基于推理路径的文档级别关系抽取软件

使用说明书

编写人：

周惠巍，李雪菲，徐奕斌，雷弼尊，姜海斌

大连理工大学

目　录

第一部分　软件简介

新闻、报刊等非结构化的长文本中蕴含着大量有价值的结构化信息，是相关领域科学研究和实际应用中知识的重要来源，从非结构化长文本中抽取结构化信息则是获取这些知识的必要途径。文档级别关系抽取作为信息抽取的关键任务，能够自动抽取长文本中众多实体间的句内和跨句关系，这对于长文本结构化信息的自动抽取具有重要意义。

本软件提出一种基于推理路径的文档级别关系抽取方法，将非结构化的长文本信息与实体对间的推理路径结合，主要包括三个模块：（1）推理路径抽取。从构建的实体交互图中抽取实体对之间的限长路径；（2）路径编码。构建出多条推理路径利用双向长短时记忆网络（BiLSTM）进行编码，结合注意力机制获得最终的推理路径表示；（3）关系抽取。将推理路径表示和实体表示拼接输入关系分类层中，得到实体间交互关系。

1.1　软件名称

中文: 基于推理路径的文档级别关系抽取软件　简称 RP-DRES V1.0
英文: Reasoning path-based document-level relation extraction software

1.2　软件适用行业与用途

适用领域：自然语言处理相关行业。
用途：基于注意力机制融合实体对间多条推理路径，获得推理路径表示，辅助抽取文档级别实体关系。

1.3　软件开发平台

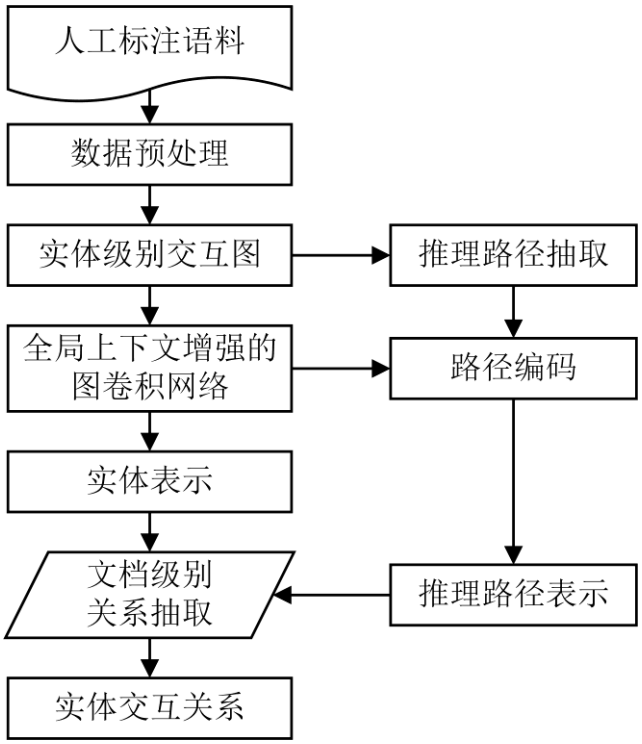
软件运行平台：Linux 操作系统
编程语言：Python3.6
版本号：　V1.0
程序量：　1628 行源代码

1.4　创作目的：

(1) 利用实体对之间的推理路径对关系抽取模型提供显式的指导，解决隐式推理无法

抽取出更明确、更具针对性的推理信息的问题，对于文档级别关系抽取具有重要的研究价值。

1.5 系统总流程图：



1.6 主要功能：

(1) 输入模块

该部分主要包括三部分：①模型训练和测试数据：这一部分需要用户根据其自身需求，自行准备。包括非结构化文档级别实体关系标注文本，即指定的实体对及其对应的上下文文本（训练语料）；待标注文本（验证、测试语料）；②初始化向量：此部分为可选项，用户可以选择采用自己训练的词向量作为初始化参数，软件根据用户输入自动为模型中相应的向量进行初始化赋值，若用户未提供此项输入，则默认为随机初始化；③模型训练超参数：此部分为模型所有超参数设置，包含迭代次数、批处理大小、图卷积层数、学习率、词向量维度、隐层维度等，用户也可以根据需求自行设定。

(2) 数据预处理模块

该部分主要包括四部分：①数据清洗：这一部分将用户给定的文档级别的数据进行文本清洗，具体方法是：首先将所有大写字母改写成小写字母，并利用空格代替文本中出现的特殊符号，实体中出现的特殊符号不做处理；然后采用 GloVE 词表和 BERT 的分词工具对文本进行分词，采用“UNK”字符串代替不在词表中的词；②实体特征抽取：

