doi:10.6043/j.issn.0438-0479.201805013

《同义词词林》的嵌入表示与应用评估

段宇光[1,2],刘 扬[1,3*],俞士汶[1,3]

(1.北京大学计算语言学教育部重点实验室,2.北京大学元培学院,3.北京大学计算语言学研究所,北京 100871)

摘要:在自然语言处理中,嵌入表示是表达语言知识的重要途径和手段,以《同义词词林》为例,提出基于知识库训练嵌入表示的伪句式构造方法,并在多项任务上测试新方法的有效性.根据《同义词词林》词义编码反映的层级结构,将这些编码扩展为多种伪句式,并据此生成不同的伪语料库,采用 word2vec 模型在伪语料库上训练义素向量及词向量,得到 CiLin2Vec 资源,并应用于词义合成、类比推理和词义相似度计算等任务.在词义合成、类比推理任务上的准确率达到 90% 以上,超过了以往在语料库上训得的结果.证明该方法可以有效地将知识库中的理性知识注入嵌入表示中,也显示了 CiLin2Vec 嵌入表示资源在应用上的巨大潜力.

关键词:《同义词词林》;嵌入表示;词义合成;类比推理;相似度

中图分类号:TP 391 **文献标志码**:A **文章编号**:0438-0479(2018)06-0867-09

在机器智能时代,自然语言的理解和分析具有重要价值.在实现途径上,大体分为基于知识库的理性方法和基于语料库的经验方法.在理性方法方面,《同义词词林》(以下简称《词林》)作为汉语知识库的一个典范代表,由语言学家对汉语中的词进行划分、归类,形成语义上的层级结构,在词义相似度计算[1-3]、实体关系抽取[4-5]、语义角色标注[6]、文本分类[7]等多种任务或应用中有广泛影响;在经验方法方面,建立在语料统计分析上的分布式表示也在不断发展,早期基于词共现矩阵获得词嵌入表示[8-10],后来通过前馈神经网络学习词嵌入的方法成为主流[11],并广泛应用于自然语言处理的多种任务或应用[12-14].

基于知识库的理性方法,解释性强,但一般情况下需要针对不同任务设计不同算法,在不同领域间的适用性较差.而基于语料库的经验方法,往往采用无监督训练,自动化程度高,获得的词向量可以适用于多种任务.因此,如何将两者的优势结合起来,采用经验方法在知识库中自动地提取词义信息,最大程度地重复使用已有的人类专家知识,获得适用于多种任务的嵌入表示,是一个较新的研究课题.

此前,有研究者注意到将理性知识注入以改善嵌入表示的潜在需求,挖掘 WordNet 图结构中简化的邻接关系信息[15]或者参照多部词典的释义条目信息[16],以此作为训练内容来获得词嵌入表示;也有人关注如何由已有的词嵌入表示获得义素嵌入表示和同义词集嵌入表示[17],以及通过建立词嵌入表示到同义词集的映射来提高词向量的解释性[18];另有一些工作希望在语料库训练中加入词义、句法知识以获得更有效的词嵌入表示[19];或者采用随机游走(random walk)方法利用知识库构建伪语料,再通过神经网络训练获得词嵌入表示[20].这些方法大多是在基于真语料库训练词向量,部分地加入或联结了知识库中的词义、句法知识.之后邻接关系信息及随机游走方法对此有所改进,不再依赖于真语料库的预训练,但在利用知识库知识构建嵌入表示或构建伪语料时采用了相对间接、繁琐的手段,其生成过程较为复杂.除此之外,针对一般的知识库资源,目前也没有相对直接的应对策略和解决方法.

本研究使用哈尔滨工业大学研发的《词林》扩展版(http://www.ltp-cloud.com/download)为知识本体,提出并展示基于知识库训练嵌入表示的伪句式构造方法.根据《词林》词义编码的层级结构,将其扩展

收稿日期:2018-05-10 **录用日期**:2018-08-06

基金项目:国家重点基础研究发展计划(973 计划)(2014CB340504);国家社会科学基金重大项目(12&ZD119);国家社会科学基金(16BYY137)

* **通信作者**:liuyang@pku.edu.cn

引文格式:段宇光,刘扬,俞士汶.《同义词词林》的嵌入表示与应用评估[J].厦门大学学报(自然科学版),2018,57(6):867-875.

Citation:DUAN Y G,LIU Y,YU S W. An embedded representation for "Tongyici Cilin" and its evaluation on tasks[J]. J Xiamen Univ Nat Sci,2018,57(6):867-875. (in Chinese)

为词义描述式并构造 3 类伪句式:义素编码句式、义素编码扩展句式、词编码句式,以此生成符合理性知识分布规律的不同的伪语料库,在此基础上使用 word2vec 训练义素向量及词向量;考察不同训练模型、不同窗口大小在不同伪语料库上的训练效果,并将获得的向量分别应用于词义合成、类比推理和词义相似度计算等自然语言处理任务上.

