文章编号：1003-0077（2017）00-0000-00

基于形态学信息的中文词嵌入方法：一种双通道视角

陶汉卿[1] 童世炜[1] 徐童[1,2] 刘淇[1,2] 陈恩红[1,2]
（1. 中国科学技术大学 计算机科学与技术学院，安徽省 合肥市 230026；
2. 中国科学技术大学 大数据学院，安徽省 合肥市 230026）

摘要： 词嵌入是自然语言处理领域的一个基础而又十分重要的课题。对于具有象形表意特性的汉语来说，如何捕捉隐藏于文字形态中的语义信息，同时使得方法具有良好的可解释性，成为一个亟待解决的问题。在该文中，我们详细阐释了汉语的形态学信息在传达语义和增强汉语词嵌入上的重要性。然后，我们提出了一个新颖的双通道词嵌入模型来实现汉字笔画序列信息和字形空间信息的联合学习，进而丰富汉语词的表示。通过两个经典词嵌入测试任务的评估，我们的模型在形态学突出的词语相似度和词义类比任务中明显优于其他的模型，同时展现出了很好的可解释性。

关键词： 词嵌入；形态学；笔画；字形

中图分类号： TP391 **文献标识码：** A

Chinese Word Embedding via Morphological Information: A Dual-channel View

Hanqing Tao[1], Shiwei Tong[1], Tong Xu[1,2], Qi Liu[1,2] and Enhong Chen[1,2]

(1. School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, 230026, China;
2. School of Data Science, University of Science and Technology of China, Hefei, Anhui, 230026, China)

Abstract: Word embedding is a basic and very important topic in the field of Natural Language Processing. For Chinese, which has the nature of pictographic representation, it is urgent to explore more interpretable strategies to capture the language patterns in which morphological information is used to convey semantics. In this paper, we elaborate that Chinese word embeddings can be substantially enhanced by the morphological information hidden in characters which is reflected not only in strokes order sequentially, but also in character glyphs spatially. Then, we propose a novel Dual-channel Word Embedding model to realize the joint learning of sequential and spatial information of Chinese characters, so as to further enrich the representation of words. Through the evaluation on both word similarity and word analogy task, our model significantly outperforms other baseline methods and shows great interpretability.

Key words: word embedding; morphology; stroke; glyph

0 引言

词嵌入是指用一个固定维度的向量去表示语言中每个单词，从而便于使用计算机对其进行数学处理。近年来，用于自然语言处理（Natural Language Processing, NLP）领域下游任务的词嵌入因其在语言表示方面的强大效用而引起了广泛的关注[1]。随着人们对自然语言认知程度的不断加深，一些研究者逐渐注意到了字词本身的形态学信息在辅助词语相似度计算和传达语义上也有着不可或缺的作用[2]。

收稿日期：201*-**-**；**定稿日期**：201*-**-**
基金项目：国家重点研发计划项目课题(No. 2016YFB1000904)；国家自然科学基金（U1605251，61703386）

所谓的形态学则是语言学的一个分支，其主要目的是研究单词的内部结构和组成方式。借助形态学传达特定语义的模式广泛存在于各种文字系统中。例如，英语中的“dog”、“dogs”和“dog-catcher”有着相当紧密的联系，而英语阅读者可以直观地在字面上利用他们的背景知识来判断这种关系[1]。而对于文字系统是基于语素文字[2]的汉语来说，由于其文字起源于具有象形性质的甲骨文，这种通过形态学信息在视觉上直观地传达语义的模式更为突出和常见。在汉语语义特征的探索过程中，学者们发现不仅汉语的字和词是重要的语义载体,而且汉字的笔画顺序也能够刻画一定的语义[3][4]。需要指出的是，一个汉语词通常由若干个汉字组成，而根据汉语字典规范，每个汉字又可以进一步地分解为一个唯一且不变的笔画序列，即“笔顺”。

