哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

自然语言处理

实验三：实体识别系统

School of Computer Science and Technology

Harbin Institute of Technology

1 实验目标

本次实验目的是对命名实体识别技术有一个全面的了解，包括数据集构建、特征提取、实体识别方法、识别结果评价等环节。本次实验所要用到的知识如下：

- 基本编程能力（文件处理、数据统计等）
- 训练数据和测试数据的使用方法
- 实体识别所需的特征构造和特征提取方法
- 最大熵方法、条件随机场方法（CRF）
- 分词、词性标注技术与实体识别技术之间的关系
- 命名实体识别评价常用指标

2 实验环境

编程语言为：C++/C#/python/Java

其他无特殊要求

3 实验内容及要求

3.1 特征提取

输入文件：训练文件

输出：特征文件

提交内容：1）特征文件(文件名：feature.txt，文件中每行对应一个样本及其对应的实体标记)

2）程序源代码

3.2 基于最大熵模型（或 CRF）的实体识别模型

输入文件：feature.txt

输出： ME_model.txt

编程要求：

- 允许调用各种最大熵工具包
- 鼓励使用工具包对应的源代码，不鼓励直接命令行调用工具包

提交内容：1）模型文件 model.txt

2）程序源代码

3.3 基于最大熵模型（或 CRF）的实体识别结果

输入文件：测试文件 模型文件 model.txt

输出：test_feature.txt result.txt

编程要求：

提交内容：1）测试文件对应的特征文件 test_feature.txt

2）result.txt（格式和训练文件格式相同）

3）程序源代码

3.4 实体识别结果评价

输入文件：测试文件对应的标准答案 result.txt

输出：评价结果 evaluation_result.txt

要求：

- 调用实验提供的评价程序（SharedTaskEval.pl）；

提交内容：1）保存评价结果的文件 evaluation_result.txt

4 实验报告

要求：字数不少于 3000 字，采用科技论文的组织方式，内容包括作者信息（姓名、学号、email）、中英文摘要、引言、实体识别相关研究工作、自己所采用的方法和特征的详细介绍、在实验给定数据集（NLPBA 2004 数据集）上的实验结果（P、R 和 F1 值）、实验结果分析、本次实验的心得收获及相关的参考文献。

实验报告要格式规范、逻辑清晰、内容完整。

5 提交方式(暂定)

截止日期：11 月 27 日

提交方式：现场提交