1 研究基础与任务简介

1.1 《词林》知识表示

《词林》是由梅家驹编撰的汉语同义词或相关词的划分、归类词库[21],经哈尔滨工业大学社会计算与信息检索研究中心扩展后,目前共包含 77 343 个词、90 102 个义项,这些义项被分为 12 个大类、95 个中类、1 428 个小类、4 026 个词群和 17 797 个原子词群. 其大类编码为 1 位大写英文字母,中类编码在之后加 1 位小写英文字母,小类编码在之后加 2 位十进制整数,词群编码在之后加 1 位大写英文字母,原子词群编码在之后加 2 位十进制整数并附 1 位符号对分类结果作特别说明;符号"="代表该词群内的不同词为同义词,"#"代表该词群内的不同词为相关词,"@"代表该词群内只有一个词. 例如,原子词群编码"Aa01A01="代表具有特定义项的同义词集{人,士,人物,…}. 《词林》结构与编码如图 1 所示. 在下文中,以词义编码来泛指以上各类编码.

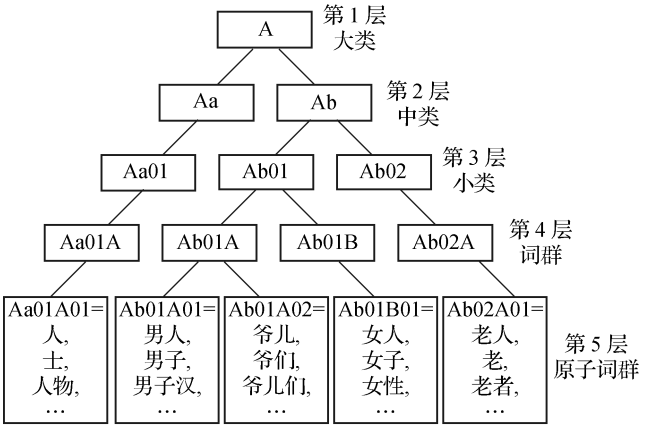

图 1 《词林》结构示意图

Fig. 1 The structure diagram of "Cilin"

1.2 分布式语义与分布式表示

分布式语义是一种数据驱动的语义分析,旨在对语料中的语义相似性进行量化和归类,它基于 Harris[22] 提出的著名的分布式假设,即"上下文相似的词,其语义也相似".

在此基础上,Hinton 等[23] 提出了词的分布式表示,又称嵌入表示(embedded representation),其思想源于如下认知看法:词义或称概念可以通过刻画它的各种属性来高效表示,而这些属性又同时与多个概念相关联,因此,一个概念可以通过这些属性的激活状态来表示. 这种表示方法显著区别于传统的独热表示(one-hot representation). 在形式上,独热表示使用向量的一个维度来表示不同的词,嵌入表示则用低维、稠密的实数向量来表示所有的词,如表 1 所示. 嵌入表示将词之间的语义关联进行了适当的编码,使近义词在大多数维度上都相近,因而比独热表示有更强的表达能力. 比如,即使只使用二值表示(即将每一维的取值限定为 0 或 1),长度为 n 的独热表示只能表示 n 个不同的概念,而嵌入表示则可以表示 2^n 个不同的概念[24].

表 1 独热表示与嵌入表示对比

Tab. 1 One-hot representation vs. embedded representation

词	独热表示	嵌入表示
猫	[0,0,…,0,1,0,0,…,0]	[0.14,…,0.61,…,−0.27]
狗	[0,0,…,0,0,1,0,…,0]	[0.18,…,0.71,…,−0.31]

1.3 词义合成任务

语义合成一直是自然语言理解关注的重点,使用有限单位组合出无限的含义,这是人类可以有效交流的重要原因[25]. 基于此,不少研究者致力于使用神经网络训得的词向量合成表示短语、句子等更大语言单位的向量[14,26-28]. 但是,建立在神经网络模型上的语义合成不易捕获和解释,这仍是计算和认知科学中反复探讨的一个未解难题[29-30].

此前,有关语义合成的研究大多将注意力放在词以上的语言单位上,鲜有学者关注更基本的语言层级上的语义合成问题. 事实上,词并不是语言中的基本意义单位,文献[31]中指出,语言中的一个基本语义单位是义位,相当于词的一个义项表达,通过分解义位可以进一步得到最小的义素单位. 比如,男人="人"×"男性"×"成年"[32],其中,"人"、"男性"、"成年"都是最小的义素单位. 基于此,本研究提出一种由词以下单位进行词义合成的任务,即由义素向量合成词向量的测试. 在本研究中,将以《词林》为例衡量该任务,其测试集由《词林》中的所有词及其词义编码构成.

1.4　类比推理任务

类比推理任务由 Mikolov 等[33]提出,目的在于用词向量来预测句法和语义的关联性. 比如,一个标准的表述形式如"男人:女人::父亲:w_i",在理想状态下,词 w_i 的词向量 $\boldsymbol{V}$ 可通过"男人""女人""父亲"的词向量的加、减运算得到,即 $\boldsymbol{V}(w_i)=\boldsymbol{V}(\text{女人})-\boldsymbol{V}(\text{男人})+\boldsymbol{V}(\text{父亲})$. 在类比推理任务中,人工预先给定 w_i 的理想答案,计算给定词的词向量与理想词向量的夹角余弦,以此评价词嵌入的实际效果.

