文本表征学习

5. Word2vec的改进方法

EE1513

宋彦

主要内容

- 改进Word2vec需要解决的问题
- 使用不同信息增强Word2vec
  - 单词内部的信息
  - 上下文中的信息
  - 其他资源中的信息
- 改进方法增强Word2vec
  - 增强学习的过程
  - 后处理
- 将Word2vec扩展到其他任务

Word2vec的特点

- Word2vec使用基于上下文的语言建模
  - 实际上并不不是语言模型
- Word2vec在没有语义指导的情况下进行学习
- 在其实现中受到限制，不容易扩展

Word2vec的问题

- 学习过程缺乏监督
  - 当前的word2vec（以及GloVe）是无监督的
  - 因此可以提供指导来增强学习过程
- 缺乏深度信息整合
  - 缺乏高层次的结构信息
  - 许多其他不同的属性可以帮助学习
- 灵活的使用方式
  - 当前的模型只进行嵌入学习
  - 可以进行许多可能的扩展

- 可能的增强方法
  - 词典？
  - 手动标注？
- 可能的增强方法
  - 结构化知识？
  - sub-word信息？
- 可能的增强方法
  - 利用单词向量结果
  - 扩展到其他任务

真的莽夫

使用词内部信息提升Word2vec

动机

- 在神经网络语言模型中，我们需要使用特殊标记“[UNK]”来处理未知单词。
- 单词具有内部结构，许多单词由sub-word组成。
  - 例如，"Motivation" = "motivate" + "tion"
  - "高大上" = "高" + "大" + "上"
  - "林" = "木" + "木"
  - ...
- 我们可以利用sub-word信息来增强未知单词和稀有单词的嵌入。

Sub-word信息

特殊性->普遍性

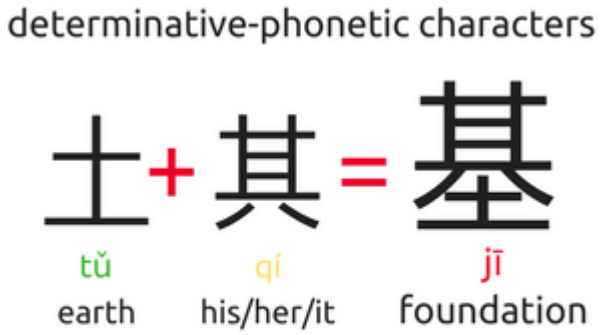
- Luong, et al. (2013)

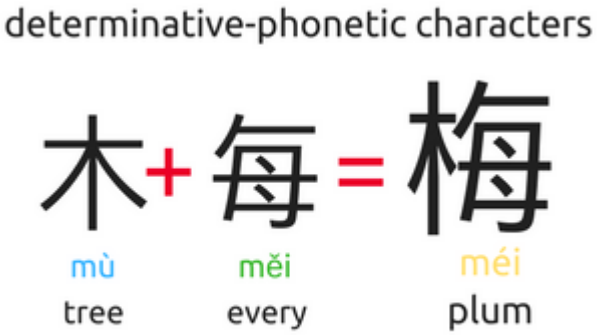
- 使用词缀和词干对单词进行分解
  - 单词通过形态分割工具进行分割，然后应用后处理以获取前缀、后缀和词干
  - 例如，un + fortunate + ly = unfortunately
- 两层结构
  - 语素级别的RNN
  - 基于单词的语言模型

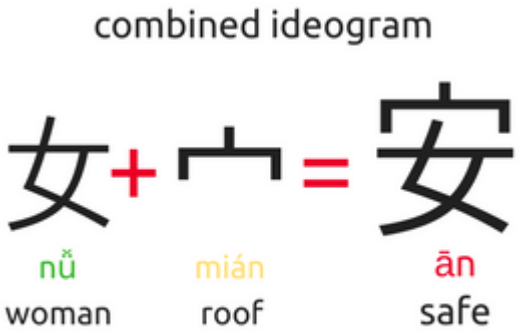
Luong, et al., Better Word Representations with Recursive Neural Networks for Morphology, CoNLL-2013

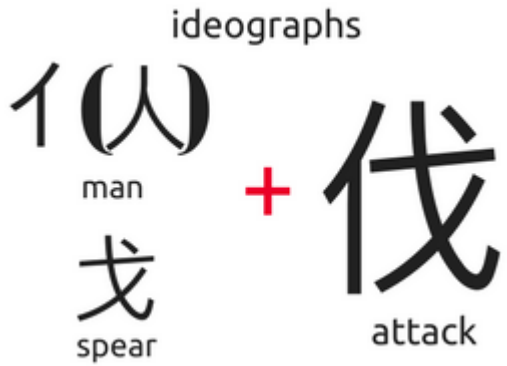
中文的sub-word信息

- Song, et al. (2018)
- 汉字可以被分解为更小的语义单位，单词也是如此。
- 这种分解过程是一种信息丰富的过程。
- 我们可以学习完整的分解链中的单词、字符和部首。



