计算机科学
COMPUTER SCIENCE

细粒度语义知识图谱增强的中文OOV词嵌入学习

陈姝睿，梁子然，饶洋辉

引用本文

陈姝睿，梁子然，饶洋辉.细粒度语义知识图谱增强的中文OOV词嵌入学习[J]．计算机科学, 2023, 50(3): 72-82.

CHEN Shurui, LIANG Ziran, RAO Yanghui. Fine-grained Semantic Knowledge Graph Enhanced Chinese OOV Word Embedding Learning [J]. Computer Science, 2023, 50(3): 72-82.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700249

细粒度语义知识图谱增强的中文 OOV 词嵌入学习

陈姝睿　梁子然　饶洋辉
中山大学计算机学院　广州 510006
(chenshr8@mail3.sysu.edu.cn)

摘　要　随着信息化领域的范围不断扩大,许多特定领域的文本语料开始涌现。这些特定领域,如医疗、通信等,由于受到安全性和敏感性的影响,其数据规模通常较小,传统的词嵌入学习模型难以获得有效的结果。另一方面,直接应用现有的预训练语言模型时会出现较多未登录词,这些词汇无法表示成向量,从而影响下游任务的性能表现。许多学者开始研究如何利用细粒度语义信息来得到较高质量的未登录词向量表示。然而,当前的未登录词嵌入学习模型大多针对英文语料,对中文词的细粒度语义信息只能进行简单的拼接或映射,难以在中文未登录词嵌入学习任务中得到有效的向量表示。针对上述问题,首先通过中文构字规则,即中文词所包含的汉字、汉字所包含的部件和拼音等,构建细粒度的知识图谱,使其不仅能涵盖汉字和单词之间的关联关系,还能对拼音和汉字、组件和汉字等细粒度语义信息之间的多元且复杂的关联关系进行表征。然后,在知识图谱上运行图卷积算法,从而对中文词的细粒度语义信息之间以及它们与词语义之间更深层次的关系进行建模。此外,文中通过在子图结构上构建图读出来进一步挖掘细粒度语义信息与词语义信息之间的组成关系,据此提升模型在未登录词嵌入推断中的精准度。实验结果表明,在面对未登录词占比较大的特定语料上的词配对、词相似任务,以及文本分类、命名实体识别等下游任务时,所提模型都取得了更好的性能。

关键词:未登录词嵌入学习;中文细粒度语义信息;细粒度知识图谱;图卷积网络学习
中图法分类号　TP311

Fine-grained Semantic Knowledge Graph Enhanced Chinese OOV Word Embedding Learning

CHEN Shurui, LIANG Ziran and RAO Yanghui
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China

Abstract　With the expansion of the scope in informatization fields, lots of text corpora in specific fields continue to appear. Due to the impact of security and sensitivity, the text corpora in these specific fields (e.g., medical records corpora and communication corpora) are often small-scaled. It is difficult for traditional word embedding learning methods to obtain high-quality embeddings on these corpora. On the other hand, there may exist many out-of-vocabulary words in these corpora when using the existing pre-training language models directly, for which, many words cannot be represented as vectors and the performance on downstream tasks are limited. Many researchers start to study how to infer the semantics of out-of-vocabulary words and obtain effective out-of-vocabulary word embeddings based on fine-grained semantic information. However, the current models utilizing fine-grained semantic information mainly focus on the English corpora and they only model the relationship among fine-grained semantic information by simple ways of concatenation or mapping, which leads to a poor model robustness. Aiming at addressing the above problems, this paper first proposes to construct a fine-grained knowledge graph by exploiting Chinese word formation rules, such as the characters contained in Chinese words, as well as the character components and pinyin of Chinese characters. The knowledge graph not only captures the relationship between Chinese characters and Chinese words, but also represents the multiple and complex relationships between Pinyin and Chinese characters, components and Chinese characters, and other fine-grained semantic information. Next, the relational graph convolution operation is performed on the knowledge graph to model the deeper relationship between fine-grained semantics and word semantics. The method further mines the relationship between fine-grained semantics by the sub-graph readout, so as to effectively infer the semantic information of Chinese out-of-vocabulary words. Experimental results show that our model achieves better performance on specific corpora with a large proportion of out-of-vocabulary words when applying to tasks such as word analogy, word similarity, text classification, and named entity recognition.

Keywords　Out-of-vocabulary word embedding learning, Chinese fine-grained semantic information, Fine-grained knowledge graph, Graph convolution network learning

到稿日期:2022-07-26　返修日期:2022-12-10
基金项目:国家自然科学基金面上项目(61972426)
This work was supported by the National Natural Science Foundation of China(61972426).
通信作者:饶洋辉(raoyangh@mail.sysu.edu.cn)

1 引言

在文本分类和聚类等自然语言处理(Natural Language Processing，NLP)任务中，如何将文本表示为数值向量具有重要意义。当前，NLP 模型通常将大规模语料上预训练好的词嵌入向量直接用于下游任务中，即在海量文本数据上，预先通过词嵌入学习方法[1-4]进行训练，得到对应的词嵌入向量，之后在下游任务中直接使用这些预训练的词嵌入向量作为输入。近年来，随着 BERT 模型[4]的提出，许多基于 Transformer 的模型[5]以及 BERT 模型的变体[6-10]被提出。这些模型在大规模语料上取得了很好的成效，可以为下游任务提供更好的初始化词嵌入向量，并且能够以更快的速度收敛，有效减少了所需的计算资源[11]。

这类预训练的词向量虽然具有较好的通用性，但由于预训练使用的大规模语料大多是从新闻、百科全书、社交媒体等通用领域产生的，直接将模型运用在医疗、通信等特定领域下游任务中往往会出现较多未登录(Out-of-vocabulary，OOV)的词汇，即出现在预训练词表外。这种情况下，下游任务在面对 OOV 词时只能统一把它们当作未知词(Unknown word)来处理，并以〈UNK〉等符号进行标记，而无法使用向量来表示未登录词的语义。此外，在这些特定领域的下游任务中，可能出现预训练的词嵌入向量与领域内该词的实际语义不一致的现象。例如，“定期”一词在医疗文本中通常反映固定时间段需要复诊的语义，而在金融语料中表示一种储蓄方式。也就是说，在特定领域内，OOV 词的上下文信息是带有领域特殊性的，且特殊领域中能够利用的文本数据规模非常小，模型无法直接利用通用领域下预训练好的上下文信息对 OOV 词进行推断或者根据小样本数据微调预训练模型。为了更好地表示特定领域下游任务中出现的词汇，许多学者提出在使用预训练语言模型得到通用词表和词嵌入后，还需要基于小样本语料对出现频次较少的词(即 OOV 词)进行嵌入学习，并对 OOV 词嵌入进行有效推断[12]，这类学习可以被视作小样本学习的一种类型[13]。

本文基于中文特殊的构词方法，在这些特定领域下，通过图神经网络拟合中文构字和构词，据此解决 OOV 词较少出现的问题，并利用细粒度语义信息构建的知识图谱来对 OOV 词嵌入进行有效推断。对于上述“定期”这个示例，该中文词的细粒度结构包括汉字“定”、部件“月”等，它们是更底层的单词组成结构，不具有领域特殊性，因此本文提出可以利用一些高质量的在通用语料上预训练好的中文词嵌入，来辅助学习汉字和部件等细粒度语义信息，进而通过细粒度语义信息对下游特定领域中的 OOV 词嵌入进行精准推断。此外，尽管在下游语料中 OOV 词的出现次数可能非常少，但是它的细粒度结构往往较为丰富，因此我们可以通过学习细粒度语义信息及其构词规则来对出现次数受限的 OOV 词进行推断。当前，细粒度语义信息的 OOV 词向量学习方法主要基于模仿学习的方式[14]，其流程如下：首先，基于通用语料预训练高质量的词嵌入；其次，通过神经网络学习细粒度结构到预训练词嵌入之间的映射关系；最后，利用上述映射关系来推断 OOV 词嵌入。然而，当前利用细粒度语义信息的模型主要以英文语料为研究对象，若是直接将这些模型应用于中文词嵌入的学习，则会丢失中文字形中本身携带的一些信息。中文作为世界上使用人数最多的语言，所产生的中文语料是自然语言处理的重要数据来源。中文含有的象形文字特点及其特殊的语法规则，使得模型较难直接通过英文词嵌入学习模型获得高质量的中文词嵌入结果，因为这些英文词嵌入学习模型无法利用中文的汉字内部结构所含有的丰富语义信息。另一方面，许多中文词嵌入学习模型仅对细粒度语义信息进行简单的求和拼接[15-16]，或是转为图像后再提取视觉信息[17-18]，它们仅简单地针对细粒度语义信息与词语义信息的关系进行建模，忽视了细粒度语义信息之间的关联关系。此外，这类模仿学习方法生成的细粒度嵌入与词嵌入处在两个不同的空间，这在实际应用中可能会导致细粒度语义信息和实际词嵌入语义信息存在没有对齐的问题。

