Evaluation methods for unsupervised word embeddings

Evaluation methods for unsupervised word embeddings review

Jul 4, 2020 • Bujie Xu • 1 min read

NLP

本文比较了各种衡量词向量的方法，并提出了一种新的评测词向量的方法。本文主要有以下贡献

- 分析了不同评判标准间的关系，表明了生成词向量的方式要和特殊任务相关联
- 提出了一种通过人为评分方式衡量直接衡量单个词向量的方法
- 提出了选择词向量（用于评价）时要考虑到选择不同词频，词性，词义的向量。保证数据的多样性
- 本文还发现了词向量包含着词频信息

需要注意的是这篇文章的目的不是去比较词向量的好坏，而是去研究评判词向量方法的差别。

1. Embeeding的准备

本文准备了以下六种生成词向量的方式用于评判：

- 基于概率预测的embeeding
  - CBOW model of word2vec (Mikolov et al 2013a)

    - C&W embeddings (Collobert et al. 2011)
- 基于反应语料中的词汇的同现关系
    - Hellinger PCA (Lebret and COllobert 2014)
    - GloVe (Pennington et al., 2014)
    - TSCCA (Dhillon et al., 2012)
    - Sparse Random Projections (Li et al., 2006)

对于C&W的词向，因为只有基于2007年的维基百科的。所以本文选取了2008-03-01日的维基百科来训练其余5中词向量。这里，所有词向量的维度为50，总共的词典大小为103647

2. Evaluation

评价词向量主要有两种方式，一种是内部评价（intrinsic evaluation），另一种是外部评价（extrinsic evaluation）。

内部评价指的是用词的词性，相关性等内部固有关系来评价生成的词向量的好坏。 外部评价指的是用生成的词向量去作为下游任务的输入，看哪种词向量可以更好的实现下游任务。

内部评价（intrinsic evaluation）

对于内部评价，本文采用的绝对的内部评价（absolute intrinsic evaluation）和相对的内部评价（comparative intrinsic evaluation），绝对内部评价有以下方法

- Relatedness：比较生成的词向量的词于词之间的余弦相似度和人类评价的相似度的关系
- Analogy：对于一个y，去找到一个x，使得x:y的关系要和a:b的关系一样
- Categorization：把生成的词向量做聚类，看聚类是否准确
- Selectional preference：确定一个词是某个动词的主语还是宾语

