计算语言学大作业说明

一、任务描述-中文语义角色标注

- 论元识别

根据中文宾州命题库(CPB)，给定某个特定的命题(/rel)，识别出句子中的该命题的论元以及其左右边界。例如在下列例句中：

我们/PN/O 希望/VV/O 台湾/NR/B-ARG0 当局/NN/E-ARG0 顺应/VV/O 历史/NN/O 发展/NN/O 潮流/NN/O ，/PU/O 把握/VV/rel 时机/NN/S-ARG1 ，/PU/O 就/P/O 两/CD/O 岸/NN/O 政治/NN/O 谈判/NN/O 作出/VV/O 积极/JJ/O 回应/NN/O 和/CC/O 明智/JJ/O 选择/NN/O 。/PU/O

例句已经完成分词和词性标注(part of speech, POS)。对于每一个词块“A/B/C”，A 是词；B 是词性信息；C 是论元标记。

在上述例句中表征命题的目标动词为“把握”，该命题有两个论元“台湾当局”以及“时机”，他们所充当的角色是 arg0 和 arg1，参评系统应能正确识别这些论元的左右边界以及所充当的角色。如果"台湾当局"只识别出来了"台湾"，是不可以算识别正确的论元。

关于 arg0~arg4 以及 argM 等语义角色的含义详见补充材料。

- 评价指标：

论元识别性能采用 P/R/F 指标加以评价，具体而言：
命题论元识别正确率(P)=系统识别正确的命题论元数/系统识别的所有命题论元数*100%
命题论元识别召回率(R)=系统识别正确的命题论元数/标准答案中所有命题论元数*100%
命题论元识别 F 值=2*P*R/(P+R)

二、数据集和评价

- 数据集

cpbtrain.txt　训练数据集
cpbdev.txt　开发数据集
cpbtest.txt　测试数据集(无正确答案)

数据集格式为每行一个句子，每个句子都有且仅有一个命题标签“/rel”，并且句子已经完成中文分词。对于每一个词块“A/B/C”，A 是该词；B 是词性信息；C 是论元标记。论元标记信息在测试数据集中没有给出，在训练和开发集中给出。

论元标记由两部分组成：位置标签和标记内容。位置标签共有四种，分别是 S-(单词论元词), B-(论元的首词语)，E-(论元的尾词) 以及 I-(论元首尾词之间的词)。

对于含有多个词的论元，我们需要正确识别出它的论元标签以及左右边界。比如：

《/PU/B-ARG1 国家/NN/I-ARG1 高新/JJ/I-ARG1 技术/NN/I-ARG1 产业/NN/I-ARG1 开发区/NN/I-ARG1 管理/NN/I-ARG1 暂行/JJ/I-ARG1 办法/NN/I-ARG1》/PU/E-ARG1

在 B-标签和 E-标签之间的所有位置标签都是 I-标签，在输出 cbptest.txt 的语义角色标注答案时也应该按照此项原则。

注意： 本数据集是从 LDC 收费数据集 Chinese Treebank 选取部分样例构成的，请勿用于大作业以外的用途。

- 评测脚本 calc_f1.py

用法：python calc_f1.py <your_predict_file> <origin_data_file>
<your_predict_file>：你预测的答案文件
<origin_data_file>：原始文件
可以用此脚本测试训练集以及开发集的表现用于调参。
测试集答案未给出，提交作业后由助教评测，模型表现作为评分依据之一。

三、作业要求

- 要求 1：

本次大作业由 1~2 人完成，推荐 2 人组队完成。

- 要求 2：

最终打包提交如下内容：
实验报告： 需要详细说明使用的方法、资源、工具和参考文献，分析实验结果，列出组员分工。报告中需列出所使用的方法在开发集上的结果。
程序源代码： 统一放入命名为 src 的文件夹内，要求程序风格良好、注释详细，可运行，并给出程序的运行方式。程序应能在 Linux 环境下运行，运行方式应简单明了（如以 makefile 的形式给出编译，并给出程序运行脚本）
测试集数据答案文件： 与 cpbtrain/cpbdev.txt 文件的格式相同，注意以 utf-8 格式编码。

- 要求 3(可选)：

课堂报告：各组如果对自己组的方法比较满意，想将方法分享给其他同学的，可以自愿选择课堂报告，形式为 oral presentation，各组做好 ppt 派代表进行报告，报告时间为 10 分钟。课堂报告的组会有相应加分。课堂报告将在最后一堂课进行，想要进行课堂报告的同学请提前报名。

最终成绩评定由测试数据指标、模型方法难度、组员分工、实验报告、代码和课堂报告综合评定。

四、作业提交方式

打包提交到如下邮箱(写明姓名学号)：
rogerliuty@foxmail.com

截止时间：
2017 年 12 月 22 日