文章编号：1003-0077（2019）00-0000-00

语言知识驱动的词嵌入向量的可解释性研究*

林星星[1,2]，邱晓枫[1,3]，虞梦夏[1,3]，祁晶[1,3]，康司辰[1]，刘扬[1]

（1.北京大学 计算语言学教育部重点实验室，北京 100871；
2.哈尔滨工业大学（深圳） 计算机科学与技术学院，广东 深圳 518055；
3.北京大学 中国语言文学系，北京 100871.）

摘要：神经网络语言模型应用广泛但可解释性较弱，其可解释性的一个重要而直接的方面表现为词嵌入向量的维度取值和语法语义等语言特征的关联状况。先前的可解释性工作集中于对语料库训得的词向量进行知识注入，以及基于训练和任务的算法性能分析，对词嵌入向量和语言特征之间的关联缺乏直接的验证和探讨。该文应用基于语言知识库上的伪语料法，通过控制注入语义特征，并对得到的词向量进行分析后取得了一些存在性的基础性结论：语义特征可以通过控制注入到词嵌入向量中；注入语义特征的词向量表现出很强的语义合成性，即上层概念可以由下层概念表示；语义特征的注入在词嵌入向量的所有维度上都有体现。

关键词：可解释性；词向量；伪词法

中图分类号：TP391 **文献标识码**：A

A Study of Interpretability of Knowledge-based Word Embedding Vectors

LIN Xingxing[1,2], QIU Xiaofeng[1,3], YU Mengxia[1,3], QI Jing[1,3], KANG Sichen[1], LIU Yang[1,2]

(1. Key Laboratory of Computational Linguistics (Ministry of Education), Peking University, Beijing 100871, China;
2. School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China;
3. Department of Chinese Language and Literature, Peking University, Beijing 100871, China)

Abstract: Multi-layer neural network models have many applications but are less interpretable. An important and direct aspect of its interpretability is the association between word embedding vectors and linguistic features. The previous work of interpretability mainly focuses on the knowledge injection to corpus-based word embedding vectors and the theoretical analysis of training models. However, there lacks direct verification and discussion on the correlation between word embedding vectors and linguistic features. In this paper, the pseudo-word method based on knowledge bases is applied to inject the linguistic features to obtain the word vectors. Some basic conclusions are obtained after experiment and analysis: it is remarkable to inject semantic features into the word embedding vectors under control; the compositionality of the word embedding vectors injected linguistic features is also significant, i.e. the upper concept can be represented by the lower concepts; the injection of semantic features is reflected in all dimensions of word embedding vectors.

Key words: interpretability; word embeddings; pseudo-word method

1 引言

机器学习的可解释性（Interpretability）一直是个核心的科学问题。可解释性没有明确的定义，有两种说法比较流行：一种认为可解释性意味着人类对模型决策背后原因的理解程度[1]；另一种强调人类预测出模型的结果需要保持一致性[2]。可解释性工作不可忽视的原因在于，模型的简单参数，在现实世界中不足以对模型进行充分描述[3]。而其背后的动机包括[3-5]：第一是为了知识的需要，人类进行科学研究的目的，一部分就是为了促进学习同时满足好奇心，并协调知识结构中的矛盾和不一致性，黑盒模型本身就是一个需要进行探索的目标；第二是希望判别并减少误差，比如 word2vec 向量存在的性别误差等，可解释性工作有助于判断误差并进行修正；第三是希望

* **收稿日期：** **定稿日期：**
基金项目：国家社科基金一般项目（16BYY137）、国家社科基金重大项目（18ZDA295，12&ZD119）

了解模型中包含和未包含的因素，从而做出正确的决策；第四是为了改进泛化能力和性能，拥有可解释性的模型通常有更好的泛化能力，可以得到更好的性能；第五是基于道德和法律原因，模型的可解释性能让人清楚模型的原理，从而具有公平性、隐私性、鲁棒性、可信度等特点，这也为多层神经网络模型在更多传统行业应用增加可能性。

机器学习中的部分算法具有很强的可解释性，比如线性回归、朴素贝叶斯、决策树等[4]。这些可解释性强的模型在一些任务上的表现一般。而多层神经网络模型在很多领域应用广泛[6, 7]，但是其可解释性不强[8]。当前备受关注的神经网络可解释性研究，从内容上主要分为计算机视觉领域和自然语言处理领域。

在计算机视觉领域，需要解决一些关键性问题[9]：神经网络的展开表示是什么，它的因素是怎么被度量和发现的；可解释的隐层节点是否是特征空间的一个平行反映；网络结构、数据以及训练条件对表示的影响与区别。之前，研究者们通过训练让模型在一些识别任务上效果更好，但是并没有全面地对隐层节点进行概念阐述。Zhou 等[9]提出了网络解剖（Network Dissection）的方法来度量视觉表示可解释性。他们通过让隐层节点和概念对实现最佳对齐，隐层节点被赋予可解释的标签，范围从颜色、材料、纹理、部件到对象和场景。这些图形图像特征是否注入是直观的，因为图形本身是可视化的。另一种增强神经网络可解释性的方法 LIME[10]用具有可解释性的简单模型去拟合多层神经网络模型，但只能进行局部拟合。

