中文信息处理
马尔科夫随机过程（模型）- 以 ngram 语言模型为例

引子 李正华

苏州大学

2015 年 10 月 7 日

随机过程

- 状态之间的转移依据一定的概率，是随机的（stochastic）

马尔科夫随机过程

Markov process

- 马尔可夫过程指的是在过程中，下一个状态只和前面几个有限的状态相关。
- 两个性质
  - 有限历史
  - 时间不变性

什么是语言模型

- 什么是建模　抽象表示的过程
- 什么是模型　抽象
- 什么是语言模型？
  - 定义 $P(S;\lambda)$; S 表示一个句子，λ 为模型参数（判别句子概率）
  - 定义 $P(w_i|w_{i-1})$; 2-gram 语言模型（语言生成）

语言模型有什么应用？

- 机器翻译
- 以英中翻译为例（I read at home）
- 根据词典生成很多翻译候选，然后排序
- word-based MT
- greedy search -> beam search

语言模型有什么应用？

- 语音识别
- 输入法
- 搜索引擎中的 query 补全和提示

语言模型的最初形式：句子的生成过程

- $P(S, \lambda) = ...$
- $P(w_i | w_1, w_2, ..., w_{i-1})$
- 存在什么问题？

语言模型的最初形式：句子的生成过程

- $P(S, \lambda) = ...$
- $P(w_i | w_1, w_2, ..., w_{i-1})$
- 存在什么问题？
- 参数估计时，存在严重的数据稀疏问题

N-gram LM（马尔科夫过程，简化）

- $P(S, \lambda) = ...$
- $P(w_i | w_{i-N+1}, ..., w_{i-1})$
- N 元模型，$N-1$ 阶模型
- $P(w_i | w_{i-1})$，$N = 2$ 元模型

N-gram LM 参数估计

- 如何估计 $P(w_i|w_{i-1})$

ngram 平滑方法

- 参考 Michael Collins 课件。

References

- 计算所宗承庆老师课件：http://www.nlpr.ia.ac.cn/cip/ZongReportandLecture/Lecture_on_NLP/Chp-05.pdf
- http://52opencourse.com/111/斯坦福大学自然语言处理第四课-语言模型（language-modeling）
- http://52opencourse.com/49/斯坦福大学自然语言处理公开课课件汇总
- http://spark-public.s3.amazonaws.com/nlp/slides/languagemodeling.pdf
- http://xuh5156.github.io/2014/03/25/languagemodel2/
- http://www.speech.sri.com/projects/srilm/manpages/pdfs/chen-goodman-tr-10-98.pdf

1 个编程作业（共 10 分）

- 给一个分好词的文件 train.conll，基于极大似然估计 + 某一种平滑方法，得到一个三元语言模型 M，储存起来（4 分）。
- 在另一个分好词的文件 dev.conll 上，计算 M 的混乱度（perplexity）（3 分）.
- 利用 M，随机生成 100 个句子（3 分）。