基于以上观察，本文提出使用知识图谱的方式保存和学习中文细粒度语义信息，包括汉字、汉字部件和拼音。该方法利用节点和边的关系建模中文单词的语义构成过程，使得细粒度节点与词节点能够在同一嵌入空间上进行语义信息的传递与聚合。构建的知识图谱除了可以建模细粒度语义信息和词语义信息之间的关系以外，还可以建模汉字和汉字部件、汉字和汉字拼音等细粒度语义信息之间的关系。然后，通过图卷积方法对知识图谱上的节点拓扑关系进行挖掘，得到更深层次的节点嵌入表示以及 OOV 词嵌入表示。

2 相关工作

2.1 词及其细粒度结构层面的嵌入学习

近年来，神经语言建模过程因在学习单词语义分布方面的有效性而引起了广泛关注。传统的词嵌入学习方法 SGNS[2]，通过预测给定目标词及其上下文词的正样例词对和随机的 k 个负样例词对来学习词的嵌入分布。而 Glove[3]通过整合全局统计信息扩展了上述模型。随着深度学习的发展，许多方法(如 BERT[4]等)都致力于充分利用语境化语义来学习单词的高质量表示。在细粒度级别的中文词嵌入学习任务上，CWE [15]使用字作为最小语义单元，它在 SGNS[2]的基础上学习字的嵌入。JWE [16]则把字的组成部分作为最小语义单元，联合单词、汉字以及汉字组成的部件进行学习。cw2vec[19]把笔画作为语义单元，使用 n-gram 的方法对笔画进行特征提取。这些方法都是对中文词不同级别的细粒度信息进行简单的特征提取与加和。

2.2 OOV 词的嵌入推断方法

现有的大多数词嵌入学习方法都需要大量的语料以获得充分的训练样本，从而实现高质量的单词语义分布学习。但是许多词嵌入学习方法会去掉语料中的低频词，从而产生了许多不在预训练词表中的词，即未登录词。在下游任务中，模型遇到未登录词时只能统一将它们当作未知词进行标记，这样的处理方式会使得语料中的文本出现一定程度的语义信息损失。面对下游语料中存在的 OOV 词，现有的工作提出了不同的嵌入推断方法，主要可以分为 3 类：1)基于词出现的上下文信息进行推断；2)基于词的细粒度信息进行推断；3)通过之前任务学习到的语义知识对新语料中的 OOV 词进行适配

性学习。Singh 等[20]提出把单词对应为其字符的多个 n-gram 单元，根据词的字符 n-gram 单元的相似程度，选择与未登录词最相似的 K 个近邻词，并用这 K 个近邻词加权求和后的嵌入向量来表示 OOV 词的嵌入向量。Bojanowski 等[21]同样对单词的字符序列进行 n-gram 特征提取，然后基于 SGNS 框架学习单词中每个 n-gram 单元的嵌入向量，最后以单词的 n-gram 嵌入向量的加和作为单词的语义嵌入向量表示。Pinter 等[14]提出了 Mimick 模型，其使用双向 LSTM 模型分别提取单词字符序列的前向、后向信息，然后经过多层感知机得到模型推断的词向量，将它与预训练好的词嵌入进行对比训练，使得推断得到的词向量尽量接近预训练的词向量。这个过程可以理解为模仿(Mimick)学习。Schick 等通过融合上下文信息和字形信息来为 OOV 词推断嵌入向量[22]，根据信息量的不同为上下文分配不同的权重[23]，并且针对 BERT 模型[4]的分词机制中将单个长词或者短语拆分为多个子词嵌入的问题，提出了采用一个词嵌入向量来近似多个子词嵌入向量的方法，以便将长词或短语的细粒度信息进行融合，进而为模仿学习提供高质量的词嵌入向量基准[24]。Fukuda 等首先提取与 OOV 词形态相似的已知词，然后通过 CNN 计算每个已知词的相似性分数，据此聚合已知词的词嵌入并将其作为 OOV 词的嵌入表示[25]。Chen 等[8]认为在模仿学习过程中，模型只根据细粒度语义信息学会了如何学习正样例，但是没有学习负样例，为此提出在对英文单词进行数据增强后，完成正样本和负样本之间的对比学习，同时在推断过程中使用位置嵌入的自注意力机制来替代传统的基于 RNN 的编码器。Hu 等除了融合上下文信息和字形信息之外，还通过 MAML 元学习算法对领域差异性进行适配学习[13]。Buck 等则在上述研究的基础上提出了 Leap 元学习算法[27]，该方法解决了 MAML 中存在的不稳定性，进一步提升了模型的效果。

上述 OOV 词嵌入推断方法都是针对英文单词设计的，而对于中文语料而言，直接使用这类方法容易丢失中文字形中的细粒度语义信息，使其很难得到有效的词嵌入表示。为了解决此问题，Chen 等[18]针对中文字形提出了 Glyph2vec 模型。与 Mimick 的学习思想类似，他们使用中文的汉字、汉字图片、仓颉汉字编码等特征作为模型输入，然后利用 GRU 模块[28]作为编码器进行模仿学习从而推断词向量。此外，还有一类方法基于 BERT 模型对中文细粒度语义信息进行提取，以改进中文词嵌入的学习效果。例如，Sun 等在 BERT 模型的基础上，融入了中文的字形信息和拼音信息[29]。Li 等利用词性标注信息进一步提升了基于字符级的中文 BERT 模型的效果[30]。但是这类方法均是基于 BERT，需要从头开始训练，计算资源消耗大，且不利于对新词进行拓展。此外，现有的模型在利用中文细粒度语义信息对 OOV 词嵌入进行推断的研究上仍有不足，无法建模汉字细粒度部件之间的语义关系。

3 方法描述

3.1 知识图谱构建与图卷积

在汉字的组成部件及其组成结构的语义研究中，现有的工作主要从汉字和部件层面对细粒度特征和词特征进行加和[15-16]，或在笔画层面通过将每个汉字编码成序列，进而提取 n-gram 特征的加和来表示语义信息[18-19,31]。但是这些方法没有考虑到部件在不同的组成结构下可能存在不同的语义信息贡献，并且忽视了部件到汉字的构造关系。本文提出的模型通过挖掘中文词中含有的汉字、部件和拼音等细粒度语义信息，并且利用知识图谱构造细粒度特征之间的关系，来辅助 OOV 词的嵌入学习。

首先，对于部件信息，其在汉字中的不同组成结构和组成关系对汉字语义的表达有着不同的影响。例如，对于两个常见字“景”与“晾”来说，虽然它们的组成部件相同，但是上下结构的“景”和左右结构的“晾”具有很大的语义差别；对于两个部件“木”和“口”来说，它们可以在同样的上下结构下，组成两个语义不同的汉字，即“杏”和“呆”。因此，对于汉字和部件之间的关系，用简单的加和关系来构建是不合适的。Zhang 等[31]考虑到了汉字结构的信息，但他们只对汉字的 13 种部件的组成结构进行统一编码，并和笔画以及拼音编码一起作为序列输入至模型，而没有将组成的部件与其对应的结构组成关系联系起来。基于以上观察，表 1 列出了汉字组成结构和组成部件之间的关系，并且本文模型将这种关系用知识图谱的方式进行表征学习。除了镶嵌结构和独体结构之外，其他组成结构至少有两种部件。通过将部件和对应组成关系联系起来，可以得到 26 种汉字与部件之间的组成关系。

