文章编号：1003-0077（2017）00-0000-00

基于知网相关概念场的中文词向量

冯煜博[1] 蔡东风[1] 宋彦[2]
（1.沈阳航空航天大学 人机智能研究中心，辽宁 沈阳 110136；
2.创新工场，广东 深圳 518061）

摘要：词向量是词的低维稠密实数向量表示，在自然语言处理的各项任务中都扮演了重要角色。目前词向量大多都是通过构造神经网络模型，在大规模语料库上以无监督学习的方式训练得到，这样的模型存在着两个问题：一是低频词词向量的语义表示质量较差；二是忽视了知识库可以对该模型提供的帮助。该文提出了利用知网相关概念场来提升词向量语义表示质量的模型。实验结果表明，在词语相似度任务（ws297s）上该模型将 GloVe 词向量的斯皮尔曼相关性系数提高了 10.29，在词语相关度任务（ws240r）上将 SG 词向量提高了 3.39；在词语类比任务上将 GloVe 词向量的准确率提高了 10.59 个百分点。

关键词：词向量；知网相关概念场；低频词；神经网络语言模型

中图分类号：TP391 **文献标识码：**A

Chinese Word Representations Based on HowNet Relevant Concept Field

Feng Yubo[1], CAI Dongfeng[1], Yan Song[2]
(1.Reseach Center for Human-computer Intelligence, Shenyang Aerospace University ,Shenyang, Liaoning 110136, China ; 2.Sinovation Ventures, Shenzhen, Guangdong 518061, China)

Abstract：Word embeddings are low-dimensional dense real number vectors of words, which play an important role in various natural language processing tasks. Traditional word embedding is derived from neural network language model trained on a large-scale unlabeled text corpus, which suffers from the quality of resulting vectors of the low frequent word is not satisfactory, and external knowledge is ignored. We proposed three models that can systematically learn embedding for all the relevant concept fields defined in HowNet, and consequently, obtain word vectors, in particular for low frequent words. Experimental results show that our model can greatly improve the performance of word embeddings. E.g. our model improved embeddings Spearman Correlation of GloVe and SG on word similarity and word relevance with 10.29 and 3.39 respectively. Our model improved accuracy of GloVe on word analogy with 10.59 percent.

Key words: word embedding; HowNet Relevant Concept Field; low frequency word

0 引言

词表示学习，目的在于得到词的低维稠密实数向量表示，词的向量表示可以作为特征，广泛地应用到很多后续的自然语言处理任务中。词表示学习是自然语言处理的重要任务之一[1][6]。目前大部分的研究都是基于分布式假说[7]，在大规模语料库上无监督地学习词的向量表示，这种学习方法倾向于把经常在相似语境中出现的词聚在一起，从而可以学习到词语的上下文相似性。但是由于低频词在语料中出现的次数较少，被训练的不够充分，所以导致低频词的词向量不能很好地表示它自身的语义。

一些研究者基于汉语的特点，巧妙地利用了汉字的笔画[8]和偏旁部首[10][12]开展研究。一些研究者利用汉语的词语是由汉字组合而成的特点，将词语拆分到字级别[13][14]，还有一些研究者借助汉语词语的语素[15]。上述工作都提高了词向量的语义表示质量，但是却忽视了专家标注的高质量知识库可能对该任务带来的帮助，缺少语义资源的指导。

收稿日期：2018-06-18；**定稿日期**：2017-08-15
基金项目：辽宁省重点研发计划（协同式多语言机器翻译云平台）；教育部人文社会科学研究青年基金（17yjczh003）

借助于人工知识库是一种高效的提升词向量质量的方法。一些研究者借助于 WordNet[16]、FreeBase[17]、PPDB[18]和 FrameNet[19]等知识库开展了研究[20][29]。但是这些工作大都是基于英文知识库开展的研究，或者只是利用了词语间的相似性关系。所以在中文领域，如何借助于知识库来提升词表示学习的质量仍然需要探索。

一些研究者基于知网[30]开展了词表示学习的研究[31][35]。其中，孙茂松等关注了低频词的词向量语义表示质量较差的问题，借助于知网义原提升了低频词的词向量语义表示质量，是一种借助于词语自身属性的方法。在本文中，我们更加侧重于探索如何借助词与词之间的相关性关系，提升词向量的语义表示质量。

为了进一步提升低频词的语义表示质量，我们借助于知网相关概念场[36]这一强大的知识库，将高频词与低频词在词向量的语义表示上建立起了连接，从而使得低频词可以参照高频词来修正自身的向量表示。

此外，当前的相关研究都未专门评价词向量的上下文语义相关度，为此我们提出了一个评价词语上下文相关度的数据集。

本文中，我们介绍了词语相似度与相关度的区别，为实验中细粒度地评价词向量的相似度与相关度确立了理论基础。然后基于知网相关概念场提出了一种提升词向量语义表示质量的词表示学习模型，并进行了词语相似度、词语相关度和词语类比实验。实验结果表明，本文提出的模型可以显著提升低频词词向量的语义表示质量。

