文章编号：1003-0077(2019)03-0033-09

基于 HowNet 的语义表示学习

朱靖雯[1]，杨玉基[2]，许 斌[2]，李涓子[2]

（1. 北京信息科技大学 信息管理学院，北京 100192；
2. 清华大学 计算机系知识工程实验室，北京 100084）

摘 要：HowNet 是一个大规模高质量的跨语言(中英)常识知识库，蕴含着丰富的语义信息。该文利用知识图谱领域的方法将 HowNet 复杂的结构层层拆解，得到了知识图谱形式的 HownetGraph，进而利用网络表示学习以及知识表示学习方法得到了跨语言(中、英)、跨语义单位(字词、义项①、DEF_CONCEPT②和义原)的向量表示，在词语相似度(word similarity)和词语类比(word analogy)任务上对中英文数据集进行了实验，实验结果显示该文提出的方法在词语语义相似度的任务上取得了最好效果。

关键词：HowNet；知识图谱；语义表示；表示学习
中图分类号：TP391 **文献标识码**：A

Semantic Representation Learning Based on HowNet

ZHU Jingwen[1], YANG Yuji[2], XU Bin[2], LI Juanzi[2]

(1. School of Information Management, Beijing Information Science and Technology University, Beijing 100192, China;
2. Knowledge Engineering Group, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: HowNet is a large-scale and high-quality cross-lingual commonsense knowledge base, containing a wealth of semantic information. This paper disassembles HowNet's complex structure and obtains HownetGraph in the form of knowledge graph. Then Network Representation Learning and Knowledge Representation Learning methods are applied to obtain cross-lingual vector representation of different semantic units, i.e., word, sense, DEF_CONCEPT and sememe. Two series of experiments (word similarity and word analogy) are conducted on Chinese and English datasets, and the results show the proposed method achieves the best results.

Keywords: HowNet; knowledge graph; semantic representation; representation learning

0 引言

近年来，词向量技术的进步极大地促进了自然语言处理领域的发展。目前大部分研究者都是基于大规模无监督语料学习词语或义项级别的语义表示，这种基于上下文的学习方法倾向于把共现较多的词语或义项聚在一起，可以学习到好的上下文相似性，却难以捕获到好的语义相似性，尤其是在语料中出现频次较低的词语和义项的语义相似性。

为了学习到词语和义项的语义相似性，我们使用了语义信息最为丰富的中英文常识知识库 HowNet[1]。在 HowNet 中，词语由一个或多个义项组成，而每个义项又由更小的语义单位(义原)和几十种动态角色组合而成。很多人基于 HowNet 开展了语义表示的研究[2-3]，但是他们或是忽略了义原和动态角色之间的复杂结构[2]，或是只能学习到词语

① HowNet 中称“义项”为“概念”，为了避免和知识图谱中的“概念”混淆，本文均用“义项”表示。

② HownetGraph 中的 DEF 单元，是介于义项和义原之间的语义单位，详见 2.2.2 节。

收稿日期：2018-09-29 **定稿日期**：2018-10-29
基金项目：国家高技术研究发展计划(863)(2015AA015401)；国家科技部重点研发计划(2018YFB100283)

级别的语义表示[3]，并未充分学习到 HowNet 中蕴含的全部语义信息。

本文中，我们将 HowNet 中定义的各种关系和定义进行拆解，构建了包含 HowNet 全部信息的知识图谱 HownetGraph。接下来，我们利用网络表示学习以及知识表示学习的方法，从 HownetGraph 中学习得到跨语言（中、英）、跨语义单位（字词、义项、DEF_CONCEPT 和义原）的语义表示，并进行了词语相似度和词语类比的实验。实验结果表明，从 HownetGraph 中学习得到的语义表示较好地捕获到了 HowNet 的语义信息。

本文的主要贡献有：

（1）将 HowNet 层层拆解，构建了一个跨语言（中、英）、跨语义单位（字词、义项、DEF_CONCEPT 和义原）的常识知识图谱 HownetGraph。

（2）据我们所知，我们第一个同时学习到了 HowNet 的词语、义项、DEF_CONCEPT 和义原等不同语义单位的跨语言（中英）的向量表示。

（3）我们根据 wordsim-353 和 wordsim-297 数据集，构建出了中文语义相似度数据集 wordsim-297-similarity，作为 wordsim-297 数据集的子数据集，细粒度评测中文词语的语义相似度。

（4）我们在词语相似度和词语类比任务上进行了实验，中英文词语相似度和词语类比任务上的实验验证了所提方法的有效性。

本文结构如下：第 1 节为相关工作部分介绍，第 2 节为 HownetGraph 的构建过程，第 3～4 节为通过实验和例子给出学习到的语义表示效果，最后是后续工作展望。

