

中科信利基于标注文本的语言模型

2.0.1版

优化训练指南

中科信利基于标注文本的语言模型2.0.1版

优化训练指南

©2004–2017北京中科信利技术有限公司版权所有。

中华人民共和国北京市海淀区北四环西路21号声学所DSP大楼

北京中科信利技术有限公司尽力保证本文档内的信息真实可靠。但由于可以理解的原因，本文档内的信息可能有错或者未及时更新。本文档内所描述的软件只在有许可或者保密协议的情况下才能提供，而且只有符合和遵守许可条款才能使用或者复制。除非许可或者保密协议明确规定，您不得使用、复制、修改或者散发这些软件。未经北京中科信利技术有限公司书面许可，不论出于何种目的，任何人不得以任何方式或途径复制或传播本文档的任何内容。

ThinkIT是在中华人民共和国境内的注册商标，属北京中科信利技术有限公司所有。本文档内包含其它公司的商标，由其各自拥有者所有。

文档历史

|  |  |  |
| --- | --- | --- |
| 版本号 | 日期 | 描述 |
| 2.0.0 | 2018-11-22 | LH建立文档，支持utf8编码文本语料训练 |
| 2.0.1 | 2021-03-05 | 肖素杰修改文档，修改build解码网络步骤说明 |
|  |  |  |
|  |  |  |

目录

[关于本文档 III](#_Toc65856272)

[读者对象 III](#_Toc65856273)

[文档组织 III](#_Toc65856274)

[相关文档 IV](#_Toc65856275)

[印刷约定 IV](#_Toc65856276)

[获得帮助 V](#_Toc65856277)

[第一章 系统要求 1](#_Toc65856278)

[硬件要求 1](#_Toc65856279)

[软件要求 1](#_Toc65856280)

[第二章 软件概述 3](#_Toc65856281)

[软件目录结构 3](#_Toc65856282)

[运行许可 3](#_Toc65856283)

[第三章 语言模型概述 5](#_Toc65856284)

[语言模型基本原理 5](#_Toc65856285)

[语言模型文件格式 6](#_Toc65856286)

[语言模型训练环境 9](#_Toc65856287)

[第四章 通用（领域）语言模型训练软件 10](#_Toc65856288)

[基本功能 10](#_Toc65856289)

[目录结构说明 10](#_Toc65856290)

[总控程序各子模块介绍 11](#_Toc65856291)

[运行步骤 13](#_Toc65856292)

[文本语料格式说明 13](#_Toc65856293)

[第五章 语言模型自适应优化训练软件 15](#_Toc65856294)

[基本功能 15](#_Toc65856295)

[目录结构说明 15](#_Toc65856296)

[总控程序各子模块介绍 16](#_Toc65856297)

[运行步骤 17](#_Toc65856298)

[第六章 语言模型性能验证 18](#_Toc65856299)

[基本功能 18](#_Toc65856300)

[目录结构说明 18](#_Toc65856301)

[运行步骤 18](#_Toc65856302)

[PPL文件格式说明 19](#_Toc65856303)

[第七章 语言模型静态解码网络的训练软件 22](#_Toc65856304)

[基本功能 22](#_Toc65856305)

[目录结构说明 22](#_Toc65856306)

[总控程序各子模块介绍 23](#_Toc65856307)

[运行步骤 23](#_Toc65856308)

[第八章 语音识别系统语言模型更新 25](#_Toc65856309)

[语言模型更新 25](#_Toc65856310)

关于本文档

本文档是中科信利语音处理环境软件文档的一部分。

本文档讨论中科信利语音处理环境语言模型优化训练过程。用户在收到语言模型训练软件包之后，请先阅读本文档。

读者对象

本文档的读者对象为使用中科信利语音识别系统集成商的开发人员和工程人员。

文档组织

本文档内容安排如下：

1. 介绍了中科信利语音识别*语言模型训练所需要的软硬件环境*。
2. 介绍了中科信利语音识别*基于标注文本的语言模型软件概述*。
3. 介绍了中科信利语音识别*语言模型概述*
4. 介绍了中科信利语音识别*通用（领域）语言模型的训练方法*。
5. 介绍了中科信利语音识别*语言模型优化自适应的训练方法*。
6. 介绍了中科信利语音识别*语言模型静态解码网络的训练方法*。

