《同义词词林》的嵌入表示与应用评估
（“自然语言处理”约稿）

段宇光[1,3]，刘 扬[1,2]，俞士汶[1,2]

1.北京大学 计算语言学教育部重点实验室, 2.北京大学 计算语言学研究所,

3.北京大学元培学院, 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，通常应用于语料库资源。本文以《同义词词林》为本体，提出一种基于知识库资源训练嵌入表示的新方法，并在多项任务上测试新方法的有效性。根据《词林》词义编码反映的词义层级结构，将每条编码扩展为多种句式，并据此生成不同的伪语料库，采用 Word2vec 模型训练义素向量及词向量资源 CiLin2vec，并应用于词义合成、类比推理和词义相似度计算等任务上。在这些任务中均取得了进展或突破，其中，在词义合成、类比推理任务上的准确率均达到 0.9 以上，超过了以往在语料库上训得的结果。证明新方法可以有效地将知识库中的理性知识注入分布式表示中去，也显示了新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，相似度计算

Creating Sememe and Word Embeddings as for CiLin

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

1. Key Laboratory of Computational Linguistics (Ministry of Education), Peking University,

2. Institute of Computational Linguistics, Peking University, 3. Yuanpei College, Peking University, Beijing 100871, China

Abstract: In Natural Language Processing (NLP), to learn distributed representation is an effective approach to capturing semantics from the language resources. However, by now, this has been much limited to using large-scale corpora, with little attention to extracting rational knowledge from knowledge bases. In this paper, based on CiLin, a famous Chinese thesaurus, we present a new method for implanting rational knowledge into distributed representation, then evaluate it on different NLP tasks. According to the hierarchical encodings for morphemic and lexical meanings in 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 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 over 0.9 for both semantic compositionality and analogical reasoning, which proves that the pieces of rational knowledge has been appropriately implanted and shows very promising prospects for adoption of the knowledge bases.

Key words: CiLin, embeddings, semantic compositionality, analogical reasoning, word similarity measurement

1 引言

在机器智能时代，自然语言的理解和分析具有重要的价值。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》）作为汉语知识库的典范代表，由语言学家对汉语中的词进行归类、划分，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]等多种任务中有广泛应用和影响。此外，建立在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也被提出并逐渐成为主流[11]，广泛应用于自然语言处理的多种任务[12-14]。

基于知识库的理性方法，反映了人类专家知识，解释性强，但需要针对不同任务人工设计不同算法，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，获得的词向量可以简单地适用于多种任务。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，最大程度地复用已有的人类专家知识，得到适用于多种任务的嵌入表示，是一个新的研究课题。此前，有个别工作注意到了这种需求，尝试使用 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。除此之外，针对一般的知识库资源，目前并没有系统的应对策略和解决方法。

本文使用《词林》扩展版[1]作为知识本体，提出并展示一种基于知识库训练义素向量及词向量的新方法，尝试生成这些理性知识的分布式表示，并将其注入到向量表达中。我们依据《词林》词义编码携带层级结构的特点，将其扩展为词义描述式并构造三类伪句式：义素编码句式、义素编码扩展句式、词编码句式，生成符合理性知识分布规律的不同伪语料库，在此基础上使用 Word2vec 训练义素向量及词向量。实验考察不同训练模型及不同窗口大小在不同伪语料库上的训练效果，并将得到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。实验结果表明，新得到的义素向量及词向量在不同任务中都取得了进展或突破，显示了该方法在应用上的巨大潜力。我们也将《词林》的嵌入表示资源 CiLin2vec 发布在网络上，以方便科研和业界使用、验证、推广。

2 相关研究基础

2.1 《词林》知识表示

《词林》是由梅家驹[17]编撰的汉语同义词和同类词划分词库，经哈工大社会计算与信息检索研究中心扩展后，目前共包含 77343 个词、90102 个义项，这些义项被划分为 12 个大类、95 个中类、1428 个小类、4026 个词群和 17797 个原子词群。其大类编码为一位大写英文字母，中类编码在之后加一位小写英文字母，小类编码在之后加两位十进制整数，词群编

基金项目：国家重点基础研究发展计划资助项目(2014CB340504)、国家社科基金一般项目(16BYY137)、 国家社科基金重大项目(12&ZD119)

[1] 本文中的数据采用哈尔滨工业大学社会计算与信息检索研究中心研发的《同义词词林》扩展版，下载自 http://www.datatang.com/data/42306/

