基于矩阵分解和矩阵变换的多义词向量研究

石昊悦
hyshi@pku.edu.cn

2018 年 6 月 9 日

Content

词向量和多义词向量

词向量

词向量是一种基于分布式语义的词义表达方式。

分布式语义假设：上下文相似的词语具有相似的含义。

词向量间的相似度（如余弦相似度）可以表达对应词间的相似度。

词向量可以由**统计方法**或**神经网络**学习得到。

词向量和多义词向量

词向量

词向量是一种基于分布式语义的词义表达方式。

分布式语义假设：上下文相似的词语具有相似的含义。

词向量间的相似度（如余弦相似度）可以表达对应词间的相似度。

词向量可以由**统计方法**或**神经网络**学习得到。

多义词向量

故名思义，多义词向量试图用多个不同的向量表达多义词的不同词义。

多义词向量一般由以下三部分构成：

- 全局向量：每个词只有一个全局向量。
- 词义向量：每个词义对应一个词义向量。
- 参数向量：帮助根据语料中的上下文选择具体词义的**参数**。

多义词向量中的“伪多义”现象

“伪多义”现象

我们定义“伪多义”表示自动挖掘词义的词向量学习算法对事实上的同一词义学习出多个词向量表示，且这些词向量表示之间相似度并不大的现象。

下表展示了使用 [1] 中模型进行训练的词向量中的伪多义现象（同色表示人工推理倾向于相同含义）。

词义	词义向量的最近邻对应词
$star_{s1}$	stars, movie, song, MVP
$star_{s2}$	stars, award, eagle, two-time
$star_{s3}$	supergiant, constellation, aurigae
$algorithm_{s1}$	hash, algorithms, quick sort, recursive
$algorithm_{s2}$	algorithms, optimization, public-key

[1] Neelakantan et al.. 2014. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In Proc. of EMNLP

工作目标

观察表明，伪多义现象在几乎所有基于神经网络的多义词向量模型上都存在且比较显著。

本工作尝试对上述伪多义现象进行检测，并在保留其他词向量相对关系的基础上进行伪多义消除。

伪多义检测

算法一：基于外部知识库的伪多义检测

分布式语义假设：在相似上下文中的词语表达相似的含义。
词向量：相似含义的词在词向量中相似度较高。

推论：对于多义词向量，一个义项的具体含义可以借助外部知识库、结合其近邻向量大致确定。

伪多义检测

算法一：基于外部知识库的伪多义检测

分布式语义假设：在相似上下文中的词语表达相似的含义。
词向量：相似含义的词在词向量中相似度较高。

推论：对于多义词向量，一个义项的具体含义可以借助外部知识库、结合其近邻向量大致确定。

外部知识库：WordNet[1]/同义词词林[2]
本工作主要利用知识库中的上下位 (hypernymy-hyponymy) 关系，所有词和上下位关系构成一张有向拓扑图。

上位词	下位词
animal/动物	cat/猫
country/国家	China/中国

[1] Miller. 1995. WordNet: a lexical database for English. Communications of the ACM
[2] 梅家驹. 1984. 同义词词林. 商务印书馆；上海

伪多义检测

算法一：基于外部知识库的伪多义检测

多义词不共上位假设：一个词的多个含义不共享直接上位词。

词	词义 1	直接上位 1	词义 2	直接上位 2
bank	银行	financial institution	河岸	slope
net	网	trap	净 (收入)	income

伪多义检测

算法一：基于外部知识库的伪多义检测

多义词不共上位假设：一个词的多个含义不共享直接上位词。

词	词义 1	直接上位 1	词义 2	直接上位 2
bank	银行	financial institution	河岸	slope
net	网	trap	净 (收入)	income

根据词义向量对应邻域的打分机制：

$$score(\mathbf{v}_{s_1}^{w}, SynH_i^{w}) = \sum_{\mathbf{v}_{s_k}^{w'} \in NN(\mathbf{v}_{s_1}^{w})} \cos(\mathbf{v}_{s_1}^{w}, \mathbf{v}_{s_k}^{w'}) isPossibleHypernym(SynH_i^{w}, \mathbf{v}_{s_1}^{w})$$