据我们所知，本文的主要贡献有：

（1）提出了借助于知网相关概念场来提升词向量质量的方法。

（2）我们根据 wordsim240（ws240）数据集，构建了侧重于评价中文词语相关度的数据集 wordsim240-relevance（ws240r），作为 ws240 的子数据集，ws240r 可以细粒度地评价中文词语之间的上下文语义相关度。

（3）我们在词语相似度、词语相关度和词语类比任务上进行了实验。通过这些实验证明了本文所提方法的有效性。

本文结构如下，第 1 节为相关工作介绍，第 2 节为基于知网相关概念场的词表示学习模型介绍，第 3、4 节为通过实验和例子给出通过模型学习得到的词向量的表示效果，最后是对未来工作的展望。

1 相关工作

表 1 知网相关概念场举例

举例	欧几里得	小行星	中国
知网相关概念场	数学家	天文馆	中国科学院
	几何图形	航天飞机	商务印书馆
	几何学	北斗星	中医
	解析几何	哈雷彗星	中药店
	法线	太空站	中餐厅
	...	...	...

1.1 知网相关概念场

董振东指出[36]：相关是指不同的概念在某种语境中共现的可能性（也是本文所遵从的词语间相关性的定义）。相关概念是指词语所代表的概念与哪些概念相关。相关概念场是一个相关概念的集合，是与一个词语的某个概念相关的所有概念的集合。相关概念场中的相关概念是由词语来表现的，这些词语所代表的概念已经是唯一的不再有歧义。

词语的相关性与词语的相似性容易混淆。词语的相关性反应的是两个词语之间互相关联的程度，可以用词语在同一语境中共现的可能性来衡量。而词语的相似性反映的则是词语之间的聚合特点，比如刘群等曾利用知网进行了词语相似性的研究[37]。

董振东等构建了一套相关性激发器[36]，这套机制中包括了 DEF 分析器、规则匹配器和构造器，从而实现了知网的相关概念场。我们将词所辖的相关概念场合并，得到了基于词的相关概念场，如表 1 举例所示。

1.2 基于知网的词表示学习

知网蕴含了丰富的关于语言和世界的知识，如何将这些知识有效地、合理地加入到词向量的学习当中，已经成为了词表示学习领域的一个重要课题。一些研究者将知网与词向量的学习进行了有机结合。例如，唐共波[31]等将语料中的单义原词替换成对应义原，然后使用词表示学习模型在语料库上训练，从而得到了义原的向量表示。孙茂松等[32]通过将预训练得到的高频词的词向量固定，并假设词向量是由词所辖的义原向量线性组合而成，首先基于高频词词向量来训练义原向量，然后基于义原向量来训练低频词的词向量。该方法使用义原来充当桥梁，低频词词向量的质量借助高频词词向量得到了提升。Niu Y 等[33]的方法与孙茂松等的方法一脉相承，两者都假设词

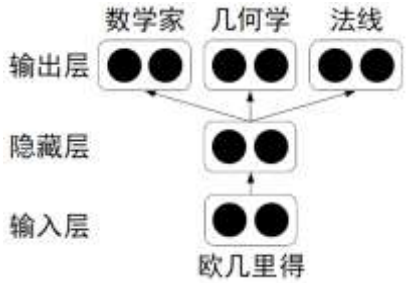


图 1 低频词加强神经网络模型

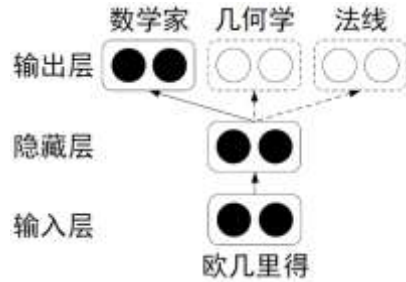


图 2 高频词加强神经网络模型

向量是由词所辖的义原向量线性组合而成，Niu Y 等的方法是令义原向量和词向量在同一向量空间中，一起从头开始学习各自的表示。陈洋等[34]基于义原独立假说，使用施密特正交化将义原向量矩阵处理成了正交单位矩阵，词向量矩阵是由义原向量矩阵乘以权重矩阵得到的，然后在语料上无监督地训练模型，不断更新权重矩阵的参数，从而得到词的向量表示。朱静雯等[35]把知网层层拆解，构建了一个跨语言、跨语义单位的常识知识图谱，然后使用图表示学习方法和知识图谱表示学习方法，得到了知网中各个实体以及词的低维稠密实数向量表示。

上述方法都只关注到了词的义原、义项等属性信息；没有关注到词与词之间的关系，比如在某种语境中经常共现的一些词对具有相关性。知网相关概念场中包含的内容，都是由专家标注的，在某种语境中可能会经常共现的词对。为了探究词语间的关系（主要是相关性）对词表示学习的影响，本文提出了基于知网相关概念场的词向量学习模型。

