Adline125's Blog

NLP Engineer, Google Developers Expert

Evaluation methods for unsupervised word embeddings

2020-07-04 | 2020-07-05 | NLP

本文是对论文Evaluation methods for unsupervised word embeddings的总结。相较于大量生成词嵌入模型的研究，评估词嵌入模型的工作相对较少。 该论文是第一篇对词嵌入评估进行深入研究的论文，发表于2015年，涵盖了广泛的评估标准和当时流行的嵌入技术。其目的并非是证明某个词嵌入方法优于其他方法，而是要对词嵌入的评估方法本身做较深入的探讨。

本文的主要内容包括：

- 评估方法概述
- 不同评估标准下模型表现比较
- 词嵌入中的词频信息
- 总结

评估方案概述

现有词嵌入评估方案分为两大类：外部评估（extrinsic evaluation）和内部评估（intrinsic evaluation）。

外部评估：词向量被用作下游任务的输入特征，好的词向量应该对使用其的任务有正面影响。但是不同的任务对词向量的偏好可能是不同的，因此外部评估是否有效一直是一个存在争论。

内部评估：直接测试单词之间的句法或语义关系。内部评估通常涉及一组预先选择的查询词和与语义相关的目标词，称为查询清单。文章中内部评估进一步分成两类：*绝对内在评估*和*比较内在*

评估。其中比较内在评估由文章提出，并应用在相关性评估和一致性评估中。

不同评估标准下模型表现比较

论文从单词相关性，一致性，下游表现三个不同评估标准入手，分析了三种标准下各词嵌入模型排序结果之间的关系。

相关性评估

绝对内在评估

绝对内在评估使用通常用作嵌入方法基准的***14***个数据集（如Table1），对***6***个词嵌入模型从如下***4***个范畴进行评估：

- 相关性(relatedness): 两个单词的词嵌入的余弦相似度应与人类相关性得分具有高度相关性
- 类比性(analogy): 对给定的单词y, 能否找到一个对应的单词x, 使得x与y的关系能够类比另外两个已知词a与b的关系
- 类别化(categorization): 对词向量进行聚类, 看每个簇与有标记数据集各类相比纯度如何
- 选择倾向(selectinal preference): 判断某名词是更倾向做某个动词的主语还是宾语, 例如一般顺序是 he runs 而不是 runs he

评估结果如下：

	relatedness						categorization			sel. prefs		analogy			
	rg	ws	wss	wsr	men	toefl	ap	esslli	batt.	up	mcrae	an	ansyn	ansem	average
CBOW	**74.0**	**64.0**	**71.5**	**56.5**	**70.7**	66.7	**65.9**	**70.5**	**85.2**	24.1	13.9	**52.2**	**47.8**	**57.6**	**58.6**
GloVe	63.7	54.8	65.8	49.6	64.6	**69.4**	64.1	65.9	77.8	27.0	**18.4**	42.2	44.2	39.7	53.4
TSCCA	57.8	54.4	64.7	43.3	56.7	58.3	57.5	**70.5**	64.2	**31.0**	14.4	15.5	19.0	11.1	44.2
C&W	48.1	49.8	60.7	40.1	57.5	66.7	60.6	61.4	80.2	28.3	16.0	10.9	12.2	9.3	43.0
H-PCA	19.8	32.9	43.6	15.1	21.3	54.2	34.1	50.0	42.0	-2.5	3.2	3.0	2.4	3.7	23.1
Rand. Proj.	17.1	19.5	24.9	16.1	11.3	51.4	21.9	38.6	29.6	-8.5	1.2	1.0	0.3	1.9	16.2

Table 1: Results on absolute intrinsic evaluation. The best result for each dataset is highlighted in bold. The second row contains the names of the corresponding datasets.

绝对内在评估在设计上有两点缺陷：

- 查询清单（query inventory）：上述的14个数据集在设计上都没能考虑以下几个方面，
  - the frequency of the words in the English language
  - the parts of speech of the words
  - abstractness vs. concreteness
- 度量指标（metric aggregation）：我们找不到一种对完全无关的单词对进行排序的方式（metric）。例如，我们如何确定（狗，猫）是否比（香蕉，苹果）更相似

