词向量的评估方法 论文总结

大纲：

论文题目：Evaluation methods for unsupervised word embeddings

时间：2015

论文地址：

https://www.aclweb.org/anthology/D15-1036.pdf
www.aclweb.org

1. 基本介绍

本论文主要内容：词向量的评估方法

现有词嵌入评估方案分为两大类：外在评估（extrinsic evaluation）和内在评估（intrinsic evaluation）。

外在评估：把词向量作为下游任务的输入

内在评估：衡量词之间的句法和语义关系，内部评估进一步分成两类：绝对内在评估（absolute intrinsic evaluation）和比较内在评估（comparative intrinsic evaluation）。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

绝对内在评估直接衡量给定两个单词之间的句法和语义关系。共有四种类型的评价：

（1）相关性(Relatedness)：对于两个单词，他们之间的余弦相似度应该和人类主观评价的得分有较高的相关性。

即评价词向量模型在两个词之间的语义相关性，如：学生与作业，中国与北京等。

具体方法由监督模式实现，首先需要一份如下的标记文件，一般可以由人工标注：

```
学生 上课 0.78
教师 备课 0.8
...
```

上述文件代表了词语之间的语义相关性，我们利用标注文件与训练出来的词向量相似度进行比较，如：词向量之间的cos距离等，确定损失函数，便可以得到一个评价指标。

但这种方法首先需要人力标注，且标注的准确性对评价指标影响非常大。

（2）类比性(Analogy)：假设给了一对单词 (a , b) 和一个单词c, task会找到一个单词d，使得c与d之间的关系相似于a与b之间的关系.

```
queen-king+man=women
```

在给定word embedding的前提下，一般是通过在词向量空间寻找离(b-a+c)最近的词向量来找到d。

（3）分类(Categorization)：把词聚类成不同的堆，看是否聚类准确

（4）选择偏好(Selected 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

评价结果如下,可以看出，绝大多数任务中，CBOW表现最好。但是个别任务里，其他词向量更好

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.

1.2 比较内在评估 comparative intrinsic evaluation

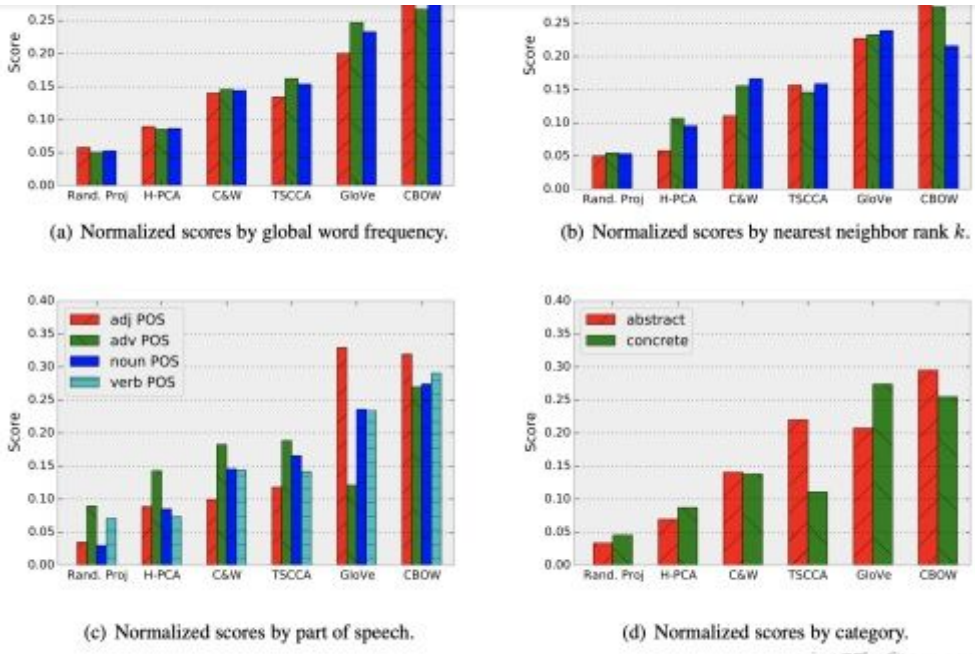
Query:	**skillfully**			
	(a)	swiftly	(b)	expertly
	(c)	cleverly	(d)	pointedly

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.