表 1 汉字组成结构与组成部件之间的关系

Table 1 Relationship between structure and components of Chinese characters

组成结构	关系	示例	组成结构	关系	示例
左右结构	左	格	下包围结构	外	凶
	右	格		里	凶
左中右结构	左	树	左包围结构	外	区
	中	树		里	区
	右	树	左上包围结构	外	店
上下结构	上	花		里	店
	下	花	右上包围结构	外	可
上中下结构	上	慧		里	可
	中	慧	左下包围结构	外	过
	下	慧		里	过
全包围结构	外	回	镶嵌结构	镶嵌	果
	里	回	独体结构	独体	工
上包围结构	外	闪			
	里	闪			

其次，对于拼音信息，汉字拥有的拼音信息可以是一种或者多种，分别对应于单音字和多音字。在多音字中，不同的发音通常会导致同一个汉字的语义完全不同。例如，“乐”在发音“yue(四声)”时，表示动听的声音，在发音“le(四声)”时往往表示心情愉悦。同样地，本文将拼音和汉字之间的构成关系看作一种组成关系，并且一个汉字可能与多个拼音具有组成关系。汉字中，一个词往往由多个字组成，同理可以把汉字和中文词之间的构成看作一种组成关系。基于以上分析，本文构建了中文词与细粒度语义信息之间的组成关系，以及不同的细粒度语义信息之间的组成关系。

上述组成关系是一种图结构，它表示细粒度语义信息

之间以及它们和词语义信息的关系。在这种异构网络的知识图谱上，本文模型还引入了 WordNet[32] 中的近义词和近义字关系，使得知识图谱含有单词之间、汉字之间的语义关系。例如，“变化”在 WordNet 中的近义词是“改变”，通过将这种近义关系加入到本文所提的知识图谱中，使得节点拥有的邻居信息更加丰富。基于上述分析，本文构建了一个如图 1 所示的知识图谱。

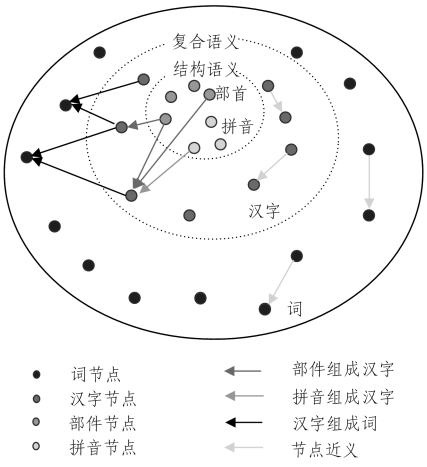


图 1 中文细粒度语义知识图谱

Fig. 1 Chinese fine-grained semantic knowledge graph

该知识图谱是异构的，即节点有多种类型，且节点之间连接的边有多种类型。其中，节点分为 4 种类型，包括汉字组成部件节点、拼音节点、汉字节点和单词节点。由表 1 可知，在连接汉字组成部件节点和汉字节点之间的边中，一共有 26 类边关系，表示部件在汉字中属于某种特定结构。比如对于部件“各”和汉字“格”，两个节点之间的边类型是左右结构中的左结构。在拼音节点和汉字节点之间，由拼音指向汉字，表示该汉字的拼音组成。最后，在汉字节点和单词节点之间，由汉字指向单词节点，表示该单词含有的汉字。

为了获得并训练图的节点表示和边表示，本文模型使用基于图卷积网络的学习方法，其中节点的信息传递使用异构图的聚合方式。参照 RGCN[33] 的思路，每种边连接类型都有自己的权重参数。另外，在大规模图网络中，为了保留一定的归纳式学习能力，即可以对新出现的节点进行快速学习的能力，本文模型在 RGCN 的网络上应用了 GraphSage[34] 的邻居节点采样模式。最后得到中心节点 i 通过采样得到的邻居节点集合 N_{nei} 和边类型集合 $R(N_{\text{nei}})$ 的更新过程，即如式(1)所示的聚合函数：

$$h_i^k=\sigma\left(\sum_{r\in R(N_{\text{nei}})}\sum_{j\in N_{\text{nei}}^r}\frac{1}{|N_{\text{nei}}^r|}\boldsymbol{W}_r^{k-1}h_j^{k-1}\boldsymbol{W}_o^{k-1}h_i^{k-1}\right) \tag{1}$$

其中，边类型集合 $R(N_{\text{nei}})$ 表示在采样得到的邻居节点中，中心节点 i 与这些邻居节点的边关系的类型集合；N_{nei}^r 表示以边关系 r 为连接边的采样邻居节点的集合；h_i^k 表示在图卷积网络的第 k 层节点 i 的隐层嵌入；$\boldsymbol{W}_r^{k-1}$ 表示在图卷积网络的第 $k-1$ 层由边关系 r 决定的权重参数；$\boldsymbol{W}_o^{k-1}$ 是中心节点保留上一层网络中自身嵌入信息的权重参数；σ 表示 sigmoid 激活函数。由于在该公式下，模型的参数量会随着边关系数量的增加而大幅增长，并且对于部分在图中出现次数较少的边关系而言，其在训练过程中很容易出现过拟合的现象。因此 RGCN[33] 采用了权重共享策略，通过用更少的 B 组基础权重参数进行加权求和的方式得到不同的 R 组边关系的权重参数，如式(2)所示：

$$\boldsymbol{W}_r^k=\sum_{b=1}^{B}a_{rb}^k\boldsymbol{V}_b^k \tag{2}$$

其中，a_{rb}^k 表示对于图卷积网络第 k 层的隐层节点，在边关系 r 下，第 b 组的基础权重参数的相关系数；$\boldsymbol{V}_b^k$ 表示对于图卷积网络第 k 层的隐层节点，第 b 组的基础权重参数的线性变化权重。

3.2 知识图谱的训练目标函数

在更新知识图谱的过程中，现有的训练方法主要分为有监督和无监督两种。其中，无监督方法的目标大多是让近邻节点之间的隐层嵌入更相似，非近邻节点之间的隐层嵌入距离更远；有监督方法的目标则是使用图隐层节点更好地拟合标签。本文使用的是无监督的学习方法。在 GraphSage[34] 的工作中，中心节点 u 的训练是通过均匀随机采样一个邻居节点作为正样本对，以及 Q 个非近邻节点作为负样本，使得在拉近正样本对的嵌入距离的同时拉远负样本对的嵌入距离。这样的训练方式存在两个问题：首先，在均匀采样下，负样本的代表性可能不足以使模型有效地进行对比学习；其次，在大规模图中进行随机游走非常耗时。在嵌入表示学习的工作中，本文模型引入了一种新的训练目标函数[35] 进行对比学习。

具体而言，在特征的嵌入空间中，不仅需要模型能够赋予相似样本更相似的特征，而且要求特征分布能够保存尽可能多的信息，前者通常被称为对齐性(Alignment)，后者则被称为均匀一致性(Uniformity)。对于现有的对比学习模型，Wang 等[35] 证明了它们的有效性在于优化了上面这两种特性，据此提出了两种指标，分别如式(3)和式(4)所示：

$$l_{\text{align}}=\mathbb{E}_{(x,x^+)\sim p_{\text{pos}}}\|f(x)-f(x^+)\|^{\alpha} \tag{3}$$

$$l_{\text{uniform}}=\log\mathbb{E}_{(x,y)\overset{i.i.d.}{\sim}p_{\text{data}}}\mathrm{e}^{-t\|f(x)-f(y)\|^2} \tag{4}$$

式(3)表示在正样本对分布 p_{pos} 中的 (x,x^+) 经过模型 f 学习之后的嵌入距离尽可能小，α 为不同范式距离的取值。式(4)表示对于所有的样本数据分布 p_{data}，任意两个样本对 x 和 y 经过模型 f 学习之后在高斯势能核上的势能值尽可能小(即两个样本在高斯核特征空间中的距离尽可能大)。通过优化式(3)和式(4)所示的两种指标，模型能够输出较好的嵌入表达。