2 基于知网相关概念场的词向量学习模型

2.1 低频词加强模型（L）

我们构建了一个如图 1 所示的神经网络模型，用来二次训练低频词的词向量。这个模型与 Mikolov 的 Skip-Gram 模型[38][39]的差别在于：Skip-Gram 模型的输入层可以是整个词表，而我们的模型要求输入层只能是词表中的低频词。Skip-Gram 模型的输出层可以是中心词的上下文，而我们的模型要求输出层只能是中心词的相关概念场。

我们首先在语料库上运行词表示学习模型得到词的向量表示，然后固定高频词的词向量不变，不断更新低频词的词向量。其基本思想是：训练低频词的词向量去逼近该词的相关概念场，使学习到的低频词的词向量可以更加准确地预测该词的相关概念场。

形式化地，给定低频词 w_i 的词向量 v_i 和该词对应的相关概念场 $R(w_i)$，训练目标如式（1）所示。

$$L=\frac{1}{T}\sum_{i=1}^{T_L}\log P(R(w_i)\mid w_i) \tag{1}$$

式（1）求和遍历了词表中的低频词（规模为 T_L）来计算低频词正确预测各自相关概念场的对数概率。我们假设词 w_i 的相关概念场可以充当该词的上下文，那么 $P(R(w_i)\mid w_i)$ 可以由式（2）定义的方式求得。

$$P(R(w_i)\mid w_i)=\prod_{w_r\in R(w_i)}P(w_r\mid w_i) \tag{2}$$

我们使用 softmax 函数定义由词 w_i 预测它的相关词 w_r 的概率 $P(w_r\mid w_i)$，如式（3）所示，其中 W 是词表。

$$P(w_r\mid w_i)=\frac{\exp(v_r^{\mathrm{T}}\cdot v_i)}{\sum_{w=1}^{W}\exp(v_w^{\mathrm{T}}\cdot v_i)} \tag{3}$$

以低频词“欧几里得”为例，我们的模型用该词的词向量作为隐藏层依次地去预测该词的相关概念场（数学家、几何学、法线等）。

从式（3）可以看出，计算预测概率时需要遍历整个词表，而词表的大小往往是比较大的（这里超过 40 万个词），因此本文使用了负采样[39]来降低计算复杂度。

2.2 高频词加强模型（L+H）

在低频词加强模型中，考虑到词的相关概念场中可能会包含低频词，而低频词的词向量语义表示质量较差，可能会污染模型的学习。所以在低频词加强模型的基础上，我们提出了如图 2 所示的高频词加强神经网络模型。给定低频词，我们只取出该词相关概念场中的高频词，构造出新

表 2 评测集中词语关系分布统计

评测集	ws240	ws240r	ws297	ws297s
词对总数	240	183	296	169
相似	22.50%	0.00%	**39.19%**	**53.57%**
相关	**57.50%**	**75.41%**	26.69%	2.98%
相似相关	1.25%	0.00%	1.69%	2.98%
无关	18.75%	24.59%	32.43%	41.07%

表 3 评测集中低频词样本分布

评测集	样本总数	低频词样本数
ws240	240	57
ws240r	183	50
ws297	296	102
ws297s	169	64
A1125	1125	693

的相关概念场 $R_{high}(w_i)$，则训练目标如式（4）所示。

$$L=\frac{1}{T}\sum_{i=1}^{T_L}\log P(R_{high}(w_i)\,|\,w_i) \quad (4)$$

上式中 $P(R_{high}(w_i)\,|\,w_i)$ 的定义与式(2)类似，如式（5）所示。

$$P(R_{high}(w_i)\,|\,w_i)=\prod_{w_r\in R_{high}(w_i)}P(w_r\,|\,w_i) \quad (5)$$

2.3 相关概念场加强模型（L+H+R）

在进行负采样的过程中，词 w_i 的负例可能会与 $R_{high}(w_i)$ 产生交集，也即 $R_{high}(w_i)$ 中的词可能在 w_i 的学习过程中既要作正例又要作负例，这就造成了模型在学习过程中的自相矛盾。

基于上述分析，我们在高频词加强神经网络模型的基础上，提出了相关概念场加强神经网络模型。在前述模型中，负采样的目标都如式（6）所示，其中 K 是负采样个数。

$$\log\sigma(v_i^{\mathrm{T}}\cdot v_r)+\sum_{j=1}^{K}E_{w_j\sim P_n(n)}[\log\sigma(-v_j^{\mathrm{T}}\cdot v_i)] \quad (6)$$

相关概念场加强神经网络模型的负采样方法与 Mikolov 的负采样方法类似，但是在负采样过程中，我们认为词 w_i 的相关概念场 $R_{high}(w_i)$ 不应作为词 w_i 的负例，所以新的负采样目标在式（6）的基础上做出了改进，具体形式如式（7）所示。

$$\log\sigma(v_i^{\mathrm{T}}\cdot v_r)+\sum_{j=1}^{K}E_{w_j\sim P_n(n)}[\log\sigma(-v_j^{\mathrm{T}}\cdot v_i)]\quad w_j\notin R_{high}(w_i) \quad (7)$$

