信息抽取

提供业界最大规模的中文信息抽取数据集，
让机器具备从海量自然语言文本中自动获取知识的能力。

报名成功

国内用户下载

train_data.json.zip 训练集 https://nlpc-du.cdn.bcebos.com/KG/train_data.json.zip	dev_data.json.zip 验证集 https://nlpc-du.cdn.bcebos.com/KG/dev_data.json.zip	all_50_schemas 50个定义好的schema https://nlpc-du.cdn.bcebos.com/KG/all_50_schemas	test1_data_postag.json.zip 测试集1 https://nlpc-du.cdn.bcebos.com/KG/test_data1_postag.json.zip
test_data_postag.json.zip 测试集2 5月13日开放下载			

海外用户下载

train_data.json.zip 训练集	dev_data.json.zip 验证集	all_50_schemas 50个定义好的schema	test1_data_postag.json.zip 测试集1
test_data_postag.json.zip 测试集2 5月13日开放下载			

数据介绍

2. 验证集：共2万个句子，包含句子中对应的SPO，用于竞赛模型训练和参数调试。

3. schema约束：共50个限定的schema，定义了关系P以及其对应的主体S和客体O的类别。

4. 测试集1：约1万个句子，不包含句子中对应的SPO，用于参赛者在平台上自助提交模型预测结果、验证效果。

5. 测试集2：本次竞赛最终测试集，约2万个句子，不包含句子对应的SPO，包含测试集1。该部分结果不能在平台上自助验证，会在竞赛结束前1周发布。

数据样本

平台提供的数据为JSON文件格式，样例如下:

```
{
  "text":"如何演好自己的角色，请读《演员自我修养》《喜剧之王》周星驰崛起于穷困潦倒之中的独门秘笈",
  "spo_list":[
      {
          "object_type":"人物",
          "predicate":"主演",
          "object":"周星驰",
          "subject_type":"影视作品",
          "subject":"喜剧之王"
      }
  ],
  "postag":[  #分词及词性信息
    {
        "word":"如何",   #word segmentati
        "pos":"r"       #POS tagging
    },
    {
        "word":"演",
        "pos":"v"
    }
  ]
}
```

*更多样例和详细数据格式说明参见数据集中包含的“数据格式说明”文档。

入门参考

基线系统: 开源的信息抽取基线系统源码

PaddlePaddle使用教程: 请参考paddlepaddle官网 http://www.paddlepaddle.org/

联系我们

任何与本次技术竞赛相关的问题，请随时联系竞赛会务组。

竞赛会务组邮箱：lic2019@126.com