

实验2: 命名实体识别

Named Entity Recognition (NER)

NER任务说明

- 输入： 一段文本
- 输出： 文本中的实体， 包括实体名称、类别、位置

示例： Volkswagen AG won 77,719 registrations , slightly more than a quarter of the total .

输出： (Volkswagen AG, organization, [0, 13])

BIO标签

示例： Volkswagen AG won 77,719 registrations

B-ORG I-ORG O O O

解码得到实体： (Volkswagen AG, organization)

标签含义

- B-ORG begin of ORG entity
- I-ORG inside of ORG entity
- O outside of entity

注意：本次实验暂不需要进行BIO 标签的解码，只需要输出每个单词的BIO-tag即可，但提交后的评价指标是预测出完整实体的f1值。

数据说明

train.txt

```
1 -DOCSTART- -X- -X- 0
2
3 EU NNP B-NP B-ORG
4 rejects VBZ B-VP 0
5 German JJ B-NP B-MISC
6 call NN I-NP 0
7 to TO B-VP 0
8 boycott VB I-VP 0
9 British JJ B-NP B-MISC
10 lamb NN I-NP 0
11 . . 0 0
12
13 Peter NNP B-NP B-PER
14 Blackburn NNP I-NP I-PER
15
```

文档分段标签行，该行不做处理

分词后的一个句子，句语句之间使用空行隔开
句子中每一行有四个元素，分别为：（原句中的单词、词性、组块标签、实体标签）

测试集每一行只包含前三个元素，需要根据前三个元素预测最后一个元素，也就是实体标签

训练集、验证集

测试集

test.txt

```
1 -DOCSTART- -X- -X- 0
2
3 SOCCER NN B-NP
4 - : 0
5 JAPAN NNP B-NP
6 GET VB B-VP
7 LUCKY NNP B-NP
8 WIN NNP I-NP
9 , , 0
10 CHINA NNP B-NP
11 IN IN B-PP
12 SURPRISE DT B-NP
13 DEFEAT NN I-NP
14 . . 0
15
```

每行只有三个元素，需要预测最后一个实体标签

数据说明

- 数据集：见QQ群（实验2-命名实体识别数据集.zip）
- 数据集大小（按句子数统计）
 - training set: 14987 dev set: 3466 test set: 3684
- 实体类别
 - persons (PER)
 - organizations (ORG)
 - locations (LOC)
 - miscellaneous names (MISC) (其他名称)

提交要求

- 结果提交网址: <https://competitions.codalab.org/competitions/27723>
- 每个队伍在指定网站注册一个账号, 提交预测结果, 每天每队最多提交6次
- 格式与训练数据一致
- 提交文件行数与test.txt一致
- 提交文件和测试集中各行顺序不能改变
- 预测结果命名为out.txt并打包为zip文件, 可以添加其他方案说明文件, 和out.txt一起打包。(压缩文件不包含目录)

```
test.pred.txt
1  -DOCSTART- -X- -X- 0
2
3  SOCCER NN B-NP B-PER
4  - : 0 B-PER
5  JAPAN NNP B-NP B-PER
6  GET VB B-VP B-PER
7  LUCKY NNP B-NP B-PER
8  WIN NNP I-NP B-PER
9  , , 0 B-PER
10 CHINA NNP B-NP B-PER
11 IN IN B-PP B-PER
12 SURPRISE DT B-NP B-PER
13 DEFEAT NN I-NP B-PER
14 . . 0 B-PER
15
16 Nadim NNP B-NP 0
17 Ladki NNP I-NP 0
```

submit.example

评测指标：Micro-F1

- 使用F1值作为评估标准，精确率和召回率是针对**实体**而非词标签。

$$Recall = \frac{TP}{FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

评分标准(实验部分15分)

■ 队内贡献得分 (上限2分)

- 组队方式: 每个队伍1-3人, 队内根据贡献度评分, 每人分数不能相同。(按照团队基础分+2、+1.5、+1的顺序)

■ 团队分 (上限13分)

- 方法分: 每支队伍至少使用两种不同的实体识别方法。(8分, 可以使用开源工具)
- 代码分: 队伍至少代码实现一种算法。(2分, 指非仅使用第三方工具的情况, 但可以参考开源代码)
 - 例如实现规则词典匹配、一种机器学习算法或自己搭建神经网络
 - 根据代码工作量、规范程度、复现情况评分
- 效果分: 根据队伍提交最高F1分数、排名以及代码复现情况打分。(3分)
 - 前25%名 3分
 - 前25%-50%名 2分
 - 后50% 1分

代码规范及提交流程

不限语言和第三方开源工具，为了方便复现，对代码目录有如下基本要求：

- team-**{teamID}**-**{methodID}**/ //项目根目录, teamID为组队编号(从1开始), methodID为队内采用的方法编号 (从1开始)
 - data/ //数据集目录, 至少包含以下三个文件
 - train.txt
 - valid.txt
 - test.txt
 - run.sh 或者 run.bat //运行入口, 根据开发环境编写sh或者bat命令脚本, 要求运行该脚本后生成out.txt
 - readme //依赖环境搭建说明, 尽可能简洁清晰 (还可以简单描述下项目方案)
 - (可选)requirements.txt //python依赖包或其他语言依赖

每种方案用一个单独的项目根目录，一个团队的多种方案压缩至team-**{teamID}**.zip

- 压缩包小于30M直接发送至changhong.he@hit.edu.cn，邮件主题为**team-**{teamID}**-代码提交**。
- 压缩包较大的单独联系助教(何长鸿)提交，最大不可超过1Gb。

评分标准(实验报告, 5分)

- 每队提交一份报告
 - 摘要（如何写好一个摘要？） (0.5分)
 - 内容完整、格式规范、排版美观 (2.5分)
 - 摘要、关键词、引言、相关工作、方法一、方法二、总结与展望、实验体会与收获、参考文献
 - 排版规范美观
 - 内容表述准确简练
 - 完整描述实验方法 (2分)
 - 算法原理介绍、优劣分析
 - 所用方法取得的识别结果
 - 自己实验中带来识别结果提升的创新点
 - 不同方法之间的比较