相关文档

印刷约定

中科信利文档采用如下印刷约定：

|  |  |
| --- | --- |
| *斜体字* | 表示变量、文件及路径名、程序名、程序参数、网站或邮箱地址，以及第一次介绍的名词，例如：  编辑配置文件*xxx.cfg* |
|  |  |
| Courier New/楷体 | 表示方法/成员函数，参数名/参数值，程序常量以及屏幕输出，例如：  设置参数loglevel为3  > xxx  在上例中“>”字符代表MS-DOS提示符或者Unix Shell提示符，该提示符之后的内容是需要输入的。 |
| *Courier New/楷体* | 表示在该处要输入的值是需要根据情况替代的，例如：  设置日期为*YYYY-MM-DD*  实际上可能要输入：  *2017-06-25* |

获得帮助

如果您有问题或对本文档有任何建议或者评论，请致电或发电子邮件：

电话：*(86)-10-82547559*

邮箱：*support@thinkit.com.cn*

# 系统要求

1

本章介绍中科信利语音识别语言模型训练所需的系统配置

硬件要求

* CPU：12物理核24超线程 e5-2620及以上
* 内存：64G以及上
* 硬盘：80G预留空间

软件要求

* 64位Debian Linux系统
* 64位ActivePerl软件
* 64位Python2.7软件
* 需要安装kaldi，需要用到openfst相关命令

# 软件概述

2

本章介绍基于标注文本的语言模型训练软件的基本结构及授权信息

软件目录结构

表1. 基于标注文本的语言模型训练目录结构

|  |  |
| --- | --- |
| 目录 | 说明 |
| Env | 语言模型软件训练环境 |
| Doc | 文档 |
| Examples | 演示示例 |
| Tools | 需要的工具 |

运行许可

本语言模型训练软件需要授权才能运行，授权过程如下：

1. 获取机器信息。

软件包“Tools/lic”目录下会有授权申请工具getinfo，执行该授权申请工具后会在该目录下生成机器信息文件，类似于\*\_machine.info的文件（如： c24\_machine.info）；

1. 将机器信息发给授权方（中科信利公司）获取授权文件，授权版本号为1；
2. 将获取的授权文件重命名为license.dat，放入“Tools/lic”目录下。

# 语言模型概述

3

本章介绍语音识别语言模型基本原理及基于标注文本的语言模型训练环境构成

语言模型基本原理

统计语言模型的训练是通过对训练语料的统计，得到模型参数的过程。假定表示某一个句子，其词序列为组成，是句子的长度，句子的概率表征了出现的可能性。将展开为：



利用条件概率公式，序列出现的概率等于每一个词出现的条件概率相乘，于是可展开为：



其中表示第一个词出现的概率，是在已知第一个词出现的前提下，第二个词出现的概率，以此类推。可见，词出现的概率取决于它前面的所有词。

从计算上来看，第一个词的条件概率很容易计算，第二个词的条件概率也还可以，第三个词的条件概率就比较难了，因为它涉及到三个变量，每个变量的可能性都是一种语言模型字典的大小。到最后一个词，它的条件概率的可能性就太多了，根本无法估计。鉴于此，俄国数学家马尔科夫提出了一种马尔科夫假设，认为任意一个词出现的概率只与它前面的个词有关，而于更前面的词无关。主要估计前个词出现的前提下，第个词出现的概率，因此，统计语言模型又称为N元文法语言模型。也即：



时称为一元文法模型（Unigram Model），时称为二元文法模型（Bigram Model），时称为三元文法模型（Trigram Model），以此类推。

语言模型文件格式

Ngram语言模型以arpa的格式存储，一个三元语言模型的arpa格式如下：

\data\

ngram 1=270928

ngram 2=133698

ngram 3=279509

\1-grams:

-0.7381738 </s>

-99 <s> -1.382782

-4.138755 a -0.7724786

…

-5.099739 连连 -0.4258974

-7.241138 连连看

-7.241138 连绵

…

\2-grams:

-4.110618 <s> a -0.5118836

…

-2.113996 八点 以前 -0.30103

-1.412258 八点 之后 -0.4259687

…

\3-grams:

-1.885562 两 g 吗

-1.812005 两 g 那个

-1.947823 两 g 那种

…

\end\

其中，语言模型的头信息为：

\data\

ngram 1=270928

ngram 2=133698

ngram 3=279509

“ngram 1=”、“ngram 2=”以及“ngram 2=”后面的数值表示一元、二元以及三元文法的条数。

“\1-grams:”是一元文法的标识符，其后面的文法为一元文法；“\2-grams:”是二元文法的标识符，其后面的文法为二元文法；“\3-grams:”是三元文法的标识符，其后面的文法为三元文法；

每一个Ngram的概率表示为：

概率 （tab键）N gram（tab键）[回退概率]

概率是必须有的，而回退概率不一定会有，且最高阶的Ngram没有回退概率。因此，第一列的数值表示概率，最后一列的数值表示回退概率。

\end\指示语言模型结束。

语言模型训练环境

中科信利语音识别语言模型训练由下面几部分构成：

通用（领域）语言模型训练软件（Env/BuildCorpusLM）：

根据采集到的文本语料，使用该软件训练语言模型。

语言模型优化自适应训练软件（Env/LMAdaption）：

使用该软件对多领域的语言模型进行基于模型层面的优化自适应训练。

语言模型静态解码网络训练软件（Env/ build\_ctc\_wfst\_checkLicense）：

使用该软件将语言模型(arpa)训练成wfst语音识别系统所需的wfst静态解码网络。

# 通用（领域）语言模型训练软件

4

本章介绍语音识别通用（领域）语言模型基本训练。

基本功能

通用（领域）语言模型训练软件是提供给中科信利语音识别系统使用者训练语言模型的软件。使用者可以根据采集到的文本语料，使用该软件训练语言模型，文本语料要求utf8编码格式。

目录结构说明

软件包目录结构（Env/BuildCorpusLM）：

表2.基于标注文本训练语言模型目录结构

|  |  |
| --- | --- |
| 目录 | 说明 |
| bin | 可执行程序文件目录 |
| data | 训练以及总控程序目录 |
| |--863.lmlist | 语言模型字典 |
| |--run.sh | 总控程序 |

总控程序各子模块介绍

1. ../bin/txt2utt.pl：语料预处理工具

功能：

去除训练语料中的标点符号；

数字转换成汉字：包括对时间、日期、电话号码等的处理；

英文大写转小写。

参数说明：

第一个参数：预训练文本

输出预处理后的文本：默认在输入参数后面添加. preprocessed后缀

1. ./../bin/segword：分词工具

功能：

基于给定字典（863.lmlist），采用适当的算法，对训练语料进行分词。

参数说明:

-wordlist:分词字典（输入）；

-in: 预处理后的文本（输入）；

-out: 分词后的文本文件（输出）；

-method：分词算法(前向最大匹配：fw；后向最大匹配：bw)；

-tb：是否进行回溯(回溯：1; 不回溯：0);

-t：线程数

1. ./../bin/ngram-count: 语言模型训练工具

功能:

基于分词后的训练语料，训练语言模型。

参数说明：

–order: 语言模型的训练阶数；

–text: 分词后的训练文本；

–vocab: 语言模型字典；

–unk: 将未登录词映射成unk；

–lm: 训练的语言模型；

–wbdiscount： 使用wb折扣算法，（默认为gt折扣算法）；

–interpolate： 查用插值平滑算法。

-gt1min, -gt2min, -gt3min: 相应ngram的count数小于此设定值的ngram不参与模型的概率计算。

1. ../bin/prune.sh： 语言模型剪枝工具

功能：

将大语言模型裁剪成小语言模型。

参数说明:

第一个参数：裁剪前的语言模型；

第二个参数: 裁剪后的语言模型的文法数；

第三个参数: 裁剪后的语言模型。

运行步骤

1. 将训练语料的文本文件（utf8编码）放于“data”目录下，如：train.txt。
2. 进入data目录，运行总控程序: ./run.sh train.txt；
3. 训练完成后生成的语言模型文件名为：final\_corpus.arpa。