简而言之，在决定每个词义向量所对应的词义时，它的近邻向量为二者可能共享上位的词义贡献值为该词义向量与近邻向量余弦相似度的分数。

伪多义检测

算法一：基于外部知识库的伪多义检测

多义词不共上位假设：一个词的多个含义不共享直接上位词。

词	词义 1	直接上位 1	词义 2	直接上位 2
bank	银行	financial institution	河岸	slope
net	网	trap	净 (收入)	income

根据词义向量对应邻域的打分机制：

$$score(\mathbf{v}_{s_1}^{w}, SynH_i^{w}) = \sum_{\mathbf{v}_{s_k}^{w'} \in NN(\mathbf{v}_{s_1}^{w})} \cos(\mathbf{v}_{s_1}^{w}, \mathbf{v}_{s_k}^{w'}) isPossibleHypernym(SynH_i^{w}, \mathbf{v}_{s_1}^{w})$$

简而言之，在决定每个词义向量所对应的词义时，它的近邻向量为二者可能共享上位的词义贡献值为该词义向量与近邻向量余弦相似度的分数。

最终，每个词义向量打分最高的词义为其对应词义；一个词的不同向量对应到了相同上位即为伪多义。

伪多义检测

算法二：基于邻域相似度的伪多义检测

考虑不依赖外部知识库，两个词的邻域相似度可以直接由其邻域中的向量来表示:

$$P_{pseudo}(\mathbf{v}^w_{s_0}, \mathbf{v}^w_{s_1}) \propto \sum_{\mathbf{v'}_{s_0} \in NN(\mathbf{v}^w_{s_0})} \sum_{\mathbf{v'}_{s_1} \in NN(\mathbf{v}^w_{s_1})} \cos(\mathbf{v'}_{s_0}, \mathbf{v'}_{s_1})$$

考虑通过观察和交叉验证法设定阈值 θ，一个词的不同词义向量的邻域相似度超过 θ 时认定为伪多义。

伪多义检测

算法二：基于邻域相似度的伪多义检测

考虑不依赖外部知识库，两个词的邻域相似度可以直接由其邻域中的向量来表示:

$$P_{pseudo}(\mathbf{v}^w_{s_0}, \mathbf{v}^w_{s_1}) \propto \sum_{\mathbf{v'}_{s_0} \in NN(\mathbf{v}^w_{s_0})} \sum_{\mathbf{v'}_{s_1} \in NN(\mathbf{v}^w_{s_1})} \cos(\mathbf{v'}_{s_0}, \mathbf{v'}_{s_1})$$

考虑通过观察和交叉验证法设定阈值 θ，一个词的不同词义向量的邻域相似度超过 θ 时认定为伪多义。

算法二存在一定问题：对于不同话题/域的词，这个伪多义的阈值很可能由于语料不均衡而不均等。而算法三通过一个更为直接的矩阵分解方式避免了这个问题。

伪多义检测

算法三：基于词内部意义对差矩阵分解的伪多义检测

通过观察，我们得到伪多义的两点性质：
(1) 系统性 (cat_{s0}-dog_{s0}，cat_{s1}-dog_{s1}) (2) 占比大 ($> 70\%$)

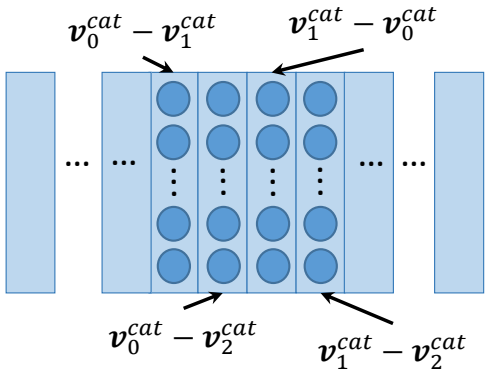
伪多义检测

算法三：基于词内部意义对差矩阵分解的伪多义检测

通过观察，我们得到伪多义的两点性质：
(1) 系统性 (cat_{s0}-dog_{s0}，cat_{s1}-dog_{s1}) (2) 占比大 ($> 70\%$)