相比计算机视觉，自然语言处理的特点和任务更为复杂。计算机视觉中处理的对象图形本身是平面化的，易于表示；而自然语言处理面对的是离散化、难以表示的语言知识：语言处理的基础单元不清晰、组织结构不明确、语义语法嵌入困难、特征注入只能通过外部任务来验证……因此，自然语言处理的可解释性问题更为艰难。在自然语言处理领域，可解释性研究的一个重要方面是词嵌入向量（在不影响歧义的情况下，以下均简称词向量）的维度取值和语法语义等语言特征的关系问题，因为词向量是各种语言模型的基础。现阶段对词向量的可解释性工作包括两种：一种是对语料库训得的词向量进行知识注入，另外一种是对词向量训练算法以及训得的词向量进行理论分析。

对语料库训得的词向量进行知识注入的工作又分为两种：一种是在前处理阶段进行知识注入来获得词向量，另外一种是在后处理阶段对词向量增加限制使其具有某种特征。在前处理阶段，Panchenko[11]将 synset 向量化表示，构造语义嵌入和 synset 之间的简单映射来对语义嵌入进行解释。除了这种建立简单映射的办法，其他绝大多数方法主要是构建联合模型来注入语言特征。Li 等[12]利用“知网”的语义知识以及《同义词词林》中的同义关系，对从不同语义向量空间得到的词向量进行叠加得到最后的词向量；Gao 等[13]利用知识图谱，选择关系知识、分类知识构造函数，并和 skip-gram 模型相结合并在任务上验证；Yu 等[14]利用字典中的同义词信息，建立 RCM（Relation Constrained Model）模型，和 CBOW 结合构成联合模型。另外的一些方法也为词向量注入了不同层面的信息，在子词层面注入的信息包括汉字部首信息[15]、汉字笔画信息[16]、语言词缀内涵[17]等。在后处理阶段，Faruqui[18]提出了一种通用型的方法，为训练后的词向量增加限制使其具有某种特征。在这种方法中，使用者可以根据自己的需要增加不同的限制。比如，Faruqui 限制意思相关的词语有相近的向量表示，而 Mrkšić[19]通过增加限制使训练后的词向量具有同义、反义信息。

对词向量训练算法以及训得的词向量进行理论分析的工作也在同步进行。Levy 和 Goldberg[20]从代数角度进行理论分析并得到结论，基于负采样的 skip-gram 模型隐含地分解了一个 PMI 矩阵；Li 等[21]则认为基于负采样的 skip-gram 模型是显式地分解了一个 PMI 矩阵；Yin 和 Shen[22]从几何角度对词向量进行解释，并给出了最优维度大小的求解方法；Gittens 等[23]从信息论的角度表明 skip-gram 模型训得的词向量最优；Peng[24]在词和句子两个层面对词向量的维度取值进行解释。

以上的可解释性工作都没有清晰、明确地指出词向量本身和语法语义信息的关联。对语料库训得的词向量进行知识注入的方法，注意到了语言知识对词向量可解释性的重要性，但没有采用

特征注入的方法，只是浅层次地利用文本知识，无法给出词向量的维度取值和语言特征之间的具体关联；对词向量训练算法进行理论分析的方法，对训练算法虽然尝试从算法层面对词向量做出解释，但因为缺少对神经网络的控制方法，只能结合部分任务对词向量进行浅层次的解说；对词向量本身进行理论分析的相关工作，只是从几何角度和信息论角度对词向量进行探讨而没有涉及任何语言特征。Peng[24]虽然尝试对词向量的维度取值进行语言特征方面的解释，但是其划分过于粗浅而且只显示了语法层面的信息，不够全面。

基于伪词法的词向量训练方法很好地解决了这一难题，并在不同的知识库上得到了验证。段[25]通过对《同义词词林》中的五层层级信息进行编码来制作伪句子，从而得到拥有丰富词义内涵的词向量 CiLin2Vec。林[26]将北大《汉语概念词典》中的语素概念知识和汉语构词法知识相结合，在子词层面构建了义素嵌入表示，并在词义的组合表达上取得了很好的效果。

在上述研究的基础上，本文以具有广泛应用和影响的 WordNet 知识库为例，进行基于语言特征的伪语料构造；通过修改词向量训练模型，使其初始化特征矩阵固定，不会对语言特征注入产生影响；在构造的伪语料上进行词向量训练得到词向量；之后，通过维度擦除方法和降维方法对词向量进行分析，在特征注入的有效性、特征注入后的向量语义合成性以及特征注入在向量维度上的体现等三个方面，对语言特征和词向量之间的直接关联进行实验和探讨。