1 相关工作

1.1 HowNet

HowNet 语义信息的丰富性引起了很多研究者的关注。刘群等提出基于 HowNet 的词汇语义相似度计算[3]，他们将 HowNet 中的义原建成树状，通过构成词义项的义原的距离得到词义相似度，实验得到的词语相似度结果与人的直觉比较符合，但由于他们对每个词都只取了最常见的义项，而不是所有义项，因此对 HowNet 的义项描述信息并没有很好利用；梅立军等通过为同义词词林的每个词集确定一个义项描述，实现 HowNet 与同义词词林的信息融合[4]；Sun J 等提出基于 HowNet 的中文问题自动分类，将问题中的词对应义项的基本义原作为问题特征进行处理[5]；Yan J 等基于 HowNet 中事件类义原层次结构，创建了一个中文情感的领域本体[6]；唐怡等提出了基于 HowNet 的中文语义依存分析，将句子转化为树状，并根据 HowNet 中的动态角色进行语义关系标注，实验结果标注比例高达 91.5%[7]；Liu J 等提出用 HowNet 实现 Word 相似度计算的混合层次结构方法[8]；向春丞等提出了 HowNet 与中文概念辞书（CCD）的映射方法，将 HowNet 中的词与 CCD 词典中的词进行映射[9]；Niu Y 等利用 HowNet 的义原提升词语的表示[2]；Zeng X 等则利用义原扩充中文 LIWC（Linguistic inquiry and word count）词典[10]。上述工作涉及自然语言处理领域下的诸多子领域，说明了 HowNet 的重要价值。在语义表示领域，大部分研究学者都只是提取了 HowNet 中的部分信息，对于 HowNet 层次化的义项定义并未能很好利用，而这却正是 HowNet 的语义核心。

1.2 表示学习

表示学习指的是将研究对象表示为稠密低维向量。本文的研究对象是 HowNet 的语义信息，故叫做语义表示学习。为了学习到语义表示，本文使用了网络表示学习和知识表示学习两类方法。

网络表示学习（network representation learning），指的是为网络中的每个节点学习到稠密低维的向量表示，从而使得在大规模网络上进行快速高效的算法成为可能。目前，效果比较好的网络表示学习模型为基于神经网络和深度学习的方法。DeepWalk 模型[11]第一次将深度学习的方法引入网络表示学习领域，借鉴 Word2Vec 方法在网络上随机游走生成序列，从而可以直接利用 Word2Vec 方法学习到节点表示；LINE 模型[12]通过对节点间的第一级相似度和第二级相似度进行概率建模，最小化该概率分布和经验分布之间的 KL 距离，得到好的节点表示；Node2Vec 模型[13]是对 DeepWalk 模型的扩展，通过改变随机游走序列生成方式来优化节点表示效果。此外，还有 Grarep 模型[14]、GCN 模型[15]、TADW 模型[16]和 Cane 模型[17]等。

知识表示学习（knowledge representation learning）指的是为知识图谱中的节点和边学习到稠密低维的向量表示。学习了向量表示后，便可以基于向量表示去做知识图谱领域内关系预测等任务。近年来随着深度学习技术的不断发展，研究者们提出了一系

列基于深度学习的知识表示学习方法。Bordes A 等提出了一种简单且易拓展的模型，把知识库中的实体和关系映射到低维向量空间中，从而计算出隐含的关系的 TransE 模型[18]。随后，Wang Z 等提出了对 TransE 进行优化后的 TransH 模型[19]，解决了 TransE 对于一对多、多对一以及多对多关系处理效果不太好的问题。TransE 和 TransH 模型默认实体和关系处于相同的语义空间，而事实上，一个实体是由多种属性组合成的综合体，不同关系关注实体的不同属性，因此，Lin Y 等提出了对头实体和尾实体投影到关系空间中再做操作的 TransR 模型[20]，在此基础上，后续又出现了 TransD 模型[21]、TranSparse 模型[22]、TransG 模型[23]、KG2E 模型[24]等。

2　从 HowNet 到 HownetGraph

2.1　HowNet 介绍

HowNet 是一个以汉语和英语的词语所代表的义项为描述对象，以揭示义项与义项之间以及义项所具有的属性之间的关系为基本内容的常识知识库。在 HowNet 中，义原是最基本的、不易于再分割的意义的最小单位，词由义项组成，义项由义原定义。在 HowNet 数据文件①中，共包含 118 347 个中文词，104 027 个英文词，212 541 个义项，2 468 个义原（包括实体、事件、属性、属性值、第二特征、专有名词和符号 7 大类）和 116 个动态角色。下面我们举例说明 HowNet 的语义组织模式。