3 实验

本节中，我们基于词语相似度、词语相关度和词语类比任务来检验本文所提方法的有效性。实验结果表明本文提出的模型：

（1）在词语相似度和词语相关度任务上能够提升与人类打分的斯皮尔曼相关度；

（2）在词语类比任务上能够提升事实常理推断的准确率（Accuracy）和平均排序（Mean Rank）。

（3）在低频词的词语相似度、词语相关度和词语类比任务上，能够提升低频词词向量的表现。

实验使用中文维基百科语料①作为训练语料库。中文维基百科语料库共包含 26.5 万篇中文维基词条文章。我们使用 gensim 工具包②从 XML 格式的文本中提取出待处理的语料，得到 32.7 万篇维基词条文章。据我们的观察，待处理的语料中包含简体中文和繁体中文，所以我们使用 opencc 工具包③将繁体中文字符全部转换成了简体中文字符。并且我们限制语料库中字符的 Unicode 编码范围为 0x4E00 至 0x9FA5，为的是仅保留语料库中的中文字符，该语料库预处理方法与主流的中文词向量语料库预处理方法保持一致[8][15]。我们使用结巴分词工具④对语料进行了分词，最终得到 1.7 亿个中文 Token，语料大小超过 1.1GB。

我们使用 HowNet2015 版本⑤作为义原、义项词典和相关概念场。经整理后，共含 2124 个义原。词向量的维度均设置为 200 维、上下文窗口设置为 5、负采样个数为 8、词频低于 10 的词舍弃。

3.1 数据集和实验设置

我们采用公开评测集 ws240[40]、ws297[41]和 ws297s[35]来评价模型学习到的词向量的相关度和相似度语义表示质量，并且手工统计了各个评测集的词对儿之间的语义关系分布，如表 2 所示。经过观察我们发现，ws240 评测集更侧重词语之间的相关性关系、ws297 评测集更侧重评价词语

①https://dumps.wikimedia.org/zhwiki/latest/
②https://radimrehurek.com/gensim/
③https://github.com/BYVoid/OpenCC
④https://github.com/fxsjy/jieba
⑤http://www.keenage.com/html/c_index.html

表 4 词语相关度和词语相似度评测结果

模型	完整				低频词			
	ws240	ws297	ws297s	ws240r	ws240	ws297	ws297s	ws240r
CBOW	51.26	57.43	64.74	51.15	51.26	57.64	64.76	51.15
义项不敏感（知网）	52.21	56.07	61.34	52.50	51.28	58.40	64.36	51.14
义项敏感（知网）	**52.23**	55.85	60.73	**52.53**	51.33	58.31	64.11	51.17
义项敏感（语料）	52.22	55.80	60.70	52.51	51.30	58.27	64.05	51.16
CBOW(L)	51.49	58.96	65.78	51.27	**52.51**	**59.61**	**66.54**	**52.48**
CBOW(L+H)	52.13	**59.46**	**66.06**	52.12	51.16	57.95	65.39	51.03
CBOW(L+H+R)	52.19	59.39	65.99	52.21	51.31	58.16	65.56	51.20
SG	51.75	57.24	66.01	51.89	51.75	57.53	65.56	51.89
SG(L)	**54.96**	**60.15**	**66.79**	**55.28**	**56.21**	**60.28**	65.58	**56.70**
SG(L+H)	54.72	59.92	66.26	54.84	52.26	58.02	**66.20**	52.72
SG(L+H+R)	54.73	59.86	66.70	54.86	52.46	57.87	66.11	53.04
DSG	54.42	59.38	66.11	54.59	51.26	57.64	64.76	51.15
DSG(L)	56.75	61.39	66.44	56.78	**57.23**	**62.27**	65.99	**57.31**
DSG(L+H)	56.86	**61.52**	**66.52**	56.92	54.42	60.20	66.43	54.59
DSG(L+H+R)	**57.14**	61.40	66.33	**57.11**	54.43	60.21	**66.46**	54.45
GloVe	48.83	39.87	42.37	48.13	48.83	41.12	42.86	48.13
GloVe(L)	51.33	**49.68**	**52.66**	50.90	**52.73**	**48.07**	**51.28**	**52.73**
GloVe(L+H)	**51.76**	44.18	44.73	**51.49**	48.89	42.26	44.34	48.14
GloVe(L+H+R)	51.72	44.42	44.76	51.11	48.87	42.12	44.18	48.19

之间的相似性关系。

此外，为了细粒度评价词向量的词语间相关性质量，我们根据 ws240 构建了它的子数据集 ws240r，一个中文领域评价上下文语义相关度的评测集，构建的方法是把 ws240 中词语间无关的词对和词语间相关的词对（总计 183 对）全部平移到 ws240r 中。在词语类比任务中，我们选择了 Chen X 等人的中文词语类比评测集(A1125)[13]。