2 数据制备方法

2.1 基于语言特征的伪句子构造

WordNet 知识库在自然语言处理中有广泛的应用和影响[27-30]。它源自实验心理学理论[31]，是一个在线词库系统，在 1991 年发布 1.0 版，现在已经更新到 3.1 版。不同于以往的词典，WordNet 将 synset 作为语义单元并构建语义体系[32]。WordNet 中包含了丰富的词汇知识，包括多种语义关系和部分句法信息等。从数据分布看，WordNet 3.1 版中的名词 synset 达 81426 个，名词 synset 占全部的 69.7%，其包含的语义关系占全部的 68.9%。其中，语义类是所有词类都拥有的基础语言特征，本文选取名词的语义类作为单一特征来开展实验。在 WordNet 中，lex_filenum 是不同领域编纂者编纂的文件编号，不同的文件编号语义类别不同，每个文件编号可以被看作是一个语义类[31]。本文用 lex_filenum 来表示语义类的编号，WordNet 中共有 45 个语义类，名词占据了 26 个，其 lex_filenum 取值范围为 03 到 28。

如何将知识库中蕴含的丰富的理性知识注入词向量中是一个难题。在知识库中，属性是和单一节点相关的独立信息，可以表示为<node-X, <att-name, att-value>>序列，它并不会影响本体的整体结构；而关系是节点和节点之间的关联信息，可以表示为< node -X, <rel-pointer, node -X’>>序列，它蕴含了节点之间的复杂结构。与此同时，词向量训练模型包括 word2vec 通常是采用线性结构数据（即句子）作为输入。因此，如何将知识库中网状结构的数据转化成线性结构数据是亟待解决的问题。

参照项目组之前的工作[25, 26]，我们将知识库中的属性和关系展开并置入伪句子模板中，以此来构造伪句子，形成伪句子语料。本文采用 WordNet 中最基础的名词语义类特征，构造伪句子模板为：“<virtual_node> <lex_filenum> <synset_offset> <lex_filenum> <virtual_node>”。其中，virtual_node 表示虚拟节点，是所有 lex_filenum 的虚拟上位，而 lex_filenum 是 WordNet 中的语义类编号，synset_offset 表示每个 synset 的偏移地址，此处用来指代 synset。表 1 是该伪句子模板生成伪句子的一个示例。其中，“00”表示虚拟节点，“03”是语义类编号，而“01816528”表示一个 synset。

表 1 伪句子模板示例

伪句子模板	<virtual_node> <lex_filenum> <synset_offset> <lex_filenum> <virtual_node>
伪句子生成	00 03 01816528 03 00

2.2 控制特征矩阵的词向量训练模型

当前常用的词向量训练模型包括 word2vec[33]和 glove[34]。Word2vec 在 2013 年由 Mikolov 等提出，包括 skip-gram 模型和 CBOW 模型，其原理简单易懂、代码便于修改，能在大规模上下文共现数据上快速训得词向量。其中，skip-gram 模型在多种任务上表现良好、能精准捕捉语法、语义等语言特征。因此，本文采用 skip-gram 模型作为词向量训练模型。

为了确保在整个训练过程中唯一的变量是定向注入的语言特征，本文对 skip-gram 模型进行控制：一个是词表顺序，另外一个是模型的初始化权重矩阵。即在整个训练过程中，使用相同的词表顺序，并且用同样的权重矩阵进行初始化。同时，为了研究注入语言特征对词向量维度取值的影响，本文在研究词向量的维度取值时，使用的是减去初始化权重矩阵后的词向量。

2.3 伪语料上的词向量训练

通过上述伪句子模板，我们在 WordNet 名词部分的 26 个语义类上共构造了 81426 条伪句子，这些伪句子形成了一个包含特定语言特征表达的伪句子语料。在获得伪句子语料库之后，就可以通过常规方法对其训练，得到包含特定语言特征注入的词向量。

我们利用控制特征矩阵后的 skip-gram 模型训练词向量，获得 81452 个词向量（严格说来，应该是排歧后的词向量，即 synset 向量。此外，也有少量语义类向量，用于词向量的语义合成性验证）。在实验中，我们依次尝试了 10、50、100 等不同的向量维度大小，情况类似，为简化描述，本文以 50 维词向量为例进行阐述。

3 数据分析方法

3.1 降维分析

很多科学问题都面临着大量的高维数据，随着而来的一个问题就是如何发现这些高维数据的紧凑表示，即降维。经过降维分析后的数据特征减少，可以尽量保证特征之间相互独立。在数据处理中，常用的降维方法包括主成分分析（Principal components analysis，PCA）、线性判别分析（Linear Discriminant Analysis，LDA）、局部线性嵌入（Locally Linear Embedding，LLE）、拉普拉斯特征映射（Laplacian Eigenmaps，LE）等。

