

中科信利基于标注文本的语言模型

2.0.1版

优化训练使用文档

中科信利基于标注文本的语言模型2.0.1版

优化训练使用文档

©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编码标注文本语料训练 |
| 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](#_Toc65856245)

[读者对象 III](#_Toc65856246)

[文档组织 III](#_Toc65856247)

[相关文档 IV](#_Toc65856248)

[印刷约定 IV](#_Toc65856249)

[获得帮助 V](#_Toc65856250)

[第一章 系统要求 1](#_Toc65856251)

[硬件要求 1](#_Toc65856252)

[软件要求 1](#_Toc65856253)

[第二章 软件概述 3](#_Toc65856254)

[软件目录结构 3](#_Toc65856255)

[运行许可 3](#_Toc65856256)

[第三章 语言模型优化训练步骤 5](#_Toc65856257)

[训练环境Env目录 5](#_Toc65856258)

[训练步骤概述 5](#_Toc65856259)

[基于标注文本训练语言模型 6](#_Toc65856260)

[语言模型优化自适应训练 6](#_Toc65856261)

[语音识别语言模型静态解码网络训练 8](#_Toc65856262)

[第四章 语音识别系统语言模型更新 9](#_Toc65856263)

[语言模型更新 9](#_Toc65856264)

[第五章 语言模型训练演示示例 10](#_Toc65856265)

[Examples目录结构 10](#_Toc65856266)

[演示示例说明 10](#_Toc65856267)

[运行步骤 11](#_Toc65856268)

[演示训练过程示意图 11](#_Toc65856269)

[第六章 语言模型训练文档指南 13](#_Toc65856270)

[文档名称和说明 13](#_Toc65856271)

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*基于标注文本的语言模型优化训练步骤。*
4. 介绍了中科信利语音识别*系统中如何语言模型更新*。
5. 介绍了中科信利语音识别*基于标注文本的语言模型训练演示示例*。
6. 介绍了中科信利语音识别*基于标注文本的语言模型训练文档指南*。

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。

# 语言模型优化训练步骤

3

本章介绍如何使用本语言模型训练软件包基于标注文本优化训练语言模型。

训练环境Env目录

表2.语言模型训练Env目录结构

|  |  |
| --- | --- |
| 目录 | 说明 |
| BuildCorpusLM | 基于标注文本训练语言模型工具 |
| LmAdaption | 语言模型优化自适应工具 |
| build\_ctc\_wfst\_checkLicense | 训练语音识别语言模型静态解码网络工具 |

训练步骤概述

基于标注文本的语言模型优化训练主要包括如下3个步骤：

1. 基于标注文本训练领域语言模型（Env/BuildCorpusLM）
2. 将领域语言模型与基线语言模型进行自适应优化训练（Env/LMAdaption）；
3. 将自适应优化后的语言模型训练成语音识别系统的语言模型静态解码网络（Env/ build\_ctc\_wfst\_checkLicense）

基于标注文本训练语言模型

1. 目录结构（Env/BuildCorpusLM）

表3.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 |

1. 运行步骤
2. 将训练语料整理成train.txt的文本形式（utf8编码）；
3. Dos2unix命令转换成linux换行符；
4. 进入data目录，运行总控程序: ./run.sh train.txt；
5. 训练完成后生成的语言模型文件名为：final\_corpus.arpa

语言模型优化自适应训练

1. 目录结构（Env/ LmAdaption）

表4.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 |

1. 修改配置文件（mix\_model.txt）

配置文件格式如下：

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

说明：

1. 第一行的base.arpa为基础语言模型，mix1.arpa为领域语言模型；

若对标注文本文件训练的语言模型进行自适应，则mix1.arpa为基于标注文本文件训练的语言模型final\_corpus.arpa

1. 第二行的（0.5 0.5）为对应第一行的各模型的权值
2. 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。
3. 运行步骤
4. 将基线模型以及自适应模型拷贝到data目录下，并确保配置文件中的模型文件名与实际的模型文件名保持一致
5. 运行./run.sh
6. 自适应训练完成后生成的语言模型文件名为：final.arpa

语音识别语言模型静态解码网络训练

1. 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 |

1. 运行步骤
2. 自适应后的语言模型（final.arpa）放入资源文件目录（NewSource）下；
3. 运行：./mkgraph.sh NewSource；
4. 训练完成后生成的解码网络存放于NewSource目录下：TLG.bin；

# 语音识别系统语言模型更新

4

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(Env/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。

# 语言模型训练演示示例

5

本章介绍使用本软件优化训练语言模型的完整演示示例。

Examples目录结构

表6.语言模型训练演示示例目录结构

|  |  |
| --- | --- |
| 目录 | 说明 |
| demo\_corpus.sh | 基于标注文本文件训练语言模型演示示例总控程序 |
| train.tx | 训练文本文件示例 |

