3
C	l	m	s	s:	t	d	q
PRR	79.6	82.0	89.5	62.1	67.0	62.7	75.7
C	c	j	z	y	r	w	f
PRR	58.7	69.1	65.8	71.4	75.4	73.1	28.2
C	ng	N	ai	ex	sil	平均识别率	
PRR	86.1	77.4	74.8	77.1	97.7	约 71 %	

图 13　单音素识别结果 (%)

Fig. 13　Recognition results of phonemes (%)

5.3　连续语识别

为了在连续语识别中评价在前几节中所讨论的各类统计语言模型以及声学模型的性能, 我们同时也进行了蒙古语连续语识别实验. 测试语句用表 2 测试集中的 6 个发话人发话的 500 个短语 (共 3 540 词).

图 14 显示了在不同统计语言模型下的单词平均识别率. 在传统的词的 2-gram 模型 LM (Language model) 的最好识别结果为 82.1 % (男), 在传统的多类 2-gram 模型 (Standard class model, SCLM) 时的识别率为 84.3 %, 而在本文提案的相似词多类 2-gramy (SWCLM) 模型时的识别率达到了 87.6 %, 相对 LM 的结果提高了 5.5 %.

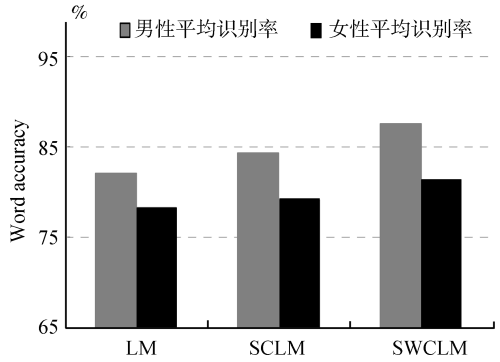

图 14 不同 2-gram 语言模型下的单词识别比较

Fig. 14 Word recognition results by different 2-gram models

5.4 多语言识别比较

图 15 是蒙古语喀尔哈发音连续语识别结果与其他大语种语言 (汉语和日本语) 识别精度的比较结果. 文本训练数据用表 3 多语言并行语料. 类数 2 000, 多类 2-gram 模型. 测试语句: 日本语、中文均用 510 个短语, 蒙古语用 500 个短语. 各语言发话人数分别为 3 人 (男). 从本次实验的结果 (图 15) 发现, 与使用大规模训练数据的日本语及中国语比较 (日本语训练集发话人数 620 名, 中国语 540 名, 而蒙古语使用语音训练集为 77 名), 虽然蒙古语使用的数据量远小于上述两个语言使用的数据量, 但实际的识别精度略接近于中国语的识别精度. 另外我们也发现, 训练数据较大, 且也有像蒙古文那样动词语尾变化特点的日本语的识别精度明显较好. 如果训练集数量增加, 数据准备 – 整理情况改善, 再利用本文提案的相似词分类多类 N-gram 模型, 估计可以进一步改善识别精度.