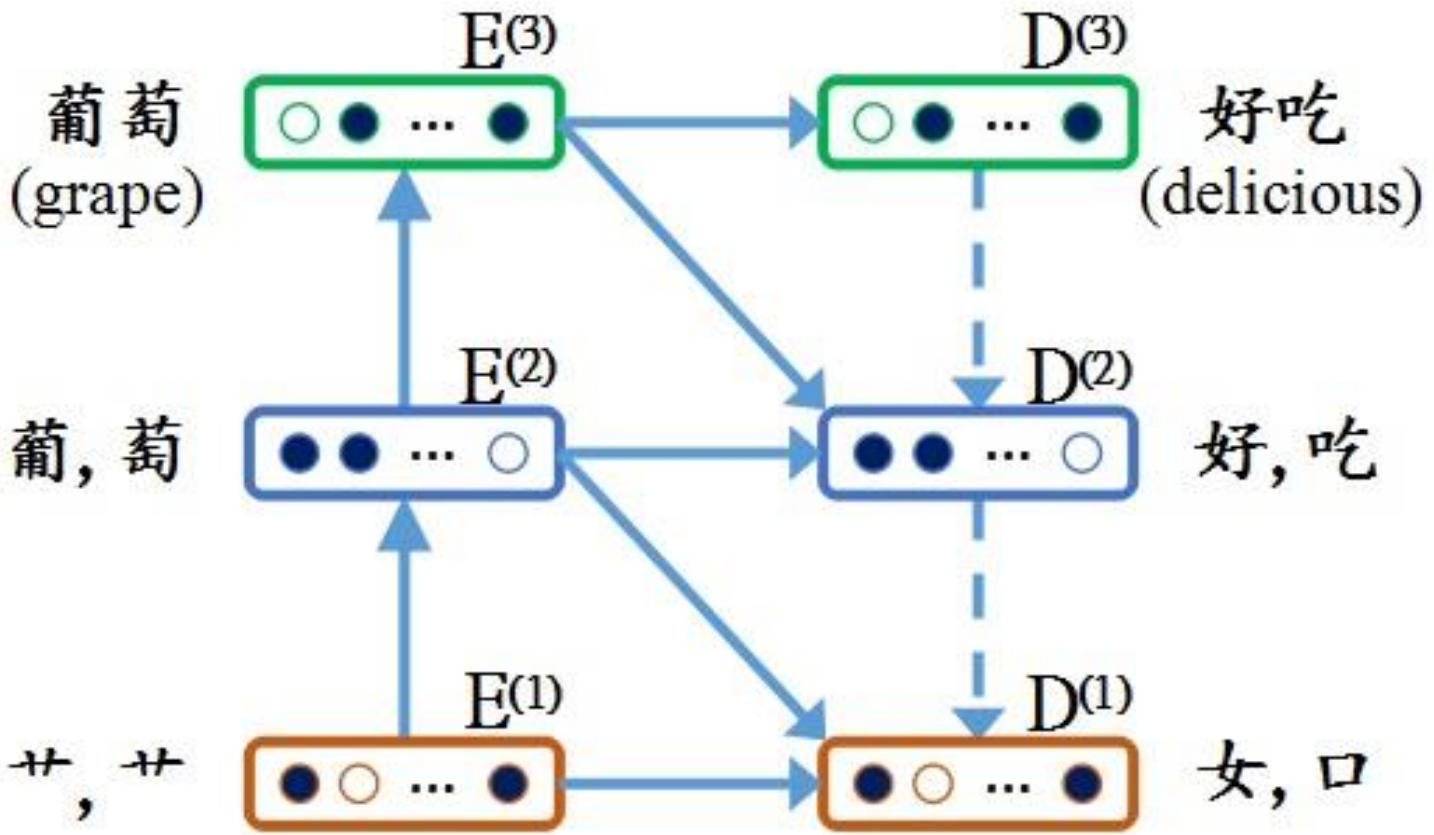






Song, et al., Joint Learning Embeddings for Chinese Words and their Components via Ladder Structured Networks, IJCAI-2018

架构



- 汉字的组成部分通过梯级连接形成了一个语义派生过程。

结果

	WS-240	WS-296
CBOW	19.22	10.94
SG	19.58	11.73
CWE	19.29	10.71
SCWE	18.88	11.03
SCWE+M	20.23	10.82
MGE	18.55	9.75
LSN (W+R)	17.79	10.30
LSN (W+C)	29.38	13.90
LSN (W+C+R)	**34.23**	**18.23**

使用小数据集训练的结果

	WS-240	WS-296
CBOW	51.25	53.82
SG	51.91	54.05
CWE	51.75	53.64
SCWE	52.11	54.20
SCWE+M	52.85	55.26
MGE	53.13	53.33
LSN (W+R)	52.01	53.44
LSN (W+C)	53.47	55.58
LSN (W+C+R)	**54.14**	**57.04**