如图 1 所示，“绿色”一词有两个义项，义项 1 指的是绿颜色，义项 2 指的是绿色环保。义项 1 的定义较为简单，只有 1 个基本义原“green|绿”，而义项 2 的定义较为复杂，有较多动态角色和义原的嵌套关系：最外层的基本义原是“PropertyValue|特性值”，动态角色“scope”表示范围，来进一步说明“PropertyValue|特性值”，“scope”的宾语是“{protect|保护：patient={Environment|情况：host={entity|实体}}}”；次外层的基本义原是“protect|保护”，动态角色“patient”表示受事，修饰义原“protect|保护”，宾语是“{Environment|情况：host={entity|实体}}”；最内层的基本义原是“Environment|情况”，动态角色“host”表示宿主，修饰义原“Environment|情况”，宾语是义原“entity|实体”。总结地说，义项 2 的定义是：绿色是一种特征值，这种特征值对应的范围是一种保护，这种保护的受事是一种情况，这种情况的宿主是实体。除了最重要的定义（DEF），每个义项还有一些其他的关系和属性，如义项 2 在数据文件里表示如图 2 所示。图 2 中，“NO.”表示这个义项的唯一标识符（id）是 103130；“W_C”表示其中文词为“绿色”；“G_C”表示其中文词词性是形容词；“S_C”表示其中文情感标识为“PlusSentiment|正面评价”；“E_C”表示词语在短语中的例子；“W_E”表示其英文词为“green”；“G_E”表示其英文词性为形容词；“S_E”表示其英文情感标识为“PlusSentiment|正面评价”；“DEF”表示用动态角色和义原的组合来表示义项的定义。

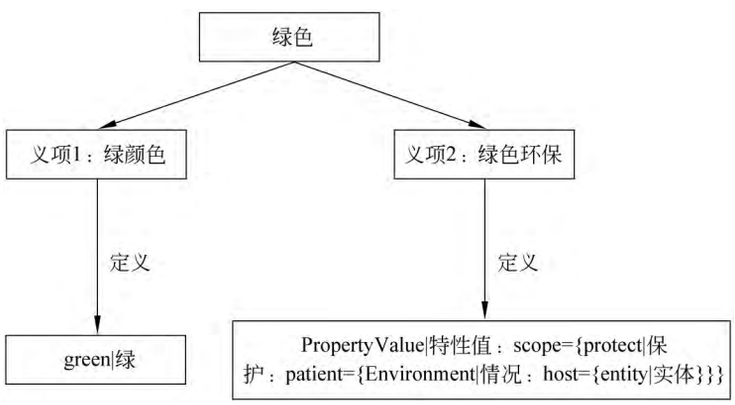

图 1　HowNet 中词语、义项和义原的例子

2.2　HownetGraph 构建

知识图谱一般使用 RDF（resource description framework）或者 OWL（web ontology language）等语言来描述，后者比前者有更强的语义表达能力。由于 HowNet 语义结构不涉及复杂推理和规则，故我们采用 RDF 来构建知识图谱 HownetGraph。

① 2012 年版本。

2.2.1 本体构建

本体构建是知识图谱的第一个步骤，可以理解为知识图谱的框架。最基本的本体包括概念、概念层次、属性、属性值类型、关系、关系定义域（domain）概念集以及关系值域（range）概念集。在HownetGraph中，我们定义概念包括中文词语、英文词语、义项、义原和DEF_CONCEPT（义原和动态角色的组合，介于义项和义原之间）；关系主要为动态角色、上下位关系和instanceof关系；HowNet中出现的每一个语义单位都是实例；属性只有rdf：label属性（其他属性被忽略）。

2.2.2 建图

根据上述本体，我们以义项为桥梁，通过层层拆解DEF，得到了包含所有语义单位的HownetGraph。如图1和图2中“绿色”的例子，则可以建成如图3所示的图结构，具体操作步骤如下。

```
NO.=103130
W_C=绿色
G_C=adj [1v4 se4]
S_C=PlusSentiment|正面评价
E_C=~食品，~奥运，~组织，~活动，~装修材料，~包装，~生命，~生活，~城市，~人生，~环境，~旅游
W_E=green
G_E=adj [3 green adj -0 prepo 1 ]
S_E=PlusSentiment|正面评价
E_E=
DEF={PropertyValue|特性值:scope={protect|保护:patient={Environment|情况:host={entity|实体}}}}
RMK=
```

