Lab2　中文分词

一.任务

(1) 阅读《Conditional random fields: probabilistic models for segmenting and labeling sequence data》论文,这篇论文是提出 CRF 模型的首篇论文,主要搞清楚 CRF 的思想和方法, 对于模型训练算法可以忽略(因为作者提出的两种算法都并不是很好,后人经过了许多改进)。

(2) 阅读Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms》论文,这是一篇训练 CRF 模型常用的算法之一,想法简单, 实现容易。

(3) 根据上述两篇论文实现CRF模型。

(4) 将CRF应用于中文分词的任务,在训练集进行训练模型,并且测试集测试模型。其中 train.utf8 是训练集、template.utf8 是特征模板、labels 是标注集合(B 表示词首字、 I 表示词中字、E 表示词尾字、S 表示单字词)。对于 template 文件可以自己进行调整以达到较佳性能。

(5) word2vec 预训练字的表示（可参考https://github.com/candlewill/Chinsese_word_vectors）。

(6) 阅读 《Bidirectional LSTM-CRF Models for Sequence Tagging》论文，实现论文中的双向LSTM+CRF的网络结构。

(7) 通过 BiLSTM + Viterbi 实现中文分词的任务。其中 train.utf8 是训练集、labels 是标注集合(B 表示词首字、 I 表示词中字、E 表示词尾字、S 表示单字词)。

(8) 提交源代码。

(9) 提交较详实的实验报告。

二.评分

- 实现CRF模型，模型能够正确运行并收敛(30%)
- 实现 BiLSTM + Viterbi 模型，模型能够正确运行并收敛(30%)
- 在另外给出的最终测试集上的性能(20%)
- 代码风格和文档(20%)

三.其他

- 实验截止时间为2018年11月25日。
- 出现抄袭现象,抄袭双方均按零分计。
- 请严格按照 deadline 提交,超出每天扣除总分的 20%
- 这次实验有一定难度,请大家努力尝试和完成，如有问题,随时向助教和任课教师询问。