基于词向量与迁移学习相结合的跨领域中文模糊限制语识别软件 V1.0

文档说明

大连理工大学

1. 主要功能：

(1)输入输出模块

首先，输入语料分为源领域训练语料S_l,目标领域训练语料D_l和目标领域测试语料D_u，然后对各部分语料进行分词，获得分词后的文本。接着对分词后的文本句法解析，获得词性标注。最后构建训练词向量语料（包括生物医学语料 6.19MB,中文维基百科语料 106MB，实验语料 4.61MB），然后使用Word2vec工具训练词向量。最后，在未标注的目标领域测试数据集上，测试基于实例迁移和基于特征迁移的跨领域识别模型，输出两者融合后的最终预测结果。

(2) 基于特征迁移的算法模块 FruDA

此模块给出在软件中需要用到的基于特征迁移算法,该算法引入源领域和目标领域的公共特征，实现源领域知识向目标领域的迁移，最终获得基于特征迁移的跨领域中文模糊限制语识别模型h_f。

(3) 基于实例迁移的算法模块 TrAdaBoost

此模块给出在软件中需要用到的基于实例迁移算法,算法选择源领域中与目标领域相似的样例，扩大目标领域训练集，辅助目标领域学习，最终获得基于实例迁移的跨领域中文模糊限制语识别模型h_t。

(4) 基于特征迁移和实例迁移相结合的算法模块 Fru-TrA

基于特征迁移和实例迁移相结合的方法 Fru-TrA，融合基于实例迁移和基于特征迁移的跨领域中文模糊限制语识别模型的识别结果，得到最终的预测结果h_l。

2. 安装过程说明

本软件是绿色软件，无须安装。运行该软件需要电脑有 VC++6.0 或以上版本的编译环境支持。

3. 代码说明

3.1 输入模块

用户输入模块主要包括三个部分：训练词向量语料、基于特征的迁移学习参数、基于实例的迁移学习参数。

构建训练词向量语料，使用 Word2vec 工具训练词向量。

基于特征的迁移学习参数包括词向量、辅助训练数据文件的路径、训练数据文件的路径、测试数据文件的路径、模糊限制语的基本特征（词语特征，词性特征，关键词特征和共现特征）。这个模块包含的文件为/Fru_DA_mul/main.cpp

基于实例的迁移学习参数包括词向量、辅助训练数据文件的路径、训练数据文件的路径、测试数据文件的路径、模糊限制语的基本特征（词语特征，词性特征，关键词特征和共现特征）、辅助领域个数、迭代次数及每次最多加入实例的个数。这个模块包含的文件为/TransferBoost/PTSVM.h

各子文件具体内容及使用说明如下：

3.1.1 /Fru_DA_mul/main.cpp 文件说明

变量名称	使用说明
sourcepath	源领域训练数据的路径
targettrainpath	目标领域训练数据的路径
targettestpath	目标领域测试数据的路径
maxF	最大的特征编号
sourcefile	源领域训练数据
targettrainfile	目标领域训练数据
testfile	目标领域测试数据

3.1.2 /TransferBoost/PTSVM.h 文件说明

变量名称	使用说明
source_file_path_input	源领域训练数据文件的路径
train_file_path_input	目标领域训练数据文件的路径
test_file_path_input	目标领域测试数据文件的路径
iter_N	迭代次数
semi_add_M	每次迭代添加的实例个数
source_num	源领域的数目
source_data_num	每个源领域训练样例的个数
train_data_num	目标领域训练样例的个数
test_data_num	目标领域测试样例的个数
source_data_label	源领域训练样例的标签
train_data_label	目标领域训练样例的标签
weight_source	源领域训练样例权重
weight_train	目标领域训练样例的权重
beit	源领域训练样例权重的调整参数

3.2 输出模块

通过训练，得到基于特征和基于实例的迁移学习模型，再在目标领域测试数据上输出两者融合的最终分类结果。

各输出文件具体内容说明如下：

输出文件	说明
source_test_data_file.test	存放源领域训练数据临时文件
train_test_data_file.test	存放目标领域训练数据临时文件
test_test_data_file.test	存放目标领域测试数据临时文件
whole_train_data_file.train	语料库训练后，带标注和加权重的包含源领域和目标领域的训练数据文件
des_train_data_file.train	语料库训练后，带标注和加权重的目标领域训练数据文件
whole_model_file.model	在包含源领域和目标领域的训练数据上训练得到的模型 Fall
des_model_file.model	在目标领域训练数据上训练得到的模型 Ft
des_result_by_whole_file.result	Fall 分类器在目标领域训练数据上的分类结果
des_result_by_des_file.result	Ft 分类器在目标领域训练数据上的分类结果
test_result_by_whole_file.result	Fall 分类器在目标领域测试数据上的分类结果
source_result_by_whole_file.result	Fall 分类器在源领域训练数据上的分类结果
temp.train	特征转换后输出的包含源领域和目标领域的训练

	数据文件
temp.test	特征转换后输出的目标领域测试数据文件
temp.model	在 temp.train 上训练得到的模型
temp.result	temp.model 在测试数据 temp.test 上的分类结果