在知识图谱的更新中，我们希望近邻节点嵌入更相似，而非近邻节点嵌入之间尽可能不相似，因此在对一个批次的目标节点进行训练时，把每个目标节点和它的邻居节点当成正样本对，在这个正样本对分布中最小化 l_{align}；把所有目标节点和邻居节点的集合当作样本数据分布，在这个数据分布上最小化 l_{uniform}。最终，在知识图谱上进行卷积的损失函数的定义如下：

$$l_{\text{loss}}=l_{\text{align}}+\alpha_{\text{unif}}l_{\text{uniform}} \tag{5}$$

其中，α_{unif} 表示模型在均匀一致性损失上的训练权重。

3.3 基于细粒度节点的聚合操作

对于传统的图结构，利用图卷积操作进行更新后，可以执行两个层面的任务，即节点层面任务和图层面任务。节点层

面上的任务通常直接利用图卷积的深层嵌入来表示节点，对节点做分类回归等。图层面上的任务则是对整个图或者部分子图进行分类或回归等，此时需要把图或者子图做图读出(Graph Readout)操作，即聚合图中的部分节点和边信息。

在3.1节和3.2节中，进行知识图谱的构建以及图卷积训练得到各节点更深层次的嵌入表达之后，模型需要通过细粒度节点的嵌入来表示词的嵌入，因此本文提出在子图结构上做图读出。大部分工作都参考图像上的池化操作来聚合图或者子图信息，对节点求最大值、求和、求平均等[36]，或者进行层次式的池化操作[37-38]，但是这些池化操作都对结点顺序不敏感。

在细粒度节点构成词的过程中，它的顺序对于聚合来说至关重要。在一个词中，汉字的序列将改变词的语义信息，如“中文”一词与“文中”一词表达的语义是不一样的。同理，部件在构成汉字时，人们通常会以从左往右、从上往下的顺序来书写汉字，因此模型需要用顺序操作来聚合细粒度语义信息。在文本领域，LSTM网络[39]在解决序列问题上有着优异表现，根据本文提出的细粒度节点构成词的特点可知，细粒度序列的信息是单向的，因此模型使用单向LSTM对节点序列进行聚合。但是LSTM网络对节点进行聚合时，无法区别各节点信息的重要程度，即无法建模词对不同细粒度节点信息的偏重情况，因此模型在LSTM的各时间步的输出中加入了注意力(Attention)机制[40]，通过在聚合过程中对各节点信息与整个节点序列提取的信息进行注意力权重系数的计算，来更好地达到节点信息聚合的效果。整个训练过程与细粒度语义信息的聚合过程如图2所示。

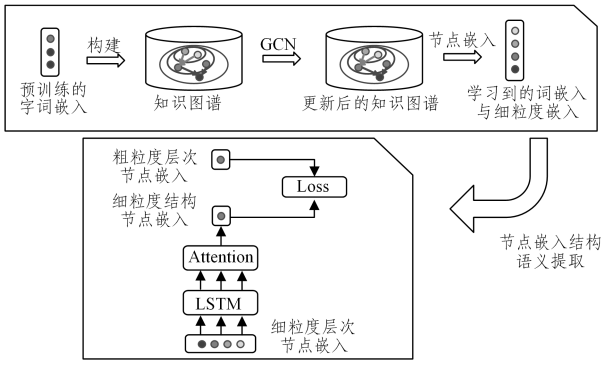


图2 模型训练过程及中文细粒度语义信息的聚合操作

Fig. 2 Model training process and aggregation operation of Chinese fine-grained semantic information

4 实验验证

4.1 实验设置

由于本文模型是基于知识图谱来捕获细粒度语义信息，从而更为精准地推断OOV词的嵌入表示。为了验证本文提出的工作的有效性，我们选择了4种利用细粒度语义信息来推断OOV词嵌入的模型作为本文实验的基准模型，分别是Mimick模型[14]、AM模型[23]、HICE模型[13]，以及Glyph2vec模型[18]。此外，本节统一采用了Skip-gram模型[2]在中文维基百科语料库上预训练好的词嵌入向量作为背景知识。

(1)Mimick模型[14]：使用预训练词嵌入作为标签，并根据英文单词的字母序列推断词的嵌入向量。首先，该模型通过双向LSTM网络对目标英文单词的字母序列进行特征提取，进而通过一层感知机预测得到嵌入，最后使用预测嵌入与预训练词嵌入之间的距离作为模型的损失进行训练。对于中文语料，把目标单词的汉字序列作为输入进行特征提取和训练。

(2)AM(Attentive Mimick)模型[23]：是在Mimick的基础上结合上下文信息来推断词嵌入的方法，其中上下文信息只由少量高质量的上下文决定。该模型融合了注意力模仿机制，采用单词在不同上下文之间的相似性作为注意力，给上下文信息分配不同的权重。与Mimick相似，AM模型使用预训练好的词嵌入作为标签，同时结合上下文信息和字形信息推断词的嵌入。

(3)HICE模型[13]：与AM模型相似，同样是利用字形信息和上下文信息来推断词嵌入，使模型可以学习如何拟合预训练好的词向量。在该方法中，上下文信息不是通过权重加和，而是对采样得到的上下文序列执行自注意力机制，然后对单词所处的不同上下文情境分别进行训练。该文还讨论了对出现次数非常少的小样本数据进行学习、训练和测试推断时，样本的分布可能存在差异，为此提出引入MAML的方法进行有效学习。

(4)Glyph2vec模型[18]：针对中文的词嵌入推断，提出在Mimick结构之前，对汉字书写的图像特征、汉字仓颉造字编码等信息进行提取，之后将其输入双向GRU模块，再根据与预训练好的词向量之间的距离损失进行学习。

(5)Skip-gram模型[2]：用目标词作为输入来预测其上下文词，并以此为目标词学习对应的词嵌入向量。

IFCWE模型：本文提出的模型，其按照中文细粒度结构与词之间的关系构建知识图谱，使用图卷积操作完成信息的传递，并且使用Attention LSTM的结构来聚合细粒度语义信息。在IFCWE模型中，知识图谱上的部分节点(即预训练词表中存在的单词)用预先训练好的词嵌入作为初始特征。

本节将在中文维基百科语料库上预训练好的背景知识作为目标向量，对上述模型进行训练，并且用各自训练出来的模型进一步推断OOV词的嵌入表示。在下游评测过程中，本节实验利用各个模型提取细粒度语义信息的能力，对不存在于背景知识中的词推断其词嵌入表示，然后将结果加入下游评测过程中，进一步对比各个模型对下游任务带来的提升效果。

在本实验中，我们统一将词嵌入向量的维度设置为300。本文模型总共训练3轮迭代，更新方法采用随机梯度下降，设置学习率为0.7，图卷积网络设置为两层，卷积网络的输出特征维度也为300，α_{unif}的值设置为1。其余模型总共训练5轮迭代，参数均采用其论文中提供的默认最优参数。

4.2 词类比任务

词类比任务衡量了预训练模型捕获两个词之间语义关系的能力。Li等[41]提出了针对中文特点的词类比任务评测数据集CA8，该评测数据集共有两种类型的类比任务，分别与字形和语义相关。字形类任务主要针对中文造词重复的特点，如“避一避”“完完全全”，还有前缀后缀的情形，如“第一”“第二”“你们”“他们”等。语义类任务则结合了中国的地理、历史、艺术、常识等。每个类型分别有4个

小类，各自代表一种主要用词情境。

在字形类的词类比任务（Morphological，Mor）上，根据中文造词的特点，对字进行形式上的重复或者规则套用，主要有以下场景：

（1）A 型。类比的词对关系主要有 AA，如“弯”和“弯弯”；A 一 A，如“让”和“让一让”；A 来 A 去，如“忙”和“忙来忙去”等；共 2554 个测试样例。

（2）AB 型。类比的词对关系主要有 AABB，如“完全”和“完完全全”；ABAB，如“湛蓝”和“湛蓝湛蓝”；A 里 AB，如“糊涂”和“糊里糊涂”等；共 2535 个测试样例。