Chen 等[34]给出了类比推理任务集(https://github.com/Leonard-Xu/CWE),其中包含 3 种类型,共计 953 组推理,包括:首都与国家 506 组,州/省与城市 175 组,亲属关系 272 组. 在《词林》中,实际包含该任务集中的 921 组,包括:首都与国家 506 组,州/省与城市 175 组,亲属关系 240 组.

1.5　词义相似度计算任务

词义相似度计算是同义词检测、歧义消解、信息抽取等任务或应用的基础,其计算方法分为 2 种[35]:一种是利用语料进行统计分析,将词频及分布等情况作为词义相似度计算的依据[36],其结果依赖于选取的语料库[37],目前常用神经网络模型获得词向量,并依据夹角余弦计算词义相似度;另一种方法是通过发掘知识库中概念之间的共性与差异性,以此来评估词义相似度[38],包括基于路径、特征、信息内容和利用概念注释等不同方法[39].

汉语中,常用的词义相似度计算任务集包括 MC30(https://github.com/huyingxi/Synonyms/blob/master/VALUATION.md)和 wordsim297(https://github.com/thunlp/SE-WRL/blob/master/datasets/wordsim-297.txt). 测试者使用计算模型对测试集中限定的词对进行相似度评分,并与人工判定标准做比较,通常使用皮尔森相关系数 r,对模型方法的有效性进行评价.

2　《词林》的嵌入表示方法

2.1　《词林》结构的调整

在《词林》层级结构中,每一层上的词义编码并没有明确标出词义的分类特征与取值. 但是,在描写词义时,每增加一层编码,客观上都会对意义表达产生进一步的约束和限定. 因此,可以将每层新增的编码信息视为构成词义的一个新增义素,而低层的词义编码中,则包含了此上各层的义素信息. 换言之,每个词义可以等价于一组义素的组合. 此外,在《词林》中,所有的词都分布在叶子节点上,其词义描写程度一样,但这并不符合语言事实. 实际上,每个词的语义颗粒度不同,颗粒度大的应位于较高层节点,而颗粒度小的应位于较低层节点. 基于以上看法,对《词林》结构进行调整.

考虑到位于群首的词往往能表征该原子词群的一般含义,其代表程度较高,颗粒度也较大,按如下方法进行《词林》结构的调整:由下至上,依次将低层中每个编码对应的首词汇集起来并挂在上一层的父节点下,从而使高层编码也有对应的词集,并通过高层词集中的所有词的共性来反映特定编码的义素信息. 最终不同抽象程度的词均获得了不同的语义颗粒度描写. 整理后的《词林》结构如图 2 所示.

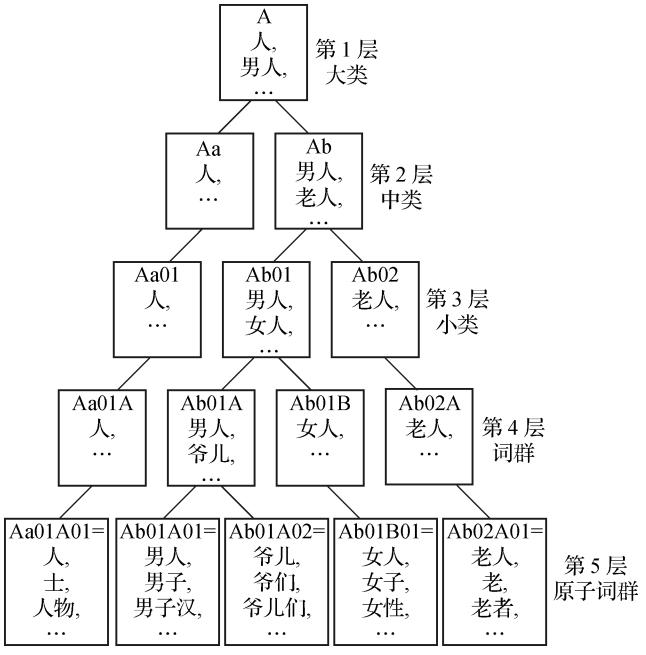

图 2　调整后的《词林》结构示意图

Fig. 2　The adjusted structure diagram of "Cilin"

2.2　基于《词林》的伪句式构造

神经网络训练依据词在上下文中的分布信息来捕捉词义,因此,使用该方法在《词林》中提取词义,就需要依据其中的知识描述来构造上下文分布合理的伪句子和伪语料库.

整理后的《词林》层级结构中共有 23 570 个节点,每个节点代表的概念都不相同. 利用每个词的词义编码信息构造 3 类伪句式,即:义素编码句式、义素编码扩展句式、词编码句式. 由于《词林》中的词义编码代表了该层上的概念含义,在造句时,依照层级结构确定编码和词的距离具有分布合理性,适合用 word2vec

模型来训练义素向量及词向量.3 类伪句式的定义如下,相应的示例如表 2 所示.