比较内在评估

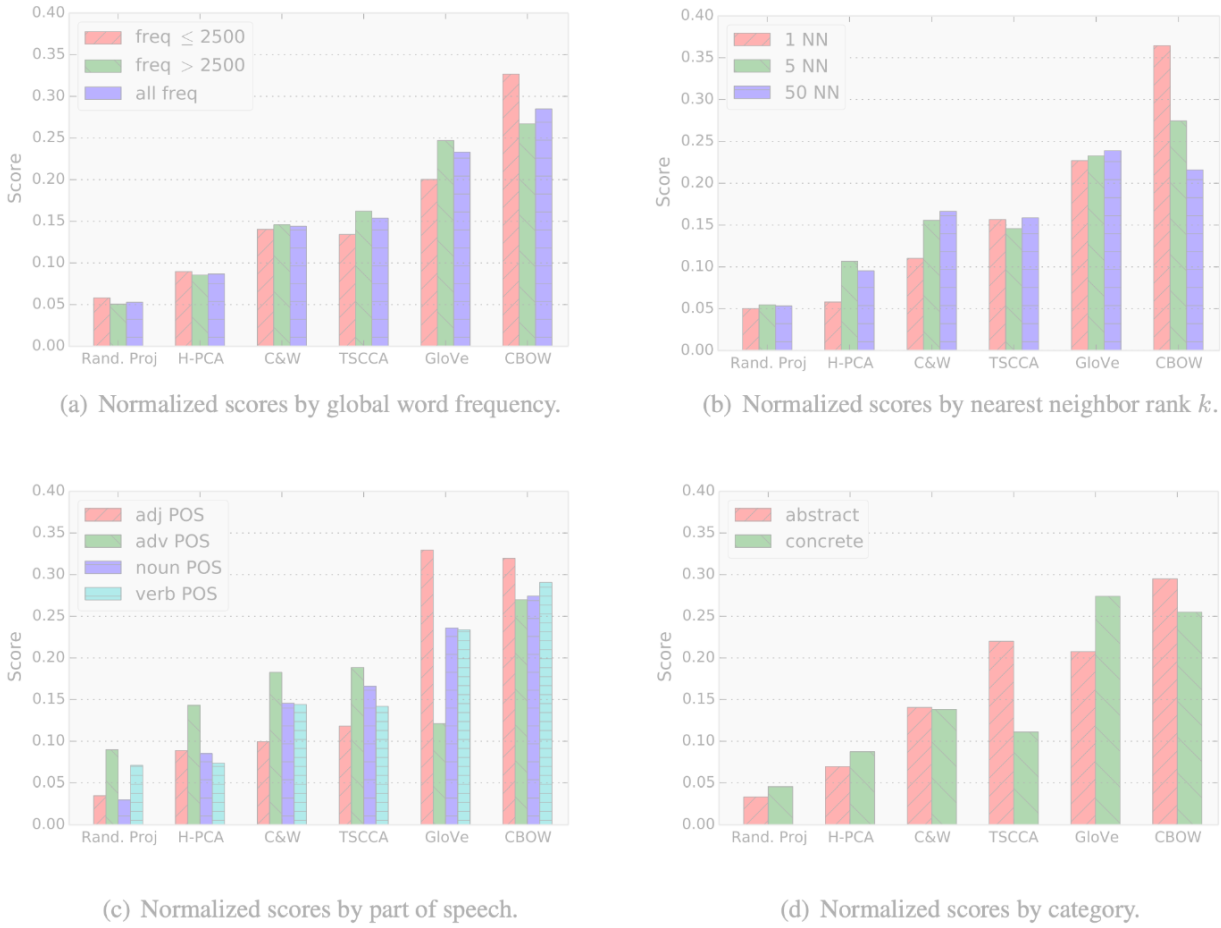
Query:	**skillfully**			
	(a)	swiftly	(b)	expertly
	(c)	cleverly	(d)	pointedly

Table 2: Example instance of comparative intrinsic evaluation task. The presented options in this example are nearest neighbors to the query word according to (a) C&W, (b) CBOW, GloVe, TSCCA (c) Rand. Proj. and (d) H-PCA.

针对绝对内在评估存在的缺陷，文章提出了*比较内在评估*方法。即，给出一个查询词，将6个词嵌入模型产生的结果呈现给用户，让用户选出最相关的，然后统计结果。如Table 2.

- 文章采用用户直接反馈的形式避免了需要定义指标（metric）的问题。
- 文章定义制作了更符合词嵌入评估任务的查询清单。考虑了词频、词性、类别、是否是抽象词四个方面。并对这四个方面分别做了评估。

比较内在评估结果如下：



(a) Normalized scores by global word frequency.

(b) Normalized scores by nearest neighbor rank k.

(c) Normalized scores by part of speech.

(d) Normalized scores by category.

Figure 1: Direct comparison task

一致性评估

(a)	finally	(b)	eventually
(c)	immediately	(d)	put

Table 3: Example instance of intrusion task. The query word is option (a), intruder is (d).