考虑词内部意义对差矩阵

$$M = \bigoplus_{w} [\oplus_{i \neq j} (\mathbf{v}_i^w - \mathbf{v}_j^w)]$$



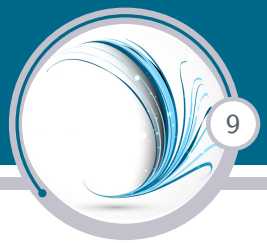
对矩阵 M 做主成分分析，所得结果应是显著的“伪多义空间”。

主成分分析

$$\min ||M - L||_F$$

$$subject\ to\ L + E = M, \mathrm{rank}(L) = d$$

其中 d 为指定的 L 的秩。



主成分分析

$$\min ||M - L||_F$$

$$subject\ to\ L + E = M, \mathrm{rank}(L) = d$$

其中 d 为指定的 L 的秩。

传统主成分分析算法没有考虑强噪声（真多义）的存在。

主成分分析

$$\min ||M - L||_F$$

$$subject\ to\ L + E = M, \mathrm{rank}(L) = d$$

其中 d 为指定的 L 的秩。

传统主成分分析算法没有考虑强噪声（真多义）的存在。

强噪声的处理：健壮主成分分析 (Robust PCA, RPCA)。

$$\min rank(L) + \lambda_1 ||S||_0 + \lambda_2 ||E||_F$$

$$subject\ to\ L + E = M, \mathrm{rank}(L) = d$$

伪多义检测

算法三：基于词内部意义对差和矩阵分解的伪多义检测

伪多义空间的秩在一个较小的范围内。
已有的 RPCA 算法 [1] 基于凸松弛后的优化问题，不能有效控制秩。

伪多义检测

算法三：基于词内部意义对差和矩阵分解的伪多义检测

伪多义空间的秩在一个较小的范围内。
已有的 RPCA 算法 [1] 基于凸松弛后的优化问题，不能有效控制秩。

基于 PCA 的迭代 RPCA 算法

初始值：令 $M^{(0)} = M, t = 0, S = 0$ 矩阵
WHILE 未收敛
　利用 PCA 计算 $L^{(t)} + E^{(t)} = M^{(t)}$
　计算 $S^{(t)} = E^{(t)}$ 的显著部分
　令 $S = S + S^{(t)}$
　令 $M^{(t+1)} = M^{(t)} - S^{(t)}$
　令 $t = t + 1$
return $L^{(t)}, S, E^{(t)}$

算法收敛性：考虑 $M^{(t)}$ 在 $L^{(t)}$ 上的投影误差，容易利用单调有界收敛定理 [2] 证得。

[1] Wright et al.. 2009. Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization. In Proc. NIPS
[2] 伍胜健. 2010. 数学分析. 北京大学出版社

基于矩阵变换的伪多义消除

给定伪多义词对，计算消除变换矩阵 T:

$$\min \ \mathcal{L}(T) = \sum_{i=1}^{m} \frac{1}{2}(|| T\mathbf{v}_{s_{i,0}}^{w_i} - \frac{\left(\mathbf{v}_{s_{i,0}}^{w_i} + \mathbf{v}_{s_{i,1}}^{w_i}\right)}{2} ||_2 + || T\mathbf{v}_{s_{i,1}}^{w_i} - \frac{\left(\mathbf{v}_{s_{i,0}}^{w_i} + \mathbf{v}_{s_{i,1}}^{w_i}\right)}{2} ||_2)$$

基于矩阵变换的伪多义消除

给定伪多义词对，计算消除变换矩阵 T:

$$\min \ \mathcal{L}(T) = \sum_{i=1}^{m} \frac{1}{2}(|| T\mathbf{v}_{s_{i,0}}^{w_i} - \frac{\left(\mathbf{v}_{s_{i,0}}^{w_i} + \mathbf{v}_{s_{i,1}}^{w_i}\right)}{2} ||_2 + || T\mathbf{v}_{s_{i,1}}^{w_i} - \frac{\left(\mathbf{v}_{s_{i,0}}^{w_i} + \mathbf{v}_{s_{i,1}}^{w_i}\right)}{2} ||_2)$$

