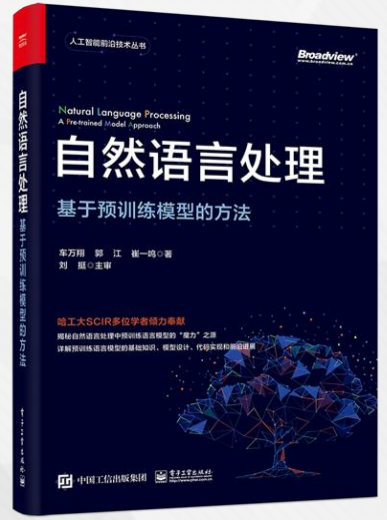


静态词向量预训练模型

车万翔、郭江、崔一鸣

社会计算与信息检索研究中心
哈尔滨工业大学

目 录

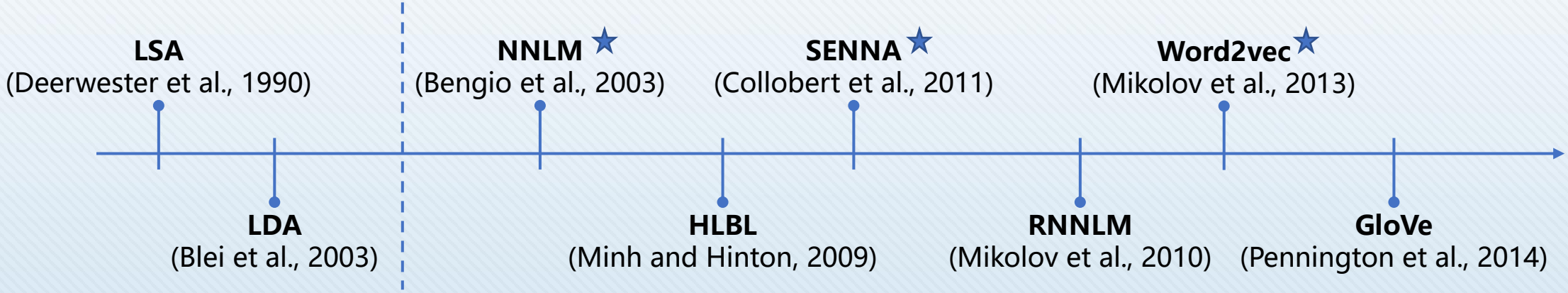
CONTENTS

目 录

CONTENTS

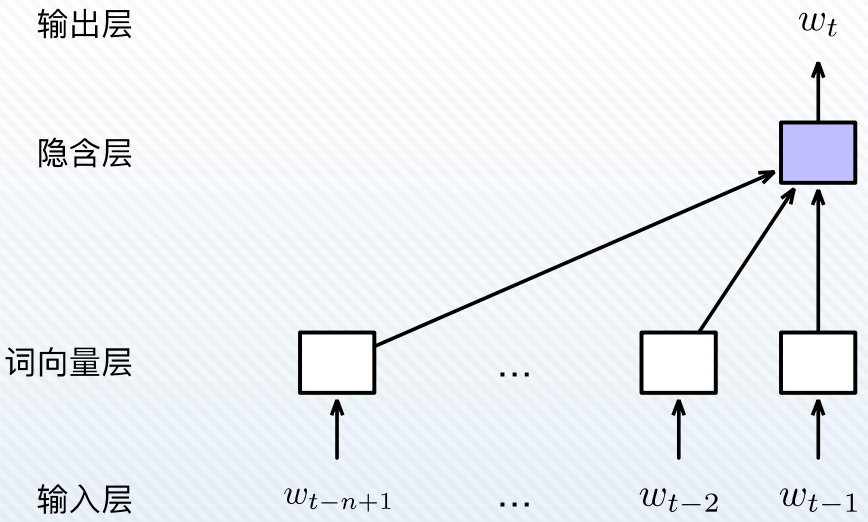
分布式（Distributed）词表示

- 分布表示的缺点
  - 训练速度慢，增加新语料库困难
  - 不易扩展到短语、句子表示
- 分布式表示直接使用低维、稠密、连续的向量表示词
  - 通过“自监督”的方法直接学习词向量
  - 也称词嵌入（Word Embedding）
- 发展历程