实际上，汉语的形态学信息正是通过笔画在两个通道上表现出来的：首先，在一维的序列性通道上，构成每个汉字的笔画序列类似于英语等字母语言的单词，这使得汉语的形态学模式一定程度上涵盖了字母语言的模式；其次，在二维的空间性通道上，笔画可以按照固有的序列进一步地在二维空间上进行拼接组合，形成类似于图像的字形结构，这使得汉语文字所含有的信息量要明显丰富于字母语言。如图 1 所示，图的上半部分阐释了不同的汉字之间存在的结构包含关系。这种包含关系不仅体现在字形上，也体现在相应的笔画序列上。即，汉字“驾”可以分解为一个八画的笔画序列，而该序列的最后三画合起来则对应于“马”这个汉字（实质上“马”在汉语字典中也是“驾”字的部首，是“驾”和“马”语义相关的核心体现），这在序列性通道上与英语单词“declare”和“clarify”共有的词根“clar”反映的形态学信息本质上是相同的，故我们往往可以从字面上直观地判断出这些汉字或英语单词之间的语义联系。但是，相同笔画序列在空间上的不同排列和组成方式也会导致迥然不同的语义。如图 1 的下半部分所示，三个汉字“入”、“八”和“人”具有一个相同的笔画序列，但由于它们的笔画在空间结构上存在着不同的组成方式，致使它们的语义也完全不同。

此外，一些生物学的研究也已经证实人脑在处理汉语文字信息时确实有两个处理通道在同时进行。具体来说，文献[5][6][7]经过实验论证了人类的左脑在处理字母语言信号上有着至关重要的作用，而文献[8]则进一步揭示了汉语的阅读者不仅会激活左脑中负责时序信号处理的区域，同时也会激活右脑中负责图像处理和空间信息的区域。因此，我们基于这些生物学和语言学相结合的研究成果得出结论，汉字的形态学信息由两部分组成，即隐藏在类似于英文词根的笔画顺序中的序列信息和隐藏在具有图像特性的字形的空间结构信息。从该角度出发，我们提出了一种新颖的汉语双通道词嵌入模型(Dual-channel Word Embedding，DWE)，用于实现对汉字含有的笔画的序列信息和字形的空间信息的联合学习，从而为汉语自然语言处理提供更具合理性和可解释性的词嵌入向量。最后，通过两个具有代表性的词嵌入任务的性能评估，实验结果表明了我们的 DWE 模型在汉语形态信息突出的词语相似度、词义类比和可解释性方面要显著优于传统词嵌入模型，也证实了汉语形态学信息在丰富汉语词嵌入上的重要作用。

1 相关工作

早期，汉语词嵌入研究在形式上大多只是仿照英文词向量的模式：先进行中文分词，然后套用基于英文开发的方法基于大规模语料进行无监督训练[3]。但是，这种模式很大程度上忽略了汉语的象形表意特性以及文字形态学上的信息。根据不同语言词嵌入方法和原理的发展历程，相关工作可以分为以下两类。

1.1 基于形态学的词嵌入研究

传统的词嵌入方法主要是基于分布式假设[9]，即上下文相似的词具有相似的语义。基于该假设，诸多词嵌入模型被提出，且其得到的词向量在自然语言处理的下游任务中都取得了不俗的效果，但是这些模型普遍存在可解释性不够强的问题，且往往不能反映出人们对语言模式的直观感受。为了构建更具有可解释性的词嵌入模型，

[1] https://en.wikipedia.org/wiki/Morphology_(linguistics)
[2] https://zh.wikipedia.org/wiki/语素文字