文章中一致性评估直接采用比较内在评估方法。一致性评估是文章提出的评估方法。将查询词本身及词嵌入模型计算出来的两个相近词再加上一个不相关的词组成一道选择题。由测试人选出不相关的词。由此判定词嵌入模型选出的词与查询词是否具有一致性。

一致性评估结果如下：

Figure 2: Intrusion task: average precision by global word frequency

到目前为止，对比$Table1, Figure1$和$Figure2$，6个词嵌入模型从一致性任务获得的排名与从相关性任务获得的排名产生了出入。

外在评估（下游表现）

外在评估评估单词嵌入模型对特定任务的贡献。

使用此类评估存在一个隐含的假设，即单词嵌入质量是有固定排名的。也就是说，嵌入模型无论在什么任务里的表现排名应该是基本一致的。因此，更高质量的嵌入将必定会改善任何下游任务的结果。

但文章发现上述假设不成立：不同的任务倾向于不同的嵌入。

在文章试验的两个任务：名词短语分块和情感分类中证实了上述观点。文章建议对词嵌入进行针对项目的训练，以优化特定目标。

	dev	test	p-value
Baseline	94.18	93.78	0.000
Rand. Proj.	94.33	93.90	0.006
GloVe	94.28	93.93	0.015
H-PCA	94.48	93.96	0.029
C&W	**94.53**	**94.12**	
CBOW	94.32	93.93	0.012
TSCCA	**94.53**	94.09	0.357

Table 4: F1 chunking results using different word embeddings as features. The p-values are with respect to the best performing method.

	test	p-value
BOW (baseline)	88.90	$7.45{\cdot}10^{-14}$
Rand. Proj.	62.95	$7.47{\cdot}10^{-12}$
GloVe	74.87	$5.00{\cdot}10^{-2}$
H-PCA	69.45	$6.06{\cdot}10^{-11}$
C&W	72.37	$1.29{\cdot}10^{-7}$
CBOW	**75.78**	
TSCCA	75.02	$7.28{\cdot}10^{-4}$

Table 5: F1 sentiment analysis results using different word embeddings as features. The p-values are with respect to the best performing embedding.

Embedding中的词频信息

文章用两个小实验证实了embedding中编码了大量的词频信息。两个小实验的情况如下：

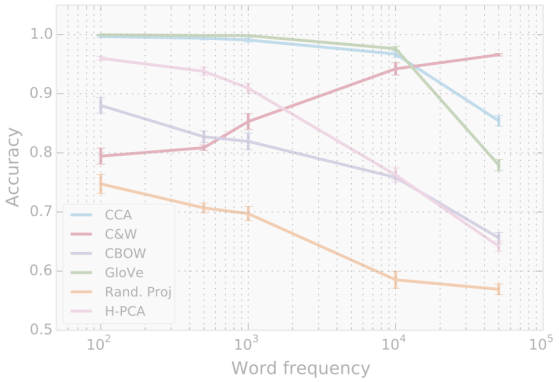


Figure 3: Embeddings can accurately predict whether a word is frequent or rare.

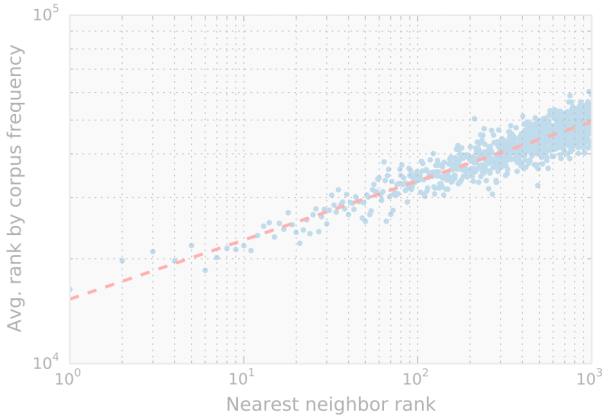


Figure 4: Avg. word rank by frequency in training corpus vs. nearest-neighbor rank in the C&W embedding space.

总结

本论文的主要贡献总结如下：

- 提出了由于使用不同的评估标准会导致嵌入效果排序的不同，对嵌入方法进行比较应在特定任务的上下文中进行，不提倡使用外部评估评测embedding。
- 证实了自动相似性评估与直接人工评估之间存在很强的相关性。说明至少在相似性任务中使用离线数据是合理的。
- 提出了一种模型驱动和数据驱动的方法来构建查询清单。
- 发现了所有词嵌入模型都编码了大量的单词频率信息，对使用余弦相似度作为嵌入空间中的相似度度量的普遍做法提出了质疑

参考文献

[Schnabel2015] Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 298-307). 2015

评估相关其他资料

[Bakarov2018] Bakarov, Amir. "A Survey of Word Embeddings Evaluation Methods." *arXiv preprint arXiv:1801.09536*. 2018

embedding evaluation

‹ WMD论文总结及代码实现: From Word Embeddings To Document Distances

Bidirectional LSTM-CRF Models for Sequence Tagging ›