（3）Pre 型。类比的词对关系是在词前加前缀，如“一”和“第一”等；共 2553 个测试样例。

（4）Suf 型。类比的词对关系是在词后加后缀，如“你”和“你们”等；共 2535 个测试样例。

在语义类的词类比任务（Semantic，Sem）上，根据使用中文的实际语境，将词通过语义关系关联起来，主要有以下场景：

（1）地理（Geo）。类比的词对关系由地理关系构成，如省和省会、省和省的简称等；共 3192 个测试样例。

（2）常识（Nat）。类比的词对关系由常识构成，如“第一”和“冠军”等；共 1465 个测试样例。

（3）历史（His）。类比的词对关系是中国历史上的实体关系，如“秦”和“嬴政”等；共 1370 个测试样例。

（4）人文（Peo）。类比的词对关系包括人物事件等关联对，如“马云”和“阿里巴巴”；共 1609 个测试样例。

此外，该任务引入了 Add[1] 和 Mul[42] 这两种度量方式，来评估模型在词类比任务上的性能。对于一个词类比（a:b，a^*:b^*），给定 a，b 和 a^*，需要在词表 V 中找到满足 Add 和 Mul 度量方式的 b^*，其中 Add 度量方式下的目标如式（6）所示，Mul 度量方式下的目标如式（7）所示。

$$\arg\max_{b^*\in V}(\cos(b^*,b)-\cos(b^*,a)+\cos(b^*,a^*)) \tag{6}$$

$$\arg\max_{b^*\in V}\frac{\cos(b^*,b)\cos(b^*,a^*)}{\cos(b^*,a)+\varepsilon} \tag{7}$$

其中，ε 为防止除 0 错误而设置的极小常数。

由于基准模型 AM 和 HICE 需要上下文信息进行词推断，而上述评测数据集中没有上下文信息，因此在词类比任务上本小节只选择 Mimick 模型和 Glyph2vec 模型与本文模型进行对比，实验结果如表 2 和表 3 所列。

表 2 不同模型的字形类词类比结果

Table 2 Morphological analogy results of different models

（单位：%）

Model	Add				Mul			
	A	*Pre*	*AB*	*Suf*	*A*	*Pre*	*AB*	*Suf*
Mimick	23.3	26.2	10.0	66.9	25.2	26.2	9.0	66.2
Glyph2vec	42.4	74.4	65.0	69.3	42.1	75.8	61.3	69.5
IFCWE	**99.9**	**99.4**	**100.0**	**99.9**	**99.9**	**99.4**	**100.0**	**99.9**

字形类词类比任务的结果如表 2 所列。由于 Mimick 模型是字符级词嵌入学习模型，其在中文细粒度语义信息的提取上有所缺失，无法捕捉更深层次的信息，因此在这个任务上 Mimick 模型取得的结果相对较差。而相比 Mimick 模型，Glyph2vec 模型进一步引入了中文的字形信息和仓颉编码等信息，以改善中文词嵌入学习上的效果，可以看到 Glyph2vec 模型的效果相比 Mimick 模型在该任务上有显著提升。而 IFCWE 模型利用部件和拼音与汉字，以及汉字与词的组成关系构建知识图谱，并在知识图谱上进行信息的深层次聚合与更新，采用较为有效的子图读出机制，最后在字形上的判别能力非常准确，几乎全部类比成功。

表 3 不同模型语义类词类比结果

Table 3 Semantic analogy results of different models

（单位：%）

Model	Add				Mul			
	Geo	Nat	His	Peo	Geo	Nat	His	Peo
Mimick	**53.9**	**36.1**	15.6	**25.0**	**54.5**	34.7	15.8	**24.4**
Glyph2vec	53.6	**36.1**	15.4	24.6	54.2	**35.1**	15.7	**24.4**
IFCWE	52.6	**36.1**	**17.3**	23.7	52.7	**35.1**	**16.5**	23.1

语义类词类比任务结果如表 3 所列，其中 Mimick 模型取得了最好的效果。通过综合字形和语义类词类比结果发现，Glyph2vec 模型虽然在先前的任务中远好于 Mimick 模型，但在词任务中准确率平均只有 10%。进一步观察该任务的词对，可以发现几乎所有的词对在字形上都没有太大的关联，Glyph2vec 模型更加侧重于字形上细粒度信息的捕捉，导致在字形信息上产生了过拟合的情况，使得其性能在这个任务中有所下滑。而 Mimick 模型对于字形信息的捕捉能力相对较差，却取得了较好的效果。虽然我们认为 Mimick 模型并没有真正地学习到词对之间的关联信息，但恰恰是因为它无法捕捉到细粒度字形信息，所以剔除了许多噪声。而 IFCWE 模型使用知识图谱可以保存和建模近义词等语义信息，因此与 Glyph2vec 模型相比，其不仅能够发掘细粒度的语义信息，还在语义类词类比任务中取得了具有竞争力的效果。

最后，为了分析 Mimick，Glyph2vec 和 IFCWE 模型在词类比任务中的总体效果，本节展示了 3 种词类比类型（即 Mor 类型、Sem 类型、全部类型）上的总体结果，这 3 种模型在 Add 和 Mul 两种度量方式上的平均指标如图 3 所示。

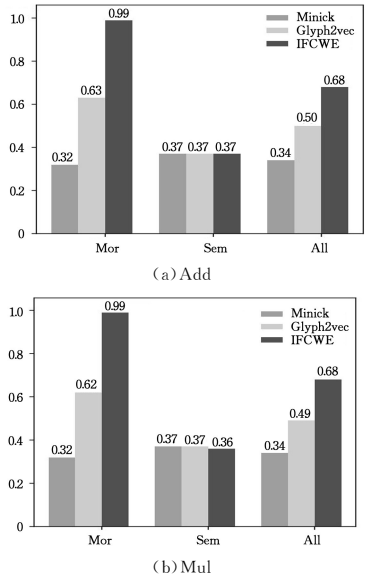


图 3 不同模型分别在类型 Mor 和 Sem 与全部类型 All 下的中文词类比结果

Fig. 3 Chinese word analogy results of models under Mor, Sem, and All

从实验结果中可以发现，在 Add 和 Mul 两种不同的计算最近目标类比词的方法上，本文提出的 IFCWE 模型在平均指标上均取得了最佳效果。

4.3 维基百科标题分类

在实际应用中，未登录词在一些特殊场景中经常出现，如学术名词、领域术语和地理位置等。维基百科标题中常常含有一些具有重要含义的特殊词，这些词往往是 OOV 词。按照 Glyph2vec 模型中的实验设置，在中文维基百科标题文本数据集[1]上进行标题分类，该数据集词表中 60.4% 的词为 OOV 词。在该数据集中，根据维基百科标题所对应的文章内容，可以将数据集中的标题文本分为 12 个类别，并利用上游预训练的词向量，在该实验中对标题文本进行分类。

参照 Glyph2vec 模型中维基百科标题分类任务的实验设置，我们选取 80% 的数据集作为训练集，并将剩下的 20% 作为测试集。另外采用 3 层全连接神经网络作为分类器，其以词嵌入向量为输入，并以测试集上的准确率(Accuracy)和 F1 值作为评测指标。

本小节通过该数据集来衡量各个模型在未登录词嵌入推断方面的能力，并对该数据集进行了如下预处理操作：首先，删除了标题文本中的标点符号，并将阿拉伯数字替换成中文字符，例如将“1”改为“一”；其次，使用 Jieba 工具包对中文标题文本进行分词。