演示示例说明

Examples目录下的demo\_corpus.sh为基于标注文本训练优化语言模型的演示示例，总共包含以下步骤：

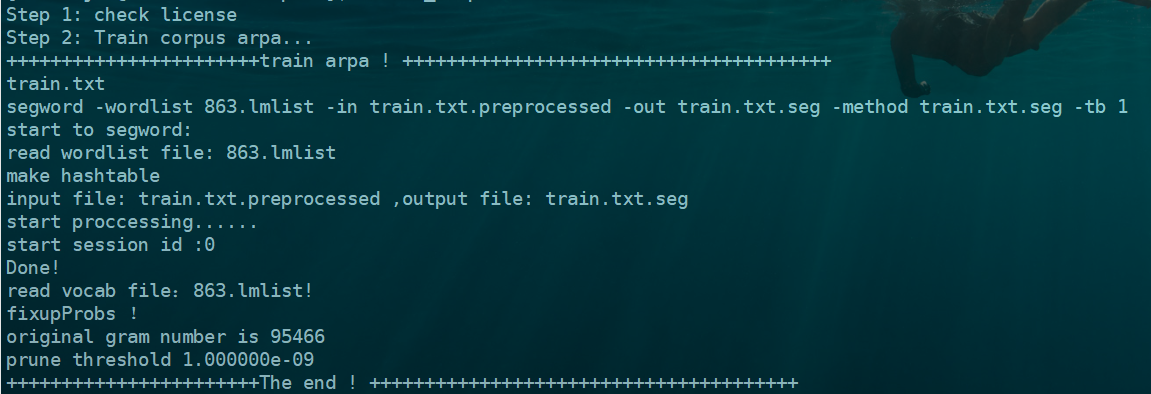
1. 检验授权是否存在： "Step 1: check license"
2. 基于标注文本训练语言模型： "Step 2: Train corpus arpa..."
3. 将训练的领域语言模型与基线语言模型自适应："Step 3: Train lm adaption..."
4. 将自适应后的语言模型训练成语音识别系统的解码网络："Step 4: Build WFST ..."

运行步骤

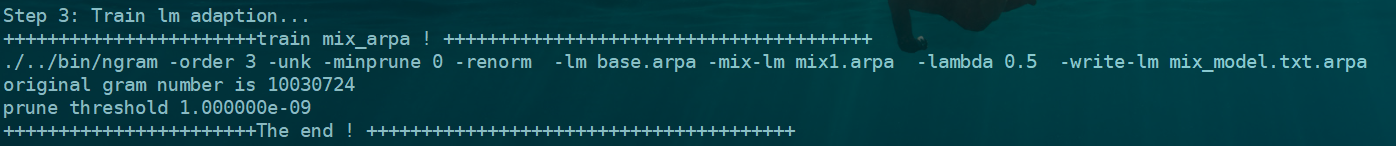
1. 进入Examples目录，运行：./demo\_corpus.sh train.txt
2. 第一个参数：train.txt输入训练文本语料

演示训练过程示意图

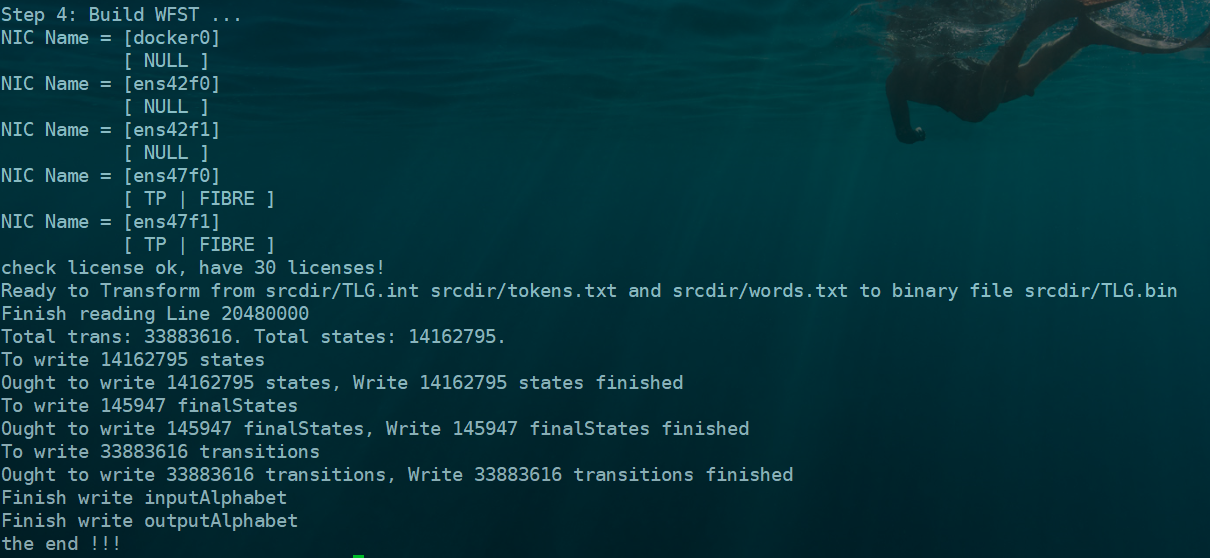
1. 开始训练语言模型：



1. 语言模型自适应



1. 语言模型训练结束，开始训练静态解码网络



1. 训练结束后，build\_ctc\_wfst\_checkLicense/NewSource /目录下会生成静态解码网络：TLG.bin



# 语言模型训练文档指南

6

本章介绍与本软件相关的文档。

文档名称和说明

表7.基于标注文本的语言模型训练软件文档说明

|  |  |  |
| --- | --- | --- |
| 目录 | 文件名称 | 说明 |
| /Doc | 基于标注文本的语音识别语言模型优化训练工具使用文档 | 语言模型优化训练快速使用入门 |
| /Doc | 基于标注文本的语音识别语言模型优化训练指南 | 语言模型优化训练详细使用指南 |