PCA 是最常用的无监督线性降维方法。它使用较少的数据维度，同时保留较多的原数据特性和关联。除此之外，PCA 方法降维速度很快。LDA 方法能根据数据所属标签进行针对性降维，其尽可能让同一个标签的向量在低维空间中靠近，而让不同标签的向量在低维空间中尽量远离。LLE 和 LE 方法利用图知识，保证在高维空间中邻点在低维空间中依然保持相同的线性关系。

在实验中，我们发现 LLE 和 LE 这两种方法在数据量很大的情况下计算速度过慢，而 LDA 方法的思想包含让同一个语义类内的向量在低维空间中靠近，而不同语义类内的向量在低维空间中尽量远离，这样的降维结果在聚类任务上有作弊的嫌疑。相比之下，PCA 方法在降维的过程中并不考虑向量从属于哪一个语义类，因此本文采用 PCA 方法进行实验结果的阐述。

3.2 维度擦除

降维方法是现在常用的降低数据维度的方法，通过数理方法简化维度并保留原始数据特征。然而这些降维方法并不区分哪些数据蕴含了哪些语言特征，这在某种程度上会破坏向量维度和语言特征之间的关联，因此我们提出了维度擦除方法。

维度擦除方法是根据维度的显著性，人为擦除向量中的显著性不高的某几维向量，从而降低向量维度的一种方法，其步骤如下：

1. 计算每个语义类的类内相似度；

2. 从 1 到 n（n 为向量维度大小），分别抹去一个维度并重复 1 的操作；

3. 找到抹去维度后类内相似度和原始值相差最大的 m（m 可自由选择）个维度；

4. 人为抹去显著度最高的 m 个维度和显著度最低的 m 个维度得到向量并做实验验证。

在维度擦除的过程中，当抹去某一维向量后的类内相似度越低，说明这一维对原始向量的类内相似度贡献越高，因此这一维的排行越靠前；相反，如果抹去某一维后类内相似度下降，说明这一维对原始向量的类内相似度贡献很低，这一维的排行就越靠后。通过这样的维度擦除方法，可以保留对原始向量的类内相似度贡献较高的维度。也就是说，进行维度擦除之后，得到的向量和注入的语义特征之间存在紧密关联：注入某一种语义特征，得到的向量在高维空间中靠近。

3.3 语义合成性分析

语义合成性是指基于概念的组成部分以及构成规则来推知概念含义的能力[35]。语义合成性是人类理解概念、构建知识的天然依据：遵循从小单位到大单位、从小概念到大概念的归约假说。在自然语言处理中，语义合成性意味着基于较小语言单位的表示来学习较大语言单位的表示，或者，基于下层概念的表示来学习上层概念的表示。语义合成性越强的语言模型越接近于人类的一般认知和思考方式，也就是说，模型的可解释性越强。因此，对于词向量来说，语义合成性是可解释性的一个重要体现。本文采用语义合成性分析来对注入特征后的词向量进行探究。

语义合成性实验的数据集，因组合的方法以及语言单位的不同而有所差异，其实验方法包括均值方法和具有丰富递归结构的方法[36]。受先前工作的启发[25, 26]，本文设计任务对 WordNet 中的名词语义类概念和其下的 synset 概念进行语义合成性验证。具体的实验细节在后文进行阐述。

4 数据结果分析

对于在伪语料上利用控制特征矩阵的训练模型获得的词向量，我们对特征注入的有效性进行验证，包括内部数据分析和聚类分析外部验证；对训得词向量的语义合成性进行验证；对特征注入在向量维度取值上的显著度进行考察。

4.1 特征注入的有效性验证

我们对每个语义类内的向量做维度上的观察，发现同一语义类内的向量在不同维度上表现出了某种模式。以 lex_filenum 为 15 的语义类内的向量为例，图 1 展示了该语义类中所有词向量的维度取值，横坐标是 50 维向量的维度编号（从 1 到 50），纵坐标是向量的取值。

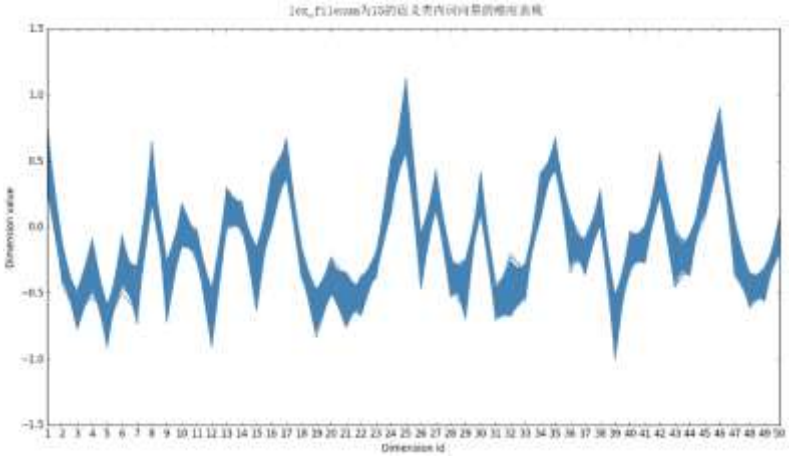

图 1 语义类（取 lex_filenum 为 15）内所有词向量的维度取值

从图中看，同一个语义类内的各个词向量在整体维度上表现出了某种一致性的模式。词向量在不同维度上的取值，和语义特征表达出某种稳定的关联，直观表明相同的语义特征成功注入到