图2 HowNet中义项“绿色”的表示

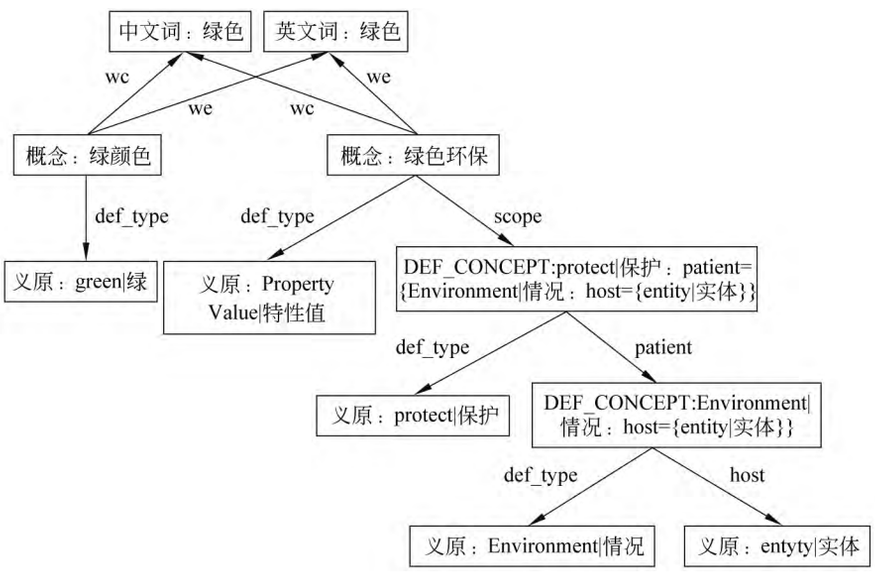

图3 HownetGraph中边表示关系

注：wc表示义项的中文词，we表示义项的英文词，def_type表示义项或DEF_CONCEPT的基本义原，scope、patient和host分别表示义项或DEF_CONCEPT的范围、受事和宿主关系。

（1）HownetGraph中只包含HowNet数据中描述语义单位的W_C、W_E、DEF部分和义原之间的所有关系，不包含G_C、E_C等内容。

（2）DEF部分是嵌套结构，为了将其转化为图结构，我们借鉴了RDF中“陈述”的结构，我们将每一层嵌套的所有内容当做一个DEF_CONCEPT实例，每一个DEF_CONCEPT实例内部再根据HowNet数据的结构拆解开，动态角色转化为关系（如patient，host和scope），义原则转化为相应的义原实例，DEF_CONCEPT实例的第一个义原前面无动态角色约束，鉴于该义原表示了DEF_CONCEPT实例的最主要语义信息，我们将这种关系定义为def_type。

（3）wc关系对应HowNet数据中的W_C部分，表示义项对应的中文词。we关系对应HowNet数据中的W_E部分，表示义项对应的英文词。

3　实验

在这一部分，我们将 HownetGraph 用网络表示学习和知识表示学习进行训练，通过词语相似度任务(word similarity)和词语类比任务(word analogy)来检验我们的语义表示效果。

3.1　数据集和实验设置

词语相似度任务，考虑到我们同时学习到了中英文的词语表示，故我们选择了数据集 wordsim-240[①](W240，中文)、wordsim-297[①](W297，中文)和 wordsim-353[②](W353，英文)。此外，为了说明我们学习到的语义表示侧重于语义相似度而非上下文相似度，我们使用了数据集 wordsim-353-similarity(W353S，英文)，并根据 W297 和 W353S 创建了中文的测试语义相似度的数据集 wordsim-297-similarity(W297S)。词语类比任务中，我们选择了 Chen X 等人的中文词语类比数据[25](A1125)检验我们学习到的词语向量的质量。值得说明的是，上述数据集中的部分词语 HowNet 中没有覆盖，具体情况如表 1 所示。

表 1　数据集统计信息

数据集	语言	词对总数量	HowNet 覆盖的词对数量
W240	中文	240	235
W297	中文	297	274
W297S	中文	169	165
W353	英文	353	335
W353S	英文	203	196
A1125	中文	1 125	1 125

考虑到图谱中有部分无效信息和重复信息，我们在进行知识表示学习和网络表示学习时，只选用了 HowNet 中的 W_C 关系、W_E 关系、DEF 部分的关系和义原之间的所有关系，若关系值为 *，则去掉该记录。我们使用了清华大学自然语言处理实验室提供的知识表示学习代码[③]和网络表示学习代码[④]。