前馈神经网络语言模型（FF-NNLM）

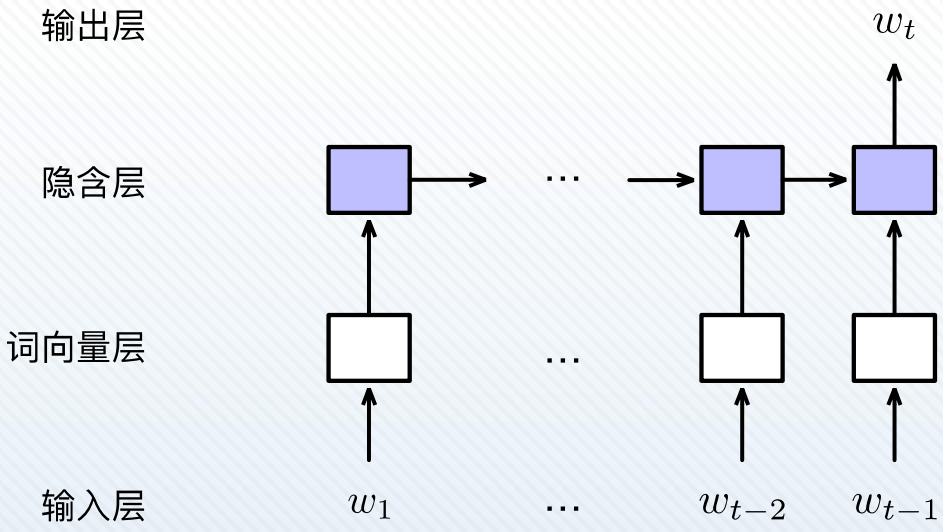
- Neural Network Language Models (Bengio et al., JMLR 2003)
  - 根据前n-1个词（历史）预测当前词，即马尔可夫假设
  - 模型结构为前馈神经网络
  - 通过查找表（Look-up Table），获得词的向量表示
    - 词向量（或词嵌入，Word Embedding）
    - 支撑图灵奖的重要工作
  - 通过梯度下降优化词向量表示



- 缺点
  - “历史”长度不可变
    - “他 喜欢 **吃** 苹果”
    - “他 **感冒** 了， 于是 下班 后 去 了 医院”

循环神经网络语言模型（RNNLM）

- Recurrent Neural Network Language Models (Mikolov et al., Interspeech, 2010)
  - 根据完整的“历史”对当前词进行预测
  - 对不定长依赖的建模能力
  - 梯度弥散/爆炸问题
    - 反向传播过程中按长度进行截断
    - 长短时记忆网络（LSTM）
- 缺点
  - “语言模型”约束
    - ：只利用了“历史”信息



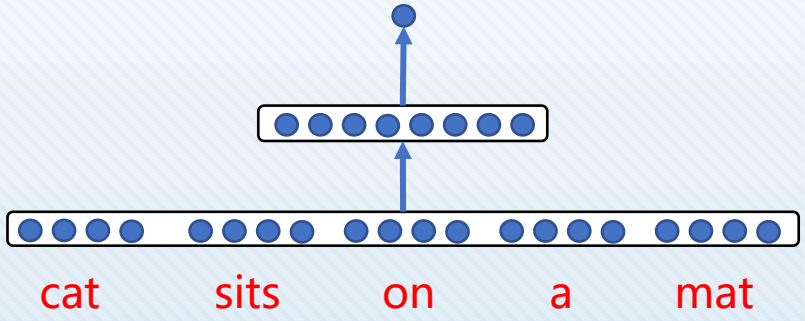
SENNA

- Semantic/syntactic Extraction using a Neural Network Architecture
  - Natural Language Processing (Almost) from Scratch (Collobert et al., JMLR 2011)
- “换词”的思想
  - 一个词和它的上下文构成正例
  - 随机替换掉该词构成负例

✚ cat sits on a mat

▬ cat sits Harbin a mat

- 优化目标
  - score(cat sits on a mat) > score(cat sits Harbin a mat)
  - score的计算方式