为了说明本文所提方法对低频词的影响，我们在词语相似度任务和词语类比任务中加入了仅针对评测集中低频词的实验，表 3 给出了各评测集的测试样本中，低频词样本的分布。

为了证明本文所提的方法可以适用于各种词向量学习模型。我们选择了三个最具代表性的模型，基于计数模型的 GloVe[42]，基于预测模型的 word2vec 和跳出了分布式假说的 DSG[43]作为基础词向量。

3.2 词语相关度与词语相似度

表 4 给出了各模型预测的词语间语义相关度和语义相似度与人工打分的 Spearman 相关性系数。我们复现了孙茂松等的“义项不敏感(知网)”、“义项敏感（知网）”和“义项敏感（语料）”模型[25]作为本文的基线系统。我们使用的 Spearman 相关性评测工具是 gensim。

实验结果表明：我们的模型不仅提升了词语间的语义相关度，还大幅提升了词语间的语义相似度。在 ws297 和 ws297s 这两个侧重于评价语义相似度的测试集上，我们的模型表现远好于基线系统。我们的方法使得模型打分与人工评价之间的 Spearman 相关性系数取得了在本文中的最大幅度的提升。分析其原因，我们认为本文提出的模型能够通过相关概念场里面的高频词来提升低频词的词向量语义表示质量，从而使得低频词的词向量语义表示质量获得了提升。

借助于知网相关概念场方法比借助于知网义原的方法更接近人类打分，因为词的义原标注更多地体现的是词的属性，而基于知网相关概念场的方法则参考了词与词之间关系。由此我们可以得出这样的结论：基于词语关系的方法比较基于词语属性的方法，在词语关系的学习上更加直接、高效。

在低频词的专项评测上，本文所提模型的平均提升幅度进一步扩大，这更加说明了本文方法对低频词质量提升的有效性。

表 5a 词语类比（完整）

模型	Accuracy				Mean Rank			
	所有	首都	城市	家庭	所有	首都	城市	家庭
CBOW	73.49	81.09	71.43	55.88	2241.20	303.33	4964.06	5312.66
义项不敏感(知网)	73.67	81.39	71.44	55.89	2120.43	288.24	4670.86	5039.82
义项敏感（知网）	73.68	81.42	71.43	55.92	2073.06	283.88	4578.61	4914.27
义项敏感（语料）	72.90	81.41	71.45	55.90	2436.38	374.98	4579.18	4922.84
CBOW(L)	76.33	84.64	74.86	56.62	1110.56	38.53	2693.73	2760.23
CBOW(L+H)	78.64	86.83	77.13	**59.20**	**676.45**	**3.29**	1843.86	**1600.84**
CBOW(L+H+R)	**78.65**	**86.85**	**77.14**	59.19	677.39	3.39	**1832.53**	1611.77
SG	77.05	83.01	86.29	56.25	1287.32	30.63	541.41	4895.10
SG(L)	81.85	89.96	87.43	58.09	661.77	1.33	102.28	2665.56
SG(L+H)	80.52	87.74	86.86	58.46	590.56	1.57	390.91	2185.01
SG(L+H+R)	**83.54**	**91.88**	**88.00**	**59.93**	**348.08**	**1.13**	**52.74**	**1401.64**
DSG	82.56	90.99	88.00	58.09	135.69	1.28	39.17	532.34
DSG(L)	**83.99**	**92.76**	**88.57**	**59.19**	**93.19**	**1.11**	**19.85**	**369.58**
DSG(L+H)	82.92	91.29	88.10	58.82	130.70	1.20	38.06	512.61
DSG(L+H+R)	83.72	92.47	88.56	58.82	118.57	1.15	33.07	465.83
GloVe	68.77	74.00	87.43	43.75	2457.44	472.97	2.76	8976.01
GloVe(L)	76.51	82.87	92.00	50.74	797.90	44.67	1.42	3149.02
GloVe(L+H)	77.76	83.90	94.28	51.84	631.04	15.88	1.11	2567.46
GloVe(L+H+R)	**79.36**	**85.38**	**94.29**	**54.78**	**310.85**	**5.56**	**1.07**	**1270.01**

在 ws240 和 ws240r 这两个侧重于评价语义相关度的测试集上，基线系统仅在基于 CBOW 模型生成的词向量上略好于我们的模型。这是因为基于 CBOW 模型生成的词向量可能更适合于类似 CBOW 模型的二次训练模型。基线系统正是一种类似于 CBOW 模型的方法。

本文所提出的三种模型其适用性各有不同，无论是在完整的评测集上还是在低频词评测集上，三种模型都能使得词向量质量获得较好的提升。但是在三种模型中，低频词加强模型（L）在总能使得词向量质量获得较大幅度提升，而理论上更加细腻的相关概念场加强模型（L+R+H）并不是总能取得最好成绩，可能是因为该模型训练时丢弃的训练样本较多，导致模型的训练不够充分。