词向量中了。下面，我们从内部和外部两个角度，对特征注入有效性进行定量考察。

4.1.1 内部验证

为了定量考察词向量维度上的表现，我们提出两个概念：类内相似度和类间相似度。类内相似度指的是同一个语义类内的词向量之间的相似程度，用来衡量同一个语义类内的词向量在不同维度上的取值是否存在较强的一致性。类内相似度越高意味着同一个语义类内的词向量一致性越高。而类间相似度指的是两个不同的语义类内的词向量在不同维度上的取值的相似程度。类间相似度用来衡量不同的语义类内的词向量在不同维度上的取值是否具有较大的差异性。类间相似度越低意味着不同的语义类内的词向量的差异性越高。本文对词向量进行相似程度的计算，采用的指标是皮尔逊系数。

在实验中，如果语义类的类内相似度越高，首先说明同一个语义类内的词向量在高维空间是相近的，同时也说明了被注入了同一种语义知识的词向量在整体维度上拥有相近的模式。也就是说，词向量的维度取值和语义特征的关联很紧密。如果类内的词向量相似度很低，说明注入同一种语义知识的向量形成的维度取值是随机的，不具有某种特定的模式。也就是说，词向量的维度取值和语义特征之间不存在特殊关联。虽然类内相似度表明了同一个语义类内的词向量具有某种模式，但是仍然需要对类间相似度进行度量。因为如果类间相似度依然很高，这说明所有的词向量，无论属于哪一种语义类，在整体维度上都显示出了某种特定的模式。这就无法说明词向量的维度取值和语义特征存在特殊关联。因此实验的基本假设是，注入同一种语义知识的向量在高维空间中靠近，而注入不同语义知识的向量在高维空间中远离。

图 2 显示了 26 个语义类的类内相似度和类间相似度，坐标轴是语义类编号（从 3 到 28）。每一行代表某个语义类和所有语义类（包括自身）的相似度，相似度越高颜色越深。从图中可以发现，对角线上的方格颜色比其他位置要深，说明每个语义类的类内相似度高，而类间相似度低。由此可见，通过在伪语料上训练出的词向量在维度取值上表现出了某种模式，这种模式在属于同一语义类的词向量上具备较强的一致性，而在不同语义类内的词向量取值上拥有显著的差性。这在一定程度上说明词向量的维度取值和语义特征之间存在关联，即语义特征成功注入到词向量中。

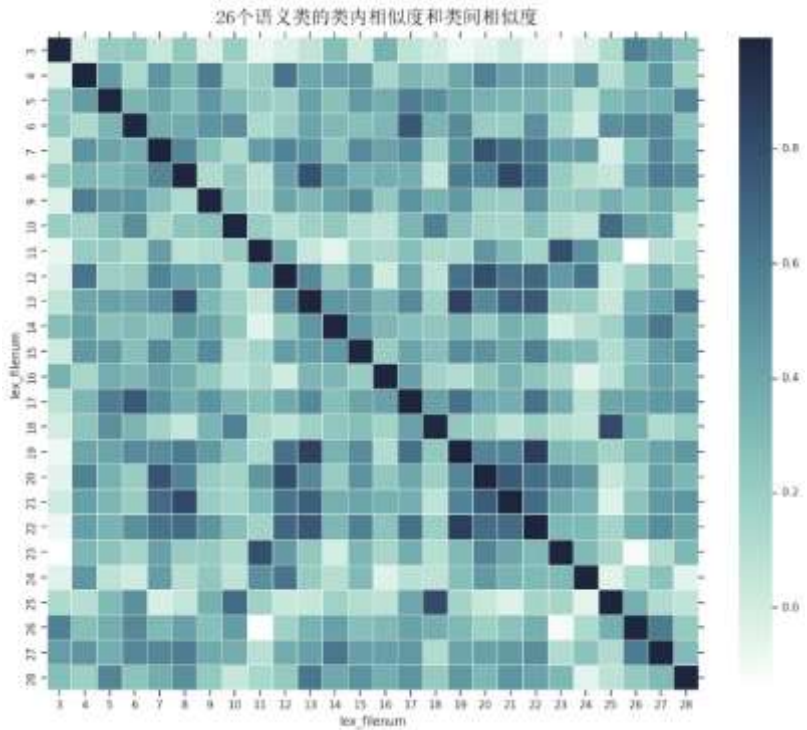

图 2 语义类（lex_filenum 范围为 3-28）两两相似度实验结果

4.1.2 外部验证

本文的外部验证采用聚类任务来对特征注入的有效性进行检验，使用 k-means 聚类方法对词向量进行聚类。实验的基本假设是如果聚类后各个类内的词向量所指向的 synset 很大程度上属于同一个语义类，那么就说明语义类特征被注入到词向量中。也就是说，利用是否被划分到原始语义类来判断聚类任务的好坏。

