小语种语音识别

石颖

2009/1/3

Contents

1 写在前面的话

16年8月份我初到语音和语言技术中心实习，当时的我只笼统的读过李航老师的撰写的《统计学习方法》，脑海中只有肤浅的机器学习的概念，更是不知深度学习为何物。然而，幸运的是，我来到实验室不到一个月的时间，就赶上了人生中的第一个deadline，ICASSP 2017。可以说我对kaldi，乃至语音识别的入门是从这个deadline开始的。Deadline结束后，恰逢中心启动了国家自然科学基金的重点项目M2ASR(Multilingual Minorlingual Automatic Speech Recognition)，于是乎王老师丢给我80小时

的哈萨克语的语音数据，让我跑个baseline…直至今日，该项目的大部分工作都是由汤博士指导，我来执行的。真不知道王老师当时哪里来的勇气，敢于把这样的任务交给我这个还不算入门的实习生来做。不过我着实应该衷心的感谢王老师对于我的信任，也许没有当时这个任务，我很可能就流落到某某公司泯然众人了。

写上一段的目的不是为了标榜自己，更不是为了吹捧我们中心，而是想告诉读者，语音识别入门以及kaldi入门，没有想象中的那么困难。甚至于深度学习中的大部分框架及理论都是较为朴素的。当然这里的朴素仅限于入门，如果恰逢某读者胸怀大志，早已将语音识别以及kaldi 的基础知识烂熟于胸，想要在语音识别领域有王老师那样的建树，请略过此章节或者直接私信王老师wangd99@tsinghua.mails.edu.cn.

2 小语种及其所面临的困境

回到本章的主题：小语种语音识别。首先介绍什么是小语种。众所周知，我国的官方用语为汉语普通话，然而我们生活在一个由56个民族组成的大家庭，因此我国有着极其丰富的语言体系，小语种与方言错综复杂，甚至于有的时候小语种和方言被混为一谈，实则不然。小语种本质上是一种独立的语言，有独立且完备的发音体系，书写方式及语法现象等。国际公认的对小语种的定义为：小语种是除联合国通用语种：英语，中文，法文，俄语，西班牙语和阿拉伯语以外的所有语种。由此可见，小语种是一门独立的语言，这是小语种与方言最本质的区别。而在具体的语音识别任务中，针对方言语音识别的解决方案也比小语种要简单的多。因为我们可以很容易的得到普通话到方言的映射，因此普通话识别用到的语言模型，词表乃至于phone表都可以直接被方言复用，在声学模型方面，我们只需要用一少部分数据对已有的普通话模型进行tuning就可以得到很好的结果。然而同样的方法在小语种上往往是行不通的。

在语音识别领域，小语种面临的两大难题是：资源上的“小”以及形式上的不规范。解决了这两个难题，我们就可以很轻松的构建一套可信的语音识别系统，然而资源的收集与规范化是一项极其费时，费力且费钱且枯燥的工作，因此无论在学术界还是在产业界，语音识别在小语种上的进展都是十分缓慢的。

Figure 2.0.1: 中国语种分布图

3 小语种中的基础数据

3.1 音频数据

上一节中，我们已经提到了小语种语音识别面临的第一个问题是“小”，这主要是因为小语种的适用人群相对较少，且小语种分布的地区信息化尚不发达。导致收集语音数据的成本偏高且质量难以控制。因此合理且充分的利用我们手头仅有的音频数据变得尤为重要，对于语音识别来讲，音频数据首先要满足的条件是尽量涵盖该语种全部的单个音素的发音。以汉语拼音为例。理论上我们在收集语音数据时，应该使得语音数据尽可能的覆盖汉语拼音所有可能的tri-phone(前文中曾提到过tri-phone的概念)，在无法满足该条件时，我们应该保证音频数据可以覆盖所有的汉语拼音中的声母与韵母。在满足了音频数据对发音的覆盖度之后，我们仍需考虑数据对场景的覆盖度，如：噪声，混响，音量高低语速快慢等。对于场景的问题，我们往往采用数据增强的方式进行弥补，即：人为的往语音数据中添加不同场景的噪音，不同距离的混响，通过信号处理的方式调整音频的音量大小及语速快慢。数据增强往往可以使数据在数量上成倍增加，在质量上也能使数据对场景的覆盖更加全面，经试验证明，该方式可以有效的提高语音识别系统的性能。

3.2 文本数据与发音词典

文本数据收集： 文本数据相对于语音数据来讲更加容易收集。文本数据在语音识别中主要被用于构建语言模型，因此文本数据同样要求场景的覆盖度。例如：文本语料如果大多是新闻稿等一些较为正式的书面内容，则该系统对于口头用语的识别效果会很差。对于语言模型我们可以以ppl为指标作为度量(srilm中提供了计算ppl的接口)。一个经久不息的爬虫可以帮我们收集大量的文本数据(BTW: 如果发现爬到的文本数据对于口头语ppl较高，可以多爬取一些论坛的评论，毕竟人们所发的牢骚大多来源于生活)。