给定伪多义空间 $U \subset \mathbb{R}^n$，计算消除变换矩阵 T:

$$\begin{cases} T\alpha = 0, & \alpha \in U \\ T\alpha = \alpha & \alpha \perp U \end{cases}$$

容易证明，这样的矩阵 T 存在且只存在一个。

基于矩阵变换的伪多义消除

给定伪多义词对，计算消除变换矩阵 T:

$$\min \ \mathcal{L}(T) = \sum_{i=1}^{m} \frac{1}{2}(|| T\mathbf{v}_{s_{i,0}}^{w_i} - \frac{(\mathbf{v}_{s_{i,0}}^{w_i} + \mathbf{v}_{s_{i,1}}^{w_i})}{2} ||_2 + || T\mathbf{v}_{s_{i,1}}^{w_i} - \frac{(\mathbf{v}_{s_{i,0}}^{w_i} + \mathbf{v}_{s_{i,1}}^{w_i})}{2} ||_2)$$

给定伪多义空间 $U \subset \mathbb{R}^n$，计算消除变换矩阵 T:

$$\begin{cases} T\alpha = 0, & \alpha \in U \\ T\alpha = \alpha & \alpha \perp U \end{cases}$$

容易证明，这样的矩阵 T 存在且只存在一个。

使用 T 对原向量空间进行变换，即可得到减低伪多义影响的新词向量

$$\tilde{V} = \{\tilde{\mathbf{v}}_s^w = T\mathbf{v}_s^w | \mathbf{v}_s^w \in V\}$$

.

词义相似度

下表展示了不同多义词向量训练的模型在词相似度任务上的表现。

模型	WS-353	SCWS
基于上下文的聚类 [1]	64.2	26.1
多义词 skip-gram (MSSG, 300D)[2]	70.9	57.3
无监督多义词 skip-gram (NP-MSSG, 300D) [2]	69.1	59.8
NP-MSSG + 算法一	68.8	62.2
NP-MSSG + 算法二	69.2	63.7
NP-MSSG + 算法三 (PCA)	69.2	65.3
NP-MSSG + 算法三 (RPCA)	69.2	**65.4**
中餐馆模型 [3]	69.5	62.4
MUSE[4]	69.4	67.9

[1] Huang et al.. 2012. Improving word representations via global context and multiple word prototypes. In Proc. of ACL
[2] Neelakantan et al.. 2014. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In Proc. of EMNLP
[3] Li and Jurafsky. 2015. Do Multi-Sense Embeddings Improve Natural Language Understanding?. In Proc. of EMNLP
[4] Lee and Chen. 2017. MUSE: Modularizing Unsupervised Sense Embeddings. In Proc. of EMNLP

下游任务测试

在这个实验中，选取 SentEval[1] 中的三个任务进行测试，它们分别是：主客观分析 (SUBJ)，问题分类 (TREC) 和段落检测 (MSRP)。选取 BoW 作为句子的特征。