实验结果如图 4 所示。在维基百科标题分类任务中，Mimick 模型作为最早的基准模型，由于只捕捉了粗粒度的字形语义信息，因此效果最差。而其他模型基本都是基于 Mimick 模型的思想在细粒度字形的捕捉或者上下文信息的融入上做了进一步的改进。AM 模型在 Mimick 模型的基础上引入了子词信息和上下文信息，因此与 Mimick 模型相比性能有所提升。同样地，HICE 模型由于在利用粗粒度的字形信息之外，还有效地利用了上下文信息进行未登录词的学习，总体性能相比前两者有所提升。但是在该任务上，维基百科标题中单篇文本的长度非常短，并不能为 AM 模型或者 HICE 模型提供充足有效的上下文信息，而从字形信息来看，Mimick 模型、AM 模型以及 HICE 模型对中文字形的细粒度信息捕获不足。Glyph2vec 模型通过引入更细粒度的字形信息，在精确率上相比以上 3 个模型都有了巨大的提升，说明在该任务中，通过捕获中文层面更细粒度的字形信息确实能有效提升模型效果。而 IFCWE 模型的结果相比 Glyph2vec 模型有了进一步的提升，说明通过引入知识图谱，IFCWE 模型相对于简单学习字形和仓颉编码的 Glyph2vec 模型来说，确实有着更强的学习中文细粒度语义信息的能力。此外，相比使用了上下文的 HICE 和 AM 模型，IFCWE 模型只利用中文字形信息就取得了较好的效果，由此可见中文的字形包含着丰富的语义信息，在该任务中中文的字形信息相对于上下文信息而言更为重要。

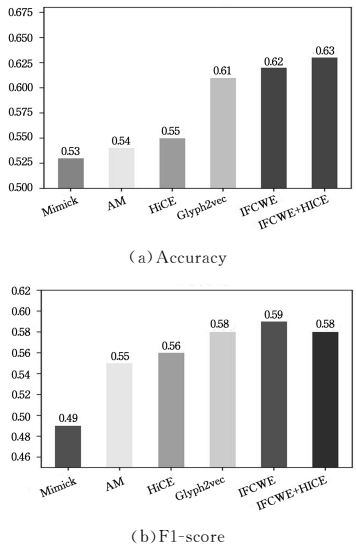


图 4 中文维基百科标题分类结果

Fig. 4 Chinese Wikipedia title classification results

为了更加直观地验证 OOV 词的上下文信息和细粒度语义信息对语义推断的效果，本实验额外利用多层 Transformer 网络来捕获 OOV 词的上下文信息，并结合由知识图谱得到的细粒度语义信息，进而推断 OOV 词嵌入，得到 IFCWE+HICE 模型。将 IFCWE+HICE 模型与 IFCWE 和 HICE 模型分别进行对比，结果表明在该任务中，由于 OOV 词的上下文信息有限，能够为 OOV 词提供的信息严重不足，在此情况下加入 OOV 词的上下文信息，反而为 IFCWE 模型引入了额外的噪音信息，因此在 F1 指标上性能略微下降，但是与 AM 和 HICE 模型相比，IFCWE+HICE 模型可以减小噪声的影响，表现出了更好的鲁棒性。

4.4 命名实体识别

在命名实体识别任务上，本文采用了医学命名实体识别数据集 Yidu。该语料是在 2019 年 CCKS 评估任务上提供的中文电子医疗记录数据集[2]，它的词表中 53.1% 的词频小于或等于 2，具有较高的 OOV 词占比，本文利用该数据集来衡量预训练中文词嵌入对于 OOV 词嵌入学习的作用。另外，本文还使用 opencc 工具包[3]将语料中的繁体字统一转为简体字以便进一步学习和处理，并且使用 Jieba 工具包对数据集进行分词操作。

在该实验中，使用基准模型对 OOV 词的嵌入向量进行推断，其余词则使用已预训练好的词嵌入向量。下游采用 BiLSTM-CRF 模型[43]来进行命名实体识别任务，它将词嵌入向量作为输入，输出则是句子中单词的预测标签。另外，在训练过程中不对预训练词嵌入进行微调和更新，并以命名实体识别的精确率(Precision)、召回率(Recall)和 F1-score 为衡量指标。

OOV 词命名实体识别的实验结果如表 4 所列，其中最优和次优的结果分别用粗体和下划线突出表示。可以看到，仅

1) https://github.com/frederick0329/Wikipedia-Title-Dataset

2) http://openkg.cn/dataset/yidu-s4k

3) https://github.com/BYVoid/OpenCC

基于细粒度语义信息的 Mimick 模型在所有基准模型中反而取得了最好的效果，说明在该任务中，OOV 词的有限上下文信息中包含了较多的噪声，因此侧重于上下文信息的模型（如 AM 和 HICE）无法学习到有效的 OOV 词嵌入。此外，利用中文细粒度语义信息的 Glyph2vec 模型在此任务中表现不佳，而本文提出的 IFCWE 模型的效果相比 Glyph2vec 模型取得了显著提升，这表明通过知识图谱和图卷积操作建模中文构字和构词规则能够更充分地利用中文的细粒度语义信息。IFCWE 模型在精确率上超越了 Mimick 模型，但是在另外两个指标上略低于后者，原因可能是 IFCWE 模型相比 Mimick 模型具有更多的参数，所以在规模较小的语料中更难收敛，而结合了上下文信息的 IFCWE＋HICE 模型的效果显著优于 Mimick 模型，说明本文提出的 IFCWE 模型不仅能更深入地挖掘中文细粒度语义信息，而且在某些上下文信息充足的下游语料中，加入 OOV 词的上下文信息可以更有效地对 OOV 词进行推断，即模型具有良好的可拓展性。

表 4　Yidu 命名实体识别任务结果

Table 4　Named entity recognition task results on Yidu

（单位：%）

Model	Precision	Recall	F1-score
AM	43.17	26.36	32.73
HICE	43.35	25.64	32.22
IFCWE＋HICE	**46.00**	**27.76**	**34.63**
Mimick	44.12	27.65	34.00
Glyph2Vec	38.35	18.37	24.84
IFCWE	44.98	25.00	32.14

4.5　消融实验

为了进一步分析 IFCWE 中每个模块的效果及作用，本节分别对图卷积模块和子图读出模块进行了拆除，以便完成消融实验分析。

IFCWE w/o RGCN：拆除图卷积模块得到基准模型，该模型在训练过程中不对知识图谱进行图卷积训练，而是直接在随机的卷积核上做子图读出。

IFCWE w/o Readout：拆除子图读出模块得到基准模型，该模型在对知识图谱进行卷积后，将子图上的节点特征加权求和后直接得到子图特征，即用平均的方式替代 Attention 机制下的 LSTM 读出操作。

在字形类（Mor）词类比任务以及在语义类（Sem）词类比任务上的实验结果分别如表 5 和表 6 所列。此外，为了更加清晰地展示在词类比任务上的总体效果，图 5 给出了在 Add 度量方式下的 3 种词类比类型（即 Mor 类型、Sem 类型、全部类型）的结果和在 Mul 度量方式下的 3 种词类比类型结果。

表 5　消融实验中字形类词类比任务结果

Table 5　Morphological analogy results of ablation models

（单位：%）

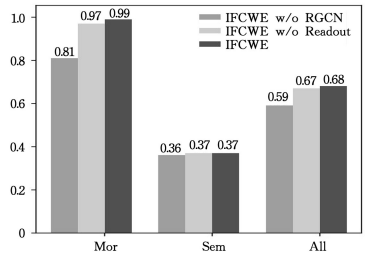
Model	Add				Mul			
	A	Pre	AB	Suf	A	Pre	AB	Suf
IFCWE	**99.9**	**99.4**	**100.0**	**99.9**	**99.9**	**99.4**	**100.0**	**99.9**
IFCWE w/o RGCN	81.8	81.0	84.4	75.4	81.2	80.7	83.9	76.2
IFCWE w/o Readout	97.3	95.8	98.7	95.1	97.1	96.0	98.6	95.6

表 6　消融实验中语义类词类比任务结果

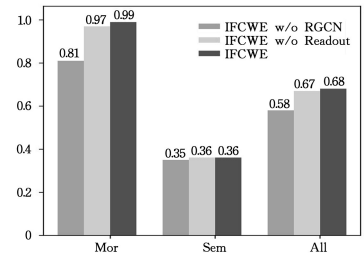
Table 6　Semantic analogy results of ablation models

（单位：%）

Model	Add				Mul			
	Geo	Nat	His	Peo	Geo	Nat	His	Peo
IFCWE	**52.6**	**36.1**	**17.3**	**23.7**	**52.7**	**35.1**	**16.5**	**23.1**
IFCWE w/o RGCN	52.5	33.6	16.7	23.2	52.6	32.2	16.4	22.1
IFCWE w/o Readout	52.5	36.0	16.6	23.5	**52.7**	35.0	16.3	23.0