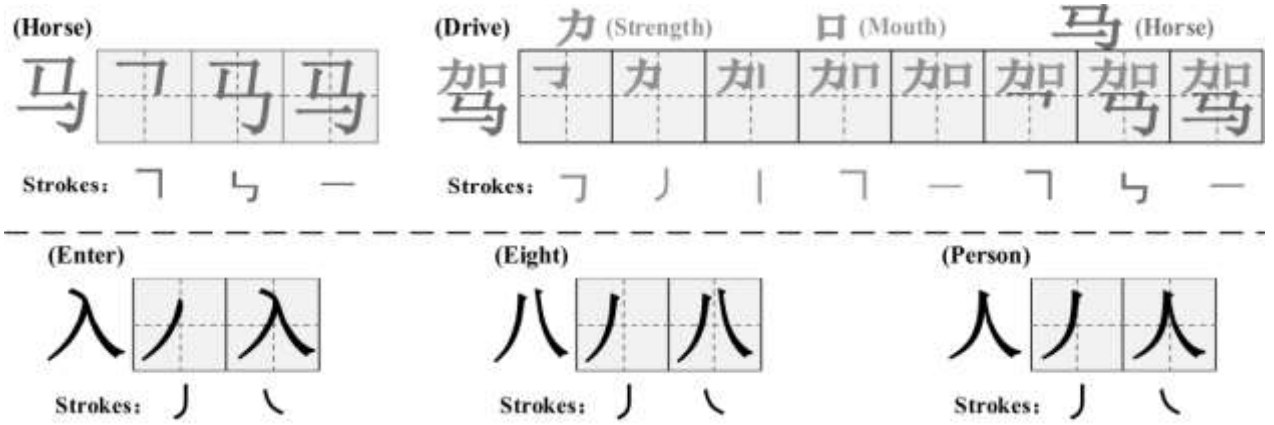

图 1 图的上半部分阐释了不同的汉字之间经常存在着的结构包含关系，这种包含关系不仅体现在字形上，也体现在相应的笔画序列上；图的下半部分反映了一个共同的笔画序列如果在字形的空间上组合方式不同，那么该笔画序列就会形成不同的汉字，从而传达迥然不同的语义。

一些国外学者逐渐将注意力集中到形态学上，他们发现词的形态学信息在传达语义上也有着不可忽视的作用，字或词本身的形态或内部结构可以帮助人们直观地获取部分语义信息[10][11]。更重要的是已有研究表明，在法语、德语、西班牙语、土耳其语等诸多语言中，引入词的形态学信息的确可以丰富词嵌入的语义表示从而提高下游任务的性能[2][12][13]。最近，Wieting 等人[14]提出使用字符级的 n-gram 向量来表示英语中的单词，以捕获包括前缀、后缀、词根等语义特征。之后，Bojanowski 等人[15]改进了经典的 Skip-gram 模型[1]，在获取词嵌入向量时考虑并结合了词根等形态学信息，并在英语、法语、德语等语言评测任务上均取得了性能的提升，这对我们开展汉语形态学研究有着十分重要的指导意义。

1.2 针对汉语的词嵌入研究

伴随着信息全球化程度的不断加深，作为世界上五分之一人口母语的汉语越来越受到世界的关注，与汉语相关的各种实际应用任务对当下的汉语自然语言处理技术提出了更高的要求。但是，因汉语的语言体系、文字系统与英语为代表的字母语言存在着巨大不同，以往基于英语开发的自然语言处理理论和方法在应对汉语文本时可能会失效，且颇失合理性（如：是否应该进行中文分词）[16]。由于词嵌入的质量直接影响自然语言下游任务的性能，因此汉语研究者们对专门适用于汉语的词嵌入进行了大量的研究。文献[17]证明了汉语词嵌入可以借助于所含汉字进行丰富和强化。进一步地，文献[18][19][20]提出汉字可以进一步地拆解为部首等内部组成结构，且通过实验验证了它们也可以增强词嵌入的表达能力。接着，考虑到汉字本身的二维图形结构，Su 和 Lee 等人[21]创造性地提出了用汉字的字形图片来增强词的语义向量表示。最近，Cao 等人[4]首次对汉字的笔画特征进行建模，提出一个中文单词可以分解成一系列类似于英语中子词（subword）的笔画特征，但是，他们的方法对于笔画标准的设定过于简单，且在获得笔画序列的时候忽略了每个词含有的不同汉字之间的边界问题。之后，Wu 等人[22]从图像处理的角度设计了一种对汉字字形图片空间结构进行建模的“田字格 CNN”，将获得的深度隐层表示视为汉字的词嵌入向量。然而该模型并没有对汉语词、字的序列特征进行建模，并且笔画的信息也没被考虑进去，并不能建模汉字和相应字形的序列-空间对应关系。

2 双通道词嵌入模型（DWE）

2.1 问题定义与形式化

词嵌入相较于传统的 one-hot 稀疏表示方法，优势在于可以将每个词表示为一个固定长度的低维向量，并使得这些向量在表达不同词之间语义相似和类比关系上有着较好的效用，从而为计算机对自然语言中的抽象符号进行较为准确的语义计算和数学处理提供了可能。

形式化地说，词嵌入任务目的在于给定任一文本数据集语料$Corpus$，通过假定文本数据$T \in Corpus$中同时出现的词具有更高的相似性，对每个词初始化之后相应的嵌入向量在$Corpus$上基于一定的损失函数$\mathcal{L}$进行学习并不断迭代更新，最终使得这些词嵌入向量能够在词语相似度任务和词义类比任务上有着较好的效果。