对聚类得到的每一个类，将类内的 synset 和 26 个语义类类内的 synset 进行配对，配对衡量的指标是聚类得到的类和语义类类内 synset 的重叠个数，重叠个数最多的语义类即和这个类进行配对。在配对结束之后，将这 26 个类类内对应上的 synset 个数相加除以总的 synset 个数，即得到聚类的准确度。采用的准确度指标为：

$$\frac{\sum_{i=1}^{n}\max\left(F\left(Q_i,P_j\right)\right),\ j\in[1,n]\ \&\ i\neq j}{T}$$

其中 n 是语义类个数，Q_i表示聚类结果的某个类，P_i表示 WordNet 中的某一个语义类，而$F(Q_i,P_j)$表示两个类内 synset 的重叠个数。T 代表所有语义类的 synset 总个数。

分别对随机初始化的词向量和注入语义特征之后的词向量进行聚类任务验证。注入语义特征之后的词向量的聚类准确度是 72.2%，而随机初始化的词向量的聚类准确度只有 14.3%，注入语义特征后的词向量的聚类准确度比随机初始化向量提升了 400%以上。这从外部验证了经过伪词法训练得到的词向量的特征注入有效性：从词向量取值的角度看，处于同一语义类中、被注入相同语义知识的 synset 在高维空间中的位置靠近，处于不同语义类中、被注入相同语义知识的 synset 在高维空间中的位置远离。

4.2 特征注入的语义合成性验证

接下来，我们对注入特征后的词向量的语义合成性进行验证。对 26 个语义类，设置实验来验证语义类向量和其下的 synset 向量之间的关联：检验上层概念的向量表示是否可以由下层概念的向量表示推得。验证是均值方法：在 synset 向量表示的均值和语义类向量表示之间计算相似度，来，结果取绝对值。

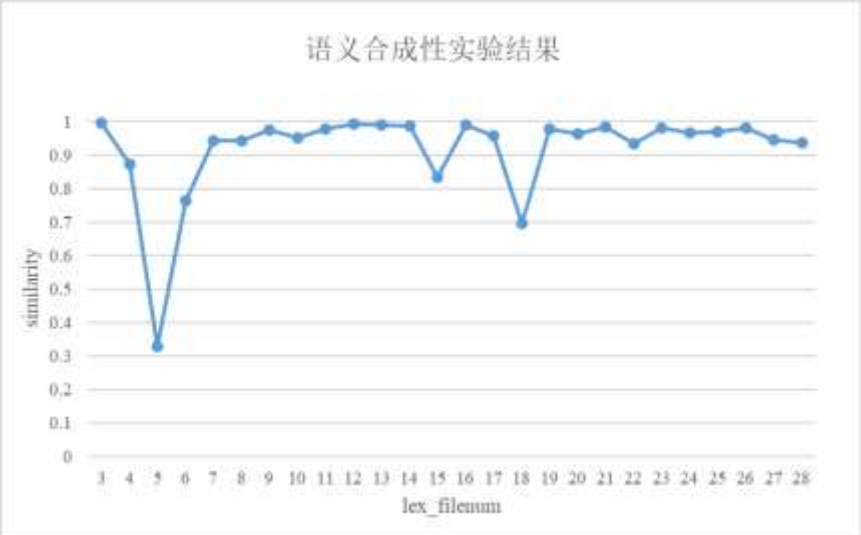

图 3 语义类（lex_filenum 范围为 3-28）与类内 synset 的语义合成性实验结果

图 3 是语义类和类内 synset 的向量相似度结果，横坐标是语义类编号（3-28），纵坐标是相似度。除少数语义类之外，多数语义类向量表示具有非常强的语义合成性：80.8%的语义类的语义合成性结果高于 90%；57.7%的语义类的结果高于 95%；语义合成性结果最高的一个语义类达到了 99.6%。这意味着在该方法下，上层的语义类向量表示和下层的 synset 向量表示之间有极强的可归约性，说明注入语义特征训得的词向量表现出了很好的特性。考虑词汇概念层级的语言学含义，这对人类而言有很好的可解释性。