(a) Add



(b) Mul

图 5　消融模型在两种类型（Mor 和 Sem）与全部类型（All）下的中文词类比任务结果

Fig. 5　Chinese word analogy results of ablation models under Mor, Sem and All

从图表中可以看出，在字形类的词类比任务上，拆除卷积模块后实验效果明显下降，这说明本文构建的知识图谱对词的细粒度语义信息的建模十分有效。而在语义类的词类比任务上，拆除卷积模块后的模型实验效果的下降程度不及在字形类任务上明显，这从侧面表明了本文构建的知识图谱含有的细粒度语义信息比词的近义语义信息多。另外，拆除子图读出模块后的模型效果也有所下降，这表明知识图谱通过图卷积学习到的各细粒度节点特征也需要使用有效的子图读出模块来聚合节点特征信息。

4.6　定性分析

为了定性地分析本文模型在细粒度语义信息建模过程中的有效性，我们选择了 3 个常见并且语义差别较明显的部件：“木”“火”和“氵”。在使用 t-SNE 降维至二维空间后，展示它们的位置。同样地，在二维图中展示了它们构成的部分汉字与词的位置。为了更好地展示这 3 个部件和它们构成汉字、词之间的语义关系，本小节采用两两分组对照的方式。

如图 6 所示，对于部件“火”和“氵”，我们将由“氵”构成的汉字、词用深色标识，由“火”构成的汉字、词用浅色标识。可以看到，大部分与“氵”相关的点分布在右上部，而与“火”相关的点分布在左下部，整体呈现出了“水火不相容”的形态。另外可以注意到，对于拥有两个“火”部件的词，它们之间的分布

比只有单个“火”部件的词更集中，这说明本文模型学习到的细粒度语义信息可以有效地进行词嵌入的推断。

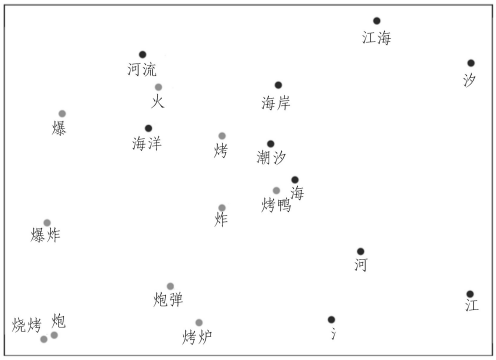


图 6 “氵”与“火”相关的字词嵌入分布

Fig. 6 Embedding distribution of words and characters related to “氵” and "火”

同样地，对于部件“火”和“木”，将由“火”构成的汉字、词在图 7 中用浅色标识，由“木”构成的汉字、词用深色标识。对于“火”和“木”来说，它们之间也存在着较大的语义差别，大部分与“木”相关的点在空间中的分布为局部集中，其中存在一个特殊词“森林”，它的部件全部由“木”构成，因此它的语义位置非常接近部件“木”。与图 6 相似的是，由两个“火”作为构成部件的词分布也比较接近。对于部件“木”和“氵”，将由“木”构成的汉字、词在图 8 中用浅色标识，由“氵”构成的汉字、词用深色标识。可以看到与“木”字结构相关的汉字和词大部分分布在左下部，“氵”则分布在右上部。

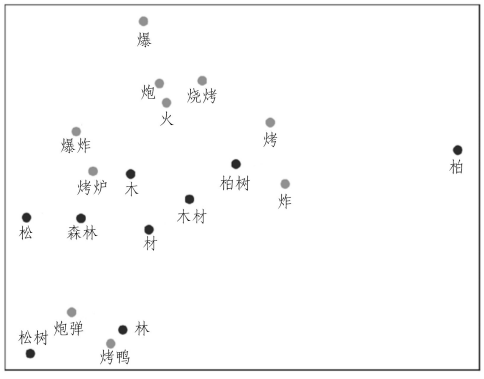


图 7 “木”与“火”相关的字词嵌入分布

Fig. 7 Embedding distribution of words and characters related to “木” and “火”

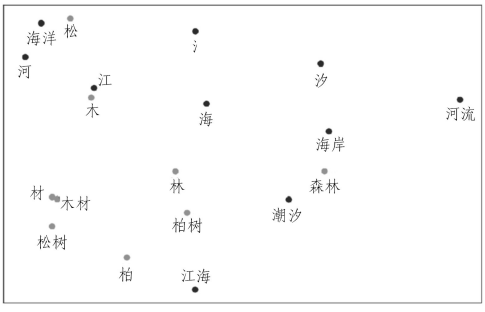


图 8 “氵”和“木”相关的字词嵌入分布

Fig. 8 Embedding distribution of words and characters related to “氵” and “木”

从这 3 个定性分析的图例中，可以发现中文的细粒度语义信息和单词语义信息之间有着较为明显的关联，并且本文提出的 IFCWE 模型可以有效地建模细粒度语义信息与单词语义信息之间的关联关系。

除了上文中描述的部件和汉字、单词之间的关系以外，为了观察通过细粒度语义信息来推断词嵌入的有效性，我们在 Yidu 数据集上选择 OOV 词“老年痴呆症”作为目标词，它的 5 个最近邻词如表 7 所列。此目标词对于预训练词表来说是未登录词，并且在选择近邻词时，只考虑对于预训练词表来说是未登录词的近邻词。从结果中可以发现，Glyph2vec 模型对“症”这个汉字的关注度过高，导致学习到了一些不相关语义；Mimick 模型在这个实例上找到的词更加不相关；而 IFCWE 模型在前五个近邻中能找到紧密相关的“老年性痴呆”一词。由此可见，本文提出的 IFCWE 模型在进行近邻词语义匹配时，也有着较好的表现。

表 7 与目标词“老年痴呆症”最近邻的 5 个近义词

Table 7 Nearest five words to target word “alzheimer”

	老年痴呆症				
Mimick	脓血	综合征	肾囊肿	前列腺炎	尿痛
Glyph2vec	腺肌症	腺肌病	癌症	胰腺癌	分化腺癌
IFCWE	同奥方克泵控	多帕菲伯尔定	错配修复	科行泽菲鲁贝	老年性痴呆

结束语 随着互联网技术和信息化进程的不断发展，越来越多的文本数据不断产生。对于自然语言处理领域来说，这些不断产生的文本数据给领域发展带来了重要作用的同时，也带来了许多挑战。按照传统人工处理方法，大规模的文本数据往往费时费力[44]，并且人为因素带来的干扰也不可忽视[45]。通常情况下，自然语言处理任务的第一步是将文本表示为计算机可处理的数值向量，其中词嵌入学习方法已经成为主流的文本表示方式[46]，它能够将词映射为低维语义嵌入向量。许多任务直接使用公开的在大规模通用语料中预训练的词向量作为模型的输入，但是这种预训练的词向量并不适用于所有任务，特别是面对下游出现的 OOV 词无法进行表征的问题时。已有的 OOV 词嵌入学习方法大多依赖于词的上下文信息，或者大部分都只针对英文单词的细粒度语义信息进行提取，它们无法在小样本中文语料上进行有效的词嵌入学习。原因在于，首先，中文作为世界上使用人数最多的语言，其文本处理任务的重要程度不言而喻；其次，中文是一种象形文字，它的细粒度构造中含有丰富的语义信息，这些细粒度的语义信息能够有效增强在下游任务中的表现[47]。如何在中文语料上，利用中文细粒度结构中语义信息丰富这一特点，学习高质量的 OOV 词嵌入向量是本文重点关注的问题。本文通过知识图谱的构建，进一步挖掘细粒度语义信息与词语义信息之间的关系，并且建模了部件、拼音、汉字、词这 4 种结构在组成过程中的联系；使用图卷积网络对知识图谱中的信息进行深层次的特征提取，然后使用 Attention 机制与 LSTM 网络做子图读出，进而根据细粒度语义信息推断词的嵌入向量。在多个真实数据集上，本文模型均取得了较好的

性能，证明了所提模型的有效性。

本文探究了从知识图谱的角度对中文细粒度语义信息进行利用。该方法是在已有的词典数据上，对词的语义信息及其对应的细粒度语义信息进行挖掘，即在判别学习模式下识别词与细粒度语义信息之间的关系。随着深度学习的发展，在人工智能领域出现了越来越多的有效的学习模式，特别是生成式模型在多种任务下取得了优异的性能。除了本文这类词典式的固定词和细粒度语义信息之间的关系发现，一些生成式的方法也值得探究。因此，在未来的工作中，针对中文细粒度语义信息的挖掘还可以利用生成式的训练方式，根据与上下文词的共现次数引入上下文信息的知识图谱模块[48]，进一步提高模型推断 OOV 词嵌入的能力。

参 考 文 献

[1] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[C]// ICLR Workshop. 2013.