表 2 《词林》中不同句式示例

Tab. 2 Examples of different sentence templates for "Cilin"

句式	示 例
义素编码句式	<BOS>A A A A A 人 A A A A A <EOS> <BOS>A Aa Aa Aa Aa 人 Aa Aa Aa Aa A <EOS> <BOS>A Aa Aa01 Aa01 Aa01 人 Aa01 Aa01 Aa01 Aa A <EOS> <BOS>A Aa Aa01 Aa01A Aa01A 人 Aa01A Aa01A Aa01 Aa A <EOS> <BOS>A Aa Aa01 Aa01A Aa01A01 = 人 Aa01A01=Aa01A Aa01 Aa A <EOS> ……
义素编码扩展句式	<BOS>A A A A A 人 A A A A A <EOS> <BOS>A A A A A 士 A A A A A <EOS> …… <BOS>A A A A A 翁 A A A A A <EOS> <BOS>A Aa Aa Aa Aa 人 Aa Aa Aa Aa A <EOS> <BOS>A Aa Aa Aa Aa 士 Aa Aa Aa Aa A <EOS> …… <BOS>A Aa Aa Aa Aa 翁 Aa Aa Aa Aa A <EOS> ……
词编码句式	<BOS>人 男人 高个儿 居民 职工 劳动者 健康人 亲戚 鼻祖 朋友 英雄 知识分子 教徒 反动派 人 反动派 教徒 知识分子 英雄 朋友 鼻祖 亲戚 健康人 劳动者 职工 居民 高个儿 男人 人 <EOS> <BOS>人 男人 高个儿 居民 职工 劳动者 健康人 亲戚 鼻祖 朋友 英雄 知识分子 教徒 反动派 人 我 你 他 自己 谁 人 谁 自己 他 你 我 人 反动派 教徒 知识分子 英雄 朋友 鼻祖 亲戚 健康人 劳动者 职工 居民 高个儿 男人 人 <EOS> ……

1) 义素编码句式:根据义素的编码构造伪句式,每个词的所有祖先节点编码构成代表该词义的义素组合,依据祖先节点在层级结构中与该词的距离,确定该祖先节点编码在句中与该词的距离.句式呈回文数形,词前后均有 5 个编码,如果编码不足 5 个,则将距离该词最近的编码复制多次进行占位处理.这样的造句方式,保证句长固定且前后对称,同时满足连续词袋(CBOW)方法和跳字(Skip-Gram)方法对窗口词形式的要求.

2) 义素编码扩展句式:依据不同的词义相似度计算方法,预先对每个词筛选出和该词相似度达到特定阈值的近义词集,并将义素编码句式中的词依次代换为其近义词集中的其他词,以此扩大伪语料库的规模.这种句式实质上是借助已有的理性方法,提升近义词在伪语料库中的分布相似度,从而使依据分布信息训得的词向量能够析取近义词.本研究中采用田久乐等[1]的、吕立辉等[2]、朱新华等[3]提出的词义相似度计算方法 .比如"人"的近义词集,如果采用田久乐等[1]的算法,在特定的相似度阈值设定下,则包括{人,士,人物,人士,人氏,人选,人类,生人,全人类,人口,口,食指,翁}等词.

3) 词编码句式:将义素编码句式中的每个义素编码替换为该编码词集中的所有词.在这种句式假定下,每个义素编码代表的义素信息可以通过该编码词集中的所有词的共性反映出来,也由此代表了该义素信息.这种句式的句长不固定,但前后依然对称.

2.3 《词林》的嵌入表示训练

Word2vec 模型基于上下文对词进行概率预测,包括 CBOW 和 Skip-Gram 方法,它可以从大量无标的语料库中学习词的嵌入表示.其中,CBOW 根据当前词 w_i 上下文的词向量表示求和或平均后,直接预测 w_i;而 Skip-Gram 则与 CBOW 对称,使用当前 w_i 预测其上下文中的每一个词.

本研究利用 gensim 自然语言处理库中的 word2vec 模块(https://github.com/RaRe-Technologies/gensim),使用 CBOW 和 Skip-Gram 方法在 3 种伪语料库上进行平行训练,完全不借助于任何其他真语料库,并考察不同窗口词大小对训练结果的影响.

值得注意的是,在构造义素编码句式和义素编码扩展句式时,《词林》中各层的词义编码在伪语料库中都有分布,并且与词的出现形成合理的分布关系,经过 word2vec 模型训练,可以同时获得针对《词林》嵌入表示的义素向量和词向量.并且本研究也在中文维基百科语料(https://dumps.wikimedia.org)上训练词向量,用于相关任务的效果对比与验证.

3 嵌入表示结果的应用评估

3.1 词义合成任务评估

由于《词林》的词义编码中包含了各层的义素信息,词义等价于一组义素的组合,理论上,可以将一个词的词向量替换成一组义素向量的归一化求和结果,

以此考察义素向量在词义合成任务上的表现.在本研究中,采取如下公式来计算义素合成的词向量:

$$w_1 = \sum_{i=1}^{n} s_i,$$

$$\boldsymbol{V}(w_1) = \sum_{i=1}^{n} \alpha_i \boldsymbol{V}(s_i),$$

$$\sum_{i=1}^{n} \alpha_i = 1,$$

其中,w_1为所要计算的词,s_i为与词义相关的义素,$\boldsymbol{V}(x)$为义素向量或词向量,α_i为权重参数.权重参数按义素所处的层级位置,采用等比递减或等比递增等不同方法,即:$\alpha_{i+1}=0.5\alpha_i$或$\alpha_{i+1}=2\alpha_i$.

