汉语语音识别的纠错处理

张全*　　张倪**　　韦向峰**

*中国科学院 声学研究所 北京 100080
**中国科学院 北京软件工程研制中心 北京 100080
E-mail: Quanzhangioa@263.net

摘要：n-gram 作为语言处理模型已经广泛应用于语音识别系统中，但处理结果中经常出现错误。本文指明影响语音识别错误产生的主要因素，制定了基于概念层次网络（简称 HNC）句类分析技术的纠错处理策略。同时构筑了纠错处理的实验系统。测试表明，该技术能够突破现有统计模型的局限，取得理想的结果，有望成为提高语音识别性能的、新的语言处理模型。

关键词：语音识别分析、语言处理模型、语句理解、句类分析、纠错处理

The Correction in Chinese Speech Recognition

ZHANG Quan*　　ZHANG Ni**　　WEI Xiangfeng*

*Institute of Acoustics, CAS , Beijing 100080
**Center of Advance Software Engineering (CASE), CAS, Beijing 100080
E-mail: Quanzhangioa@263.net

Abstract: The N-Gram is used widely in Speech Recognition Systems, however, the errors are often occurred in the output. In this paper, the main issues which cause the errors are discussed., and the correcting tactics based on the Hierarchical Networks of Concepts (HNC) Analysis of Sentence Category (SCA) technology are described. Meanwhile an experiment system based on the tactics have been built. The result of the system indicated that, the SCA is able to break through the limits of the statistical models, to get a better result. It is hopeful that the SCA will be used as a new linguistic processing model which is able to promote the performance of speech recognition system.

Keyword: Analysis of speech recognition, the model of linguistic processing, sentence understanding, sentence category analysis, correction processing.

1 引言

目前，语音识别系统中一般采用 n-gram 作为语言处理方法。n-gram 应用于语音识别系统，大大提高了识别系统的稳定性和正确率，促进了语音识别系统的实用化，同时由于自身

的缺陷，也阻碍了语音识别系统上述两指标的进一步提高。[1]中指出“（面对多种模糊）大脑的语言感知应付裕如，表现了强大的解模糊能力，自然语言处理技术当前无从望其项背”，“一言以蔽之，它（自然语言传统分析模式，含统计）不是描述语言感知过程的适当模式”。

来看一个例子。

例 1　　他身上穿着一件淡褐色倡议。（她身上穿着一件淡褐色绸衣。）

可以说，即使这个句子的最后一个词识别的结果与人脑的正常语言感知结果相去甚远，但仍然可以肯定地说这样的识别结果已经相当不错了：系统语音的识别已经相当准确了，同时也给出了正确率很高的汉字。但如果要提升处理性能，让计算机产生类似于人脑语言感知处理的结果，就需要建立起一种模拟人脑语言感知过程的、适合于计算机处理的语言处理模型。这一模型的首要特点是，具有从语义深层理解语句的能力，语句理解定位于概念联想脉络运作全过程的激活[2]。HNC 句类分析技术恰好具有这一特点。

将句类分析技术应用于纠错处理的主要思路是：测试识别系统，获取系统处理能力的基本数据，获取系统错误识别音节的混淆音集合。前者是对处理系统策略的支持；后者测得的数据将用于纠错处理。当纠错处理发现识别系统产生的结果是一个有错误的句子，首先根据线索确定错误的位置，由于识别系统只能给出汉字结果不提供中间结果，所以回到有错汉字对应的音，看是否能找到正确的词语；如果无法找到，考虑混淆音组成的词。混淆音为纠错处理提供更多的候选，使其尽可能找到符合概念联想脉络的正确词语。

2 错误分析

纠错是在识别系统产生结果的基础上进行，首先需要了解识别系统性能，对识别系统的各种错误情况有一个总体的认识。

使用三个发音人、5 万字的声音语料对语音识别系统进行了测试。下文对识别结果中的错误从音节识别错误和文字结果错误两个方面分别进行了分析，明确了混淆音数据的获取策略和纠错处理的要点。