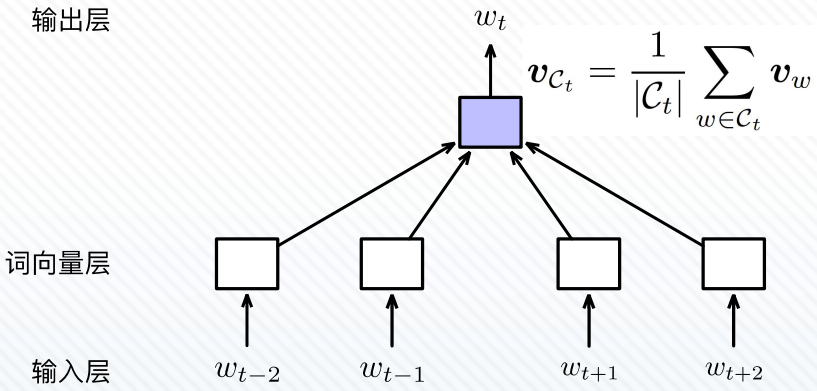
- 训练速度慢，在当年的硬件条件下需要训练**1个月**

目 录

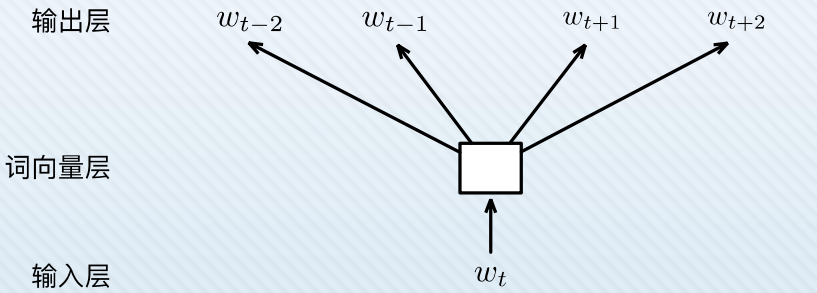
CONTENTS

Word2vec

- https://code.google.com/archive/p/word2vec/
  - Mikolov et al., ICLR 2013
- CBOW (Continuous Bag-of-Word)
  - 根据周围词（上下文）预测中间词
  - 如何计算上下文表示：词向量取平均
- Skip-Gram
  - 根据中间词**独立地**预测周围词（上下文）
- 训练速度快
  - 可利用大规模数据
  - 弥补了模型能力的不足



CBOW模型示意图



Skip-gram模型示意图

参数估计

- 优化目标
  - 对目标词进行预测（Softmax），优化分类损失
    - CBOW

$$P(w_t|\mathcal{C}_t) = \frac{\exp(\boldsymbol{v}_{\mathcal{C}_t} \cdot \boldsymbol{v}'_{w_t})}{\sum_{w' \in \mathbb{V}} \exp(\boldsymbol{v}_{\mathcal{C}_t} \cdot \boldsymbol{v}'_{w'})}$$

    - Skip-Gram

$$P(c|w_t) = \frac{\exp(\boldsymbol{v}_{w_t} \cdot \boldsymbol{v}'_{c})}{\sum_{w' \in \mathbb{V}} \exp(\boldsymbol{v}_{w_t} \cdot \boldsymbol{v}'_{w'})}$$

    - 缺点：当词表较大且计算资源有限时，概率（归一化）计算效率较低
  - 负采样（Negative Sampling）
    - 对（词，上下文）进行二元分类，1表示在给定上下文内**共现**，0表示**不共现**
    - 与SENNA思想近似，通过“换词”构造（词，上下文）负例

$$\log \sigma(\boldsymbol{v}_{w_t} \cdot \boldsymbol{v}'_{w_{t+j}}) + \sum_{i=1}^{K} \log \sigma(-\boldsymbol{v}_{w_t} \cdot \boldsymbol{v}'_{\tilde{w}_i})$$

$\tilde{w}_i \sim P_n(w)$（负采样分布）

注意：词与上下文分别使用不同的向量矩阵

目 录

CONTENTS

GloVe

- GloVe: Global Vectors for Word Representation (Pennington et al., EMNLP 2014)
  - 利用“词-上下文”共现信息
    - Word2vec：局部共现，只考虑当前样本中是否共现
    - GloVe：利用全局统计信息，即共现频次
  - 利用词向量对“词-上下文”共现矩阵进行预测（或回归）
    - 构建共现矩阵：共现“强度”按照距离进行衰减

$$M_{w,c} = \sum_i \frac{1}{d_i(w,c)}$$

    - 回归目标：$\boldsymbol{v}_w^\top \boldsymbol{v}'_c + b_w + b'_c = \log M_{w,c}$
  - 参数估计