4.3 特征注入和向量维度取值的关联

前面对特征注入的有效性进行了验证，并表明注入特征之后得到的词向量有极强的语义合成性。我们继续对注入特征和向量维度取值的关联进行讨论。

对于语言特征在维度取值上的体现，一般存在两种不同的先验假设：1、特征在所有的维度上都有所体现；2、特征只在特定维度上有所体现。从 skip-gram 模型算法角度分析，在进行梯度下降的过程中，每个节点都进行了梯度下降的操作。因此，本文的实验假设是注入特征在所有的维度上有所体现而不只是在某些特定维度上有体现。本文采用反证的方法，先假设特征只在某些特定维度上有所体现，然后证伪。我们通过维度擦除方法进行人工抹去特定维度、随机抹去维度和 PCA 降维处理后得到的向量在聚类任务上进行检验向量的信息保留能力，指标采用前文提到的聚类评价办法。基于本文提出的实验假设，三种方法得到的向量的信息保留能力没有明显的差距，即聚类结果相近。

本文实验的对象是 50 维的词向量，在维度擦除方法人工抹去特定维度时 m 设为 10，可以获得抹去特征不显著的 10 维向量后剩余的 40 维向量和抹去特征显著的向量对应的 40 维向量。前面两者作为聚类任务的实验组。对照组是抹去随机挑选的 10 个维度得到的 10 组 40 维向量。以通过 PCA 降维得到的 40 维向量也作为参照。

图 4 是实验结果，可以看出处理后的 40 维向量和原始向量的 50 维相比聚类结果较差，这是符合直觉的，因为维度减少意味着信息的丢失。与此同时，可以看出对照组 10 组 40 维向量的结果没有明显地落在两个实验组结果中间。因此，可以认为语义特征注入后并不是在某些特定维度上体现，而是在所有的维度上都有所体现。其原因有二：1、维度降低后的向量都丢失了部分语义信息，从而得到比原始向量差的聚类结果，若特征注入只体现在某些特定维度，那么抹去其他不体现特征的维度应该对实验结果没有影响；2、抹去某些被认为显著和不显著的维度和随机抹去一些维度的实验结果，没有明显的规律可循。

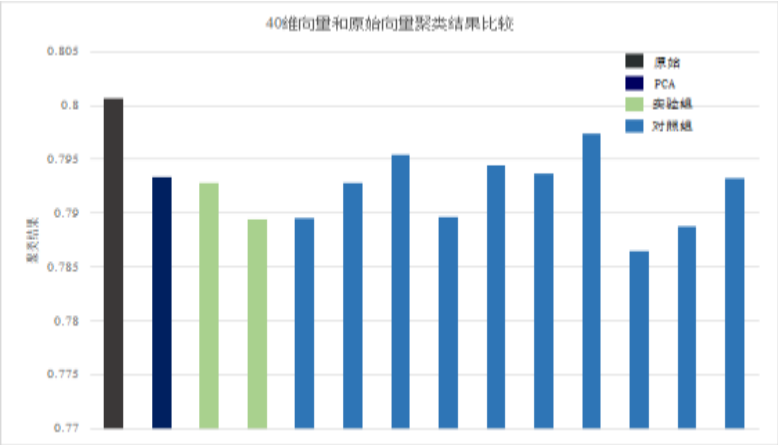

图 4 40 维向量和原始向量的聚类结果比较

5 结语

在机器学习中，可解释性问题因为知识需要、减少误差、判断因素、改善泛化等需求一直备受关注。目前流行的多层神经网络模型有较好的性能，但可解释性不强。在计算机视觉领域，神经网络的可解释性问题表现为隐藏变量和视觉概念的相关性，大量研究相继开展并率先获得成果。在自然语言处理领域，面对字符流序列的非显性化表述的文本知识，可解释性工作很难开展。

之前的研究多从两个角度进行词向量的可解释性阐述：一是对语料库训得的词向量进行知识注入的方法，这种方法注意到了知识库对词向量可解释性的重要性，但由于没有采用特征注入的方法，只是简单利用了知识库，无法表明词向量在维度取值和语言特征上的具体关联；二是对词向量训练算法及其训得的词向量进行理论分析，尝试从算法上对词向量做出解释，但因为缺少对神经网络的控制方法，只能结合部分任务对词向量进行简单解说。

本文基于单一语言特征构造伪语料并控制特征矩阵做词向量训练，对词向量和语言特征的关系做相对直接的考察，并设计实验从多个方面对涉及词向量可解释性的关键性问题进行研究：语义特征是否可以成功注入到词向量中；注入语义特征后的词向量的语义合成性如何；特征注入后是在词向量所有的维度上体现还是在某些维度上体现。通过一系列新的实验手段的开展，对语言知识驱动的词嵌入向量，取得了一些存在性的基础性结论：1、语义特征可以通过控制注入到词向量中；2、注入语义特征的词向量表现出很强的语义合成性，即上层概念可以由下层概念直接表示；3、语义特征的注入在词向量的所有的维度上都有体现。

在这些基础性结论之后，我们也在展开针对多种语言特征联合注入的研究，同时，也计划在其他知识库和其他词向量训练模型上做进一步的验证。

参考文献

[1] Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences[Z]. 2017.