3.3 词语类比

A1125 词语类比数据集包含三大类：首都、城市和家庭关系。假设 v_1、v_2、v_3、v_4 分别是 4 个词 w_1、w_2、w_3、w_4 所对应的词向量，若 $v_2-v_1=v_4-v_3$ 则说明 w_2 与 w_1 的关系和 w_4 与 w_3 的关系类似，那么我们就可以通过 w_1、w_2 和 w_3 来推断 w_4 的近似向量 v_4'，推断的方法如式（8）所示：

$$v_4'=v_3-v_1+v_2 \quad (8)$$

然后我们可以通过 v_4 和 v_4' 的余弦相似度来评估学习到的词向量的质量。本文采用两种评估指标：（1）Accuracy，假设近似向量 v_4' 所对应的词 w_4' 就是我们要找的 w_4，记录近似推断成功的次数，重复上述过程。Accuracy 值就是在 A1125 中所有测试样例 $w_4=w_4'$ 的频率。(2)Mean Rank，在每一次近似推断中，首先将词表中所有词向量与 v_4' 计算余弦相似度，然后根据余弦相似度进行逆序排序，记录 w_4 在每次近似推断中的排位。Mean Rank 就是 w_4 在整个 A1125 数据集中的平均排位。

词语类比任务的完整实验结果由表 5a 所示，词语类比任务低频词专项实验结果由表 5b 所示。本文所提模型在所有类别上都取得了比基线系统更好的效果。尤其是在首都类别下，本文所提模型的 Mean Rank 指标达到了 3.29，远好于基线系统。经过分析，我们发现知网的相关概念场中，

对于首都类别的词汇，作者标注了大量经常

表 5b 词语类比（低频词）

模型	Accuracy				Mean Rank			
	所有	首都	城市	家庭	所有	首都	城市	家庭
CBOW	72.10	80.98	71.43	55.88	2631.21	400.62	4964.06	5312.66
义项不敏感（知网）	72.28	81.35	71.42	55.87	2489.47	380.76	4670.86	5039.82
义项敏感（知网）	72.32	81.40	71.45	55.90	2433.83	374.97	4578.61	4914.27
义项敏感（语料）	72.31	81.37	71.44	55.89	2436.38	374.98	4579.18	4922.84
CBOW(L)	**76.18**	85.69	**77.14**	**57.72**	**1095.30**	18.40	**2133.39**	**2446.59**
CBOW(L+H)	75.97	**86.08**	76.05	56.99	1156.23	**15.04**	2397.83	2497.16
CBOW(L+H+R)	75.44	85.69	76.10	55.88	1275.03	16.49	2458.50	2873.38
SG	76.80	84.51	86.29	56.25	1509.53	36.09	541.41	4895.10
SG(L)	80.88	89.41	88.00	60.29	**7.80**	1.24	2.69	**23.40**
SG(L+H)	81.30	90.00	**88.57**	**60.30**	8.85	1.23	**2.29**	27.35
SG(L+H+R)	**81.71**	**91.18**	88.10	59.93	28.65	**1.15**	4.57	95.71
DSG	81.40	91.57	88.00	58.09	159.12	1.23	39.17	532.34
DSG(L)	**85.27**	92.94	**92.00**	**66.54**	**1.74**	1.11	1.62	**3.01**
DSG(L+H)	84.12	93.73	88.57	63.24	1.84	1.10	**1.34**	3.54
DSG(L+H+R)	85.16	**93.92**	**92.00**	64.34	2.83	**1.09**	2.13	6.53
GloVe	68.23	74.71	87.43	43.75	2833.23	528.32	2.76	8976.01
GloVe(L)	76.28	**83.53**	90.86	53.31	196.62	7.16	1.22	677.58
GloVe(L+H)	**77.01**	84.51	92.05	**53.41**	**187.95**	6.59	**1.19**	**648.15**
GloVe(L+H+R)	75.86	82.94	**92.10**	52.21	405.76	**3.73**	1.29	1419.78

表 6 实例观察

词语	中文维基百科词频	实例	关系	GloVe	GloVe(L+H+R)
慌	100	心慌意乱	相似性	0.45	0.72
		晕头转向	相关性	0.50	0.70
		心惊胆战	相关性	0.42	0.70
逃之夭夭	100	不见踪影	相似性	0.61	0.71
		杳无音信	相似性	0.48	0.65
		音信全无	相关性	0.42	0.63
必修课	99	选修课	相似性	0.54	0.76
		外语课	相似性	0.20	0.71
		课表	相关性	0.01	0.68