使用全部数据训练

总结：词内部信息增强词向量

- 从内往外的方向来看
  - 利用单词中的内在知识
- 特别适用于英语以外的一些语言
  - 当组成部分具有语义意义时
- 在训练数据有限的情况下有效
  - 冷启动场景下的一个很好的解决方案
- 对于形态分析是一种有用的方式
  - 一个可能的帮助语言学家的方法

使用上下文信息增强word2vec

动机

- 在CBOW和SG中，目标词和上下文词是从一个固定的窗口中选择的，忽略了句子的结构
- 在CBOW和SG中，忽略了上下文词的词序
- 在CBOW和SG中，所有上下文词都被平等对待

- 我们可以通过使用句法结构信息、词序信息以及有选择地利用不同的上下文词来改进CBOW和SG。

结构知识

- Levy and Goldberg (2014)

- 从依赖关系边学习
  - 通过依赖关系边连接的单词被视为上下文
- 自动地结合句法和谓词-论元关系
- 使用标准的SG算法来学习单词嵌入

Levy, et al., Dependency-Based Word Embeddings, ACL-2014

结构知识

- Komninos and Manandhar (2016)
- 在 Levy and Golberg's (2014) 方法的基础上提出拓展的依存 Skip-gram (EXT)

- 结合窗口选择的上下文词和依赖关系中选择的上下文词，从两种类型的上下文中获益
- 将依赖关系类型添加为上下文：
  - 对于每个节点，依赖关系上下文成为目标，并且预测该节点的其他依赖关系上下文

Alexandros Komninos and Suresh Manandhar. Dependency Based Embeddings for Sentence Classification Tasks. NAACL-2016

词序信息

- Ling et al. (2015a)

- 考虑结构信息（单词顺序）
  - CWindow
  - 结构化Skip-gram（SSG）
- 扩大的投影层
- 有助于解决句法问题

Ling, et al., Two/Too Simple Adaptations of Word2Vec for Syntax Problems, NAACL-2015

带方向的Skip-Gram (DSG)

- Song et al. (2018)

- 现有的CWindow和Structured SG过于复杂。
- 我们只需要单词顺序的信息。
- 学习单词顺序与上下文一起。
- 使用类似负采样的算法。

Song et al. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. NAACL-2018

带方向的Skip-Gram（DSG）

- 区分上下文词是在目标词的左侧还是右侧是很重要的。
  - 例如，在“merry Christmas”和“Christmas eve”中，“merry”和“eve”都经常与“Christmas”同时出现。
  - 但是，“merry”倾向于出现在“Christmas”的左边，而“eve”倾向于出现在“Christmas”的右边。
- 我们提出了一个softmax函数，通过引入一个新的向量δ来衡量上下文词w_{t+i}与w_t在其左侧或右侧上的关联。

$$g(w_{t+i}, w_t) = \frac{exp({\delta_{w_{t+i}}}^\top v_{w_t})}{\sum_{w_{t+i} \in V} exp({\delta_{w_{t+i}}}^\top v_{w_t})}$$

Song et al. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. NAACL-2018

带方向的Skip-Gram（DSG）

$$g(w_{t+i}, w_t) = \frac{exp({\delta_{w_{t+i}}}^\top \upsilon_{w_t})}{\sum_{w_{t+i} \in V} exp({\delta_{w_{t+i}}}^\top \upsilon_{w_t})}$$

- 函数 g 的更新方式类似于负采样的更新范式。

$$\upsilon_{w_t}^{(new)} = \upsilon_{w_t}^{(old)} - \gamma(\sigma(\upsilon_{w_t}^\top \delta_{w_{t+i}}) - \mathcal{D})\delta_{w_{t+i}}$$
$$\delta_{w_{t+i}}^{(new)} = \delta_{w_{t+i}}^{(old)} - \gamma(\sigma(\upsilon_{w_t}^\top \delta_{w_{t+i}}) - \mathcal{D})\upsilon_{w_t}$$

$$\mathcal{D} = \begin{cases} 1 & i < 0 \\ 0 & i > 0 \end{cases}$$

- σ: sigmoid; γ: 学习率; D: 表明w_{t+i}与w_t的相对位置标签
- 最终模型是 g 加上原始的 SG 目标 $p(w_{t+i}|w_t)$

Song et al. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. NAACL-2018

不同模型的比较

- 结果

	MEN-3k	SimLex-999	WS-353
CBOW	70.96	34.32	69.25
CWin	**74.28**	36.06	72.21
SG	71.90	34.35	70.11
SSG	71.26	31.80	69.46
SSSG	72.07	33.62	70.90
DSG	73.76	**36.10**	**72.60**

Table 3: Word similarity results ($\rho \times 100$) from embeddings trained on the large corpus.