2.1 音节识别错误的分析

考察语音识别系统对音节的识别情况。从一个句子的角度来看，识别输出的音节个数可能会多于实际句子音节的个数，这种错误本文称为添音；也可能少于实际句子的音节个数，这种错误称为吞音。还有一种情况就是识别音节个数与实际句子的个数相同，但是有的音节是错误的，这种情况称为错音。

错音	78.2%
添音	13.3%
吞音	8.6%

对错音的纠错处理，需要音节的混淆音数据。形成混淆音数据可以考虑两种方式。一种是使用大量的测试语料收集每个音的混淆音数据，形成音节混淆音数据库。这种方法得到纯粹的经验数据，比较真实全面地反映系统的实际情况。困难在于要使用大量的、多个发音人

的语音语料，以便对每个音节都能得到稳定的统计结果。工作量是非常大的。另一种方法是，考虑汉语的音节由声母和韵母构成，先测试声母和韵母的混淆矩阵，然后根据声韵组合，就可以构造出每个音节的混淆音。这种方法可以大大减少工作量，如果识别系统采用的是声韵分别识别，最后给出音节（汉字）的方式，这种方式与识别机理一致，应当可以获得较好的结果。但识别系统究竟是否采用这种语音识别方式并不知道。在这种情况下，本文采用的方法是，首先根据识别的结果构筑声韵混淆矩阵，然后验证音节混淆音是否覆盖正确的结果。并根据验证情况，对一些音节进行调整。最终使得根据错误音节的混淆音给出的正确音候选集前十名包括的正确音大于95%。

2.2 识别结果的错误分析

上面对音节识别的错误进行了分析，这里对语音识别系统输出的错误结果从语言处理的角度进行分析。需要说明的是，输出文字错误，反映在音节上有两种可能，一种是音节识别的结果是错误的；另一种是音节识别正确而给出的汉字错误。

这两种错误的示例和分布如下，计算均以句为单位：

例2 音正确　海浪充饥(音对字错)成大片的珊瑚沙滩。（正句：海浪冲击成大片的珊瑚沙滩。）

例3 音错误　他们的行程即将姐夫（音错字错）。　　（正句：他们的行程即将结束。）

音正确	16.2%
音错误	83.8%

上述数据再次表明，纠错处理的重点是错音。

为了建立纠错处理方案，从语言处理的角度对错误进行分析，依据不同的语言处理模型有不同的分析方法。HNC句类分析技术是本文纠错处理的核心技术，因此需要根据HNC理论的句类和语义块概念考察句子错误出现的位置、分类与分布。

在输出的文字结果中，错误分布在多种种位置上。造成这种结果的原因在于：语音识别系统采用的n-gram语言模型，n-gram的处理结果给出的只是一个概率最优的结果，从HNC句类观点观察到错误出现的位置是随机的，错误可以出现在句类一语义块构成成分的任何位置上。

按出现位置对错误进行分类。对各类错误进行统计，下面是各种错误所占的比例。

特征语义块一广义对象语义块搭配错误 广义对象语义块核心错误	28.3%
特征语义块一广义对象语义块搭配错误 特征语义块核心错误	30%
其他	41.7%

错误的分布情况非常有趣，一半以上的错误出现在有间隔的词语搭配上，即特征语义块

一广义对象语义块搭配错误。产生这种现象的原因是，n-gram 能够通过训练数据得到相邻或固定间隔的词语的搭配知识，而对于跨越一个不定距离的远距离搭配就失灵了。

3 处理策略

句类分析系统提供了语句级的自然语言理解处理手段，根据概念的“同行”与句类知识，考察语句中出现的词汇是否合理，是否在句中担任合理的角色[3,4]。如果句中出现不合理的词汇，和它前后的词语没有任何关联，这种词汇可以形象的称之为“孤魂”，系统可以明确的知道，这样就完成了纠错处理的第一步——发现孤魂。