会在同一上下文语境中出现的词汇，这其中就包括了很多国家和首都。完全基于知网义原的方法在本任务中未能取得较好效果，分析其原因是因为知网的义原标注有如下特点：知网的义原标注并没有关注城市和省份之间的关系，举例而言，“南京”、“上海”等 605 个城市的 DEF 是一样的；类似的，“奶奶”“娘”“娘亲”“后娘”等义项的 DEF 也是一样的。

观察表 5a 可知，本文所提出的相关概念场加强模型（L+H+R）可以提升词向量在词语类比任务上的表现。观察表 5b 可知，在本文提出的模型更适合语义表示质量较低的词向量的二次训练（比如 GloVe），可能是由于各种模型之间在高频词词向量的质量上都相差不多，决定整体质量的因素集中在了低频词上，而我们的模型解决低频词问题，所以我们提升的空间就很大。这也从侧面证明了本文提出的模型不容易使得模型过拟合。对于高质量的词向量（比如 DSG）本文提出的模型对此有小幅度地提升。但是对于低质量的词向量（比如 GloVe），本文的模型可以大幅度地提升其语义表示质量。

4 低频词实例观察

本节中，我们给出若干实例，通过观察基线模型和本文所提模型对实例的打分，加深对实验结果的感性认识。如表 6 所示，我们给出了语料库中一部分相似词和相关词，使用余弦相似度近似语义相似度和语义相关度。观察得知，本文提出的模型所预测的余弦相似度具有更高的合理性。

经过进一步地观察和分析，我们发现经典的 GloVe 模型对于低频词的训练效果较差，而本文提出的模型通过借助于知网相关概念场，使得低频词能够参照高频词来提升自身的向量表示质量。在中文维基百科语料库中出现频次小于等于 100 的词有 315866 个，占总体的 75.95%。可见我们的模型使得绝大多数词的获益是巨大的。

5 结束语

本文中，我们提出了一种借助于知网相关概念场来提升词向量质量的方法，该方法可以对基于计数模型和预测模型等多种框架生成的词向量进行二次训练。针对低频词词向量语义表示质量较差的问题，我们提出了三种模型，并且在词语相关度、词语相似度和词语类比任务上，验证了本文所提出模型的有效性。我们还深入到词向量当中，使用余弦距离来评价语义相似词或语义相关词的表示质量，进一步地证明了本文所提出方法的有效性。

此外，我们还提出了一个侧重于评价词语上下文语义相关度的数据集 ws240r，细粒度地评价词向量的相关性语义表示质量。

在后续的研究中，我们将会借助于外部评价手段，也即在词向量的下游任务上评测我们的方法的有效性，比如文本分类任务、词性标注任务等等。与此同时，我们不仅要考虑词语间的语义关系，也要把词的属性信息结合进来。我们会探索把知网义原和知网相关概念场同时加入到词表示的学习当中来，使得两者能够有机结合，从而迸发出更强大的知识支撑的力量。

参考文献

[1] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[2] Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning[C]. international conference on machine learning, 2008: 160-167.

[3] Turney P D, Pantel P. From frequency to meaning: vector space models of semantics[J]. Journal of Artificial Intelligence Research, 2010, 37(1): 141-188.

[4] Collobert R , Weston J , Bottou, Léon, et al. Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.

[5] Mikolov T, Le Q V, Sutskever I. Exploiting Similarities among Languages for Machine Translation[J]. Computer Science, 2013.

[6] Weston J, Bordes A, Chopra S, et al. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks[J]. international conference on learning representations, 2016.

[7] Harris Z S. Distributional Structure[J]. WORD, 1954, 10(2): 146-162.

[8] Cao S, Lu W, Zhou J, et al. cw2vec: Learning Chinese Word Embeddings with Stroke n-gram Information[C]. national conference on artificial intelligence, 2018: 5053-5061.

[9] 赵浩新，俞敬松，林杰. 基于笔画中文字向量模型设计与研究[J]. 中文信息学报, 2019, 33(05):22-28.

[10] Li Y , Li W , Sun F , et al. Component-Enhanced Chinese Character Embeddings[J]. Computer Science, 2015.

[11] Song Y, Shi S, Li J, et al. Joint Learning Embeddings for Chinese Words and their Components via Ladder Structured Networks[C]. international joint conference on artificial intelligence, 2018: 4375-4381.

[12] Yu J, Jian X, Xin H, et al. Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components.[C]. empirical methods in natural language processing, 2017: 286-291.

[13] Chen X, Xu L, Liu Z, et al. Joint learning of character and word embeddings[C]. international conference on artificial intelligence, 2015: 1236-1242.

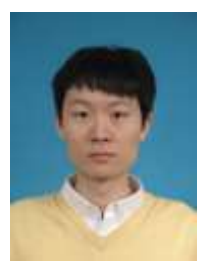
[14] Xu J, Liu J, Zhang L, et al. Improve Chinese Word Embeddings by Exploiting Internal Structure[C]. north american chapter of the association for computational linguistics, 2016: 1041-1050.

[15] Lin Z, Liu Y. Implanting Rational Knowledge into Distributed Representation at Morpheme Level[J]. national conference on artificial intelligence, 2018.

