

第十五届全国计算语言学会议 (CCL 2016)
山东 烟台



自然语言处理国际前沿动态综述

词法分析

邱锡鹏

复旦大学

2016年10月16日

<http://nlp.fudan.edu.cn/xpqi>

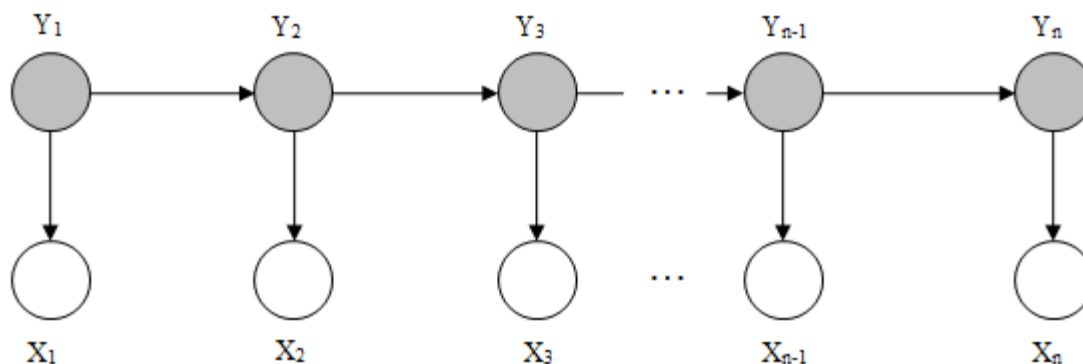


词法分析

- ▶ 词法分析是将构成句子的字符序列转换为词的序列，并对每个词加上语法或语义标记。
 - ▶ 分词
 - ▶ 词性标注
 - ▶ 命名实体识别
 - ▶ 词义消歧

传统模型

- ▶ **序列标注**是词法分析的主要使用模型。



- ▶ 模型：HMM、MEMM、CRF
- ▶ 特征：
 - ▶ 离散特征 → **特征工程问题**
 - ▶ 基于窗口方法 → **长距离依赖问题**

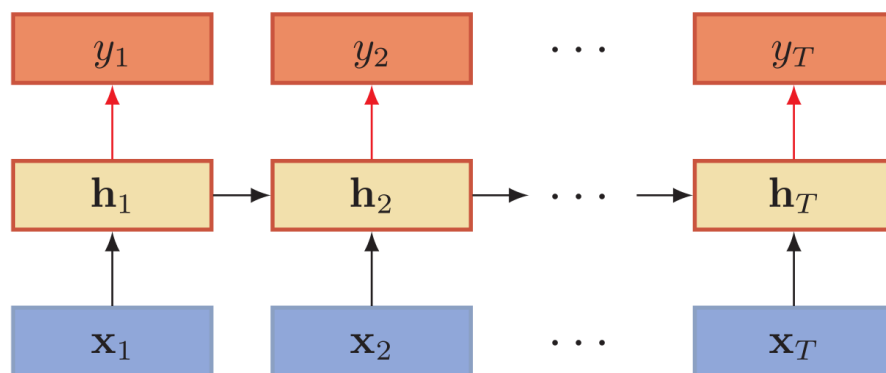
深度学习方法替代了传统方法

► 2014年以前

- → 特征工程问题
- 用分布式语义表示来代替传统的离散特征

► 2015年

- → 长距离依赖问题
- 循环神经网络 EMNLP 2015





深度学习的方法替代了传统方法

▶ 2016年

▶ 模型：LSTM成为基准模型

▶ LSTM、BLSTM、LSTM+CRF, LSTM+CNN+CRF, ...

▶ 表示：字、词的混合表示

▶ 英文

□ 引入字母表示

▶ 中文

□ 引入词表示



问题依旧

- ▶ OOV问题 out of vocabulary
- ▶ 领域迁移问题
- ▶ 语义理解问题
 - ▶ 涉及语义理解的歧义情况，仍然无法解决
- ▶ 标准问题
 - ▶ 以分词为例，
 - ▶ 标准差异：CTB、PKU、MSR等
 - ▶ 粒度差异
- ▶ 评价问题
 - ▶ P、R、F

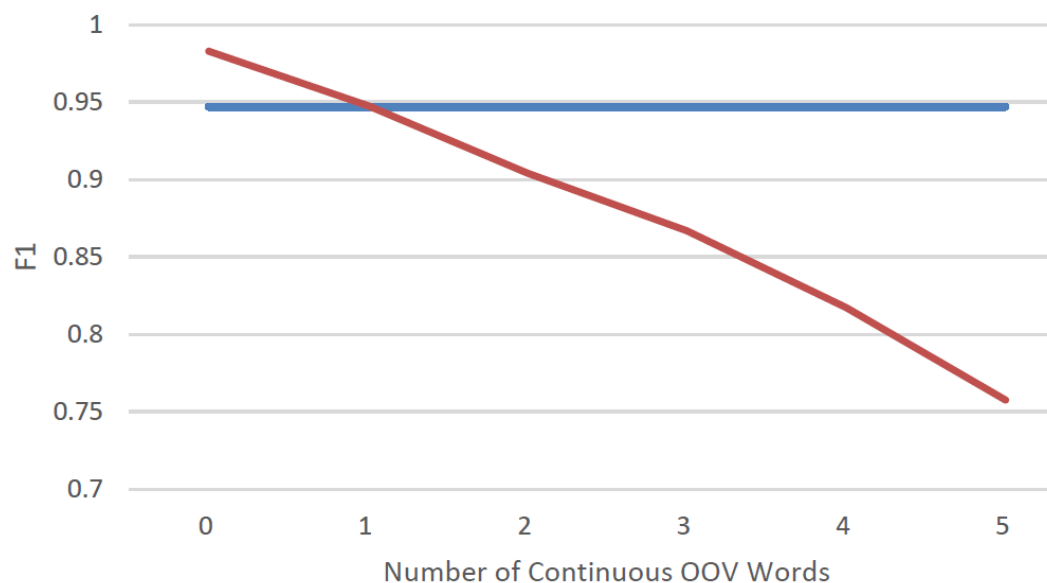
评价问题

▶ 传统PRF的不足

- ▶ 容易切分或不重要的词占了很大比例
- ▶ 和人类感受不一致

▶ 改进方法

- ▶ 引入OOV Recall





一个新的评价标准

- ▶ 根据词的难易程度对其进行加权
 - ▶ 正确切分一个难的词获得额外的奖励
 - ▶ 错误切分一个容易的词获得额外的惩罚
- ▶ 难易程度是由100个具有差异性的“专家”来投票决定
 - ▶ 专家：用不同特征、数据集训练出来的分类器



展望

► 战术

- ▶ 数据
- ▶ 半监督、无监督学习
- ▶ 更复杂的模型

► 战略

- ▶ 是否需要**显示**的词法分析?
- ▶ 词法信息是否可以**隐式**地被建模?



谢 谢