码在之后加一位大写英文字母，原子词群编码在之后加二位十进制整数并附一位符号对分类结果进行特别说明，“=”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{人, 士, 人物, …}的编码为“Aa01A01=”，代表一个具有特定义项的所有词的集合。《词林》的结构与编码如图 1 所示。在后文中，我们以词义编码来泛指以上各类编码。

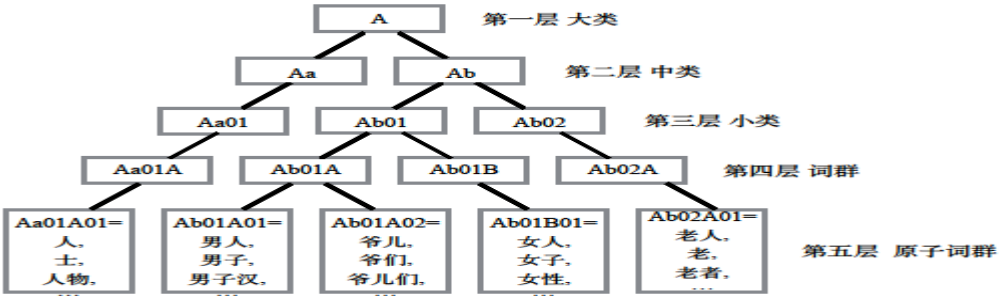

图 1 《词林》结构示意

Fig. 1 The structure of CiLin

2.2 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[18]提出的著名的语言分布式假设，即“上下文相似的词，其语义也相似”。

在此基础上，词的分布式表示的想法由 Hinton 等人[19]提出，其思想源于如下认知看法：一个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，这样，一个概念可以通过这些属性的激活状态来表示。独热表示使用向量的一个维度来表示不同的词，分布式表示则用低维的稠密实数向量来表示词，其区别如表 1 所示。分布式表示将词之间的语义关联进行编码，使近义词在大多数维度上相近，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，而分布式表示则可以表示 2^n 个不同的概念[20]。

表 1 独热表示与分布式表示对比

Tab. 1 One-hot representation vs. distributed representation

词	独热表示	分布式表示
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]

2.3 词义合成任务

语义合成性一直是自然语言处理关注的重点，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[21]。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短语、句子等更长语言单位的含义[22-25]。但是，建立在神经网络模型上的语义合成不宜捕获和解释，这仍是计算和认知科学中反复探讨的一个未解难题[26-27]。

此前，有关语义合成性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语言层级上的语义合成问题。事实上，词并不是语言中的最基本意义单位，瑞典语言学

家 Noreen[28]指出，语言中的一个基本语义单位是义位，相当于词的一个义面表达，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”[29]，其中，“人”、“男性”、“成年”都是最小的义素单位。基于此，我们提出一种由词以下单位进行词义合成的任务，即义素合成词义的测试。在本文中，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。

2.4 类比推理任务

类比推理任务由 Mikolov 等人提出[30]，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 的词向量可通过“男人”、“女人”、“父亲”的词向量的加、减运算得到，即 $vec(w_i) = vec(\text{“女人”}) - vec(\text{“男人”}) + vec(\text{“父亲”})$。在类比推理任务中，人工预先给定 w_i 的理想答案，通过计算给定词的词向量与理想词向量的夹角余弦，评价词嵌入的实际效果。

Chen 等人[31]给出了类比推理任务集[2]，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。

2.5 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，其计算方法分为两种[32]：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[33]，其结果依赖于选取的语料库[34]，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[35]，包括基于路径、基于特征、基于信息内容、利用概念注释等不同方法[36]。

汉语中，常用的词义相似度计算任务集包括 MC30[3]和 wordsim297[4]。测试者使用计算工具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*100，对工具方法的有效性进行评价。

3 《词林》的嵌入表示

3.1 《词林》结构的调整

在《词林》中，每一层上的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，都会对意义产生进一步的约束和限定，因此，我们可以将每层新增编码信息，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结合。此外，在《词林》中，所有的词都

[2] https://github.com/Leonard-Xu/CWE
[3] https://github.com/huyingxi/Synonyms/blob/master/VALUATION.md
[4] https://github.com/thunlp/SE-WRL/blob/master/datasets/wordsim-297.txt

分布在叶子节点上，其词义描写程度是一样的，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，我们对《词林》结构进行调整。

考虑到位于群首的词往往能代表该词群的一般含义，颗粒度大、抽象程度高，我们按如下方法进行《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，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。这样不同抽象程度的词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。