[16] Miller G A. WordNet: a lexical database for English[J]. Communications of The ACM, 1995, 38(11): 39-41.

[17] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for

structuring human knowledge[C]. international conference on management of data, 2008: 1247-1250.
[18] Ganitkevitch J, Van Durme B, Callisonburch C, et al. PPDB: The Paraphrase Database[C]. north american chapter of the association for computational linguistics, 2013: 758-764.
[19] Baker C F, Fillmore C J, Lowe J B, et al. The Berkeley FrameNet Project[C]. meeting of the association for computational linguistics, 1998: 86-90.
[20] Faruqui M, Dodge J, Jauhar S K, et al. Retrofitting Word Vectors to Semantic Lexicons[J]. north american chapter of the association for computational linguistics, 2015: 1606-1615.
[21] Song Y, Lee C, Xia F, et al. Learning Word Representations with Regularization from Prior Knowledge.[C]. conference on computational natural language learning, 2017: 143-152.
[22] Kiela D, Hill F, Clark S, et al. Specializing Word Embeddings for Similarity or Relatedness[C]. empirical methods in natural language processing, 2015: 2044-2048.
[23] Yu M, Dredze M. Improving Lexical Embeddings with Semantic Knowledge[C]. meeting of the association for computational linguistics, 2014: 545-550.
[24] Fried D, Duh K. Incorporating Both Distributional and Relational Semantics in Word Representations.[J]. international conference on learning representations, 2014.
[25] Liu Q, Jiang H, Wei S, et al. Learning Semantic Word Embeddings based on Ordinal Knowledge Constraints[C]. international joint conference on natural language processing, 2015: 1501-1511.
[26] Nguyen K A, Walde S S, Vu N T, et al. Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction[J]. meeting of the association for computational linguistics, 2016: 454-459.
[27] Wieting J, Bansal M, Gimpel K, et al. From Paraphrase Database to Compositional Paraphrase Model and Back[J]. Transactions of the Association for Computational Linguistics, 2015, 3(1): 345-358.
[28] Xu C, Bai Y, Bian J, et al. RC-NET: A General Framework for Incorporating Knowledge into Word Representations[C]. conference on information and knowledge management, 2014: 1219-1228.
[29] Bian J, Gao B, Liu T, et al. Knowledge-Powered Deep Learning for Word Embedding[C]. european conference on machine learning, 2014: 132-148.
[30] 董振东,董强.知网[C].//计算语言学文集.北京:1999:58～63.
[31] 唐共波,于东,荀恩东.基于知网义原词向量表示的无监督词义消歧方法[J].中文信息学报,2015,29(06):23-29.
[32] 孙茂松,陈新雄.借重于人工知识库的词和义项的向量表示:以 HowNet 为例[J].中文信息学报,2016,30(06):1-6+14.
[33] Niu Y, Xie R, Liu Z, et al. Improved Word Representation Learning with Sememes[C]. meeting of the association for computational linguistics, 2017: 2049-2058.
[34] 陈洋,罗智勇.一种基于 Hownet 的词向量表示方法[J].北京大学学报(自然科学版),2019,55(01):22-28.
[35] 朱靖雯,杨玉基,许斌, 等.基于 HowNet 的语义表示学习[J].中文信息学报,2019,33(3):33-41.
[36] 董强,董振东.基于知网的相关概念场的构建[C].//语言计算与基于内容的文本处理.北京:2003:364-370.
[37] 刘群,李素建. 基于《知网》的词汇语义相似度计算[A].台北:[出版者不详],2002:59-76.
[38] Mikolov T, Chen K, Corrado G S, et al. Efficient Estimation of Word Representations in Vector Space[J]. international conference on learning representations, 2013.
[39] Mikolov T , Sutskever I , Chen K , et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.
[40] 汪祥, 贾焰, 周斌, et al. 基于中文维基百科链接结构与分类体系的语义相关度计算[J]. 小型微型计算机系统, 2011, 32(11):2237-2242.
[41] Jin P, Wu Y. SemEval-2012 Task 4: Evaluating Chinese Word Similarity[C]. joint conference on lexical and computational semantics, 2012: 374-377.
[42] Pennington J, Socher R, Manning C D, et al. Glove: Global Vectors for Word Representation[C]. empirical methods in natural language processing, 2014: 1532-1543.
[43] Song Y, Shi S, Li J, et al. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings[C]. north american chapter of the association for computational linguistics, 2018: 175-180.



冯煜博（1993—），硕士研究生，主要研究领域为自然语言处理。

E-mail：networkprogramming@yeah.net

蔡东风（1958—），通信作者，博士，教授，主要研究领域为人工智能、自然语言处理、信息检索等。

E-mail：caidf@vip.163.com

宋彦（1981—），博士，主要研究领域为文本表征学习、自然语言处理、机器学习等。

E-mail：clksong@gmail.com