[2] Kim B, Khanna R, Koyejo O O. Examples are not enough, learn to criticize! Criticism for Interpretability[M]. Lee D D, Sugiyama M, Luxburg U V, et al, Curran Associates, Inc., 2016, 2280-2288.

[3] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning[J]. arXiv preprint arXiv: 1702.08608, 2017.

[4] Christoph M. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable[J]. 2018: 180.

[5] LARS HULSTAERT. Interpreting machine learning models[EB/OL]. (2018). https://towardsdatascience.com/interpretability-in-machine-learning-70c30694a05f.

[6] LECUN Y, KAVUKCUOGLU K, FARABET C. Convolutional networks and applications in vision[C]//Proceedings of 2010 IEEE International Symposium on Circuits and Systems. IEEE, 2010: 253–256.

[7] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]// Advances in neural information processing systems. 2012: 1097-1105.

[8] Levy O, Goldberg Y. Dependency-based word embeddings[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014, 2: 302-308.

[9] Zhou B. Interpretable representation learning for visual intelligence[Z]. Massachusetts Institute of Technology, 2018.

[10] Ribeiro M T, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016: 1135-1144.

[11] Panchenko A. Best of Both Worlds: Making Word Sense Embeddings Interpretable[C]. 2016.

[12] Li W, Wu Y, Lv X. Improving word vector with prior knowledge in semantic dictionary[J]. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

2016, 10102(2015): 461-469.

[13] Gao B, Bian J, Bai Y, et al. RC-NET: A General Framework for Incorporating Knowledge into Word Representations [J]. 2014: 1219-1228.

[14] Yu M, Dredze M. Improving Lexical Embeddings with Semantic Knowledge[J]. 2015(1): 545-550.

[15] Sun Y, Lin L, Yang N, et al. Radical-Enhanced Chinese Character Embedding[M]. Springer, Cham, 2014, 279-286.

[16] Cao S, Lu W, Zhou J, et al. cw2vec: Learning chinese word embeddings with stroke n-gram information[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[17] Bojanowski P, Grave E, Joulin A, et al. Enriching Word Vectors with Subword Information[J]. Transactions of the Association for Computational Linguistics. 2017, 5: 135-146.

[18] Faruqui M, Dodge J, Jauhar S K, et al. Retrofitting word vectors to semantic lexicons[J]. arXiv preprint arXiv:1411.4166. 2014.

[19] Mrkšić N, Séaghdha D Ó, Thomson B, et al. Counter-fitting Word Vectors to Linguistic Constraints[J]. 2016: 142-148.

[20] Levy O, Goldberg Y. Neural Word Embedding as Implicit Matrix Factorization[J]. Proc. of NIPS. 2014: 2177-2185.

[21] Li Y, Xu L, Tian F, et al. Word embedding revisited: A new representation learning and explicit matrix factorization perspective[C]//Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.

[22] Yin Z, Shen Y. On the Dimensionality of Word Embedding[J]. 2018(NeurIPS).

[23] Gittens A, Achlioptas D, Mahoney M W. Skip-Gram - Zipf + Uniform = Vector Additivity[J]. 2017: 69-76.

[24] Peng K. Evaluation method of word embedding by roots and affixes[J]. arXiv preprint arXiv:1606.07601. 2016.

[25] 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(0438-0479): 867-875.

[26] Lin Z, Liu Y. Implanting Rational Knowledge into Distributed Representation at Morpheme Level[C]// Thirty-Third AAAI Conference on Artificial Intelligence. 2019.

[27] Resnik P. Disambiguating noun groupings with respect to WordNet senses[M]. Springer, 1999, 77-98.

[28] Kamps J, Marx M, Mokken R J, et al. Using WordNet to measure semantic orientations of adjectives[C]//LREC. 2004, 4: 1115-1118.

[29] SUSSNA M, MICHAEL. Word sense disambiguation for free-text indexing using a massive semantic network[C]//Proceedings of the second international conference on Information and knowledge management - CIKM '93. New York, New York, USA: ACM Press, 1993: 67–74.

[30] Fragos K, Maistros Y, Skourlas C. Word sense disambiguation using wordnet relations[C]//First Balkan Conference in Informatics, Thessaloniki. 2003.

[31] Miller G A, Beckwith R, Fellbaum C, et al. Introduction to WordNet: An on-line lexical database[J]. International journal of lexicography, 1990, 3(4): 235-244.

[32] Miller G A. WordNet : A Lexical Database for English[J]. 1995, 38(11): 39-41.

[33] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems.: 3111–3119.

[34] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[35] Frege G, Geach P T, Black M. Translations from the philosophical writings of Gottlob Frege[M]. Philosophical Library, 1952.

[36] Bowman S R, Gauthier J, Rastogi A, et al. A fast unified model for parsing and sentence understanding[J]. arXiv preprint arXiv:1603.06021. 2016.