	MEN-3k	SimLex-999	WS-353
CBOW	58.23	26.67	64.40
CWin	59.68	25.19	62.82
SG	60.19	27.14	65.23
SSG	55.42	24.00	61.95
SSSG	62.70	26.55	66.10
DSG	**63.18**	**27.51**	**66.71**

Table 4: Word similarity results ($\rho \times 100$) from embeddings trained on the small corpus.

Song et al. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. NAACL-2018

不同模型的比较

- 结果

	WSJ		ARK	
	Dev	Test	Dev	Test
CBOW	96.86	97.01	89.36	88.36
CWin	96.98	97.25	90.03	89.94
SG	96.95	97.12	89.26	88.77
SSG	97.08	97.31	90.05	90.15
SSSG	97.01	97.19	89.83	89.78
DSG	**97.16**	**97.37**	**90.12**	**90.43**

Table 5: POS tagging results (accuracy) on WSJ and ARK datasets.

Song et al. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. NAACL-2018

DSG

- 中文上的结果
  - Tencent Embedding
  - 包括从大规模互联网数据（新闻、小说、百科全书等）中提取的800万个词的词库。
  - 在覆盖范围、时效性和准确性方面，显著超过了现有的中文词嵌入。
  - 它在2018年被评为世界前十的人工智能开源资源之一。

https://ai.tencent.com/ailab/nlp/en/embedding.html

Song et al. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. NAACL-2018

DSG

- 例子

Input	新冠病毒	煮酒论英雄	流金岁月	刘德华	自然语言处理
Top similar words	新冠肺炎病毒 新型冠状病毒 新冠状病毒 肺炎病毒 covid-19病毒 新冠 新型病毒 冠状病毒	青梅煮酒论英雄 曹操煮酒论英雄 青梅煮酒 关羽温酒斩华雄 桃园三结义 温酒斩华雄 三英战吕布 桃园结义	半生缘 大江大河2 你迟到的许多年 风再起时 情深缘起 外滩钟声 亲爱的自己 了不起的女孩	华仔 张学友 张国荣 梁朝伟 谭咏麟 周润发 刘天王 古天乐	自然语言理解 计算机视觉 自然语言处理技术 nlp 机器学习 语义理解 深度学习 nlp技术

https://ai.tencent.com/ailab/nlp/en/embedding.html

Song et al. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. NAACL-2018

总结：利用上下文信息

- 精确建模单词之间的关联关系
- 从句法信息中学习是一种解决方案，但代价较高
- 词序可能是在文本中除了引入外部知识之外唯一可以获得的指导
  - 单词的相对位置很重要
  - 虽然词序知识提供的信息有限但仍然有效
- 不同的上下文单词对目标单词的表示具有明显的贡献。我们需要仔细建模。