通过计算义素向量合成的词向量和原词向量的余弦相似度,可以评价词义合成任务的有效性.由于多义词有多种义素编码表达式,进而生成多种义素合成的词向量,在任务评估时,对每个词对,需取得和原词向量余弦相似度最高的一组义素合成的词向量.在该任务上,本研究使用 CBOW 和 Skip-Gram 方法在不同句式、不同窗口词大小下的训练结果如表 3 所示.

其中,扩展句式取 $\alpha_{i+1}=0.5\alpha_i$ 和 $\alpha_{i+1}=2\alpha_i$ 两种权重分配中得分较高的一种,最优模型均采用 $\alpha_{i+1}=0.5\alpha_i$.依据 3 种相似度计算算法的不同特点,扩展句式的相似度阈值 ρ 分别定为:$\rho_{[1]}=0.89$,$\rho_{[2]}=0.65$,$\rho_{[3]}=0.84$,训练中迭代次数为 5,词向量维度为 300,最小词频为 0,其他参数取默认值.

从实验结果可以看出,Skip-Gram 训练效果普遍优于 CBOW,较合适的窗口大小为 3～4.使用义素编码句式效果最优,达到 95.84%,表明义素信息实现了成功注入,可以有效地用义素合成的词向量来表征原词向量.这也说明《词林》对词义的分层描述具有一定的合理性,生成的伪句子在分布上依然保持了这种性质,经训练获得的义素向量和原词向量之间存在合成关系.在加入了理性算法后,合成效果反而有所下降,可能原因是扩展的句子采用了近义词,给语料带来了噪音,这也反过来说明理性算法与知识确实被注入进去了.

总体来说,《词林》知识的采用,在语义合成任务上具有显著优势.

3.2　类比推理任务评估

对于类比推理任务,使用 CBOW 和 Skip-Gram 方法在不同句式、不同窗口词大小下训练得到的结果如表 4 所示.

其中,扩展句式取 $\alpha_{i+1}=0.5\alpha_i$,扩展句式的相似度阈值 ρ 同上,训练模型参数同上,维基百科语料训练中的最低词频为 3,其他模型参数和伪语料库相同.

可以看出,Skip-Gram 效果更好,最佳句式为义素编码句式,使用该句式的义素合成的词向量成绩达到 94.37%,其效果明显优于原词向量.该结果进一步说明,《词林》知识的采用,可以有效实现词义合成,并将义素合成的词向量应用于其他任务上.在模型参数相同的条件下,在伪语料库上训得的词向量的效果优于在维基百科上训得的词向量.可能原因在于:与词单位相比,知识库中的义素单位不存在歧义,且有不重不漏的特性;此外,《词林》中的词义描述式格式整齐,在此基础上生成的伪句式分布具有规范性,句式生成过程中可以人为地控制信息分布,减少噪音,而语料库往往带有无法消解的歧义和噪音问题.

总体来说,使用新方法获得的《词林》嵌入表示在类比推理任务上具有显著优势,且普遍优于 Chen 等[34]报道的 72.99%的最好效果.

3.3　词义相似度计算任务评估

对于词义相似度计算任务,上述不同来源的词向量在 MC30、wordsim297 测试集上的相似度评分,以及与人工判定标准比较的皮尔森相关系数 r 评分,分别如表 5 和表 6 所示.

表 3　词义合成任务评估:义素合成的词向量与原词向量的余弦相似度

Tab.3　Evaluation results on semantic compositionality task: cosine similarty between sememe-vector-combined word vector and original word vector　　%

模　型	CBOW					Skip-Gram				
	3	4	5	6	7	3	4	5	6	7
义素编码句式	85.62	87.70	**88.31**	87.66	87.24	95.01	**95.84**	95.82	95.62	95.37
义素编码扩展句式[1]*	**73.61**	70.67	69.25	68.29	65.73	**82.21**	80.71	78.11	77.86	77.14
义素编码扩展句式[2]*	**64.31**	63.32	60.95	58.98	58.29	**78.62**	76.78	75.01	74.29	74.33
义素编码扩展句式[3]*	62.86	**63.29**	60.68	58.83	58.46	**79.60**	78.02	75.73	75.39	75.01

注:表中 3～7 表示窗口词大小,* 分别表示采用相应文献中的词义相似度计算方法进行计算,下同.

表 4 类比推理任务评估:推理词向量与标准词向量的余弦相似度

Tab. 4 Evaluation results on analogical reasoning task: cosine similarty between analogical word vector and correct word vector %

模 型	CBOW					Skip-Gram				
	3	4	5	6	7	3	4	5	6	7
维基百科语料	80.83	81.39	81.67	81.91	**81.93**	80.60	80.99	81.42	81.58	**81.84**
义素编码句式	64.62	65.04	**66.24**	63.46	64.65	92.99	92.89	92.93	**93.32**	93.16
义素合成的词向量	**95.12**	93.76	92.87	93.28	93.74	**94.37**	92.95	93.07	92.43	92.43
义素编码扩展句式[1]*	57.20	60.54	56.92	**61.14**	58.56	79.43	79.28	80.27	**81.04**	80.89
义素编码扩展句式[2]*	**85.23**	83.80	81.24	80.93	81.27	92.00	**92.58**	92.04	91.97	92.05
义素编码扩展句式[3]*	83.36	**86.18**	84.19	84.89	85.09	93.30	93.40	93.33	93.07	**93.46**
义素合成的词向量[1]*	91.02	**91.92**	91.74	89.53	90.95	92.00	**92.24**	91.96	91.70	92.05
义素合成的词向量[2]*	**92.64**	89.85	89.58	88.66	91.78	91.06	91.93	**92.39**	91.53	92.31
义素合成的词向量[3]*	**90.31**	90.30	88.76	88.73	89.50	92.43	91.50	92.37	92.54	**92.67**
词编码句式	77.88	78.19	**79.89**	78.04	76.64	74.66	76.12	74.88	76.60	**77.04**