2.2 方法描述

如前文所述，无论是直观上来说还是就生物学研究的证明上来看，同时在两个通道上（即，笔画的 n-gram 一维序列性通道和字形的二维空间性通道）对汉语词嵌入进行学习都是合理的，也是必要的。但是，传统的基于分布式假设的词嵌入方法不能很好地捕获汉语的形态学特征。为此，我们提出要将汉语词的表示和字的表示结合起来，并额外对笔画的序列特征和字形的空间特征加以建模，以获得更细粒度、更全面、更具可解释性的汉语词嵌入表示。我们的“双通道词嵌入模型”（DWE）如图 2 所示。为实现词与字的联合建模，我们提出了“双通道字嵌入模块”（Duel-channel Character Embedding, DCE），用于对词含有的字进行形态学信息抽取。每一个双通道字嵌入模块包含了序列性处理通道和空间性处理通道。具体来说，对于一个任意的汉语词w（如“驾车”），它首先会被分解为若干汉字（“驾”和“车”），并且每个汉字会进一步地在“双通道字嵌入模块”中得到处理，以提取出相应的两种通道的形态学信息。

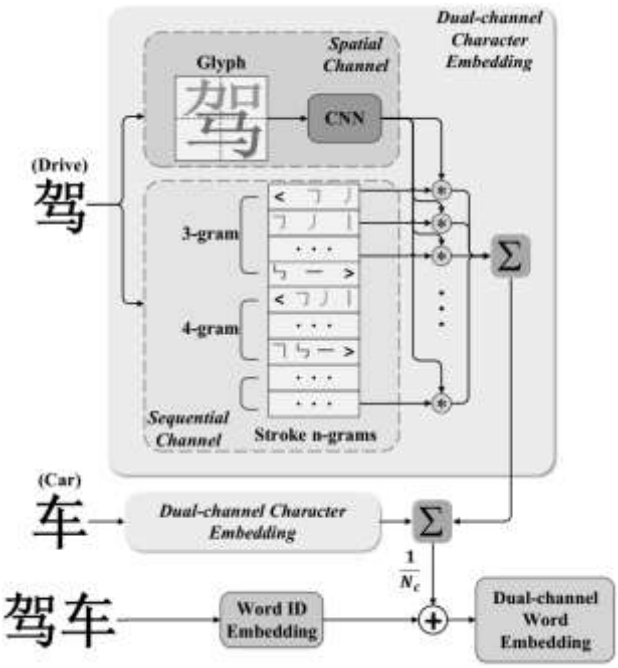

图 2 双通道词嵌入模型（DWE Model）示意图

在序列性通道中，每个汉字会依据汉语书写规范按图 1 所示方式分解为一个唯一的笔画序列。之后，我们在每个字对应的笔画序列的开头和结尾加上特殊的边界符号“<”和“>”，并效仿文献[4][3]采用“笔画 n-gram”方法来提取每个字相应的笔画向量。具体而言，我们首先扫描训练语料库中的每个字，得到一个笔画 n-gram 字典$\boldsymbol{G}$。然后，我们用$\boldsymbol{G}(c)$表示w中每个字c的笔画 n-gram 向量的集合。与此同时，在空间性通道中，为了捕捉隐藏在字形中的语义，我们为每个汉字c生成了位图形式的字形图片I_c，并应用著名的 CNN 结构 LeNet[23]对每个字的字形进行处理，这使得 DWE 模型有着区分鉴别具有相同字形结构的不同汉字的能力。

然后，我们将词的表示和字的表示结合起来，并按如下形式定义了w的词嵌入表示：

$$\boldsymbol{w} = \boldsymbol{w}_{ID} \oplus \frac{1}{N_c}\left(\sum_{c\in w}\sum_{\boldsymbol{g}\in\boldsymbol{G}(c)} \boldsymbol{g} * CNN(I_c)\right), \#(1)$$