为了证明 HownetGraph 结构表示语义的有效性，我们选取了较为简单的模型进行学习。网络表示学习使用了 Node2Vec、DeepWalk 和 LINE 三种方法，知识表示学习使用了 TransE、TransH 和 TransR 三种方法。语义表示的向量维度选取了 50、100 和 200 三个维度。

3.2　词语相似度

由于目前义原相似度分析和义项相似度分析的任务较少，所以我们仅通过词语相似度分析的任务来评估所提方法的学习质量。

3.2.1　评估方法

词语相似度任务一般是通过比较模型学习到的词对的余弦距离和标准数据集词，对标定数值的皮尔逊系数来判断词向量学习的质量，我们继承了该方法。中文数据集上我们选择了 Niu Y 等[2]的 4 种方法(包括最优的方法)作为比较，英文数据集上我们选择了 Neelakantan A 等[26]的 4 种方法(包括最优的方法)作为比较。实验结果如表 2 所示。

3.2.2　实验结果

通过实验结果，我们发现无论是在中文还是英文的语义相似度数据集上，我们的方法都达到了最好效果。其中 W297S 数据集上，DeepWalk(100 维)模型上达到了 67.0；W353S 数据集上，Node2Vec(200 维)模型达到了 71.1。这充分说明了 HownetGraph 可以较好地捕获到 HowNet 的语义信息。例如，W297S 中词对“冠军赛，锦标赛”给出的分数为 2.66/5=0.532，而 HownetGraph 学习到的分数为 0.879/1=0.879，从语义上考虑，HownetGraph 学习到的分数更符合人们的直观。

此外，我们发现知识表示学习的效果普遍不好，这是因为 HownetGraph 的关系较少(只有几十种)，因此在关系向量的牵引下，实体向量倾向于聚集而失去了区分度。

在 W297 和 W353 数据集上，HownetGraph 学习到的词语表示均不理想，这是因为这两个数据集中包含有较多基于上下文相似性的词语，基线方法均使用了无监督的大规模语料，可以很好地学习到词语之间上下文的相似性，但 HownetGraph 中几乎没有上下文的信息。

3.3　词语类比

3.3.1　评估方法

A1125 类比数据包含三大类：首都、城市、家庭关系。假设 w_1、w_2、w_3、w_4 分别是 4 个词，$\boldsymbol{E}_1$、$\boldsymbol{E}_2$、$\boldsymbol{E}_3$、

① https://github.com/Leonard-Xu/CWE/tree/master/data

② http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

③ https://github.com/thunlp/KB2E

④ https://github.com/thunlp/openne

E_4 是对应的词向量，若 w_1 和 w_2 的关系与 w_3 和 w_4 的关系相似，那么，$E_2-E_1=E_4-E_3$，即得知 E_1、E_2、E_3，我们便可以通过 $E_3-E_1+E_2$ 得到向量 E_4'，通过 E_4 和 E_4'的 cos 值来评估学习到的词向量的质量。

表2 中英文词语相似度任务实验结果

模型	维度	W240	W297	W297S	模型	维度	W353	W353S
CBOW		57.7	61.1	/	Huang et al-M		64.2	/
GloVe		59.8	58.7	/	MSSG 300d-M		**70.9**	/
Skip-gram		58.5	**63.3**	/	NP-MSSG 50d-G		61.5	/
SAT		**61.2**	**63.3**	/	NP-MSSG 300d-M		68.6	/
TransE	50	1.6	14.6	19.5	TransE	50	2.2	7.3
	100	1.8	22.4	33.6		100	5.2	12.3
	200	−11.0	20.2	30.4		200	12.5	18.8
TransH	50	−3.5	23.4	29.8	TransH	50	19.5	30.7
	100	−4.8	15.4	25.5		100	17.1	29.4
	200	4.2	29.0	38.1		200	12.1	23.0
TransR	50	19.9	38.2	49.9	TransR	50	28.5	44.0
	100	15.2	37.1	50.9		100	30.9	50.0
	200	25.2	41.4	50.0		200	30.9	48.0
LINE	50	17.1	28.4	40.0	LINE	50	35.6	46.4
	100	19.5	39.3	46.3		100	31.4	43.2
	200	29.8	40.4	48.9		200	35.1	47.5
DeepWalk	50	39.8	54.0	62.2	DeepWalk	50	53.8	66.0
	100	**42.9**	**57.5**	**67.0**		100	55.1	68.3
	200	37.3	55.2	65.7		200	54.5	68.4
node2vec	50	41.8	57.4	65.0	node2vec	50	**59.0**	70.5
	100	40.3	57.5	66.7		100	57.6	68.9
	200	39.3	56.4	66.7		200	57.3	**71.1**