给出一个查询词query word，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

- 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

同样可以看出，没有一种词向量是在所有任务中都表现最好的

1.3 一致性评估 coherence

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

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

一致性评估是文章提出的评估方法。将查询词本身(a finally)及词嵌入模型计算出来的两个相近词（b eventually c immediately）再加上一个不相关的词（d put）组成一组query。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：

可以看出不同词向量的生成方法，对于不同词频的单词，所得到的结果是不同的

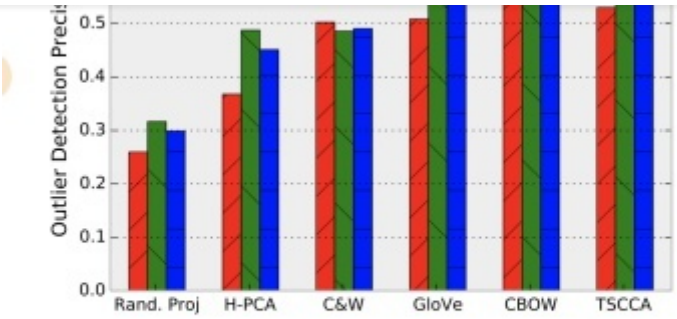


Figure 2: Intrusion task: average precision by global word frequency.

1.4 外在评估（下游表现）

外在评估是指评估单词嵌入模型对特定任务的贡献。 比如词性标注、命名实体识别、句法分析、句子分类等，将词向量作为输入，衡量下游任务指标性能的变化。

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。

本文选取了以下两种下游任务来评判

- Noun phrase chunking：名词分块
- Sentiment classification：情感分类

结果如下，对于下游任务，同样的，没有一种词向量可以在所有下游任务中都表现最好，所以对于不同下游任务，我们应该尝试不同词向量的表示

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.

1.5 Embedding中的词频信息

用两个小实验证实了embedding中编码了大量的词频信息。

结果如下,可以看出，我们可以通过词向量来较好的预测单词的词频，其中GloVe和CCA中包含了较多的词频信息。另外单词的词频于其在语料库里的词频排名也有很强的相关性

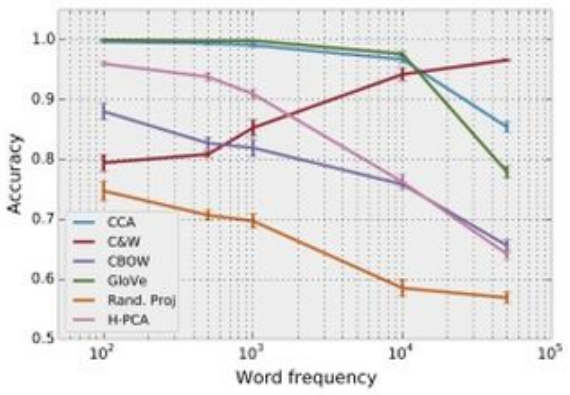


Figure 3: Embeddings can accurately predict whether a word is frequent or rare.

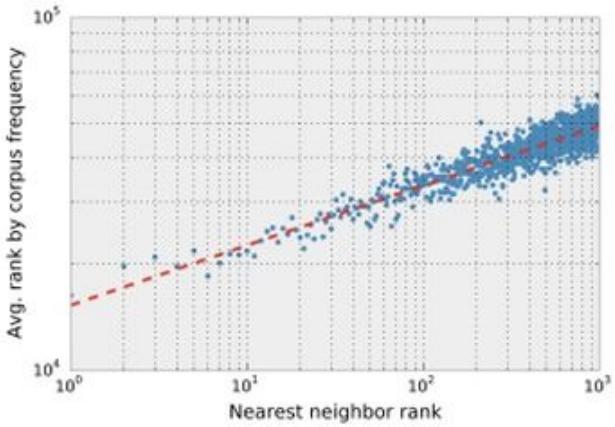


Figure 4: Avg. word rank by frequency in training corpus vs. nearest-neighbor rank in the C&W embedding space.

词向量可以去预测词的词频

词向量的neighbor排序和词频有关