北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2018.061

一种基于 Hownet 的词向量表示方法

陈洋[1]　罗智勇[2,†]

1. 北京语言大学信息科学学院，北京 100083; 2. 北京语言大学语言信息处理研究所，北京 100083;
† 通信作者, E-mail: luo_zy@blcu.edu.cn

摘要　针对基于预训练得到的词向量在低频词语表示质量和稳定性等方面存在的缺陷，提出一种基于 Hownet 的词向量表示方法(H-WRL)。首先，基于义原独立性假设，将 Hownet 中所有 *N* 个义原指定为欧式空间的一个标准正交基，实现 Hownet 义原向量初始化；其次，根据 Hownet 中词语与义原之间的定义关系，词语向量表示可以视为相关义原所张成的子空间中的投影，并提出学习词向量表示的深度神经网络模型。实验表明，基于 Hownet 的词向量表示在词相似度计算和词义消歧两项标准评测任务中均取得很好的效果。

关键词　词向量表示; Hownet; 词语相似性计算; 词义消岐

A Word Representation Method Based on Hownet

CHEN Yang[1], LUO Zhiyong[2,†]

1. College of Information Science, Beijing Language and Culture University, Beijing 100083; 2. Institute of Linguistic Information Processing, Beijing Language and Culture University, Beijing 100083; † Corresponding author, E-mail: luo_zy@blcu.edu.cn

Abstract　Word embedding method based on pre-training still has some defects in the stability and the quality of low-frequency words. The authors proposes a new word embedding method based on Hownet. First, based on the sememe independence assumption, all sememes of Hownet are specified in an Euclidean Space's standard orthogonal basis to initialize all sememe vectors. Secondly, utilizing the relationship between word and sememe defined in the Hownet, each word vector representation can be regarded as a subspace projection by related sememes. Finally, a deep neural network model is put forward to learn word representations. The experimental results indicate that proposed word embedding method based on Hownet obtained comparable results in the two standard evaluation tasks including the word similarity computation and the word sense disambiguation.

Key words　word embedding; Hownet; word similarity computation; word sense disambiguation

1 引言

近年来基于深度学习的神经网络模型在自然语言处理领域取得突破的进展，而词向量通常作为模型输入的标准配置。词向量的思想基于分布假说[1]，其形式是一个稠密连续的实数向量。目前训练词向量使用最多的方法包括 Mikolov 等[2–3]提出的 word2vec 和 Pennington 等[4]提出的 Glove 等，这些方法在大规模语料中通过目标词和上下文词语的共现训练得到词向量,但这种纯粹基于数据驱动进行训练获取词向量的方法存在性能不稳定[5]、低频词词向量质量不高[6]等不足。如何将已有的人工知识库结合到词向量的训练以提升词语向量表示质量成为值得关注的问题。

由于 Hownet 等[7]工知识库是基于离散符号表示的，而词向量是连续稠密的。本文提出一种以 Hownet 义原向量为基础的词向量表示方法，将离散形式的人工知识恰当地转化为连续向量表示形式。Hownet 中义原定义为一个最基本的、不易于再分割的意义最小单位，每个词都可以由若干个义原来组合表示。假定目前 Hownet 中所有义原之间相

收稿日期: 2018–04–15; 修回日期: 2018–08–18; 网络出版时间: 2018-08-22 17:38:35

互独立且完备，则每个词语都可视为由其构成的义原向量所张成的子空间内的一个投影，每个词语的向量表示可以转化为这个词语管辖的所有义原向量的加权平均。通过这种方式，我们就可以将义原这种离散符号形式表示转换为连续稠密的数学向量，而且这种指定义原向量、通过计算得到词向量的方法是自然高效的，不需要进行训练的初始词向量就可以取得与现有训练方式的词向量相当的效果，而进一步在大规模语料中进行训练又可以学习到文本中特有的语言现象，得到更好的词向量表示。