使用其他资源提升word2vec

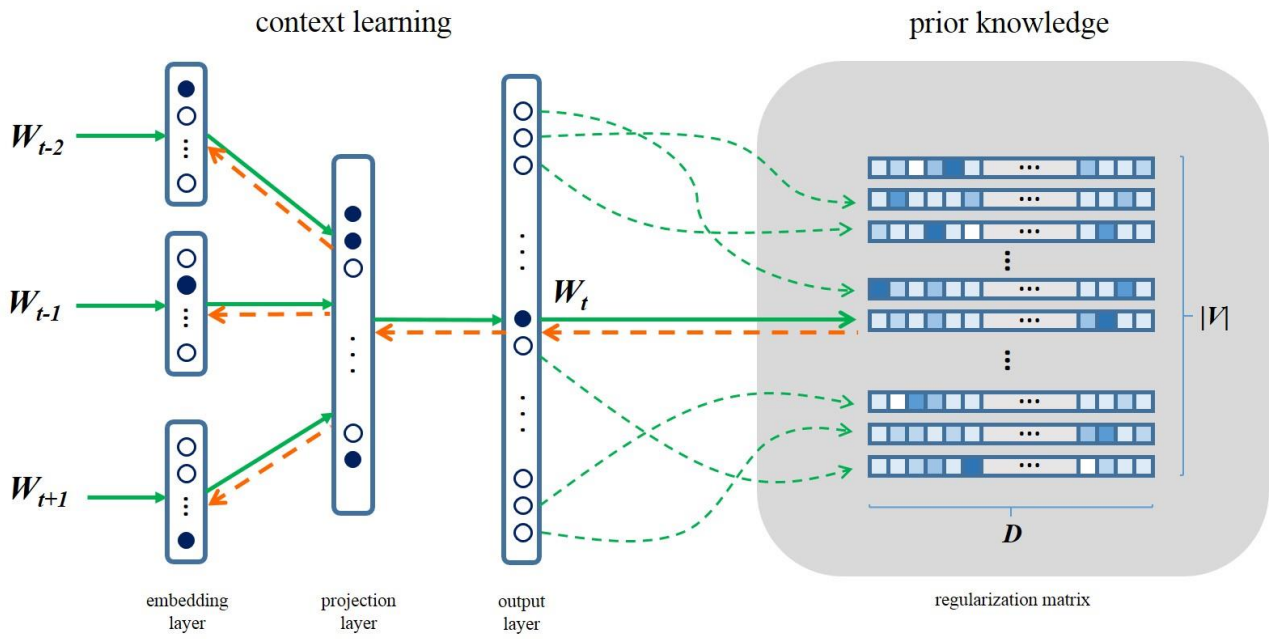
动机

- 为了说明单词的意义以及它们之间的关系（例如，同义词和反义词），人们创建了许多资源。
- 我们可以利用这些资源来改进词嵌入。
- 在我们没有太多数据的情况下，这将更加有帮助。

词向量与知识的组合

- Song et al. (2017)

- 上下文 + 外部知识
- 以预测的方式
- 适用于CBOW和SG
- 结合
  - 监督知识
  - 无监督知识

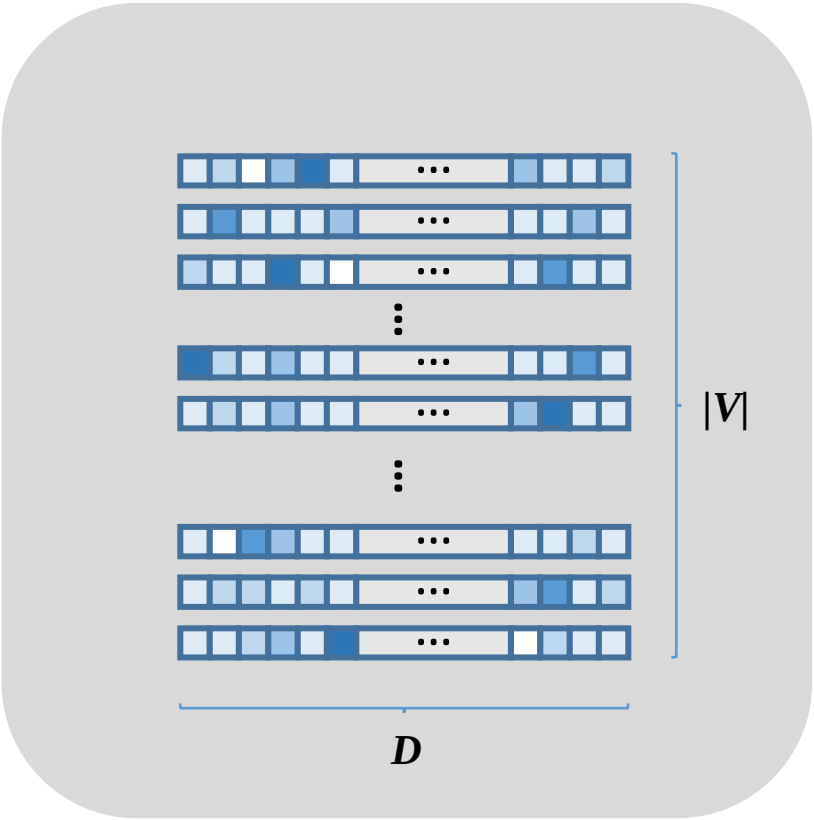


Song, et al., Learning Word Representations with Regularization from Prior Knowledge, CoNLL-2017

主要的想法

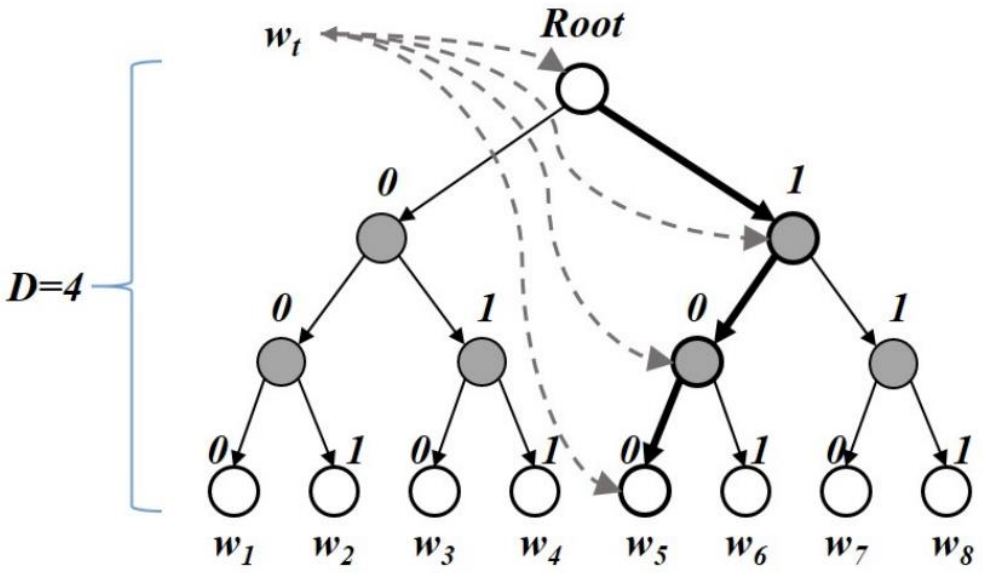
- 一种更好的学习和增强词嵌入的方法是结合以下因素：
  - 上下文
  - 外部知识
- 使用一个正则化器的通用框架，可以考虑以下内容：
  - 标注的知识，如词典、词库
  - 未标注的知识，如自动聚类的单词