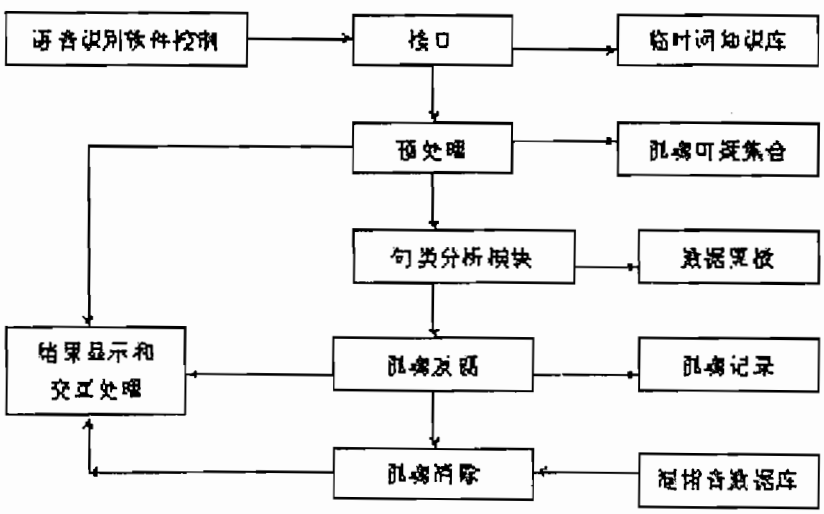
如何纠呢？

由于句类分析系统可以判断句中的词汇是否合理，因此只要在认为有错误的位置上给出可以替换的候选集，纠错实际上就转化为模糊消解。通过尝试候选集中的词汇，就可以找到合乎概念联想脉络要求的词汇，完成纠错处理。

这里有两个问题。首先，判断孤魂是根据句子中的其他词汇，如果这些“其他”词汇也是错误的，即孤魂成片出现，形成“孤群”，怎么办？对这个问题如果能将句类分析系统和语音识别系统统合起来，可能会找到解决的途径。但是另一方面，值得庆幸的是，测试表明本文使用的语音识别系统在一般情况下有比较高的正确率[5]，识别的错误多数属于孤魂而非孤群。

另一个问题是如果给出的候选集不能覆盖正确的结果，怎么办？从深入研究的角度看句类分析与语音识别的统合系统会较好解决这个问题。目前，只能交给用户修改。

下面具体来看看处理的流程。



以一个具体的例子说明处理过程：

例4　纠错处理例句　　我国今后每年都要今后许多小麦。

纠错处理系统完成预处理后发现孤魂可疑集合中词语有：

我国，今后（第二个）

句类分析发现：

这个句子没有可以能够假设为全局特征语义块的动态概念，同时也没有迹象表明句子属于无特征语义块的句类。

孤魂发现：

由于“今后（第二个）”前面有时间概念（“今后每年”），同时有特征语义块要素的逻辑说明概念（“都要”），这两类概念合起来经常充当特征语义块要素的前修饰部分。表明这个位置应当出现一个动态概念。“今后（第二个）”更可能是孤魂。“我国”是孤魂的可能性要小于“今后（第二个）”。孤魂发现只是根据可疑集合和数据黑板的数据判断孤魂的可能性，不是作定案的判断。孤魂消除如果失败，还可以再回到这里，根据现场的数据重新进行孤魂发现。

孤魂消除：

根据孤魂发现给出的孤魂记录结果，以及对该位置上概念的预期情况，寻找在这个位置上的候选词汇集合，得到“浸透”“进口”两个动态概念。运用它们对应句类的句类知识对句子进行检验，否定掉“浸透”，认可“进口”。同时，确定句子属于物转移句句类，有孤魂嫌疑的“我国”充当了物转移动作的发出者，表明与它相关联的词语是隔开的，解除了它的孤魂嫌疑。

最后完成纠错处理，给出结果显示。

对纠错系统的性能进行了测试，将纠错处理的结果分为三类：

I 型语句：发现句中有错误，纠错正确；

II 型语句：发现句中有错误，纠错不正确；

III 型语句：未能发现句中有错误，纠错不正确。

目前系统的纠错处理性能如下：

错误发现率 ＝（I 型语句句数＋II 型语句句数）÷ 总句数	78%
纠错正确率 ＝ I 型语句句数 ÷（I 型语句句数＋II 型语句句数）	71.8%

上述数据表明目前纠错处理系统已经具有了较强的纠错处理能力。错误产生的主要原因是：句类分析系统还有待完善；知识库中的知识项填写有待提高。这两点会随句类分析技术的不断完善而迅速得到改变。有理由相信在此基础上，纠错处理系统的能力将很快得到提升。