文本语料格式说明

语言模型训练的文本格式为：一行一句话，尽量将输入文本中标点符号去掉，数字转换成汉字，大写字母转换成小写字母。文本语料要求utf8编码。

如：

您好 很高兴为您服务

今天天气怎么样

有什么可以帮您的

今天在下雨

# 语言模型自适应优化训练软件

5

本章介绍基于各领域语言模型的优化自适应训练。

基本功能

语言模型优化自适应训练软件提供对多领域的语言模型进行基于模型层面的语言模型优化自适应训练（Adapation）。

目录结构说明

软件包目录结构（Env/LMAdaption）：

表3.语言模型优化自适应目录结构

|  |  |
| --- | --- |
| 目录 | 说明 |
| bin | 可执行程序文件目录 |
| data | 训练以及总控程序目录 |
| |-- mix\_model.txt | 自适应模型配置文件 |
| |--run.sh | 总控程序 |

总控程序各子模块介绍

1. ../bin/compute-best2mixlm.pl：语言模型自适应工具

功能：

读取语言模型配置文件（mix\_model.txt）中的各模型以及插值系数，并根据插值系统设置各模型权重，进行模型自适应，并输出自适应后的模型。

参数说明:

配置文件（mix\_model.txt）格式如下：

C:\Users\dell\AppData\Roaming\Tencent\Users\379241446\QQ\WinTemp\RichOle\E_5M]HZO}CJNB7PT(Q1YLED.png

第一行的base.arpa为基础语言模型，mix1.arpa为领域语言模型（buildLM生成的final.arpa）；

第二行的（0.5 0.5）为对应第一行的各模型的权值。

如果有多于1个的领域语言模型，直接在第一行后面继续添加模型名，并修改第二行括号中语言模型的权值分别对应第一行中的各语言模型（基础语言模型的权值保持不变），并确保所有语言模型的权值之和为1。

1. ../bin/prune.sh： 语言模型剪枝工具

第一个参数：裁剪前的语言模型（输入）；

第二个参数: 裁剪后的语言模型的文法数；

第三个参数: 裁剪后的语言模型（输出）。

1. ../bin/utf82gbk.p： utf8编码转gbk编码转换工具

第一个参数: utf8编码文件（输入）；

第二个参数：gbk编码文件（输出）。

运行步骤

1. 将基于通用语言模型训练软件（BuildCorpusLM）训练的各领域语言模型存放于“data”目录下。
2. 进入data目录，配置“语言模型配置文件（mix\_model.txt）”。
3. 运行./run.sh
4. 自适应训练完成后生成的语言模型文件名为：final.arpa。

# 语言模型性能验证

6

本章介绍如何验证语言模型性能

基本功能

语言模型的性能主要通过混淆度（PPL）的大小来评估，PPL值反映了信源熵的大小，表示对信源的不可知程度。直观上，困惑度可以理解为在给定的语言模型中，某个词后面可能接的词的平均数。困惑度越小，语言模型对上下文的约束能力就越强，模型就越好。

目录结构说明

语言模型PPL的计算可在语言模型自适应训练环境中进行，见（Env/ LmAdaption）。

运行步骤

计算语言模型的PPL值，需要一个与语言模型训练数据相匹配的验证集，反映语言模型预期的性能改进。验证集的格式与语言模型训练语料的格式一致，具体计算过程如下：

1. 进入Env\ LmAdaption\data目录下，将验证集 (validate.txt)放入当前路径。
2. 对验证集进行预处理：

运行：perl ../bin/txt2utt.pl validate.txt validate.txt.preprocessed

1. 对预处理后的语料进行分词：

运行：./../bin/segword -wordlist 863.lmlist -in validate.txt.preprocessed -out validate.seg -method bw -tb 1 -t 1

其中：863.lmlist是语言模型的字典；validate.txt.preprocessed是验证集预处理后的文本文件；validate.seg是验证集分词后的文本文件

1. 用语言模型在分词后的验证集validate.seg上计算PPL：

运行：./../bin/ngram -lm final.arpa -unk -ppl validate.seg -debug 2 > final.arpa.validate.seg.ppl