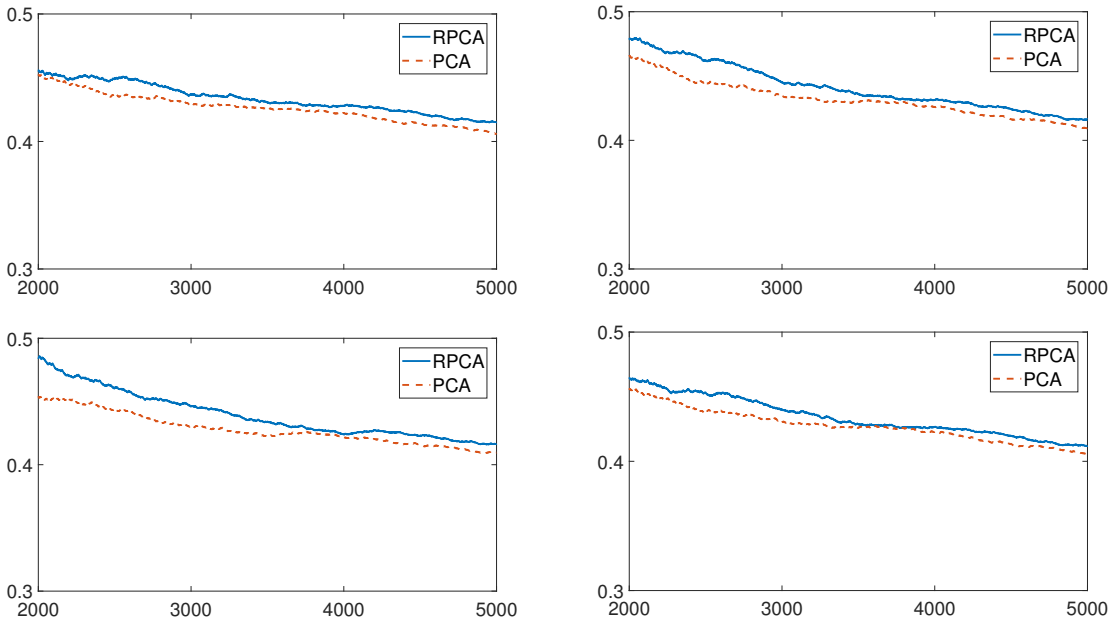
模型	SUBJ	TREC	MSRP
原空间 (NP-MSSG)[2]	91.0	78.4	70.0
+ 算法一	91.2	83.2	70.6
+ 算法二	91.0	84.2	70.0
+ 算法三 (RPCA)	**92.3**	**85.4**	**71.0**

实验表明，消除原空间中的伪多义有助于提升下游任务的表现，其中算法三起到了最佳效果。

[1] Conneau and Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In Proc. of LREC
[2] Neelakantan et al.. 2014. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In Proc. of EMNLP

PCA VS. RPCA: WordNet Synset 召回

使用 PCA 中的词对差向量-L 投影比率从小到大排序。
使用 RPCA 中的对应 $||S||_2$ 从大到小排序。
利用算法一的方法对每个词对进行 WordNet 义项确认，如不同则记为一次正确，下图分别展示了 $\mathrm{rank}(L) = 1, 2, 3, 5$ 的词对-准确率曲线。



PCA VS. RPCA: 伪多义方向代表向量对

下表展示了 $\mathrm{rank}(L) = 3$ 的 PCA 和 RPCA 提取出的伪多义空间正交基的代表（投影比率最高）词对。

	#	代表词对	特征
RPCA	1	after$_{1,2}$, eventually$_{0,1}$, whilst$_{0,1}$, again$_{1,2}$, finally$_{1,2}$	副词
	2	although$_{1,2}$, well$_{2,3}$, initially$_{1,2}$, more$_{1,4}$, both$_{0,2}$	副词
	3	Brian$_{1,2}$, February$_{0,2}$, Daniel$_{1,2}$, September$_{2,7}$, Frank$_{0,2}$	专有名词
PCA	1	income$_{2,4}$, campaigns$_{1,5}$, age$_{6,7}$, development$_{4,5}$, goals$_{2,6}$	政治话题
	2	Berlin$_{0,6}$, Martin$_{3,4}$, Greek$_{0,3}$, Jan$_{0,4/0,6}$, name$_{1,3}$	专有名词
	3	quarterback$_{3,9}$, playoff$_{3,9}$, NBA$_{0,1}$, Houston$_{1,3}$, mayor$_{0,6}$	体育话题

RPCA 倾向于更加泛化的特征，而 PCA 则倾向于具体的话题特征。

总结

- 本工作提出了伪多义的概念和三种检测算法，其中的健壮主成分分析算法可以扩展应用到其他领域中
- 本工作提出了伪多义消除算法，并证明了其在下游任务上的有效性
- 本工作基于前人的工作，尝试从数学角度对于语言学中词义的概念给出一个基于序关系的“软”定义，为无监督词义发现提供了一个新的思路

谢谢！欢迎提问！