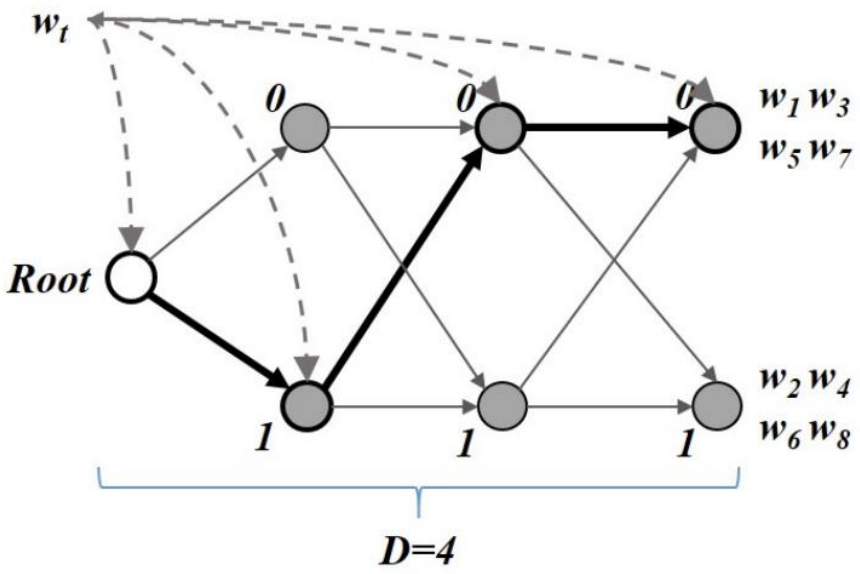
正则化



- 正则化是在表示外部知识的矩阵上进行的，其中：
  - 每个单词在该矩阵中由一个向量（行）表示
  - 每个向量都被归一化为相同的长度

降低时间复杂度





将时间复杂度降低到 $O(2 \cdot D \cdot d)$

Song, et al., Learning Word Representations with Regularization from Prior Knowledge, CoNLL-2017

总结：使用外部的语义指导

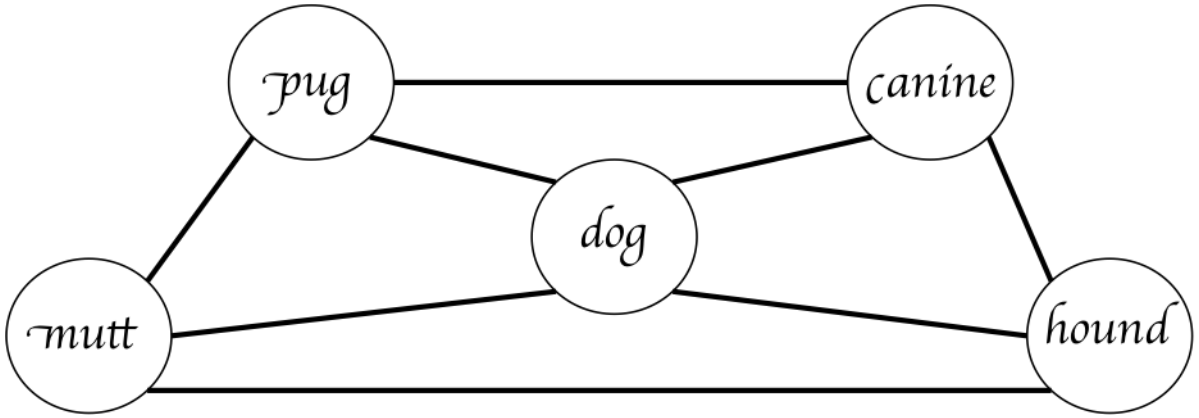
- 将上下文学习与人类知识相结合
- 有几种方法可以添加此类知识
  - 在学习词向量后（后期改进）
  - 联合学习
    - 从X方面
    - 从Y方面
- 知识的质量很重要