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

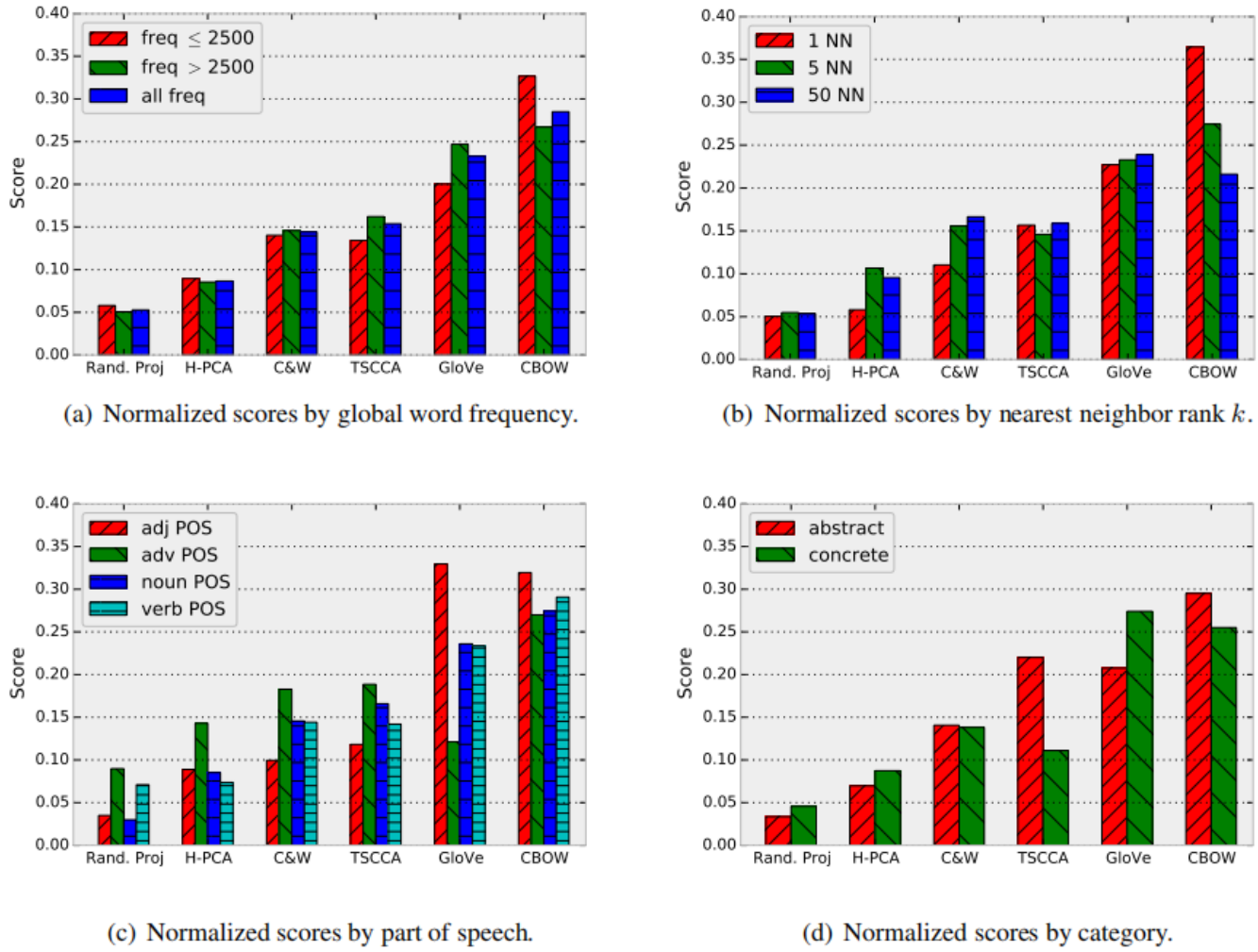
	relatedness						categorization			sel. prefs		analogy			
	rg	ws	wss	wsr	men	toefl	ap	esslli	batt.	up	mcrae	an	ansyn	ansem	average
CBOW	**74.0**	**64.0**	**71.5**	**56.5**	**70.7**	66.7	**65.9**	**70.5**	**85.2**	24.1	13.9	**52.2**	**47.8**	**57.6**	**58.6**
GloVe	63.7	54.8	65.8	49.6	64.6	**69.4**	64.1	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.

在相对内部评价中，用户直接来判断词向量的好坏。作者的具体做法如下，

- 选取了词频，词性和词义不同的100个单词（选择10种类别的词，每种类别里有一个形容词，一个动词，4个名词，4个动词）
- 找出每个词的n nearest neighbors, 选取rank为1，5，50的neighbor。所以对于6中词向量，对于每一个词，我们分为计算出rank为1，5，50的neighbor。
- 让人类来分别评价6中词向量中，rank1，5，50的neighbor里哪个于选定词最近。

结果如下,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



(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

在相似度（relatedness）的比较中，我们对于任意一个单词，我们只找了一个相近的单词，这并不理想（因为每个单词都有很多近义词）。所以作者提出了一种新的衡量方式：Coherence。对于每一个单词，事先选出两个近义词和一个不相关的词，看用生成的词向量能否辨别出不想关的词。