表 5 MC30 词义相似度计算任务评估:r

Tab. 5 Evaluation results on word similarity measurement task(MC30): r %

模 型	CBOW					Skip-Gram				
	3	4	5	6	7	3	4	5	6	7
维基百科语料	65.99	64.24	64.41	**67.09**	65.72	**67.08**	65.50	66.09	65.07	66.20
义素编码句式	22.31	11.02	27.88	24.13	**39.19**	42.88	60.82	70.92	**74.39**	70.71
义素合成的词向量	67.50	59.47	62.64	67.43	**69.99**	74.35	70.36	76.10	**77.65**	77.02
义素编码扩展句式[1]*	20.12	2.630	**33.32**	14.95	−0.05	63.83	63.17	77.05	75.60	**77.42**
义素编码扩展句式[2]*	**46.23**	42.03	22.16	19.16	32.72	**73.20**	65.09	62.70	70.88	70.41
义素编码扩展句式[3]*	26.08	29.83	**40.52**	18.80	13.82	77.97	**84.73**	80.60	82.89	79.18
义素合成的词向量[1]*	74.59	70.47	76.65	73.83	**77.39**	84.86	81.39	84.60	82.04	**84.95**
义素合成的词向量[2]*	75.26	62.01	64.13	74.62	**76.46**	81.23	77.88	73.19	80.59	**82.08**
义素合成的词向量[3]*	67.41	78.59	63.38	71.64	**79.27**	68.84	82.00	81.99	81.19	**82.50**
词编码句式	13.59	37.83	38.49	36.86	**44.70**	63.11	69.18	**78.85**	75.54	69.61

其中,扩展句式取 $\alpha_{i+1}=2\alpha_i$,扩展句式的相似度阈值 ρ 同上,训练模型参数同上,维基百科语料训练的模型参数同上.《词林》中包含 wordsim297 中的 277 个词对,最后评分以这 277 个词对为标准,受最低词频限制,维基百科训练结果中仅包括 wordsim297 中的 277 个词对,表 6 中相应为这 277 个词对上的得分.

在词义相似度的计算任务上 Skip-Gram 效果更好,最佳窗口大小是 7.义素合成的词向量比原词向量的表现要好,再次证明应用《词林》中的词义合成性可以提高相关任务的性能.加入理性算法的扩展句式后进一步提升了其性能,其中,r 最高的是加入了田久乐等[1]提出的相似度计算算法,其义素合成的词向量达到了 84.95%,表明理性方法在训练过程中被成功注入,在近义词的嵌入表示中得到了体现.考查初始的理性方法,文献[1-3]中的词义相似度计算方法在 MC30 上的 r 分别为 49.39%,74.03% 和 79.24%,在 wordsim297 上的 r 分别为 35.53%,34.11% 和 42.22%,新方法获得的《词林》嵌入表示的效果普遍更好,优于传统的知识库理性方法并可能接近《词林》知识表示的能力上限.

和维基百科训练结果相比,在迭代次数等模型参数相同的情况下,新方法获得的《词林》嵌入表示在

表 6　wordsim297 词义相似度计算任务:r

Tab. 6　Evaluation results on word similarity measurement task(wordsim297):r　　%

模　型 窗口大小	CBOW					Skip-Gram				
	3	4	5	6	7	3	4	5	6	7
维基百科语料	60.76	62.32	62.77	63.92	**64.53**	58.09	58.93	**59.86**	59.79	59.25
义素编码句式	7.22	10.38	14.71	13.97	**18.76**	26.22	32.03	32.60	**37.72**	33.19
义素合成的词向量	32.00	**32.28**	29.71	28.03	29.44	32.74	32.28	32.97	**33.62**	32.53
义素编码扩展句式[1]*	**21.32**	5.40	8.64	4.38	2.05	30.80	34.07	**40.27**	38.65	38.29
义素编码扩展句式[2]*	**16.36**	15.86	12.18	1.04	5.74	37.07	39.22	**39.42**	36.57	39.05
义素编码扩展句式[3]*	**23.39**	16.88	6.06	3.77	7.74	37.85	**39.22**	38.09	33.01	38.29
义素合成的词向量[1]*	**35.34**	28.95	34.15	31.76	32.01	36.33	38.97	38.71	**41.75**	38.26
义素合成的词向量[2]*	36.25	32.70	38.26	38.23	**39.45**	40.66	42.01	40.71	**42.74**	41.91
义素合成的词向量[3]*	34.25	35.88	34.58	38.50	**39.40**	39.40	41.17	40.73	38.71	**43.48**
词编码句式	27.70	27.23	28.17	25.18	**32.46**	34.61	36.40	39.02	**39.19**	38.62

MC30 测试集上超过了维基百科,而在 wordsim297 上则落后于维基百科. 此外,有意思的是,和 MC30 相比,wordsim297 上理性方法计算得到的结果与新方法得到的结果都表现出随迭代次数大致相同的下降趋势,这或许与 MC30 选词特殊及样本过小等因素有关.