我们采用两种评估指标：(1) Accuracy，假设和 E_4'的 cos 值最大的向量对应的词为 w_4'，Accuracy 值即为所有测试样例中 $w_4'=w_4$ 的频率值。(2) Mean Rank，按照 E_4'和词对应向量的 cos 值由大到小排列，得到词序列 $S(w)$，Mean Rank 值即所有测试样例中 w_4 在 $S(w)$中的位置的平均值。基线方法采用的是 Niu Y 等[2]的 4 种方法(包括最优的方法)。

3.3.2 实验结果

由实验结果，我们发现我们的模型在“首都”类别上 Mean Rank 值上达到了最好效果：3.4，远远好于基线方法。但是城市类别和家庭关系类别效果较差。

经过分析后发现，HowNet 的数据特征和 HownetGraph 表示学习的上述结果有着直接的因果关系。

(1) 首都类别效果好，因为首都类义项较为单一，并且 DEF 部分清晰地说明了和国家的关系。

(2) 城市类别效果不好，因为 HowNet 中并没有关注城市和省份之间的关系，因此，从我们学出的向量表示中并不存在这样的类比。例如，南京、上海等 605 个城市的 DEF 是一样的，均为 DEF={place|地方：PlaceSect={city|市}，belong="China|中国"，modifier={ProperName|专}}。

(3) 家庭关系类别效果不好，原因是 HowNet 中存在较多标注不一致的情况，例如，家庭类别的第一个数据，“男孩、女孩、兄弟、姐妹”，“女孩”有一个

义项是“daughter”，但是“男孩”却没有“son”的义项。此外，HowNet 同样没有关注到家庭角色之间的差异，例如，家庭关系中的“奶奶”“娘”“娘亲”“后娘”等 48 个义项的定义也都是一样的，均为 DEF＝{human|人：belong＝{family|家庭}，modifier＝{female|女}{lineal|直系}{senior|长辈}}。

4　示例

本节我们分别针对词语相似度任务和词语类比任务给出示例，分别选取了在实验部分效果较好的 DeepWalk 和 Node2Vec 方法，如表 3 所示。

表 3　A1125 数据集实验结果

模型	维度	Accuracy				Mean Rank			
		首都	城市	家庭关系	所有类别	首都	城市	家庭关系	所有类别
CBOW		49.8	85.7	**86.0**	64.2	37.0	1.2	62.6	37.6
GloVe		57.3	74.3	81.6	65.8	19.1	1.7	3.6	12.6
Skip-gram		66.8	93.7	76.8	73.4	137.2	1.1	3.0	83.5
SAT		**82.6**	**98.9**	80.1	**84.5**	**14.8**	**1.0**	**1.7**	**9.5**
TransE	50	0.0	0.0	0.0	0.0	35 693.6	59 032.9	43 894.2	41 311.9
	100	0.0	0.0	0.0	0.0	44 385.2	46 997.8	38 601.0	43 392.2
	200	0.0	0.0	0.4	0.0	34 284.4	47 764.4	18 952.1	32 672.8
TransH	50	0.0	0.0	0.0	0.0	51 414.5	61 803.4	52 332.1	53 254.0
	100	0.0	0.0	0.0	0.0	53 299.8	58 005.3	41 186.3	51 101.0
	200	0.0	0.0	0.0	0.0	43 089.6	56 424.4	43 943.0	45 372.3
TransR	50	0.6	0.0	2.2	0.9	877.0	10 178.6	2 928.6	2 821.7
	100	1.0	0.0	0.8	0.8	318.5	5 671.4	4 104.3	2 068.0
	200	0.3	0.0	0.8	0.4	1 402.5	10 065.6	8 137.5	4 381.1
LINE	50	0.0	0.0	0.0	0.0	19 902.1	17 101.3	17 594.4	18 907.6
	100	0.0	0.0	0.0	0.0	19 334.6	21 452.4	17 693.5	19 267.2
	200	0.0	0.0	0.8	0.4	12 605.0	14 520.5	11 771.3	12 701.5
DeepWalk	50	6.5	0.0	0.7	4.1	87.5	1 092.2	370.2	312.3
	100	10.3	0.0	1.5	6.6	55.9	670.1	256.6	200.1
	200	30.4	0.0	1.7	18.8	17.4	718.1	217.2	174.9
node2vec	50	72.9	0.0	1.4	**44.2**	**3.4**	**486.7**	206.2	**127.8**
	100	61.4	**1.1**	**2.5**	37.8	4.4	928.9	**189.7**	193.2
	200	51.4	**1.1**	1.1	31.4	9.2	900.4	200.5	194.3