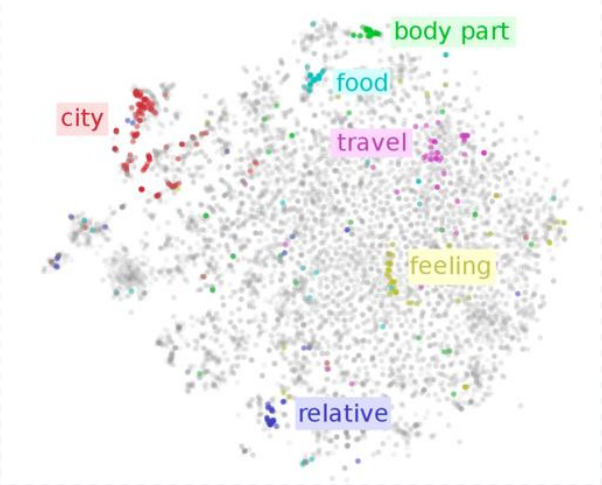
$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{M}) = \sum_{(w,c)\in\mathbb{D}} f(M_{w,c})(\boldsymbol{v}_w^\top \boldsymbol{v}'_c + b_w + b'_c - \log M_{w,c})^2$$

样本权重

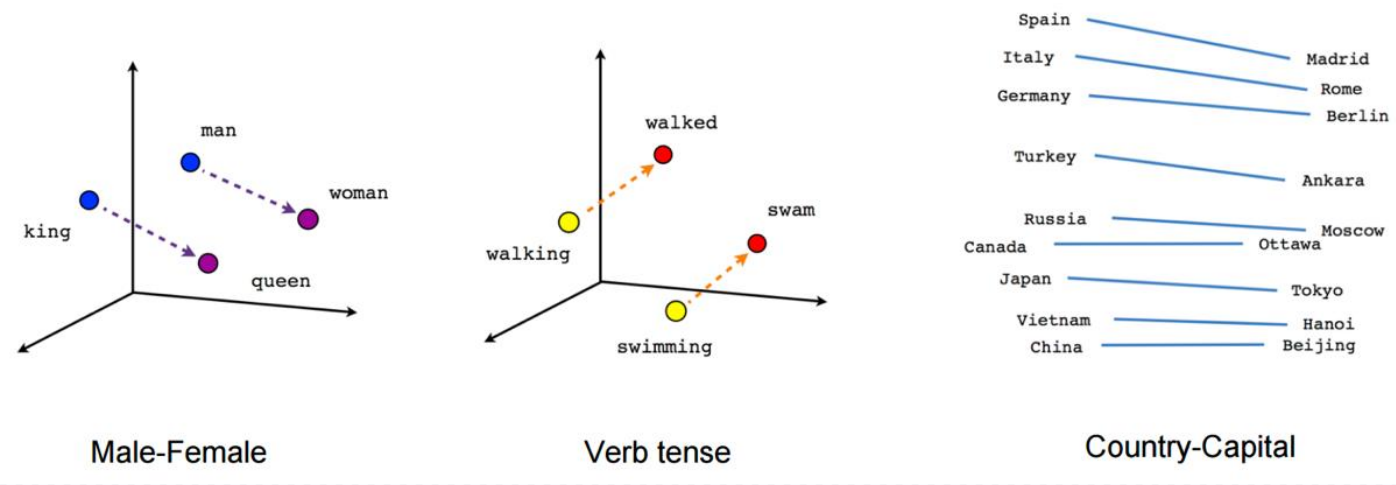
目 录

CONTENTS

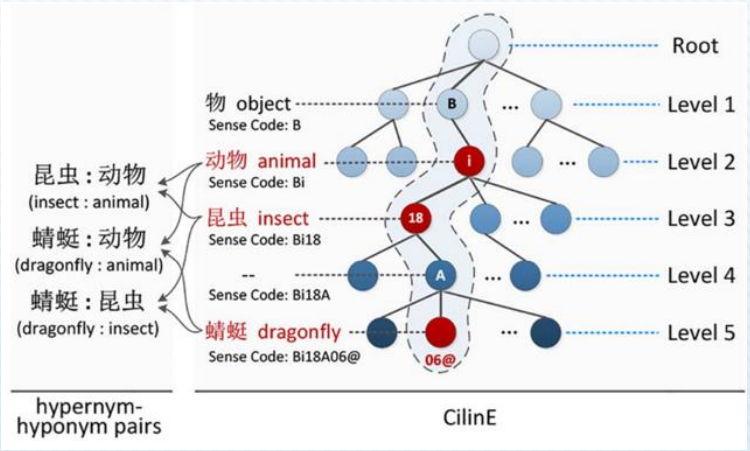
词向量的评价与应用



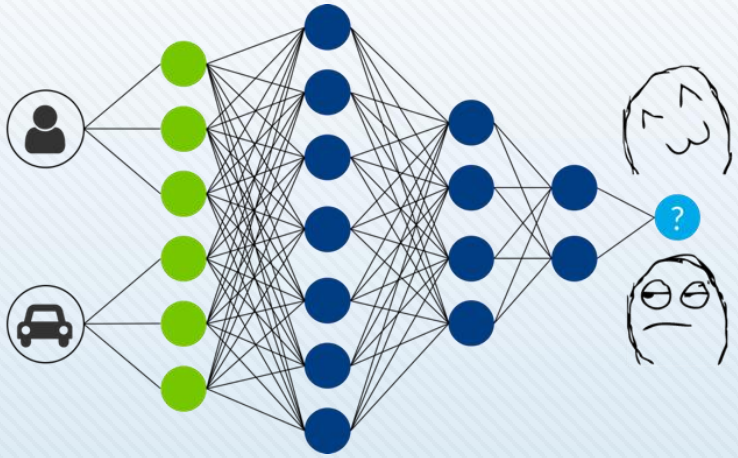
词义相似度计算