本文从 3 个方面对词向量进行了评测，首先选取 3 组高频词和 3 组低频词，对比基于 Hownet 的词向量表示和 word2vec 词向量表示的最近邻词语，实验结果表明，基于 Hownet 的词向量表示不仅在高频词方面和 word2vec 的效果是可比的，在低频词方面也更加稳定。然后对词相似度计算和中文词义消歧等标准评测任务进行实验，并将实验结果与同类型的研究进行比较，发现本文的方法可以得到更好的效果，进一步表明对 Hownet 义原的使用方式、对人工知识与词向量表示的结合是切实有效的。

2 相关工作

2.1 词向量表示

最初的词向量表示为 One-hot 形式，即每一个词向量只有一个位置为 1，其它均为 0，且长度与词表等长。这种表示方法存在两个重大的缺陷：第一个是语义鸿沟即无法通过向量之间的运算来表示语义上的相似度；第二个是向量的维度往往与词表长度成正比，所以维度往往很大，对计算和存储都是一种压力和挑战。

针对上述问题，Rumelhart 等[8]1988 年提出将词向量映射到一个低维稠密的语义空间，每个词向量是一个固定维数的浮点数向量。这种思想很好地弥补了 One-hot 表示中语义鸿沟和维度过大的缺陷。基于这种思想，又出现诸多训练词向量的模型，最著名的为 Bengio 等[9]2003 年提出的神经网络语言模型(NNLM)和 Mikolov 等[2−3] 2013 年提出的两种词向量训练模型 CBOW 和 Skip-gram。这两种模型都是基于分布假说，即通过核心词与上下文之间的关系进行建模，不同之处在于 NNLM 为 n-gram 模型，将一个预设窗口内的词语的词向量拼接用来预测目标词；CBOW 是将窗口内上下文的词向量相加用于预测目标词，Skip-gram 是通过目标词去预测上下文。CBOW 和 Skip-gram 为了解决伴随语料规模的急速增长的训练效率问题，都简化为一种词袋模型即没有将词语间的顺序考虑到建模中，是更为高效的词向量训练方法。

目前，词向量表示逐渐成为基于深度学习的语言信息处理模型输入的标准配置。

2.2 Hownet 义原及相关研究

Hownet 是一个中文语义知识库，Hownet 体系中一个很重要工作就是义原的归纳和总结。义原的定义是一个最基本的、不易于再分割的意义的最小单位，每个义原或词语都可以由一个或若干个义原来表示，同一个词的不同义项也由不同的义原组合来表示，故一个词的表示可以转化为用若干个义原对这个词进行表示。经过多年发展，Hownet 已经非常精简，我们采用的 Hownet 版本中义原仅为 2176 个，能够表示的总词数为 118343 个，义原总数仅占词表词总数的 1.8%，如果能够有效地使用这些义原，性价比将会非常可观。

在引入 Hownet 义原的研究中，如何更好的对义原进行表示是的关键。唐共波等[10]将义原引入语言模型的训练中，针对大规模语料库中的单义原词语，将语料库中的单义原词语替换为其义原，然后通过词向量训练模型进行训练，得到义原向量表示，但只得到了部分义原向量。孙茂松等[6]提出义项敏感模型，这种方法得到了 Hownet 中全部义原的向量表示。其做法是用 CBOW 模型训练得到的词向量根据 Hownet 义原标注得到义原向量，对于语料中的低频词，用其管辖的义原向量加和平均来表示，这种方法有效地改善了低频词词向量训练不足的现象。Niu 等[11]提出 SE-WRL 模型，在义项敏感模型的基础上加入注意力机制，通过上下文的向量表示，自动选取当前词最可能义项，然后对此义项的义原向量进行调整。以上研究都是通过 Hownet 的义原标注，然后借助大规模语料训练或根据训练得到的词向量反过来得到义原向量表示，是一种自上而下的思想。本文首先对义原向量进行建模，然后通过义原向量表示得到词向量表示，处理策略是自下而上。