其中，$\oplus$和$*$是组合运算符[4]，$\boldsymbol{w}_{ID}$是w依据 one-hot 的 ID 进行查表获取的初始化词嵌入向量，会于模型训练过程中不断优化。N_c 是词w含有的字的数量，$\boldsymbol{g}$是汉字c的笔画 n-gram 向量。依据该公式我们可以看出，通过对一个汉语词w在N_c个字以及含有的所有笔画上的拆解分析，词w在最基本的 ID 词嵌入之上融合了N_c个字的额外形态学信息，这些信息包括了每个字含有的笔画 n-gram 向量$\boldsymbol{g}$、每个字的字形图片I_c经过 CNN 特征抽取得到的向量$CNN(I_c)$。最重要的是，借助于组合运算符$*$，我们可以实现笔画g与字形特征I_c之间相互关系的建模，这也是实现序列特征和空间特征得到联合学习的关键。不仅如此，由于汉语中词的概念很模糊，且有限的汉字理论上可以排列组合出近乎无限的词，对于那些不断涌现的未知的词，我们模型的优势就体现了出来，对形态学的感知能力使得 DWE 有着不断学习的能力，且模型的容量和泛化能力均可因此得到提升。

为了评估词嵌入的质量，我们与该领域普遍使用的方法一致[4][21]，通过定义一个相似度得分函数来计算当前中心词w与它的上下文窗口词e之间的相似性：

$$s(w,e) = \boldsymbol{w}\cdot\boldsymbol{e}, \#(2)$$

其中，$\boldsymbol{w}$和$\boldsymbol{e}$分别是词w和e的词嵌入向量。此外，为了对模型进行训练，我们定义损失函数如下：

$$\mathcal{L} = -\sum_{w\in D}\sum_{e\in T(w)} log\sigma\big(s(w,e)\big) + \lambda\mathbb{E}_{e'\sim P}[log\sigma(-s(w,e'))], \quad (3)$$

[3] 需要注意的是，相较于这篇文章定义的过于简单的笔画标准（5 种笔划），我们参照《GB13000.1 字符集汉字折笔规范》采用了更丰富的笔划标准（32 种笔划）

[4] $\oplus$和$*$可以表示多种运算方式，即$\oplus$可表示加法、拼接操作等，而$*$可表示乘法和逐项点积等。在本文中，我们采用$\oplus$表示加法运算，用$*$表示逐项点积运算

表 1 不同模型在词语相似度和词义类比任务上的实验结果。词嵌入的维度统一设置为 300，词语相似度任务采用 Spearman 相关系数ρ作为评价标准，词义类比任务采用预测准确率作为评价标准。

Model	Word Similarity		Word Analogy			
			$3CosAdd$		$3CosMul$	
	wordsim-240	wordsim-296	Capital	Family	Capital	Family
CWE[17]	0.5035	0.4322	0.1846	0.1875	0.1713	0.1583
GWE[21]	0.5531	0.5507	0.5716	0.2417	0.5761	0.2333
JWE[20]	0.4734	0.5732	0.1285	0.2708	0.1492	0.2500
cw2vec[4]	0.5529	0.5992	0.5081	0.2941	0.5465	0.2721
GloVe[28]	0.4981	0.4019	0.6219	0.3167	0.5805	0.2375
CBOW[1]	0.5248	0.5736	0.6499	0.3750	0.6219	0.2904
sisg[15]	0.5592	0.5884	0.4978	0.2610	0.5303	0.2206
Skipgram[1]	0.5670	0.6023	**0.7592**	0.3676	**0.7637**	0.3529
DWE	**0.6105**	**0.6137**	0.7120	**0.6250**	0.6765	**0.6140**

其中，$T(w)$是中心词w的上下文窗口词集合，λ是每个中心词w的负样本数量，e'是负采样得到的负样本噪声词，$\mathbb{E}_{e'\sim P}[\cdot]$是期望函数项。对于训练语料库$D$中的每个词$w$来说，其相应的$\lambda$个负样本是依据概率分布$P$进行采样得到的，而$P$通常被设置为负样本词的一元模型分布（unigram distribution），σ则是 sigmoid 函数：

$$\sigma(x)=\frac{1}{1+e^{-x}}.\#(4)$$

依据式(3)可见，当中心词w和窗口内的背景词e的相似度$s(w,e)$越大、中心词w和未出现在窗口内的负样本噪声词e'的相似度$s(w,e')$越小时，则目标损失$\mathcal{L}$越小，即词w的语义表征越准确。

3 实验与分析

在本小节中，我们首先介绍了实验所采用的数据集以及相关实验工具和环境配置，然后介绍了所采用的模型评价任务和方法，并给出了相应的参数设置。进一步地，我们简明扼要地介绍了相关的对比方法和各自的优势，最后对实验结果进行了探讨和分析。