总体来说,在词义相似度计算任务上,语料库上的训练结果更加稳定.《词林》嵌入表示在 wordsim297 上表现不佳,有可能是因为《词林》知识表示与数据本身存在先天的局限性,比如在颗粒度表达问题或者语义分类不合理等.

4　结论及展望

本研究以《词林》为知识本体,提出并展示了基于知识库训练嵌入表示的伪句式构造方法,考察不同训练模型、不同窗口大小在不同伪语料库上的表现,并分别应用于词义合成、类比推理和词义相似度计算等自然语言处理任务上. 实验结果表明,新获得的义素向量及词向量资源 CiLin2Vec 在不同任务上都取得了进展或突破. 其中,在词义合成和类比推理任务上表现突出,准确率达到 90% 以上,显示该方法在应用上的巨大潜力. 本研究也将《词林》的 CiLin2Vec 嵌入表示资源发布在网络上(https://github.com/ariaduan/CiLin2Vec),以方便科研和业界验证、使用、推广.

在性质上,该方法有效复用已有的知识库资源,利用句式构造控制向嵌入表示中注入的理性知识,并借鉴已有的计算方法进行预处理,发掘理性知识和计算方法结合的最优方式,这些做法易于理解,有很强的解释性;该方法在训练过程中完全不使用真语料库,基于知识库生成伪语料的方式更加直接、简便,降低了获得嵌入表示的复杂度,极大地缩短了训练周期.

在未来,针对其他各类知识库,希望探究该方法的通用模型与一般特征,考察知识库上训得的词向量与语料库上训得的词向量的联合应用,并由此形成对不同资源的知识表示及数据特点的评价. 这些观点和方法,也将支持用于描述汉语语素及构词意义的北京大学《汉语概念词典》的研究与开发.

参考文献:

[1] 田久乐,赵蔚. 基于同义词词林的词相似度计算方法[J]. 吉林大学学报(信息科学版),2010,28(6):602-608.

[2] 吕立辉,梁维薇,冉蜀阳. 基于《词林》的词相似度的度量[J]. 现代计算机(专业版),2013,1:3-6.

[3] 朱新华,马润聪,孙柳,等. 基于知网与《词林》的词语义相似度计算[J]. 中文信息学报,2016,30(4):29-36.

[4] 刘丹丹,彭成,钱龙华,等.《同义词词林》在中文实体关系抽取中的作用[J]. 中文信息学报,2014,28(2):91-99.

[5] 徐庆,段利国,李爱萍,等. 基于实体词义相似度的中文实体关系抽取[J]. 山东大学学报(工学版),2015,45(6):7-15.

[6] 李国臣,吕雷,王瑞波,等. 基于同义词词林信息特征的语义角色自动标注[J]. 中文信息学报,2016,30(1):101-108.

[7] 王东,熊世桓. 基于同义词词林扩展的短文本分类[J]. 兰州理工大学学报,2015,4:104-108.

[8] DEERWESTER S, DUMAIS S T, FURNAS G W et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391- 407.

[9] SCHÜTZE H. Dimensions of meaning[C]// Proceedings of the 1992 ACM/IEEE Conference on Supercomputing. California: IEEE, 1992: 787-796.

[10] LUND K, BURGESS C. Producing high-dimensional semantic spaces from lexical co-occurrence[J]. Behavior Research Methods, Instruments, & Computers, 1996, 28(2): 203-208.

[11] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning[C]// International Conference on Machine Learning. Helsinki: ACM, 2008: 160-167.

[12] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.

[13] TURNEY P D. Domain and function: a dual-space model of semantic relations and compositions[J]. Journal of Artificial Intelligence Research, 2012, 44: 533-585.

[14] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation[C]// Conference on Empirical Methods on Natural Language Processing. Doha: Association for Computational Linguistics, 2014: 1532-1543.

[15] BARTUSIAK R, AUGUSTYNIAK Ł, KAJDANOWICZ T, et al. WordNet2Vec: corpora agnostic word vectorization method[J]. Neurocomputing, 2017. doi: 10. 1016/j. neucom. 2017. 01. 121.

[16] TISSIER J, GRAVIER C, HABRARD A. Dict2vec: learning word embeddings using lexical dictionaries[C]// Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 254-263.

[17] ROTHE S, SCHÜTZE H. AutoExtend: extending word embeddings to embeddings for synsets and lexemes[EB/OL]. [2018-04-20]. http: // arxiv. org/pdf/1507. 0112701. pdf.

[18] PANCHENKO A. Best of both worlds: making word sense embeddings interpretable[C]// Edition of the Language Resources and Evaluation Conference. Portorož: ELRA, 2016: 2649-2655.

[19] YANG L, SUN M. Improved learning of Chinese word embeddings with semantic knowledge[M]// Chinese computational linguistics and natural language processing based on naturally annotated big data. Switzerland: Springer, 2015: 15-25.