这一部分是抽取用户给定的文档级别的数据的额外特征，如类别特征、共指特征和相对距离特征；③实体交互图构建：这一部分是根据用户给定的文档级别的数据，以实体为节点，以实体对的上下文为边构建实体交互图；④数据封装：这一部分是将处理后的数据进行封装保存，训练、测试样例以列表的形式存储，预训练词向量和实体特征以 numpy 矩阵的形式存储，实体交互图以 networkx 为工具构建存储，便于模型的训练与测试，所有输入数据将被封装为 pkl、npy 以及 json 文件，存放在用户指定的目录。

(3) 模型定义模块

该部分定义了基于推理路径的文档级别关系抽取模型（Reasoning path-based document-level relation extraction model, RP-DRE），其中包括了全局上下文增强的图卷积网络与基于注意力机制的路径融合，用户可以根据需求在上述输入模块中修改模型参数。

(4) 模型训练模块

该部分将封装后的数据导入基于推理路径的文档级别关系抽取模型（RP-DRE），并进行训练，在每次训练完毕后，会对模型性能进行评价，并保存评价结果最好的模型及其参数。用户可以在上述模型定义模块中修改训练参数。

(5) 输出模块

在每轮迭代后，该部分将对模型进行测试，即用模型对用户提供的验证集与测试集进行预测，并输出预测标签，输出文件将存放在用户指定文件目录下。

第二部分　安装说明

2.1　软件运行环境

软件运行的硬件环境的最低配置： CPU 主频 3.60 GHz，4GB RAM，4GB 硬盘空间

为提高计算速度，改善用户体验，硬件建议配置：CPU 主频 2.4 G Hz 以上，内存容量 4 G，本软件需使用 cuda 加速，须保证显存容量 21GB 以上，硬盘剩余容量 2 G 以上，屏幕分辨率 1024×768。

软件运行的软件环境：Linux 操作系统。

2.2　安装过程说明

本软件是绿色软件，无须安装。

2.3 备注

未解技术问题，可联系

E-mail: zhouhuiwei@dlut.edu.cn

第三部分 使用手册

用户在使用此软件时需首先提供训练数据，用户可以根据任务需求提供文档级别的数据或直接提供训练、验证、测试样例。针对文档级别的数据（数据格式在 3.1 输入模块中进行说明），本软件需要先进行数据清洗、实体特征抽取和实体交互图构建，按照 1.6 软件简介-主要功能中所述方法处理输入数据，进行封装：

$ python3 ./ gen_bert_data_extend_graph_all_simple_path.py

然后进行模型训练，模型将在每轮训练结束后，自动对验证集进行评价，并保存在验证集中评价结果最优的模型：

$ python3 ./train.py

最后进行模型测试，保存模型的预测结果：

$ python3 ./test.py

下面将分别介绍用户需涉及的各模块文件及变量的具体含义及使用方法。

3.1 输入模块

输入模块用于提供模型训练所需数据与模型的超参数设置，主要包括三部分：用户提供的文档级别关系抽取数据；用户提供的预训练词向量（可选项）；模型参数设置，如训练论数、学习率、批尺寸、向量维度等。

用户提供文档级别的数据，需先进行数据清洗、实体特征抽取以及实体交互图构建。

参数设置文件中存储了包括输入、输出文件路径、预训练资源路径等路径参数，批尺寸、训练轮数、向量维度等模型超参数。文件名 Config_bert_all_path.py。

各子文件具体内容及使用说明如下：

3.1.2 Config_bert_all_path.py 文件说明

Config_bert_all_path.py 中定义了 Config 类，用于存储训练与测试过程中所涉及的所有超参数，其成员变量信息如下表所示：

成员变量名称	使用说明
train_prefix	文档级别训练数据（训练集）的文件名称
test_prefix	文档级别验证数据（验证集）的文件名称
data_path	文档级别全部数据的文件路径
max_length	文本数据最大长度
relation_num	总关系数量

GCN_layernum	图卷积神经网络层数
learn_rate	学习率
max_epoch	训练轮数
checkpoint_dir	训练获得的模型的输出路径
fig_result_dir	实体交互图的文件路径
result_dir	模型预测结果文件路径
pretrain_model	预训练模型参数的文件路径
word_size	词表大小
batch_size	批尺寸
model_name	选取模型的名称
save_name	训练获得的模型的名称
from_list_to_tensor()	抽取节点对的全部简单路径函数
load_train_data()	加载训练数据（训练集）
load_test_data()	加载测试数据（验证集、测试集）
train()	模型训练函数
test()	验证集模型测试函数
testall()	测试集模型测试函数

3.2 输出模块

输出模块主要用于存放训练获得的文档级别关系抽取模型参数、预测结果等。训练获得的模型默认存放在./checkpoint/目录下，模型预测结果存放在./result/目录下。各输出文件具体内容说明如下：

输出文件	说明
result.json	验证集的预测结果，{r}表示最终预测的关系
RP_DRES_model	训练获得的 RP-DRES 模型