

单词的表示

同义词词典

如WordNet

- 定义同义词和层级结构（上下位，部分整体）
- 使用单词网络，计算单词相似度
- 缺点
 - 人工更新，成本
 - 无法表示同义词之间的微妙差异

计数（统计）

减少人工干预
从数据自动获取

语料库corpus

- 预处理：分词（正则表达式）
- 单词标上ID，将单词表示为向量

分布式表示

- 将单词表示为固定长度的向量
- 例如RGB[201,30,30]-->红色系。
向量能够体现含义，有意义的计算

分布式假设：单词的
含义周围的单词形成

- 周围：窗口大小
- 共现矩阵：对周围出现什么单词计数（n*n）
- 余弦相似度：(向量内积)/(L2范数)：直观表示两个向量在多大程度上指向同一方向，同向1反向-1
- 改进：①PMI指标减少无意义词（类似the）的相关性。②将大稀疏向量SVD降维小密集向量

结果：含义和语法相似的单词在向量空间位置相近

矩阵大，SVD耗费资源
NN基于mini_batch反复更新权重

概要

- 从模型视角出发
根据周围单词（上下文）预测mask：完形填空
可以学习到单词的出现模式
- 单词转化为固定长度：one-hot
相当于给单词加ID
固定神经元输入层
- 与权重乘积相当于提取对应行向量

推理(NN)

word2vec

- CBOw
连续词袋
BOW不考虑顺序
 - 根据上下文预测目标词
 - 输入测权重W就是单词的分布式表示
 - 中间层神经元数量少，产生密集向量：相当于编码工作
从中间层获得结果：解码
 - 后验概率： $P(w_t|w_{t-1},w_{t+1})$
- 与cbow对称，从中间词预测上下文
- skip-gram
 - 一个输入层，多个输出层
各输出层loss叠加
 $p = P(w_{t-1},w_{t+1}|w_t) = P(w_{t-1}|w_t)P(w_{t+1}|w_t)$
 $L = -\log(p) = -(\log p_1 + \log p_2)$
 - 能够理解类推
king+man=woman+queen
- 高速化
 - Win：Embedding层：单词ID对应行向量的层
 - Wout：Negative Sampling替代softmax
负采样：二分类拟合多分类
Loss正1负0，根据概率采样
- 优点
 - 内嵌单词含义
 - 求解类推问题
 - 迁移学习能力

GloVe:融合计数和推理