3 基于 Hownet 的词向量表示方法

3.1 Hownet 义原向量指定

首先，随机初始化一个义原矩阵 $\boldsymbol{M}_{\text{sem}}$:

$$M_{sem}=\begin{bmatrix}\alpha_1\\ \alpha_2\\ \dots\\ \alpha_N\end{bmatrix}=\begin{bmatrix}a_{1,1} & \dots & a_{1,N-1}\\ a_{2,1} & \dots & a_{2,N-1}\\ \dots & \dots & \dots\\ a_{N,1} & \dots & a_{N,N-1}\end{bmatrix},$$

指定($\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_N$)为 N 个义原所对应的义原向量，每个义原向量为 N–1 维，故义原矩阵 $\boldsymbol{M}_{\text{sem}}$ 是一个 $N*(N-1)$的矩阵。我们假设 Hownet 中的义原相互独立，然后根据式(1)对 $\boldsymbol{M}_{\text{sem}}$ 进行施密特正交化:

$$\begin{aligned}&\gamma_1=\alpha_1,\\ &\gamma_2=\alpha_2-\frac{\langle\alpha_2,\gamma_1\rangle}{\langle\gamma_1,\gamma_1\rangle}\gamma_1,\\ &\gamma_3=\alpha_3-\frac{\langle\alpha_3,\gamma_1\rangle}{\langle\gamma_1,\gamma_1\rangle}\gamma_1-\frac{\langle\alpha_3,\gamma_2\rangle}{\langle\gamma_2,\gamma_2\rangle}\gamma_2,\\ &\dots,\\ &\gamma_N=\alpha_N-\frac{\langle\alpha_N,\gamma_1\rangle}{\langle\gamma_1,\gamma_1\rangle}\gamma_1-\frac{\langle\alpha_N,\gamma_2\rangle}{\langle\gamma_2,\gamma_2\rangle}\gamma_2-\dots-\frac{\langle\alpha_N,\gamma_{N-1}\rangle}{\langle\gamma_{N-1},\gamma_{N-1}\rangle}\gamma_{N-1}。\end{aligned}\tag{1}$$

通过式(2)对($\boldsymbol{\gamma}_1,\dots,\boldsymbol{\gamma}_N$)进行单位化:

$$\beta_1=\frac{\gamma_1}{\|\gamma_1\|},\beta_2=\frac{\gamma_2}{\|\gamma_2\|},\beta_3=\frac{\gamma_3}{\|\gamma_3\|},\dots,\beta_N=\frac{\gamma_N}{\|\gamma_N\|},\tag{2}$$

得到义原正交单位矩阵 $\boldsymbol{M}_{\text{semOrth}}$:

$$M_{semOrth}=\begin{bmatrix}\beta_1\\ \beta_2\\ \dots\\ \beta_N\end{bmatrix}=\begin{bmatrix}b_{1,1} & \dots & b_{1,N-1}\\ b_{2,1} & \dots & b_{2,N-1}\\ \dots & \dots & \dots\\ b_{N,1} & \dots & b_{N,N-1}\end{bmatrix},$$

其中($\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_N$)为 N 组标准正交单位基，是基于义原独立假设得到的新的义原向量表示。N 个义原与 N 组标准正交单位基一一对应。对 N 个义原建立“sem-id”索引，对于每一个义原 sem 可以得到对应的义原 id，再通过 id 进行 look-up 操作得到与 $\boldsymbol{M}_{\text{semOrth}}$ 行号对应的义原向量。至此，对 Hownet 中义原向量的表示完成。

3.2 基于 Hownet 义原向量的词向量表示

根据 Hownet 中每个词语标注好的对应义原，每个词语视为由对应的义原向量在这个义原子空间的投影。词语的词向量表示可以由对应的义原向量的加和平均来表示，如式(3)所示:

$$W_{vector}=\frac{1}{l}\sum\beta_i,\tag{3}$$

其中 $\boldsymbol{\beta}_{\text{i}}$ 为当前词对应的义原向量，l 为当前词对应义原的个数，$\boldsymbol{W}_{\text{vector}}$ 即为通过义原向量得到的词向量表示。通过“指定+计算”方法得到的词向量是自然且高效的，即使未训练，其效果与 word2vec 也是可比的。这得益于，这是因为在大规模语料上进行训练后，可以使 Hownet 义原标注中的语言学知识与文本中的语言学现象完美结合，提高了词向量表示的性能。词向量训练模型如图 1 所示。