3.1 数据集与实验环境

我们从用于 NLP 研究的大规模中文自然语言处理语料[5]下载了一定量的中文维基百科文章作为数据集。为了实现中文分词和停用词[6]过滤，采用了 jieba[7]工具包对数据集进行了分词处理，最后得到了 11,529,432 个词。与文献[17]中的做法一致，所有 Unicode 编码处于 0x4E00 和 0x9FA5 区间中的字符均被归类为中文字符。此外，我们从互联网上的在线新华字典[8]爬取了所有的 20,402 个汉字，并借助 Python 的 Pillow[9]工具包为每个汉字生成了 28x28 的灰度字形图片用于 DWE 模型在空间上对汉字进行形态学分析。

3.2 评价方法与参数设置

为了对词嵌入的性能进行验证和评估，我们选取了 adagrad [24]作为模型的优化算法，并设置$batch$大小为 4,096，学习率为 0.05，笔画 n-gram 的滑动窗口大小n的取值范围设置为$3\le n\le 6$，每个词的负样本个数$\lambda=5$。为了公平地评估每个模型词嵌入的优劣，我们将不同模型的词嵌入维数一致设置为 300 维，并使用两个测试任务来评估不同模型的性能：一个是词语相似度（word similarity），另一个是词义类比（word analogy）。

对于词语相似度任务来说，一个词语相似度测试是由多个词对和人类标注的相似度分数组成。形式化来说，对于任意两个词x和y，它们的相似度被定义为其向量化表示$\boldsymbol{x}$和$\boldsymbol{y}$的夹角余弦值，即余弦相似度：$\frac{x^T y}{\|x\|\|y\|}\in[-1,1]$，公式上部即为相似度函数 $s(x,y)=\boldsymbol{x}\cdot\boldsymbol{y}$。一个高质量的词嵌入向量应该使得计算出的相似度与人类标注的分数有很高的相关性，这通常是由 Spearman 相关系数ρ来衡量的[25]。

而词义类比问题的形式就像等式“国王”:“女王”=“男人”:“？”中，“？”处“女人”是最合适且最恰当的回答。也就是说在这个任务中，每给定三个单词a、b和c，目标是推断出第四个单词d，使得“c与d之间的语义关系”相似于“a与b之间的语义关系”，而一个模型对该相似度计算的越准确，说明该模型越能够准确捕获和建模自然语言中的逻辑和语义。针对词义类比这一任务，我们采用$3CosAdd$[1]和$3CosMul$[26]函数来计算相似度。为了公平地对比模型在词语相似度计算任务上的性能，我们使用了和文献[4][17]中相同的评价方法和两个有着人工标注的数据集，即 wordsim-240、wordsim-296[27] 和一个

[5] https://github.com/brightmart/nlp chinese corpus
[6] https://github.com/YueYongDev/stopwords
[7] https://github.com/fxsjy/jieba
[8] https://bihua.51240.com/
[9] https://github.com/python-pillow/Pillow

three-group 数据集[10]。

3.3 对比实验方法

为了充分验证 DWE 模型的优势和有效性，我们复现了诸多先进的词嵌入模型，并对相应的论文与代码进行了引用。借助于 gensim[11]，我们复现了经典的 CBOW 和 Skipgram 模型。借助于作者公布的代码，我们对另一个广泛使用的词嵌入模型 GloVe[28][12]进行了复现。为了充分评估本文所提模型在中文词嵌入建模上的有效性，我们也对同样致力于解决中文词嵌入问题的 CWE[13]、JWE[14]、GWE[15]模型进行了复现。

由于文献[4]的作者并没有公开 cw2vec 模型的代码，故我们遵循其论文中描述的细节在 mxnet[16]框架下实现了 cw2vec 模型，并利用该框架实现了 sisg 模型[15][17]以及我们的 DWE 模型。

3.4 实验结果与分析

实验结果如表 1 所示。通过观察比较可以发现我们的 DWE 模型在融入了汉字笔画和字形信息后，于数据集 wordsim-240 和 wordsim-296 的相似性任务上都取得了预期的最佳效果。同时我们也可以注意到，DWE 模型在词义类比任务的“Family”（家族关系）类别上相较于其他模型有着绝对压倒性的优势，但在“Capital”（首都）的词义类比任务上略逊色于基于分布式假设的词嵌入方法。为了探讨这其中的原因并给出直观的解释，我们通过结合相关工作的探索深入到汉语的特点与数据当中，给出了如下的分析。