其中：-lm 后的final.arpa是训练的语言模型；-unk指示语言模型中包含<unk>；-ppl 后的validate.seg是验证集分词后的文本文件；final.arpa.validate.seg.ppl 是计算得到的ppl文件。

PPL文件格式说明

PPL文件的基本格式如下：

<s> 您好 很高兴 为您服务 </s>

p( 您好 | <s> ) = [2gram] 0.0184669 [ -1.73361 ]

p( 很高兴 | 您好 ...) = [3gram] 0.228959 [ -0.640243 ]

p( 为您服务 | 很高兴 ...) = [3gram] 0.801359 [ -0.0961731 ]

p( </s> | 为您服务 ...) = [3gram] 0.997261 [ -0.00119099 ]

1 sentences, 3 words, 0 OOVs

0 zeroprobs, logprob= -2.47121 ppl= 4.14766 ppl1= 6.66405

<s> 嗯 你 </s>

p( 嗯 | <s> ) = [2gram] 0.0213846 [ -1.6699 ]

p( 你 | 嗯 ...) = [3gram] 0.00605899 [ -2.2176 ]

p( </s> | 你 ...) = [3gram] 0.109779 [ -0.959481 ]

1 sentences, 2 words, 0 OOVs

0 zeroprobs, logprob= -4.84698 ppl= 41.2724 ppl1= 265.149

……

file validate.seg: 10 sentences, 75 words, 0 OOVs

0 zeroprobs, logprob= -203.66 ppl= 248.889 ppl1= 519.365

其中，文件的前面部分是验证集中每句话对应的PPL：

如：p( 很高兴 | 您好 ...) = [3gram] 0.228959 [ -0.640243 ]中[3gram]表示在语言模型中存在3gram，0.228959表示此3gram的概率对数值，[ -0.640243 ] 表示此3gram的概率。

每句话对应最后的一行：0 zeroprobs, logprob= -2.47121 ppl= 4.14766 ppl1= 6.66405，其中logprob=后面是这句话整体的概率，ppl=表示这句话整体的PPL值；概率值越大，PPL值越小，反映模型的在这句话上的性能越好。

同理，整个PPL文件的最后一行“0 zeroprobs, logprob= -203.66 ppl= 248.889 ppl1= 519.365”，反映模型在整个验证集上的性能，概率值越大，PPL值越小，模型的性能越好。

# 语言模型静态解码网络的训练软件

7

本章介绍语音识别语言模型静态解码网络训练。

基本功能

语言模型静态解码网络训练软件将语言模型(arpa)训练成wfst语音识别系统所需的wfst静态解码网络，从而将语言模型用于wfst语音识别系统中。

目录结构说明

软件包目录结构（Env/ build\_ctc\_wfst\_checkLicense）：

表5. 语言模型静态解码网络训练目录结构

|  |  |
| --- | --- |
| 目录 | 说明 |
| srcdir | 资源文件存放目录 |
| utils和fstbin  NewSource | 可执行程序存放目录  arpa存放路径 |
| mkgraph.sh  path.sh | 总控程序  需要配置Kaldi安装环境目录 |

总控程序各子模块介绍

1. mkgraph.sh 执行脚本

功能:

训练语言模型静态解码网络

参数说明：./mkgraph.sh NewSource

NewSource为arpa存放路径，最后生成的语言模型解码网络TLG.bin也会生成在这个目录

1. srcdir 资源文件目录

vocab.eteh.txt 端到端单子字典

运行步骤

1. 将解码使用的语言模型（final.arpa）放入资源文件目录（NewSource）下，命名为final.arpa。
2. 进入根目录，运行：./mkgraph.sh NewSource
3. 训练完成后生成的解码网络存放于NewSource目录下：TLG.bin（此文件可直接用于wfst语音识别系统）。

# 语音识别系统语言模型更新

8

本章介绍如何将优化后的语言模型应用于语音识别系统中。

语言模型更新

将优化训练出来的语言模型静态解码网络(build\_ctc\_wfst\_checkLicense/NewSource /TLG.bin)替换语音识别系统中对应的TLG.bin (语音识别系统根目录下的Env/model/cts\_src目录下)，更新后即可使用新的语言模型进行语音识别。