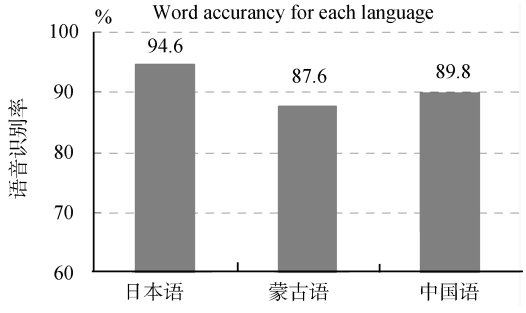

图 15 不同语言类 2-gram 单词识别率

Fig. 15 Word recognition results by multi-class 2-gram models

6 结论与展望

本研究是在日本 ATR 研究开发的多语言语音自动翻译系统中追加蒙古语部分而设置的研究. 针对缺少实验数据的语言 (如中国少数民族各语言), 如何快速实现语音识别以及人机通讯系统, 并提高实时识别精度, 本文以蒙古语作为讨论的对象, 重点考察了黏着性语言文本相似词分类多类 N-gram 语言模型的方法. 本次旅游会话口语连续语识别实验结果显示, 比较常用的词的 N-gram 语言模型及直接用词类 N-gram 模型方法, 通过提案的方法可使识别精度有所改善, 可以确认这种方法是可行的. 同时也发现, 通过少量发话人语音种子语音模型引导大数据自动切分建模有助于那些语言资源缺少的少数民族语言速建连续语音识别 (Continuous speech recognition, CSR) 系统. 另外还发现, POS tagging 辞典的规模、精度以及相似词分类的正确率均会影响系统精度. 如果训练集数量增加, 数据准备 – 整理情况改善并且使用多层次词类 N-gram 语言模型, 估计识别精度会进一步改善, 这是我们下一步工作的重点.

致谢

蒙古国国立大学语言学院阿勒泰语及语言学教研室的 SH · Choimaa 教授为首的各位老师对于蒙古文语音 – 文本数据的收集、整理方面给予了大量的支持和帮助, 日本 NICT/ATR 语音翻译组的各位学者为本系统的设计 – 实现提供软件并给予了热心的建议和意见, 在此表示感谢.

References

1 Young S J, Evermann G, Gales M J F, Hain T, Kershaw D, Moore G. *The HTK Book, Version3.4*. Berlin: Springer, 2006

2 Kawahara T. Participants′ areas of research and technical work [Online], avaiable: http://www.julius.scorceforge.jp/, March 17, 2009

3 Dawa I, Lu Xu-Gang, Shimizu T, Sagisaka Y, Nakamura S. A continuous speech recognition system considering language characteristics in linguistics and text structures. In: Proceedings of the 10th National Conference on Man-Machine Speech Communication. Lanzhou, China: Xinjiang Normal University Publisher, 2009. 57
(伊 · 达瓦, 卢绪刚, 清水徹, 中村哲. 蒙古语连续语音识别在不同结构语言模型下的精度讨论. 第十届全国人机语音通讯学术会议. 兰州, 中国: 新疆师范大学出版社, 2009. 57)

4 Tao Mei, Silamu W, Tursun N. The uyghur acoustic model based on HTK. *Journal of Chinese Information Processing*, 2008, **22**(5): 56−59
(陶梅, 吾守尔斯拉木, 那斯尔江. 基于 HTK 的维吾尔语连续语音声学建模. 中文信息学报, 2008, **22**(5): 56−59)

5 Schultz T, Waibel A. Experiments on cross-language acoustic modeling. In: Proceedings of the 7th European Conference on Speech Communication and Technology. Aalborg, Denmark: ISCA, 2001. 567−570

6 Lee C H. Attribute-based universal phone modeling for multilingual speech recognition. In: Proceedings of the Interna-

tional Conference on Speech Databases and Assessments. Kyoto, Japan: NICT, 2008. 1−28

7 Dawa I, Okawa S, Shirai K. Inquiry onto common acoustic model to realize Mongolian dialectal speech recognition. *Journal of the Central University for Nationalities (Humane and Social Sciences)*, 2001, **28**(4): 114−121
(伊 · 达瓦, 大川茂树, 白井克彦. 蒙古语多方言语音识别及共享识别模型的探索, 中央民族大学学报 (哲学社会科学版), 2001, **28**(4): 114−121)

8 Yamamoto H, Isogai S, Sagisaka Y. Multi-class composite N-gram language model using multiple word clusters and word successions. *IEIC Technical Report*, 2001, **101**(156): 13−18

9 National Bureau of Statistics of China. Report of the population [Online], avaiable: http://www.stats.gov.cn/enGliSH/, July 22, 2006

10 Dawa I, Nakamura S. A study on cross trasformation of mongolian language. *Journal of National Language Processing*, 2008, **15**(5): 3−21

11 Shagdarsun T. *Mongolyn Utga Soyolyn Tovchoo*. Ulaanbaatar: Science Publisher Pvvs, 1991

12 Sambuudorj O. *The Phoneme System of Oirat Dialect of Mongolian*. Ulaanbaatar: Culture and Education Publisher, 2000

13 Dawa I, Sambuudorj O, Arai Y, Eredenbat D. *A Brief introduction to the Mongolian Languages*. Los Angeles: Word Scientific, 2008