[2] MIKOLOV T，SUTSKEVER I，CHEN K，et al. Distributed representations of words and phrases and their compositionality [C]// NIPS. 2013：3111-3119.

[3] PENNINGTON J，SOCHER R，MANNING C D. Glove：Global vectors for word representation [C] // EMNLP. 2014：1532-1543.

[4] DEVLIN J，CHANG M，LEE K，et al. BERT：Pre-training of deep bidirectional transformers for language understanding [C]// NAACL-HLT. 2019：4171-4186.

[5] DAI Z，YANG Z，YANG Y，et al. Transformer-XL：Attentive language models beyond a fixed-length context [C] // ACL. 2019：2978-2988.

[6] YANG Z，DAI Z，YANG Y，et al. XLNet：Generalized autoregressive pretraining for language understanding[C]// NeurIPS. 2019：5754-5764.

[7] LIU Y，OTT M，GOYAL N，et al. RoBERTa：A robustly optimized BERT pretraining approach[J]. arXiv：1907. 11692，2019.

[8] LAN Z，CHEN M，GOODMAN S，et al. ALBERT：A lite BERT for self-supervised learning of language representations [C] // ICLR. 2020.

[9] ZHANG Z，HAN X，LIU Z，et al. ERNIE：Enhanced language representation with informative entities[C]// ACL. 2019：1441-1451.

[10] SUN Y，WANG S，LI Y K，et al. ERNIE 2. 0：A continual pretraining framework for language understanding [C] // AAAI. 2020：8969-8975.

[11] AJI A F，BOGOYCHEV N，HEAFIELD K，et al. In neural machine translation，what does transfer learning transfer? [C]// ACL. 2020：7701-7710.

[12] JIN D，JIN Z，ZHOU J T，et al. Is BERT really robust? A strong baseline for natural language attack on text classificationand entailment[C]// AAAI. 2020：8018-8025.

[13] HU Z，CHEN T，CHANG K，et al. Few-shot representation learning for out-of-vocabulary words [C] // ACL. 2019：4102-4112.

[14] PINTER Y，GUTHRIE R，EISENSTEIN J. Mimicking word embeddings using subword RNNs [C] // EMNLP. 2017：102-112.

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

[16] YU J，JIAN X ，XIN H ，et al . Joint embeddings of Chinese words，characters，and fine-grained subcharacter components [C]// EMNLP. 2017：286-291.

[17] SU T R，LEE H Y. Learning Chinese word representations from glyphs of characters[C]// EMNLP. 2017：264-273.

[18] CHEN H，YU S，LIN S. Glyph2vec：Learning Chinese out-of-vocabulary word embedding from glyphs[C]// ACL. 2020：2865-2871.

[19] CAO S，LU W，ZHOU J，et al. cw2vec：Learning Chinese word embeddings with stroke n-gram information[C]// AAAI. 2018：5053-5061.

[20] SINGH M，GREENBERG C，OUALIL Y，et al. Sub-word similarity based search for embeddings：Inducing rare-word embeddings for word similarity tasks and language modelling [C]// COLING. 2016：2061-2070.

[21] 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.

[22] SCHICK T，SCHÜTZE H. Learning semantic representations for novel words：Leveraging both form and context[C]// AAAI. 2019：6965-6973.

[23] SCHICK T，SCHÜTZE H. Attentive mimicking：Better word embeddings by attending to informative contexts[C]// NAACL-HLT. 2019：489-494.

[24] SCHICK T，SCHÜTZE H. Rare words：A major problem for contextualized embeddings and how to fix it by attentive mimicking[C]// AAAI. 2020：8766-8774.

[25] FUKUDA N，YOSHINAGA N，KITSUREGAWA M. Robust backed-off estimation of out-of-vocabulary embeddings [C] // EMNLP Findings. 2020：4827-4838.

[26] CHEN L，VAROQUAUX G，SUCHANEK F M. Imputing out-of-vocabulary embeddings with LOVE makes language models robust with little cost[C]// ACL. 2022：3488-3504.

[27] BUCK G，VLACHOS A. Trajectory-based meta-learning for out-of-vocabulary word[J]. arXiv：2102. 12266，2021.

[28] CHO K，VAN MERRIENBOER B，GÜLÇEHRE Ç，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// EMNLP. 2014：1724-1734.

[29] SUN Z，LI X，SUN X，et al. ChineseBERT：Chinese pretraining enhanced by glyph and pinyin information[C]// ACL/IJCNLP. 2021：2065-2075.

[30] LI L，DAI Y，TANG D，et al. MarkBERT：Marking word boundaries improves Chinese BERT[J]. arXiv：2203. 06378，2022.

[31] ZHANG Y，LIU Y，ZHU J，et al. Learning Chinese word embeddings from stroke，structure and pinyin of characters[C]// CIKM. 2019：1011-1020.

[32] MILLER G A. Wordnet：A lexical database for English [J]. Commun. ACM，1995，38(11)：39-41.

[33] SCHLICHTKRULL M S, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]// ESWC. 2018: 593-607.
[34] HAMILTON W L, YING Z, LESKOVEC J. Inductive representation learning on large graphs[C]// NIPS. 2017: 1024-1034.
[35] WANG T, ISOLA P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere [C]// ICML. 2020: 9929-9939.
[36] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for quantum chemistry[C]// ICML. 2017: 1263-1272.
[37] MA Y, WANG S, AGGARWAL C C, et al. Graph convolutional networks with eigenpooling[C]// KDD. 2019: 723-731.
[38] DIEHL F. Edge contraction pooling for graph neural networks [J]. arXiv: 1905. 10990, 2019.
[39] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computer, 1997, 9(8): 1735-1780.
[40] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// NIPS. 2017: 5998-6008.
[41] LI S, ZHAO Z, HU R, et al. Analogical reasoning on Chinese morphological and semantic relations[C]// ACL. 2018: 138-143.
[42] LEVY O, GOLDBERG Y. Linguistic regularities in sparse and explicit word representations[C]// CoNLL. 2014: 171-180.
[43] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv: 1508. 01991, 2015.
[44] BAO C, QIAO J, LI H, et al. Research on long text classification method based on fusion features[J]. Journal of Chongqing University of Technology(Natural Science), 2022, 36(9): 128-136.
[45] DONG L, YANG D, ZHANG X. Large-scale semantic text overlapping region retrieval based on deep learning[J]. Journal of Jilin University (Engineering and Technology Edition), 2021, 51(5): 1817-1822.
[46] YANG Q. Hybrid semantic similarity calculation of knowledge ontology and word vector based on reinforcement learning[J]. Journal of Chongqing University of Technology (Natural Science), 2022, 36(1): 128-135.
[47] HOU Y, ABULIZI A, ABUDUKELIMU H. Advances in Chinese pre-training models[J]. Computer Science, 2022, 49(7): 148-163.
[48] CAO X, NIU Q, WANG R, et al. Distributed representation learning and improvement of Chinese words based on co-occurrence[J]. Computer Science, 2021, 48(6): 222-226.

CHEN Shurui, born in 1998, postgraduate. Her main research interests include word embedding learning and graph convolution neural network.

RAO Yanghui, born in 1986, associate professor, is a member of China Computer Federation. His main research interests include text data mining, representation learning and emotion detection.

(责任编辑:何杨)