首先，对于词语相似度任务来说，汉语相较于英语在文字系统上有着天然的象形和表意优势，其中汉字的部首就是一种十分重要的语义单元。文献[29]指出，汉字的部首有着悠久的字形起源，具有相同部首的不同汉字在语义上有着天然的联系。文献[21]也指出，汉字是由包括部首在内的若干组件组成， 具有相同组件的不同汉字具有相似的语义或发音，当汉语阅读者在见到一个陌生汉字时，根据构成该字的部首等组件猜测其语义和发音是本能的，因此理解汉字的图形组件并将它们与语义相关联有助于人们学习和理解中文。例如:“蝇”（fly),“蚊”(mosquito),“蜂”（bee),“虱”（louse）和“蚁”（ant）五个汉字均含有部首“虫”，故我们可以直观地从形态学上知悉它们均属于“昆虫”，而 DWE 可以很准确地捕获到这些汉字共同含有的部首的笔画序列信息以及其字形空间信息，这使得 DWE 学习到的汉语词嵌入在表征形态学信息时有着天然的优势，从而在相似性任务上表现出了最佳的性能。

其次，对于词义类比任务来说，汉语中与家族关系相关的字词相较于英语等字母语言来说有着十分明显的形态学特征，因此我们得到这些实验结果并不是偶然的。即，与女性亲属相关的汉字大部分均带有字形组件“女”，与男性亲属相关的汉字大多有着字形组件“父”、“男”等，与晚辈家庭成员相关的汉字则多数含有字形组件“儿”、“子”等。为了给出直观验证，在词义类比对“兄弟”：“姐妹”=“儿子”：“女儿”中，我们可以很轻易地发现“兄弟”与“儿子”有着相同的字形组件“儿”，同时“姐妹”和“女儿”有着相同的字形组件“女”，而这些形态学信息对于我们的 DWE 模型来说可以被轻易捕捉到，因此 DWE 在“Family”（家族关系）类别上有着十分优越的性能。反观“Capital”（首都）的词义类比任务，其中表达国家与城市名字的词几乎全为音译词，导致它们的词义与所含汉字的形态学信息几乎毫无关联，这与文献[21]给出的结论也是一致的。例如：在词义类比对“西班牙”(Spain)：“马德里”(Madrid)=“法国”(France)：“巴黎”(Paris)中，因为它们相应的汉语词是基于英文音译过来的，我们无法从字面上获取他们之间的任何语义关系，也就导致 DWE 所提取的形态学信息在语义分析上作用略小，从而基于分布式假设的 CBOW 和 Skipgram 等模型在该类型的词义类比任务上略胜一筹。

总结来说，汉语中除了音译词之外，形态上相似的词大多具有十分相似的语义，同时对汉字的笔画序列和字形空间信息进行建模的确可以有效改进汉语词嵌入的语义表征。

4 结论与展望

在这篇文章中，我们分析了英语等字母语言与汉语在形态学辅助语义传递和理解上的共同点和不同点，并结合我们的直观感受和生物学上的研究成果，详细探讨了我们利用形态学信息丰富汉语词嵌入表征的动机和合理性。然后，我们提出了适用于汉语的“双通道词嵌入模型”（DWE），通过学习汉字笔画的序列通道信息和字形的空

[10] Capitals of countries, (China) provinces of cities, and family relations.

[11] https://radimrehurek.com/gensim/

[12] https://github.com/stanfordnlp/GloVe

[13] https://github.com/Leonard-Xu/CWE

[14] https://github.com/hkust-knowcomp/jwe

[15] https://github.com/ray1007/GWE

[16] https://mxnet.apache.org/

[17] http://gluon-nlp.mxnet.io/master/examples/word embedding/word embedding training.html

间通道信息，实现了对汉语形态学信息的建模和挖掘，经过“词语相似度”和“词义类比”两种任务的评估，我们的模型展现出了其在捕获汉语形态学信息上的强大能力，证明了汉语的形态学信息在丰富汉语词嵌入表征上的有效性，为汉语词嵌入研究提供了一定的新见解。在将来的工作中，我们将进一步地挖掘并研究汉语词嵌入的相关问题，在汉语形态学信息的建模上做出更多的探索。

参考文献

[1] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[2] Cotterell R, Schütze H. Morphological word-embeddings[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1287-1292.

