20.1-20.3

计算词汇语义学.

- 词义排歧 WSD (word sense disambiguation) ← 监督学习
  - ① 词汇采样任务 (lexical sample)
    用训练好的分类器, 进行词汇标注.
  - ② 全词排歧任务 (all-word)
    使用semantic concordance语料

和目标词语相关的上下文特征, 需要丰富的特征信息.

特征向量 (feature vector)
- ① 搭配特征 (collocational features) → 包含特定位置
- ② 词袋特征 (bag-of-words) → 不包含特定位置
  - 不使用停用词 (stopword) 作特征.

→ 第23章: 词袋技术形成了搜索引擎中"向量空间模型"的基础 (vector space model)

训练语料 → 特征 → 意义分类器
- ① 朴素贝叶斯分类器 (naive Bayes classifier)
- ② 决策表分类器 (decision list classifier)
  - 会用到时效以然比

with the following formulation of a **naive Bayes classifier for WSD**:

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} P(s) \prod_{j=1}^{n} P(f_j|s) \quad (20.6)$$

→ 计算每种词义的概率, → 然后返回 argmax 值
最大似然估计.

各个特征概率是 naive 假设为独立的, 所以可以连乘:

$$P(\vec{f}|s) \approx \prod_{j=1}^{n} P(f_j|s)$$

☆ 那么, 如何评价 WSD (词义排歧)?
精确匹配词义的准确率. (sense accuracy)

↓
评价的框架为 SENSEVAL | SEMEVAL.
提供了全词排歧和词汇采样任务的意义清单 (sense inventories)
↓
通过: ① 基准线 (baseline) 合格指标.
② 上限 (ceiling) 最佳性能.

20.4

WSD: 字典方法和同义词库方法.

- 也有非监督方法:
  - ① Lesk算法 (Lesk algorithm)
- e.g. 比如 如何排歧 Bank?
  - ① Bank' ... deposits ... mortgage ...
  - ② Bank² ... river ...

  如果, 某句话里 Bank 和 deposits 出现 (词义重叠)
  则, Bank 是 Bank'

  如果, 停用词重叠算不算呢?
  则, 算, 但停用词可以改为另一种对"功能词打折"的方法 → IDF
- e.g. 通过"选择关联性"对选择限制. (概率改变.)
  选择优先度 (selection preference strength) ← 分布差异
  - ① 相对熵 (relative entropy)
  - ② Kull-Leibler 散度距离 (Kullback and Leibler)

20.6 词语相似度 { ① 语义词典方法 (thesaurus-based)
② 分布方法 (distributional)

- 词语相似度 ≠ 词语相关度
e.g. 反义词 (低相似度，高相关度)

{ ① 两个词语之间的路径越短，越相似
② 相似层次结构：最低公共包含点的信息量 —估计→ 共同的信息量.
③ 扩展的注释交集法：两个义项注释包含相同的词语，则把它们视相似.

+20.7
20.6 词语相似度分布方法

- 为什么不用语义字典方法，还要用这个？
- 很难比较名词、动词之间的相似度.

↓ 可以用分布方法，自动生成语义词典.
(automatic thesaurus generation)

↓ 如何评价分布式词语的相似度？

- 最佳方法：① 与检验集对关联度进行加权.
② 用用Dice或Jaccard算法无度量向量相似度.

20.8 下位关系和其他词语关系. (略)

20.9 语义角色标注. (略)

20.10 高级主题：无监督语义排歧.