[20] GOIKOETXEA J, SOROA, AGIRRE E. Random walks and neural network language models on knowledge bases[C]// Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL. San Diego: ACL, 2015: 1434-1439.

[21] 梅家驹, 竺一鸣, 高蕴奇, 等. 同义词词林[M]. 上海: 上海辞书出版社, 1983: 1-362.

[22] HARRIS Z. Distributional structure[J]. Word, 1954, 10(2): 146-162.

[23] HINTON G E, MCCELLAND J L, RUMELHART D E. Distributed respresentations[M]// RUMELHART D E, MCCLELLAND J L. Parallel distributed processing: explorations in the microstructure of cognition(volume 1). Cambridge: MIT, 1986: 77-109.

[24] 孙飞, 郭嘉丰, 兰艳艳, 等. 分布式单词表示综述[J]. 计算机学报, 2016, 39: 1-22.

[25] CHOMSKY N. Three models for the description of language[J]. IRE Transactions on Information Theory, 1956, 2(3): 113-124.

[26] YESSENALINA A, CARDIE C. Compositional matrix-space models for sentiment analysis[C]// Conference on Empirical Methods on Natural Language Processing. Edinburgh: Association for Computational Linguistics, 2011: 172-182.

[27] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrixvector spaces[C]// Conference on Empirical Methods on Natural Language Processing. Jeju Island: Association for Computational Linguistics, 2012: 1201-1211.

[28] GREFENSTETTE E, DINU G, ZHANG Y Z, et al. Multi-step regression learning for compositional distributional semantics[EB/OL]. (2013-01-29)[2018-04-01]. http: // cn. arXiv. org/abs/: 1301. 6939.

[29] FODOR J A, PYLYSHYN Z W. Connectionism and cognitive architecture: a critical analysis[J]. Cognition, 1988, 28(1/2): 3-71.

[30] GERSHMAN S, TENENBAUM J B. Phrase similarity in humans and machines[C]// Proceedings of the 37th Annual Conference of the Cognitive Science Society. Cambridge: MIT, 2015: 776-781.

[31] VAKULENKO S. The notion of sememe in the work of Adolf Noreen[J]. Henry Sweet Society for the History of Linguistic Ideas Bulletin, 2005(44): 19-35.

[32] LYONS J. Linguistic semantics[M]. Cambridge: Cambridge University Press, 1996.

[33] MIKOLOV T, YIH W T, ZWEIG G. Linguistic regularities in continuous space word representations[C]// Proceeding

of the 2013 Conference of the North American Chapter of the ACL. Atlanta: Association for Computational Linguistics,2013:746-751.

[34] CHEN X,XU L,LIU Z,et al. Joint learning of character and word embeddings[C]// Proceedings of IJCAI. Buenos Aires:AAAI,2015:1236-1242.

[35] 葛斌,李芳芳,郭丝路,等. 基于知网的词汇语义相似度计算方法研究[J]. 计算机应用研究,2010,27(9):3329-3333.

[36] 石静,吴云芳,邱立坤,等. 基于大规模语料库的汉语词义相似度计算方法[J]. 中文信息学报,2013,27(1):1-6.

[37] LI Y, BANDAR Z A, MCLEAN D. An approach for measuring semantic similarity between words using multiple information sources[J]. IEEE Transactions on Knowledge and Data Engineering,2003,15(4):871-882.

[38] 梅立军,周强,臧路,等. 知网与同义词词林的信息融合研究[J]. 中文信息学报,2005,19(1):64-71.

[39] TAIEB M A H, AOUICHA M B, HAMADOU A B. Ontology-based approach for measuring semantic similarity [J]. Engineering Applications of Artificial Intelligence, 2014,36:238-261.

An Embedded Representation for "Tongyici Cilin" and Its Evaluation on Tasks

DUAN Yuguang[1,2], LIU Yang[1,3*], YU Shiwen[1,3]

(1. Key Laboratory of Computational Linguistics (Ministry of Education), Peking University,
2. Yuanpei College, Peking University, 3. Institute of Computational Linguistics, Peking University, Beijing 100871, China)

Abstract: In natural language processing (NLP), to learn embedded representation is an effective approach of capturing semantics from language resources. At present, however, this approach has been much limited to using large-scale corpora, with little attention to extracting rational knowledge from knowledge bases. In this paper, based on "Tongyici Cilin", a famous Chinese thesaurus, we present a method for implanting rational knowledge into embedded representation, then evaluate it in terms of different NLP tasks. According to the hierarchical encodings for morphemic and lexical meanings in "Tongyici Cilin", we design multiple templates to create instances as pseudo-sentences from these pieces of knowledge, and apply word2vec to obtain CiLin2Vec, the sememe and word embeddings of new kinds as for "Tongyici Cilin". For evaluation, tasks of semantic compositionality, analogical reasoning and word similarity measurement are taken into consideration. We make progress and breakthrough on the tasks, reaching an accuracy of over 90% for both semantic compositionality and analogical reasoning, demonstrating that the pieces of rational knowledge have been appropriately implanted, with very promising prospects for adoption of the knowledge bases.

Key words: "Tongyici Cilin"; embedded representation; semantic compositionality; analogical reasoning; similarity