14 Dawa I, Okawa S, Shirai K. Acoustic features analyses of Mongolian dialects vowels by computer. *Acta Acustica*, 1999, **24**(1): 94−97
(伊 · 达瓦, 大川茂树, 白井克彦. 蒙古语七个元音声频特性计算机分析. 声学学报, 1999, **24**(1): 94−97)

15 Furui S. *Sound and Speech Technology Introduction*. Tokyo: Kindai Science Publisher, 1996

16 Dawa I, Sagisaka Y, Nakamura S, Shirai K. A data driven approach to rescuing dangerous languages. *Journal of the Western Mongolian Studies*, 2009, (1): 44−55
(伊 · 达瓦, 匂坂芳典, 中村哲, 白井克彦. 一种基于语料库的濒危文化拯救方法, 西部蒙古论坛, 2009, (1): 44−55)

17 Yamamoto H, Kikui G, Nakamura S, Sagisaka Y. Speech recognition of foreing out-of-vocabulary words using a hierarchical language model. In: Proceedings of the International Conference on Speech Communication Association and International Conference on Spoken Language Processing. Pittsburgh, USA: ACM, 2006. 267−270

18 Masek W, Paterson M. A faster algorithm computing string edit distances. *Journal of Computer and System Sciences*, 1980, **20**(1): 18−31

19 Dawa I, Sagisaka Y, Nakamura S. Modeling characteristics of agglutinative languages with multi-class language model for ASR system oriental. In: Proceedings of the International Conference on Speech Database and Assessments. Washington D. C., USA: IEEE, 2009. 104−109

20 Dawa I. Processing of Mongolian by computer. *Journal of Chinese Information Processing*, 2006, **20**(4): 56−62
(伊 · 达瓦. 蒙古语语言 – 文字的自动化处理. 中文信息学报, 2006, **20**(4): 56−62)

21 Dawa I, Liu Y, Yue Y M, Cheng B S, Arai Y, Mitsunaga M. Multilingual text-speech corpus of mongolian. In: Proceedings of the International Symposium on Chinese Spoken Language Processing. Kent Ridge, Singapore: ISCA, 2006. 759−770

22 Advanced Telecommunications Research Institute International Automatic Speech Recognition, Acoustic Model Creating Tools, 2007

23 Advanced Telecommunications Research Institute International Automatic Speech Recognition, Speech Recognition Engine Memu, 2004

伊 · 达瓦 日本独立行政法人信息通信技术研究所 (NICT) 高级研究员. 2000 年获得早稻田大学博士学位. 主要研究方向为语言信息处理及语音翻译. 本文通信作者.
E-mail: dawa.idomuco@gmail.com
(**I · Dawa** Senior researcher at National Institute of Information and Communications Technology (NICT). He received his Ph. D. degree from Waseda University, Japan in 2000. His research interest covers spoken language processing and speech translation. Corresponding author of this paper.)

匂坂 芳典 日本早稻田大学国际信息通信研究科教授, 日本国际电气通信基础技术研究所 (ATR) 研究员. 主要研究方向为语音合成技术, 语言识别技术, 语音信号处理以及语言信息处理技术.
E-mail: yoshinori.sagisaka@nict.go.jp
(**SAGISAKA Yoshinori** Professor at Global Information and Telecommunication Institute (GITI), Waseda University, and researcher at Advanced Telecommunications Research Institute International (ATR). His research interest covers speech synthesis, speech recognition, speech signal processing, and language information processing.)

中村 哲 日本国际电气通信基础技术研究所 (ATR) 上席研究员. 主要研究方向为自由发话语音处理技术, 语音识别, 语音合成技术, 以及多语言语音翻译技术.
E-mail: satoshi.nakamura@nict.go.jp
(**NAKAMURA Satoshi** ATR (Advanced Telecommunications Research Institute International) fellow. His research interest covers spoken language processing, speech recognition, speech synthesis, and multilingual speech translation.)