4.1　词语相似度

表 4 给出了几个 DeepWalk 方法学习到的语义表示示例。通过观察示例，我们可以发现：①该方法较好地捕获到了不同语义单位之间的语义表示，结果比较合理；②输入中文词和英文词，查询到的最相近的词大多都是该词语的义项，比如“钱”和“Money”这两个示例，这个比较好理解，因为在 HownetGraph 中，词语只和义项有连接，所以在随机游走的时候，词语便只会和它的义项共现，也因此和义项的相似度最高。

表 4 DeepWalk 方法(100 维)学习到的语义表示示例

输入词	距离输入词距离最小的 10 个结果(距离升序)
语义[1]	Semantic[2]，语义 | semantic[3]，语义 | meaning[3]，意义 | meaning[3]，meaning[2]，谛 | meaning[3]，含义 | meaning[3]，涵义 | meaning[3]，意味 | meaning[3]，义 | meaning[3]
Money[2]	款子 | money[3]，钱币 | money[3]，铜钿 | money[3]，钱财 | money[3]，钱 | money[3]，币 | money[3]，货币 | money[3]，银根 | money[3]，财 | money[3]，银钱 | money[3]
月亮 | moon[3]	月亮[1]，太阴 | moon[3]，moon[2]，月亮 | lunar[3]，月宫 | moon[3]，lunar[2]，月球 | moon[3]，太阴[1]，卫星 | moon[3]，太阴 | lunar[3]
女 | female[3]	女[1]，女性 | female[3]，female[2]，女 | woman[3]，雌性 | female[3]，坤 | female[3]，女 | girl[3]，雌 | female[3]，牝 | female[3]，母 | female[3]
flower | 花[4]	鲜花 | flower[3]，蕤 | flower[3]，flower[2]，鲜花[3]，FlowerGrass | 花草[4]，草本植物 | herbaceous plant[3]，杏花 | apricot flower[3]，花 | flower[3]，仙人掌 | cactus[3]，矢车菊 | knapweed[3]

注：右上角序号代表含义：1 表示中文词，2 表示英文词，3 表示义项，4 表示义原。

4.2 词语类比

根据我们用知识图谱方式处理 HowNet 得到的词向量是符合语义的，因此在词语类比上我们也有明显的优势，以下是 Node2Vec(100 维)的结果。

表 5 给出了几个 Node2Vec 方法学习到的语义表示示例。通过观察示例，我们可以发现，该方法同样较好地捕获到了不同语义单位之间的语义表示，正确的答案总会在输出的前几个。

表 5 Node2Vec 方法(100 维)学习到的词语类比示例

输入词	距离输入词距离最小的 10 个结果(距离升序)
雅典[1]，希腊[1]，巴格达[1]	巴格达 | BAGDAD[3]，巴格达 | Baghdad[3]，巴格达 | Bagdad[3]，伊拉克 | Iraqi[3]，伊拉克[1]，Iraq | 伊拉克[4]，British Commonwealth of Nations[2]，立陶宛共和国[1]，spana[2]，Balkan | 巴尔干半岛[4]
北京[1]，中国[1]，东京[1]	日本[1]，东洋[1]，东瀛[1]，中国 | PRC[3]，日本 | Nipponese[3]，Grenada[2]，东洋 | Japanese[3]，日本 | Japanese[3]，锡金 | Sikkim[3]，日本 | Nippo[3]
侄子[1]，侄女[1]，爸爸[1]	妈 | ma[3]，岳父母 | father-in-law and mother-in-law[3]，侄女 | brother's daughter[3]，老婆婆 | mother-in-law[3]，senior | 长辈[4]，妈妈[1]，妈[1]，grandmother[2]，mother[2]，collateral | 旁系[4]

注：右上角序号代表含义：1 表示中文词，2 表示英文词，3 表示义项，4 表示义原。

5 结束语

本文通过将 HowNet 中复杂的语义结构建成知识图谱 HownetGraph，将较难处理的语义嵌套递归结构转化为易处理的图结构，可以使用基于神经网络和深度学习的方法学习到 HowNet 的语义表示，希望能为其他研究者提供借鉴。本文还使用网络表示学习和知识表示学习的模型为 HownetGraph 中的每个语义单位学习到了稠密低维的向量表示，并通过实验证明了学习到的语义表示的质量，也表明了知识图谱表示 HowNet 的有效性。

Niu Y[2]等使用 HowNet 的语义相似度去增强词向量表示的工作，和 Xie R[27]等使用词向量的相似性来预测 HowNet 义原的工作，都是比较好地结合了上下文相似性和语义相似性的工作。后续研究中，我们会探索两个方向：一是如何基于 HowNet 学习到更好的语义表示；二是如何把基于 HowNet 得到的偏语义相似性的表示和基于大规模无监督语料学习到的偏上下文相似性的表示结合起来。希望借此提升自然语言处理领域相关任务的效果。

参考文献

[1] 董振东，董强. 知网和汉语研究[J]. 当代语言学，2001，3(1)：33-44.