使用后处理提升词向量

词向量改造

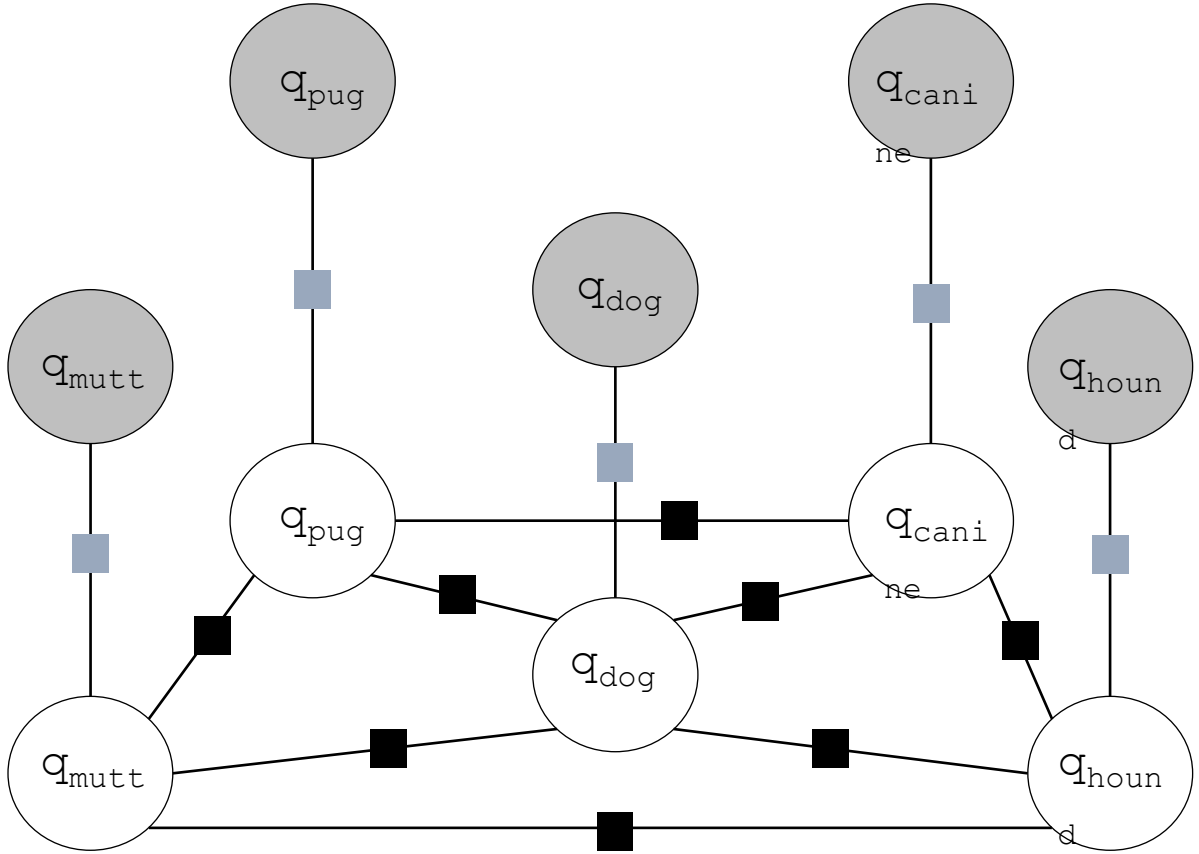
- Faruqui et al. (2015)

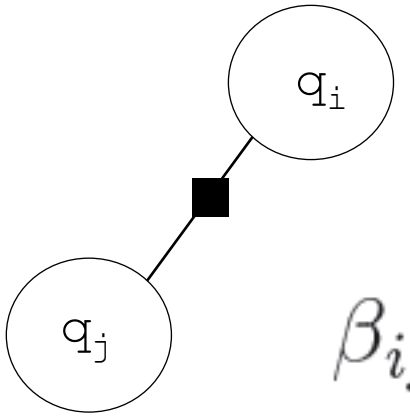
- 后处理词向量
- 根据语义词典进行限制
- 在线更新



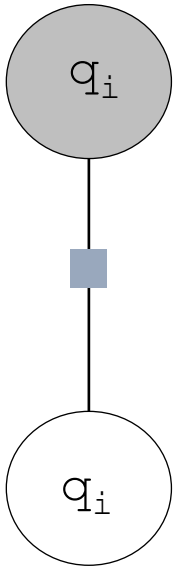
Faruqui, et al., Retrofitting Word Vectors to Semantic Lexicons, NAACL-2015

词向量改造





$$\beta_{ij}\|q_i - q_j\|^2$$



Euclidean Distance!

$$\alpha\|q_i - \hat{q}_i\|^2$$

使用新架构提升 word2vec

CBOW和SG的互补学习

- CBOW和SG实际上是互补的，它们从不同的角度解决同一个问题，而一个任务的输出可以作为另一个任务的输入。
- 我们提出了一种增强的架构，以一种加强的方式同时考虑这两种范式作为互补任务。
  - 我们从SG模型中获取输出，然后将其与现有上下文组合形成CB模型的输入。
  - 相应模型的参数根据我们设计的奖励函数的期望进行更新。
  - 我们进一步提出了一种直接的抽样策略，旨在确保将额外的信息引入到学习过程中。

