实验二描述文档

实验目的

本实验要求以给定的英文文本数据集为基础，实现一个信息抽取系统。

任务描述

信息抽取系统的主要功能是从文本中抽取出特定的事实信息，我们称之为实体。然而，在大多数的应用中，不但要识别文本中的实体，还要确定这些实体之间的关系，我们称其为实体关系抽取。在本次实验中，关系抽取为必做内容，实体识别为可选做内容。

本实验利用经过特定处理的英文数据集作为训练数据，数据集中包括训练所用的文本、文本中所包含的实体及实体间的关系。例如

> “The system as described above has its greatest application in an arrayed configuration of antenna elements.”
> Component-Whole(elements, configuration)

其中第一句是文本信息，这些句子都包含一对以上的实体，无需考虑句子中只包含单实体的情况。第二句Component-Whole(elements, configuration)是得到的结果，提取到的实体对是(elements, configuration)，实体之间的关系是Component-Whole。实验的任务是训练关系抽取模型，即给定一个句子输入，模型可以输出该句子中最有可能所包含的关系。

在本次实验中，你需要：

- 对于文档进行适当的预处理，例如，去除标点符号与停用词等。
- 选取合适的模型对文本进行建模，并在训练集上进行关系抽取模型训练。例如，将关系类别作为文本的标签，将问题形式化为文本分类任务，并使用相应的模型进行处理。
- 在在线平台提交结果验证关系抽取模型的准确率。

除此之外，可选做的内容包括：

- 进行命名实体识别任务，设计模型提取句子中的所有实体，如有必要可自行进行数据标注。
- 设计一种模型，该模型同时完成命名实体识别与关系抽取任务。即给定一个句子，模型可以同时输出句子中的所包含的实体与实体间的关系。

对于实现的优化我们将视优化效果给予酌情加分。

数据集描述

数据集由6400条训练数据与1600条测试数据组成。训练数据样例如下：

```
1 "The system as described above has its greatest application in an arrayed configuration of antenna elements ."
Component-Whole(elements,configuration)
2 "The child was carefully wrapped and bound into the cradle by means of a cord."
Other(child,cradle)
3 "The author of a keygen uses a disassembler to look at the raw assembly code."
Instrument-Agency(disassembler,author)
4 "A misty ridge uprises from the surge ."
Other(ridge,surge)
5 "The student association is the voice of the undergraduate student population of the State University of New York at Buffalo."
Member-Collection(student,association)
```

其中每条数据包括一条英文句子与句子中的关系与实体。

测试数据样例如下：

```
6401 "The body of her nephew was in a suitcase under the bed."
6402 "The drama unfolded shortly after 7pm last Tuesday (December 22), when Glyn saw that smoke was coming from a bonfire ."
6403 "Prior to the 4004, engineers built computers either from collections of chips or from discrete components."
6404 "The effective utilization of a cluster of workstations for the implementation of a scientific application requires a highly flexible software environment."
6405 "A player manipulates a keyboard , a mouse, or a joystick as a game scene is displayed on a video monitor, or the like."
```

其中每条数据只包括给定的英文句子。

数据中的实体类型包括：
Cause-Effect、Component-Whole、Entity-Destination、Product-Producer、Entity-Origin、Member-Collection、Message-Topic、Content-Container、Instrument-Agency、Other。当句子中实体之前不满足前九种关系时，将标签设置为Other。

评价指标

详细评分算法如下：

$$Accuracy = \frac{S_c}{S}$$

其中S_c为关系判断正确的个数，S是数据总数。Accuracy值越高，代表判断的越准确。

提交要求

要求提交exp2文件夹，其中包括存放源代码的src文件夹，以及实验报告pdf文件与README文件。实验报告中需介绍你所使用的算法及所做的优化，并给出最终模型在测试集上的评价结果。如果是多人组队，请在实验报告中注明所有组成员的学号和姓名。README 文件中包含你的源代码的运行环境，编译运行方式，以及对关键函数的说明。同时，对于所作要求之外的文件，也请在 README 中注明这些文件的含义。

时间安排

自2020年11月20日起，为期六周，两人一组，延迟提交将根据延迟提交扣分。时间安排如下：

- 第1周：发布训练数据，进行程序设计和算法准备
- 第2周：发布测试数据（无标签）和提交样例，正式开始实验
- 第3~5周：指定平台，提交实验结果（总提交次数不超过10次，取最佳的一次）
- 第6周：提交实验报告

截止日期：2020年1月1日23:59