文本数据处理： 小语种文本处理最令人头疼的问题是，技术人员看不懂。使用拉丁文作为载体的语言我们尚可分析一二，然而对于维语、哈萨克语、柯尔克孜语等，以阿拉伯字母作为书写载体的语言往往最令我们左右为难(每个字母看起来都差不多)。所以我们建议读者在处理小语种文本时，将小语种的文本映射到拉丁字母上，便于我们分析结果。如果没有语言学家为我们制定映射表或映射规则，我们可以直接将小语种的中的文字转化为其对应的UTF-8编码，这样不仅保证了唯一性，同时也为结果的分析提供了可能。

发音词典lexicon：发音词典是语音识别的核心组成部分。然而对于极其濒危的语种，不排除没有可靠发音词典的可能。在以往的工作中，我们尝试过使用鲁棒性较强的语音识别模型，如：汉语语音识别模型。对小语种的音频做强制对齐(force alignment)以达到使机器自动生成发音词典的目的。遗憾的是这项工作并没有成功。原因也很容易解释，汉语的发音空间并不能代表人类发音的全空间。即汉语模型达到了汉语识别的局部最优，无法对小语种进行无偏估计。如果我们能够构建人类发音的全空间，即学习到一组参数可以表征人类的全部发音，则发音词典的构建将不再依赖语言学家。在今后的工作中，我们将在此方向上做更多的尝试。

4 对小语种声学模型的探索

4.1 迁移学习：transfer learning

上一节我们提到了使用数据增强(Data augmentation)的方式人为的增加语音数据的总量以及覆盖度，这种方式可以解决一部分问题，然而，通过数据增强得到的语音信号，与纯正的语音信号还是存在着相当的差距。例如：我们无法使用数据增强的方式，使数据覆盖更多的说话人。数据增强模拟的噪音数据有限且无法覆盖更多的信道。

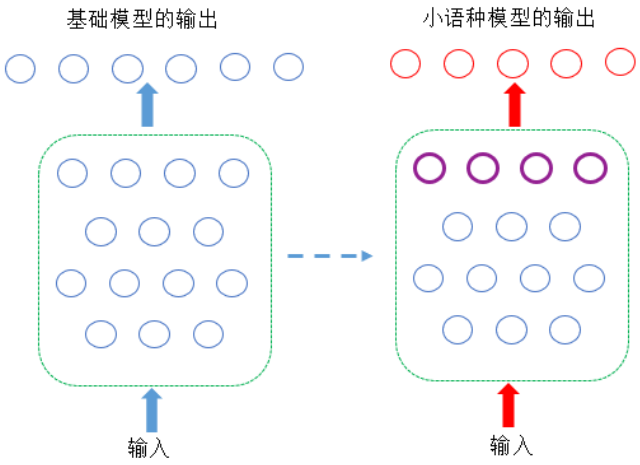

Figure 4.1.2: 迁移学习

因此我们想到了迁移学习(Transfer learning)，虽然两种语种的发音空间不完全相同，但是噪声，信道，说话人等无关信息的模式是可以相同的。另外，两种不同的语言的发音空间应该有可共享的部分，例如：汉语拼音中的“i”与英语中的“I”基本无异。而迁移学习的本质恰恰就是让待训练的模型站在“巨人”的肩膀上。具体的做法是如图4.1所示，使用将基础模型的输出层，替换为小语种的输出。同时保持基础模型较低的数层不随网络的训练而更新(图中蓝色部分为不更新部分，紫色部分为可更新的节点。具体固定层数由小语种的数据量决定，完全不固定会产生训练数据与模型参数量不匹配的问题。从而导致欠拟合)。进过我们实验的论证，迁移学习对于数据量较少的小语种识别有着很好的效果。

4.2 MaR:Map and relabel

迁移学习确实可以在一定程度上解决小语种的问题，然而当我们遇到数据极缺的濒危语种，比如我们只有1小时的标注数据。此时迁移学习也会显得捉襟见肘，1小时的标注数据甚至无法训练出一个较好的高斯混合模型(GMM),为了应对这样的问题，我们在2018年初，提出了一种新的方法。Map and relabel(MaR)。通过该方法，我们可以使用极少数的带音素级别标注的语音，生成大量的带标注的语音，从而解决数据稀缺的问题。该方法主要分为两步。