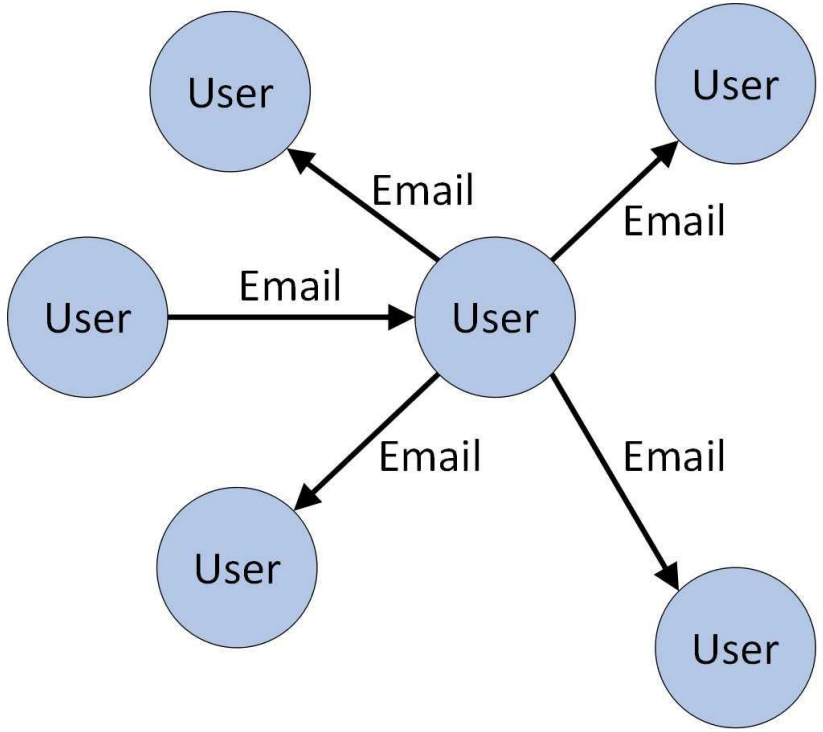
Yan Song, and Shuming Shi. Complementary Learning of Word Embeddings. IJCAI-2018.

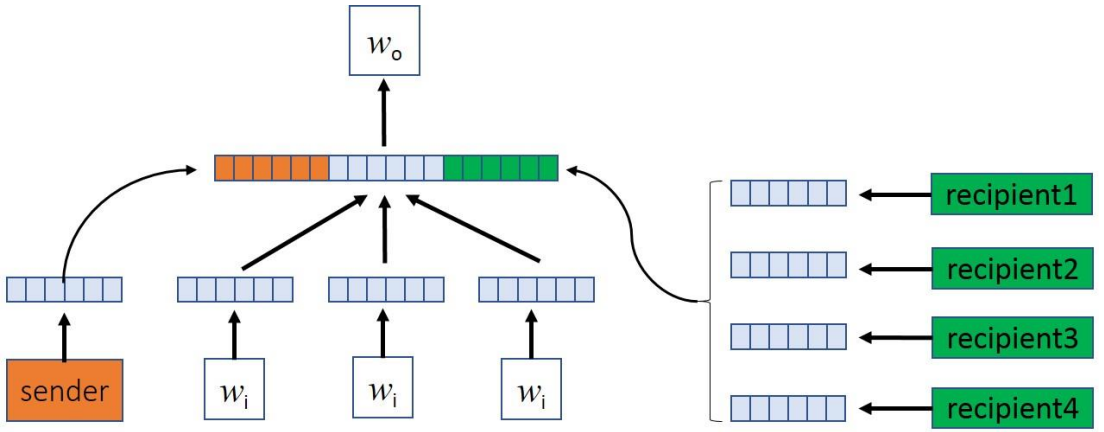
拓展word2vec到其他任务

用户向量

发送者、接受者、权重

- Song and Lee (2017)





$$y = Xh(w_i, ..., w_{i+n};\ W, s, r_1, ..., r_m; U) + b$$

$$h = v_s \oplus \sum_{j=i}^{i+n} v_j \oplus \frac{1}{m} \sum_{r=1}^{m} v_r$$

Song and Lee, Learning User Embeddings from Emails, EACL-2017

拓展

- 应用类似于word2vec的模型有许多可能的选择。
  - 许多数据或任务具有类似语言建模的特性。
- 嵌入不仅是一个结果，还可以作为学习其他语言单位的标签。
- 短语/句子对于自然语言理解很重要。
- 在过去几年中，基于（为）任务进行学习是一个趋势。