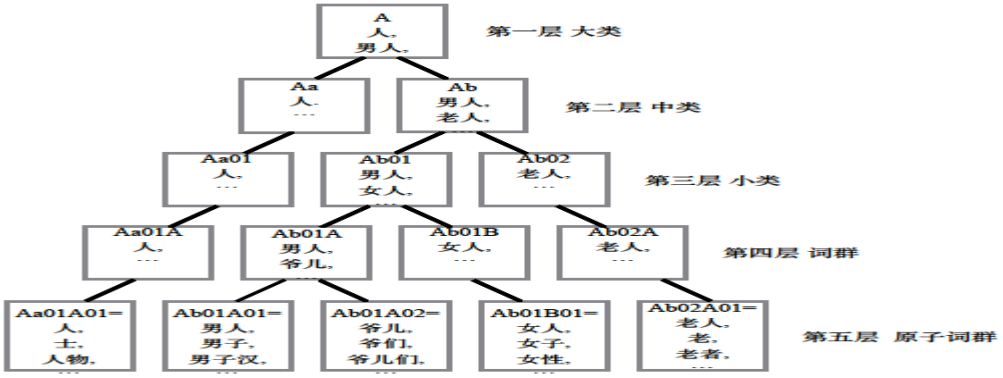

图 2　调整后的《词林》结构示意
Fig. 2　The adjusted structure of CiLin

3.2 基于《词林》的句式生成

神经网络训练是依据词在上下文中的分布信息捕捉词义的，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语料库。经过整理后的《词林》层级结构上共有 23570 个节点，每个节点代表的概念都不相同。我们利用每个词的结构编码信息构造三类伪句子，即：义素编码句式、义素编码扩展句式、词编码句式，它们的定义如下而示例见表 2。由于《词林》中的每层编码代表了该层的概念含义，在造句时，依照层级结构确定每个编码和词的距离具有分布合理性，适用于 Word2vec 模型来训练义素向量及词向量。

义素编码句式：根据义素的编码构造伪句子，每个词的所有父节点编码构成代表该词义的义素组合，按照父节点在层级结构上和该词的距离，确定该父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多次进行重复占位处理。这样的造句方式，保证句长固定且前后对称，同时满足 CBOW 和 Skip-Gram 两种模型对窗口词形式的要求。

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内的其他词，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，提高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练得到的词向量能够提取近义词信息。本文采用田久乐

（2010）、吕立辉（2013）、朱新华（2016）等三种词义相似度计算方法 。比如“人” 的近义词集，采用田久乐（2010）算法，在特定的相似度阈值设定下，就包括了{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。

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。这种句式假定，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。

表 2 《词林》生成句式示例

Tab. 2 Examples of different sentence templates for CiLin

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> ……
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> ……

3.3 嵌入表示的训练

Word2vec 模型基于上下文对词进行概率预测，包括 CBOW 和 Skip-gram 两种方法，可以从大量无标注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 预测其前后上下文中的每一个词[20]。

本文使用 gensim 自然语言处理库中的 Word2vec 模块[5]，使用 CBOW 和 Skip-Gram 两种方法在以上三种伪语料库上进行平行训练，并考察不同窗口词大小对训练结果的影响。同时，也在中文维基百科语料[6]上训得了词向量，用于相关任务的效果对比与验证 。

4 嵌入表示的应用评估

4.1 词义合成任务评估

[5] 代码参考 gensim: https://github.com/RaRe-Technologies/gensim

[6] 语料来源 wikimedia: https://dumps.wikimedia.org

由于《词林》的词义编码中包含了各层级的义素信息，词义等价于一组义素的结合，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中任务中的表现。

在构造义素编码句式和义素编码扩展句式时，《词林》中各层级的义素编码都在伪语料库中有所分布，并且与词形成合理分布关系，经过 Word2vec 模型的训练，同时得到了该知识库中的所有的义素向量。在本文中，我们采取如下公式来计算义素合成的词向量：

$$w_1 = \Sigma(s_1,\ldots,s_i,\ldots,s_n)$$

$$Vec(w_1) = \Sigma(\alpha_1 * Vec(s_1),\ldots,\alpha_i * Vec(s_i),\ldots,\alpha_n * Vec(s_n))$$

$$\Sigma(\alpha_1,\ldots,\alpha_i,\ldots,\alpha_n) = 1$$

其中， w_1 为所要计算的词，s_i 为与词义相关的义素，$Vec(x)$ 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：$\alpha_{i+1} = \alpha_i * 0.5$ 或 $\alpha_{i+1} = \alpha_i * 2$。

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性能。由于多义词有多种义素编码表达式，进而得到多种义素合成的词向量，在任务评估时，对每个词对，我们取使得和原词向量余弦相似度最高的一组义素合成的词向量。在该任务中，我们使用 CBOW 和 Skip-Gram 在不同句式、不同窗口词大小下训得的结果如表 3 所示。