词类比关系计算



知识图谱补全

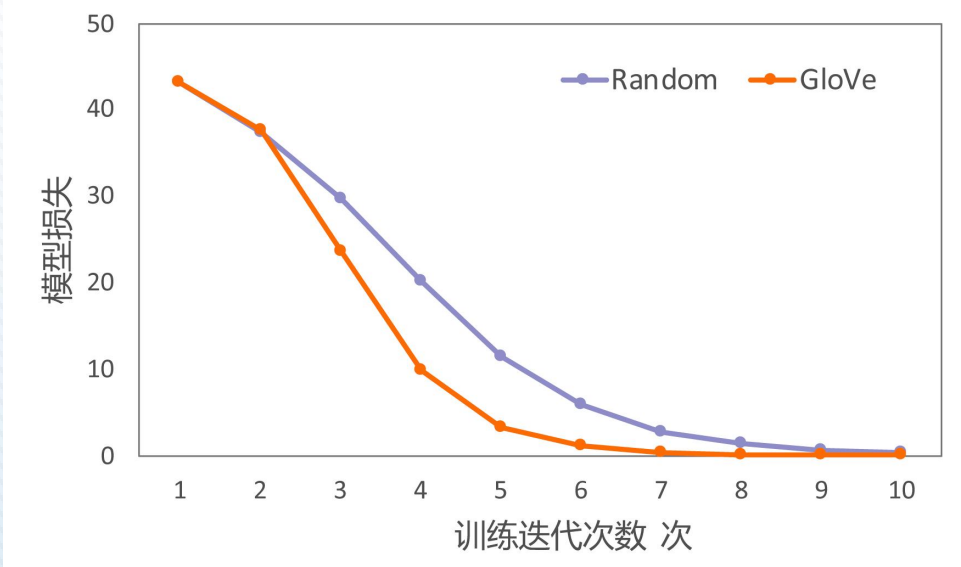


推荐系统

词向量的评价与应用

- 作为词的特征用于下游任务（如文本分类、NER）的输入
- 作为词表示在下游任务训练过程中进行学习（精调）

- 实验设置
  - 数据来源：NLTK sentence_polarity
  - 数据大小：正负各1000个样本



一词多义现象

- 静态词向量假设一个词由**唯一**的词向量表示
  - 无法处理一词多义现象

我 喜欢 吃 土豆
我 刚刚 在 土豆 看 视频

“上下文”

如何获得词义信息？

我 喜欢 吃 土豆#2
我 刚刚 在 土豆#1 看 视频

谢 谢!

理解语言，认知社会

以中文技术，助民族复兴

长按二维码，关注哈工大SCIR

微信号：HIT_SCIR