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

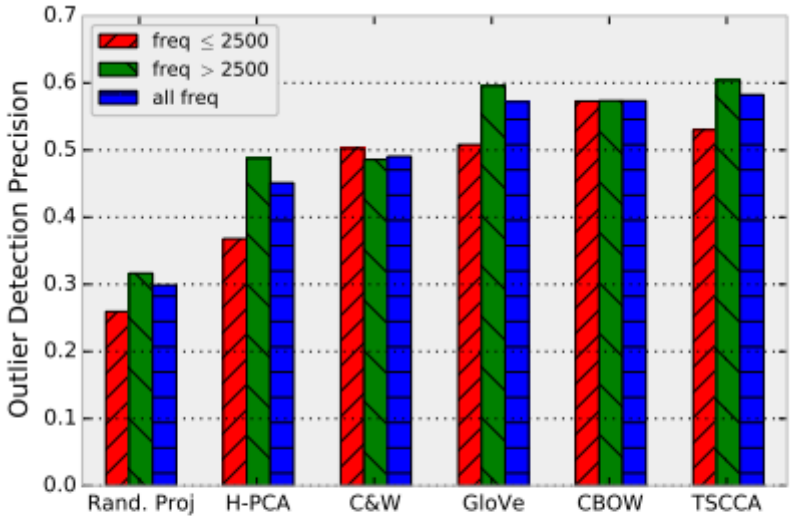


Figure 2: Intrusion task: average precision by global word frequency.

外部评价（Extrinsic evaluation）

外部评价主要用来测量词向量对于下游任务的贡献。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

	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.

3. Frequency information

最后，作者通过以下两种实验发现了词向量里面包含词频信息。

- 用词向量来预测单词在语料中词频
- 对于所有在WordSim-353数据集的单词，研究其K=1000 nearest neighbors和他们在语料中词频的大小排序。

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

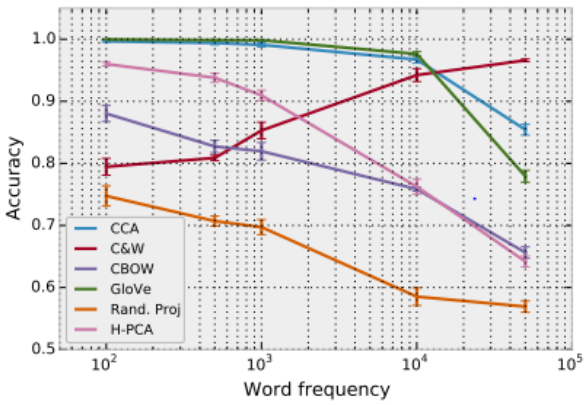


Figure 3: Embeddings can accurately predict whether a word is frequent or rare.

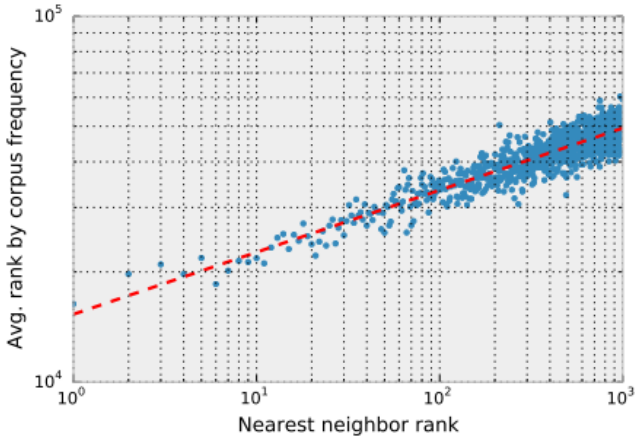


Figure 4: Avg. word rank by frequency in training corpus vs. nearest-neighbor rank in the C&W embedding space.

4. 思考

通过本文，我们发现没有任何一种词向量可以在所有任务中都表现的最好，所以每个单词应该不存在一种绝对正确的词向量。那么，词向量是否是用来表示单词的最好方式呢，我对此表示疑问。以后很有可能会发现一种新的表示单词的方式。