4 讨论

语音识别系统处理的语音信号具有两个特点：一个是存在环境噪声，而且环境噪声的声学特性与人的语音相近；另一个是不同的发音人语音具有较大的差异，即使是同一个，不同时间的语音都存在一定的差异。因此仅凭单纯的信号处理很难提高语音识别系统的正确率。

n-gram 在语音识别系统中的应用，提供了利用统计已有语料来获取语言数据的方法，结合语音识别的隐马尔可夫模型，目前已经有商业系统面市。但是，n-gram 自身有难以克服的缺陷：

1、 语言中常用词语和非常用词语在实际语料中出现的频度差别很大，一些非常用词可能在语料中不出现或出现的次数很少，造成统计数据稀疏。[8]中对此作了详细的分析，认为实际上统计只能得到常用词的数据。但无论如何，语句中很难避免对非常用词的使用。

2、 语句中词语之间的关联不单是紧邻词语间的搭配关系，不是一个线性连续的关系，而是一个层级的关系。n-gram 对此无法具体区分，都简单的用概率给予表达。因此，尽管 n-gram 提供了便捷的处理算法，但无法给出符合概念联想脉络的词语，也无法判断给出的结果是否正确，只能得到一个概率最优的结果。

已经有人尝试将语法分析融合到语音识别系统中[9]，并在特定受限的领域内比一般的识别方法正确率提高 8%。更进一步的研究要深入到语言深层－－语义，建立起概念联想脉络和自然语言的句义约束模式－－HNC 称之为句类，并在处理中综合使用概念间的关联知识和句类知识。本文利用已有的 HNC 句类分析在这方面作了有益的尝试，取得了预期的效果。

本文中使用的语音识别系统是一个商品系统，纠错处理是在语音识别系统最终的输出结果上进行的。如果能将句类分析技术与语音识别系统融合，至少可以带来两个好处从而大大提高系统的性能和效率：其一，直接从音节识别的候选集中选择词语，可以避免“识别—>字词—>音节（混淆音）—>字词（纠错结果）”过程中的错误传播；其二，利用现场的数据形成句类假设，反过来又可以根据句类知识产生预期，指导语音处理部分利用丰富的中间结果有重点地进行识别，减少信息的损失。这正是本文进一步努力的方向。

参考文献

[1] 黄曾阳 HNC 理论概要 《中文信息学报》Vol.11(4)，1997。

[2] 黄曾阳 HNC 理论与自然语言语句的理解 《中国基础科学》1999 年 2－4 期 1999/4。

[3] 黄曾阳 HNC 理解处理系统的基本框架 《HNC（概念层次网络）理论》黄曾阳 著 第 60～79 页 清华大学出版社 1998/11。

[4] 晋耀红 基于 HNC 理论的句类分析系统的设计与实现 同[3] 第 442～478 页。

[5] 孙伟峰 基于句类的指代解析及其在语音识别中的应用 附件 中科院北京软件工程研制中心 硕士学位论文 2000 年 6 月。

[6] 黄曾阳 自然语言语义网络的基本结构及其特性 同[3] 第 17～43 页。

[7] 黄曾阳 自然语言的深层结构及句类分析 同[3] 第 17～43 页。

[8] 关毅 等 现代汉语计算语言模型中语言单位的频度－频级关系 《中文信息学报》Vol.13(2)，第 8～15 页，1999。

[9] 赵力 等 汉语连续语音识别中语音处理和语言处理统合方法的研究 《声学学报》 Vol.26(1)，第 73～78 页，2001/1。

注：本文得到本文得到“973”项目 G1998030506，国家“九五”重点科技攻关项目 98-779-02-04 和中科院声学所知识创新项目资助。