[2] Niu Y，Xie R，Liu Z，et al. Improved word representation learning with sememes [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017(1)：2049-2058.

[3] 刘群，李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学，2002，7(2)：59-76.

[4] 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.

[5] 孙景广，蔡东风，吕德新，等. 基于知网的中文问题自动分类[J]. 中文信息学报，2007，21(1)：90-95.

[6] Yan J，Bracewell D B，Ren F，et al. The creation of a Chinese emotion ontology based on HowNet[J]. Engineering Letters，2008，16(1)：166-171.

[7] 唐怡，周昌乐，练睿婷. 基于 HowNet 的中文语义依存分析[J]. 心智与计算，2010 (2)：109-116.

[8] Liu J，Xu J，Zhang Y. An approach of hybrid hierar-

chical structure for word similarity computing by HowNet[C]//Proceedings of the 6th International Joint Conference on Natural Language Processing, 2013: 927-931.

[9] **向春丞，穗志方，詹卫东**. HowNet **与** CCD **映射方法研究**[J]. **中文信息学报**, 2015, 29(3): 44-51.

[10] Zeng X, Yang C, Tu C, et al. Chinese LIWC lexicon Expansion via Hierarchical classification of word embeddings with sememe Attention[C]//Proceedings of AAAI 2018, 2018.

[11] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations [C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 701-710.

[12] Tang J, Qu M, Wang M, et al. LINE: Large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015: 1067-1077.

[13] Grover A, Leskovec J. node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 855-864.

[14] Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. ACM, 2015: 891-900.

[15] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv: 1609.02907, 2016.

[16] Yang C, Liu Z, Zhao D, et al. Network representation learning with rich text information[C]//Proceedings of the 24th IJCAI., 2015: 2111-2117.

[17] Tu C, Liu H, Liu Z, et al. Cane: Context-aware network embedding for relation modeling[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, 1: 1722-1731.

[18] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of the 27th ALL on Neural Information Processing Systems, 2013: 2787-2795.

[19] Wang Z, Zhang J, Feng J, et al. Knowledge gGraph embedding by translating on hyperplanes[C]//Proceedings of the 14th AAAI conference on Artifical Intelligence, 2014(14): 1112-1119.

[20] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion [C]//Proceedings of the 29th AAAI Conference on Artifical Intelligence 2015(15): 2181-2187.

[21] Ji G, He S, Xu L, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, 1: 687-696.

[22] Ji G, Liu K, He S, et al. Knowledge graph completion with adaptive sparse transfer matrix[C]//Proceedings of the 30th AAAI Conference on Artifical Intelligence, 2016: 985-991.

[23] Xiao H, Huang M, Hao Y, et al. TransG: A generative mixture model for knowledge graph embedding [J]. arXiv preprint arXiv: 1509.05488, 2015.

[24] He S, Liu K, Ji G, et al. Learning to represent knowledge graphs with gaussian embedding [C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. ACM, 2015: 623-632.

[25] Chen X, Xu L, Liu Z, et al. Joint learning of character and word embeddings[C]//Proceedings of IJCAI, 2015: 1236-1242.

[26] Neelakantan A, Shankar J, Passos A, et al. Efficient non-parametric estimation of multiple embeddings per word in vector space[J]. arXiv preprint arXiv: 1504.06654, 2015.

[27] Xie R, Yuan X, Liu Z, et al. Lexical sememe prediction via word embeddings and matrix factorization [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI Press, 2017: 4200-4206.

朱靖雯(1997—)，**本科生，主要研究领域为知识图谱、数据挖掘。**

E-mail: zhujingwen221@foxmail.com

杨玉基(1994—)，**硕士研究生，主要研究领域为知识图谱、数据挖掘。**

E-mail: yangyujiyyj@gmail.com

许斌(1973—)，**通信作者，博士，副教授，博士生导师，主要研究领域为知识图谱、数据挖掘。**

E-mail: xubin@tsinghua.edu.cn