表 3 词义合成任务中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				
窗口词大小	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(2010)	**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(2013)	**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(2016)	62.86	**63.29**	60.68	58.83	58.46	**79.60**	78.02	75.73	75.39	75.01

其中，扩展句式取 $\alpha_{i+1} = \alpha_i * 0.5$ 和 $\alpha_{i+1} = \alpha_i * 2$ 两种权重分配中得分较高的一种，根据三种算法相似度计算的不同特点，扩展句式的相似度阈值 ρ 分别定为：$\rho_{2010}=0.89$，$\rho_{2013}=0.65$，$\rho_{2016}=0.84$，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 训练效果普遍好于 CBOW，最合适的窗口大小是 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》对意义的分层描述具有一定的合理性，且句子生成在分布设计上保持了这种性质，经训练得到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，在语义合成任务中具有显著优势。

4.2 类比推理任务评估

对于类比推理任务，使用 CBOW 和 Skip-gram 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。

表 4 类比推理任务中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				
窗口词大小	3	4	5	6	7	3	4	5	6	7
Wikipedia 语料	80.83	81.39	81.67	81.91	**81.93**	80.6	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(2010)	57.2	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(2013)	**85.23**	83.8	81.24	80.93	81.27	92	**92.58**	92.04	91.97	92.05
义素编码扩展句式(2016)	83.36	**86.18**	84.19	84.89	85.09	93.3	93.4	93.33	93.07	**93.46**
义素合成的词向量(2010)	91.02	**91.92**	91.74	89.53	90.95	92	**92.24**	91.96	91.7	92.05
义素合成的词向量(2013)	**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(2016)	**90.31**	90.3	88.76	88.73	89.5	92.43	91.5	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	76.6	**77.04**

其中，扩展句式取 $\alpha_{i+1} = \alpha_i * 0.5$，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，使用该句式的义素向量合成词向量达到 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：相比于词单位，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子分布具有规范性，句式生成过程中可以人为控制信息分布，减少噪音，而语料库往往带有无法消解的歧义性和噪音问题。

总体来说，使用新方法得到的《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（2015）报告的最好效果 72.99。

4.3 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 训练集上的相似度评分，以及与人工判定标准比较的皮尔森系数 r*100 评分，分别如表 5、表 6 所示。

表 5 MC30 词义相似度计算任务：皮尔森系数 r*100

Tab. 5 Evaluation results on word similarity measurement task (MC30): Pearson r*100

模型	CBOW					Skip-gram				
窗口词大小	3	4	5	6	7	3	4	5	6	7
Wikipedia 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（2010）	20.12	2.63	**33.32**	14.95	-0.05	63.83	63.17	77.05	**75.60**	77.42
义素编码扩展句式（2013）	**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（2016）	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（2010）	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（2013）	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（2016）	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

表 6 wordsim297 词义相似度计算任务：皮尔森系数 r*100

Tab. 5 Evaluation results on word similarity measurement task (wordsim297): Pearson r*100

模型	CBOW					Skip-gram				
窗口大小	3	4	5	6	7	3	4	5	6	7
Wikipedia 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（2010）	**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（2013）	**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（2016）	**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（2010）	**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（2013）	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（2016）	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

其中，扩展句式取 $\alpha_{i+1} = \alpha_i * 2$，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 276 个词对，表 6 中相应为这 276 个词对上的得分。

在该任务中 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。加入了理性算法的扩展句式进一步提升了性能，其中， r*100 最高的是加入了田久乐（2010）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，田久乐（2010）、吕立辉（2013）、朱新华（2016）在 MC30 上的 r*100 分别为：49.39、74.03、79.24，在 wordsim297 上的 r*100 分别为：35.53、34.11、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接近《词林》知识表示的上限。

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，新方法得到的《词林》嵌入表示在 MC30 测试集上超过了维基百科，而在 wordsim297 上则落后于维基百科，这说明，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》嵌入表达在 wordsim297 中表现不佳，有可能是因为《词林》知识表示与数据本身存在局限性，比如在颗粒度表达上的问题或者语义分类不合理。因为，相较于 MC30，纯理性方法计算得到的结果与新方法得到的结果表现出相同的下降趋势，这或许与 MC30 的选词特殊与样本过小有关。