[3] **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**, 33(5): 17-23.

[4] Cao S, Lu W, Zhou J, et al. cw2vec: Learning chinese word embeddings with stroke n-gram information[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018: 5053-5061.

[5] Springer J A, Binder J R, Hammeke T A, et al. Language dominance in neurologically normal and epilepsy subjects: a functional MRI study[J]. Brain, 1999, 122(11): 2033-2046.

[6] Knecht S, Deppe M, Dräger B, et al. Language lateralization in healthy right-handers[J]. Brain, 2000, 123(1): 74-81.

[7] Paulesu E, McCrory E, Fazio F, et al. A cultural effect on brain function[J]. Nature neuroscience, 2000, 3(1): 91.

[8] Tan L H, Spinks J A, Gao J H, et al. Brain activation in the processing of Chinese characters and words: a functional MRI study[J]. Human brain mapping, 2000, 10(1): 16-27.

[9] Harris Z S. Distributional structure[J]. Word, 1954, 10(2-3): 146-162.

[10] Luong T, Socher R, Manning C. Better word representations with recursive neural networks for morphology[C]//Proceedings of the Seventeenth Conference on Computational Natural Language Learning. 2013: 104-113.

[11] Qiu S, Cui Q, Bian J, et al. Co-learning of word representations and morpheme representations[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014: 141-150.

[12] Sak H, Saraclar M, Güngör T. Morphology-based and sub-word language modeling for Turkish speech recognition[C]//2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010: 5402-5405.

[13] Soricut R, Och F. Unsupervised morphology induction using word embeddings[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1627-1637.

[14] Wieting J, Bansal M, Gimpel K, et al. Charagram: Embedding words and sentences via character n-grams[J]. arXiv preprint arXiv:1607.02789, 2016.

[15] 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.

[16] Meng Y, Li X, Sun X, et al. Is Word Segmentation Necessary for Deep Learning of Chinese Representations?[J]. arXiv preprint arXiv: 1905.05526, 2019.

[17] Chen X, Xu L, Liu Z, et al. Joint learning of character and word embeddings[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015: 1236-1242.

[18] Sun Y, Lin L, Yang N, et al. Radical-enhanced chinese character embedding[C]//International Conference on Neural Information Processing. Springer, Cham, 2014: 279-286.

[19] Yin R, Wang Q, Li P, et al. Multi-granularity chinese word embedding[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 981-986.

[20] Yu J, Jian X, Xin H, et al. Joint embeddings of chinese words, characters, and fine-grained subcharacter components[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 286-291.

[21] Su T R, Lee H Y. Learning chinese word representations from glyphs of characters[J]. arXiv preprint arXiv:1708.04755, 2017.

[22] Wu W, Meng Y, Han Q, et al. Glyce: Glyph-vectors for Chinese Character Representations[J]. arXiv preprint arXiv:1901.10125, 2019.

[23] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[24] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research, 2011, 12(Jul): 2121-2159.

[25] Zar J H. Significance testing of the Spearman rank correlation coefficient[J]. Journal of the American Statistical Association, 1972, 67(339):

578-580.
[26] Levy O, Goldberg Y. Linguistic regularities in sparse and explicit word representations[C]//Proceedings of the eighteenth conference on computational natural language learning. 2014: 171-180.
[27] Jin P, Wu Y. Semeval-2012 task 4: evaluating chinese word similarity[C]//Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2012: 374-377.
[28] 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.
[29] Tao H, Tong S, Zhao H, et al. A Radical-aware Attention-based Model for Chinese Text Classification[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 2019: 5125-5132.

陶汉卿（1995—），中国科学技术大学计算机学院博士研究生。主要研究领域为自然语言处理，包括文本挖掘、文本嵌入和建模等。
E-mail：hqtao@mail.ustc.edu.cn

陈恩红（1968—），通信作者，博士，教授，博导，中国科学技术大学大数据学院常务副院长、计算机学院副院长、教授，国家杰出青年科学基金获得者。主要研究领域为机器学习与数据挖掘、信息检索与文本挖掘等。
E-mail：cheneh@ustc.edu.cn