清华大学深圳研究生院

计算智能实验室 周报

姓名	研究方向	报告覆盖时间
李晨辉	NLP（法律文本分类）	2010.3.11-2019.3.17

本周主要完成的工作

语言模型的学习

语言模型及概率图模型

1. 语言模型

在统计自然语言处理中，一个语言模型通常构建为字符串 s 的概率分布 p(s),这里 p(s)试图反应的是字符串 s 作为一个句子出现的概率。

N 元文法所刻画的语言模型是基于马尔可夫假设（一个词语出现与否只与其前面出现的词有关），语言模型评估一般使用根据模型及计算出的测试数据概率，或者使用交叉熵和困惑度。一般的，交叉熵和困惑度越低越好。

基于 n 元语法的语言模型还存在一个比较常见的 0 概率问题，即某些词的组合在训练数据中可能并未出现，导致对应的测试数据概率为 0.解决上述问题的办法是数据平滑：

主要有以下几种方法：

- 加法平滑：假设每种词的组合最少会出现 k 次
- 古德-图灵(Good-Turing)估计法：对于任何一个出现 r 次的 n 元语法，都假设它出现了(r+1)n(r+1)/n(r),其中 n(r)表示训练预料中出现 r 次的 n 元语法的数目。经过该方法平滑之后 p(r)的概率之和不为 1，需要做归一化处理。另外，这种方法不能直接用于 n(r)=0 的计算，也无法实现高阶模型与低阶模型的结合。（高低阶对应窗口大小）
- Katz 平滑：主要思想是对出现频次大于 k 的事件运用最大似然估计的方法进行减值，对于出现频次小于 k 的事件可以用其低阶模型作为替代高阶模型的后备（back-off）。换句话说就是讲减值节省下来的概率按照低阶语法模型的分布情况分配给未出现的事件。
- Jelinek-Mercer 平滑：使用低阶的 n 元模型向高阶的 n 元模型插值
- Witten-Bell 平滑：可以看做是 Jelinek-Mercer 方法的一种实例
- 绝对减值法：同样是使用低阶的 n 元模型向高阶的 n 元模型插值，但对于高阶模型的减值不是采用乘法因子，而是对每个非零计数减去一个固定值 D<=1。
- Kneser_Ney 平滑：扩展的绝对减值算法，对所有单词的低阶分布进行插值而不是仅对那些在高阶分布中计数为 0 的单词插值。

- 修正的 Kneser_Ney 平滑：不是对所有非零计数都减去相同的 D，而是对计数分别为 1、2 和大于等于 3 的三种情况采用不同的 D1,D2,D3.
- 总结：
  - 平滑算法比较：
    1. 对于二元和三元语法而言：Kneser_Ney 平滑和修正的 Kneser_Ney 平滑效果最好；
    2. 在稀疏数据的情况下，Jelinek-Mercer 平滑由于 Katz 平滑；在有大量数据的情况下则相反；
    3. 平滑算法的相对性能与训练预料的规模，n 元语法模型阶数和训练语料本身有较大的关系。

下周工作主要安排

(1) 使用 TensorFlow 尝试实现一下算法；

(2) 自动分词、命名实体识别、词形标注等算法的学习；