5 结语

本文以《词林》为例，提出并展示了一种基于知识库训练义素向量及词向量的新方法，考察不同训练模型及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，新得到的义素向量及词向量资源 CiLin2vec 在不同任务中都取得了进展或突破。其中，在词义合成和类比推理任务中表现突出，准确率达到 0.9 以上，显示该方法在应用上的巨大潜力。

在性质上，该方法有效复用已有的知识库资源，采用伪句式方式控制向词向量中注入的理性知识，并借鉴已有的理性方法进行预处理，发掘理性知识和理性方法结合、利用的最优方式，有很强的理论解释性；此外，相较于传统的语料库训练方法，新方法的占用内存小、训练周期短，更为简单快捷。

在未来，针对其他各类知识库，我们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融合应用，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，也将支持用以描述汉语语素及构词意义的北京大学《汉语概念词典》（英文 Chinese Object-Oriented Lexicon，简称 COOL）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, 06:602-608.

[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, 01:03-09.

[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, 04:29-36.

[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, 02:91-99.

[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, 06:07-15.

[6] 李国臣, 吕雷, 王瑞波 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, 2016, 01:101-114.

[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, 04:104-108.

[8] Scott Deerwester, Susan T. Dumais, George W. Furnas et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6):391– 407.

[9] Hinrich Schütze. Dimensions of meaning[J]. In Proceedings of IEEE Conference on Supercomputing, 1992.

[10] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence[J]. Behavior Research Methods, Instruments, & Computers, 1996, 28(2):203–208.

[11] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning[J]. ICML, 2008.

[12] Ronan Collobert, Jason Weston, Léon Bottou et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research (JMLR), 2011.

[13] Peter D. Turney. Domain and function: A dual-space model of semantic relations and compositions[J]. Journal of Artificial Intelligence Research (JAIR), 2012.

[14] Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation[J]. In Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP), 2014.

[15] Bartusiak R, Augustyniak Ł, Kajdanowicz T, et al. WordNet2Vec: Corpora Agnostic Word Vectorization Method[J]. 2016.

[16] Julien Tissier, Christophe Gravier, and Amaury Habrard. 2017. Dict2vec: Learning word embeddings using lexical dictionaries. In Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), pages 254–263.

[17] 梅家驹. 《词林》[M]. 上海: 上海辞书出版社, 1983.

[18] Harris, Zellig. Distributional structure. Word, 1954.

[19] Geoffrey E. Hinton, James L. McClelland, David E. Rumelhart. Distributed Representations. //David E. Rumelhart, James L. McClelland, CORPORATE PDP Research Group. Parallel Distributed Processing: Explorations in the Microstructure of Cognition[M]. Cambridge, USA: MIT Press, 1986, 1:77–109.

[20] 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.

[21] Wilhelm von Humboldt. Gesammelte Schriften(Cited as GS): Ausgabe Der Preussischen Akademie Der Wissenschaften[M]. GS 7:98—9; 1836, p.122.

[22] A. Yessenalina and C. Cardie. Compositional matrix-space models for sentiment analysis[J]. EMNLP, 2011.

[23] R. Socher, B. Huval, C. D. Manning et al. Semantic compositionality through recursive matrixvector spaces[J]. EMNLP, 2012..

[24] E. Grefenstette, G. Dinu, Y.-Z. Zhang, M. Sadrzadeh, and M. Baroni. Multi-step regression learning for compositional distributional semantics[J]. IWCS, 2013.

[25] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, and A. Ng. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank [J] EMNLP, 2013.

[26] Fodor, J. A., Pylyshyn, Z. W. Connectionism and cognitive architecture: A critical analysis[J]. Cognition, 1988, 28, 3–71.

[27] Gershman, S., & Tenenbaum, J. B. Phrase similarity in humans and machines[J]. In Proceedings of the 37th Annual Conference of the Cognitive Science Society, 2015.

[28] Noreen, Adolf. Inledning till modersmålets betydelselära[M]. Uppsala: Almqvist & Wiksell, 1901.

[29] John Lyons. Linguistic Semantics[M]. Cambridge: Cambridge University Press, 1996.

[30] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations[J]. HLTNAACL, 2013.
[31] Xinxiong Chen, Lei Xu, Zhiyuan Liu et al. 2015. Joint learning of character and word embeddings[J]. In Proceedings of IJCAI.
[32] 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, 27:3329-3333.
[33] 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, 27:01-06.
[34] Yuhua Li, Zuhair A. Bandar, David McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15:871-882.
[35] 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, 19:63-70.
[36] Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Abdelmajid Ben Hamadou. 2014. Ontology-based approach for measuring semantic similarity. Engineering Applications of Artificial Intelligence 36, pages 238–261.