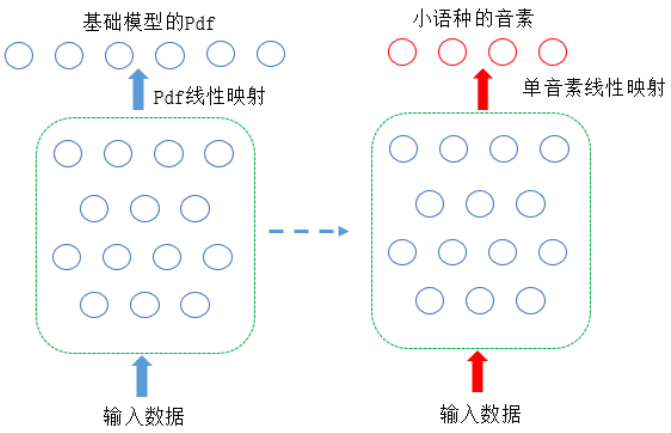

Figure 4.2.3: Map

Map: 受迁移学习启发，我们可以让模型站在巨人的肩膀上，不同的是，在Map步我们希望模型能够使用极少数据，学到输入特征到音素的映射。常规的语音识别模型的输出为Pdf(前文中有相关概念的介绍)。Pdf是tri-phone的子单元。当训练数据极少时我们无法对数量较为庞大的Pdf建模，因此，我们可以退而求其次，一个语言的音素数量是远少于Pdf个数的。我们可以通过迁移学习的方式，使用基础模型对目标的语种的单音素进行建模。如图所示4.2。我们替换基础模型最后的线性映射，使模型的输出对应到小语种的单因素，同时固定模型的其他隐藏层，此时整个模型可更新的参数只有最后一层，而将输出改为单音素之后这一层的参数会变得很小。因此我们只需要少量的数据就可以将这一层调整很好。

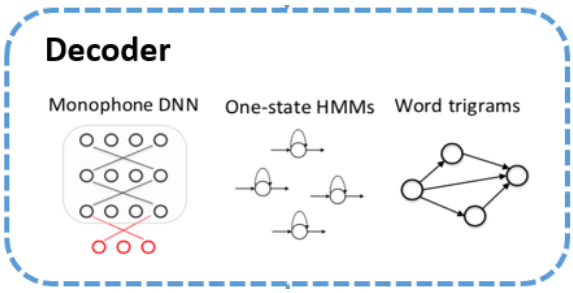

Figure 4.2.4: Map

Relabel: Map完成后，我们会得到一个基于迁移学习的分类器，该分类器可以将输入特征分类为不同的单音素，然而使用DNN进行强制对齐并

不是一个很好的选择，上文提到过基础模型无法对小语种进行无偏估计。因此Map步得到的分类器，是基于单音素的弱分类器。强制对齐的输出会掺杂大量的毛刺，例如连续的音素串“a a a a a”被识别为“a b a c a”。幸运的是一个词中包含了音素的上下文信息，语言模型中又包含了词的上下文信息，因此语音识别的解码器天然具备平滑机制，我们只需要将传统的基于HMM(前文有HMM相关的介绍)三状态的解码器，调整为基于HMM单状态的解码器即可。其中每一个转态对应一个音素。如图4.2所示。通过使用解码器对输出进行平滑，我们很容易就可以将一条语音较准确的标注到音素级别。

基于MaR我们可以将大量的无标注数据，标注到音素级别，这种方式可用于极其濒危的语种的拯救。

5 对小语种语言模型的探索

在小语种语音识别中，我们采用的语言模型，均为基于统计的n-gram模型，这种模型对于oov(out of vocabulary)词汇完全没有抵抗力，而且这种基于统计的语言模型，在词表较大的时候非常浪费存储与计算资源。所以，我们建议读者多做一些基于RNN的语言模型的尝试，惭愧的是，我们目前并没有开展过多的关于RNN语言模型的研究，在此，我们将基于ngram统计语言模型，向大家分享我们在小语种语言模型上的探索。

在之前的工作中，我们主要面对的是阿尔泰语系的小语种，诸如，维语，哈萨克语，柯尔克孜语等，这些语言有个共同的特点，均属于黏着语。及一个单词是由一个词根加多个词缀组成的(据新疆的小伙伴称，维语有个词是由一个词根加50多个词缀构成的)，这种特性带了两个问题，1.黏着语的词表极其庞大。2. 一个较长的词如果开头没有识别出来，那么整个词都无法被识别出来。为了解决这个问题，我们采用了基于词根词缀的解码，将一个单词拆分成一个词根加数个词缀，这种做法的最直接的好处是，大大减小了词表的体积。同时，也可以避免一个单词因为一部分识别错误而导致整个词都被识别错误的情况。(被识别成词根词缀的单词可以通过后处理的方式还原为整个单词)

6 小结

本章节中，我们主要向大家阐述了小语种语音识别所面临的困难以及我们为了应对这些困难所做一些尝试。总体来说可以分为两方面，数据层面，在收集数据整理数据的同时，我们要尽可能的让手中的数据发挥最大

的价值，因此数据增强是一个不错的选择。在模型方面，我们始终贯穿一个主线：没有就借。所以有了迁移学习在小语种上的应用以及后来的借助基础模型以MaR的方式生成数据的方法。在今后的工作中，我们可能会更多的投入到对RNN语言模型的探索中去。

目前市场上还没有太多关于小语种语音识别的成熟产品，这并不代表着，小语种语音识别没有应用前景与市场价值，我国少数民族人口众多，仅维语就有1300万使用人群，所以小语种的应用前景还是非常广阔的，然而，由于小语种数据稀缺，变种复杂且缺乏规范性，因此，无论是学术还是产业在小语种上的进展都十分缓慢，所以我们呼吁那些拥有小语种数据的研究机构或者公司能更我们一起进行数据共享。共同推动小语种语音识别的发展。