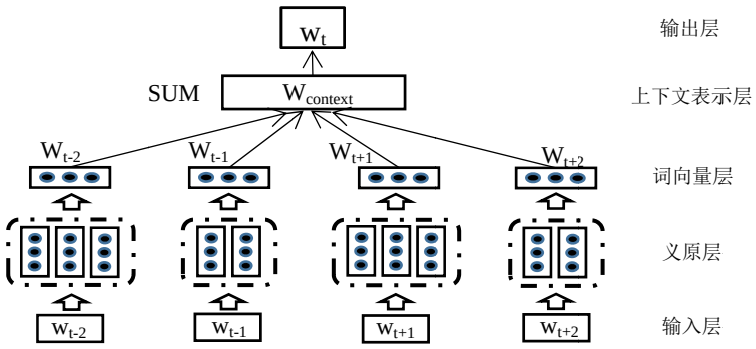

图 1 基于 Hownet 的词向量表示模型架构

Fig. 1　Word embedding representation model based on Hownet

为了更加准确表述我们的数据结构和模型，设词表长度为 V，义原数量为 N，最终的词向量维度为 D，上下文窗口大小为 k。根据 Hownet 义原标注体系，可以构建词–义原索引矩阵 $\boldsymbol{M}_{\text{word-sem}}$:

$$M_{word-sem}=\begin{bmatrix}X_1\\ \cdots\\ X_i\\ \cdots\\ X_V\end{bmatrix}=\begin{bmatrix}x_{1,1},\ldots x_{1,j},\ldots,x_{1,N}\\ \cdots\cdots\\ x_{i,1},\ldots,x_{i,j},\ldots x_{i,N}\\ \cdots\cdots\\ x_{V,1},\ldots,x_{V,j},\ldots x_{V,N}\end{bmatrix}。$$

对 $\boldsymbol{M}_{\text{word-sem}}$ 中的元素，有如式(4)的约束:

$$\begin{cases}if\left(sem_i\in word_j\right),x_{i,j}=1\\ else,x_{i,j}=0\end{cases},i\in[1,V],j\in[1,N],\tag{4}$$

其中，$\boldsymbol{M}_{\text{word-sem}}$ 是 $V*N$ 的矩阵。上一节得到的义原正交单位矩阵 $\boldsymbol{M}_{\text{semOrth}}$ 为 $N*(N–1)$的矩阵，对于句子中每个词语 w_i，可以通过对应索引得到 $\boldsymbol{M}_{\text{word-sem}}$ 中的一行 $\boldsymbol{X}_{\text{i}}=[x_{\text{i,1}},\ldots,x_{\text{i,j}},\ldots x_{\text{i,N}}]$，再通过式(5)得到每个词的向量化表示 $\boldsymbol{W}_{\text{vector}}$:

$$W_{vector}=X_i\cdot M_{semOrth},\tag{5}$$

$\boldsymbol{W}_{\text{vector}}$ 是$(N–1)$维向量。根据实际需要将词向量投影至指定维度 D，如式(6)所示:

$$W_{embedding}=W_{vector}\cdot M_{project},\tag{6}$$

$\boldsymbol{M}_{\text{project}}$ 是维度为$(N–1)*D$ 的投影矩阵。采用随机初始化，最终的 $\boldsymbol{W}_{\text{embedding}}$ 为 D 维的词向量表示，这样即可得到目标词上下文中所有词的词向量表示。然后通过式(7)得到上下文向量表示 $\boldsymbol{W}_{\text{context}}$:

$$W_{context}=\frac{1}{k-1}\sum_c W_c,\tag{7}$$

其中，$\boldsymbol{W}_{\text{c}}$ 为上下文的各个词的词向量表示，$\boldsymbol{W}_{\text{context}}$ 是$1*D$的矩阵，然后将 $\boldsymbol{W}_{\text{context}}$ 作为全连接层的输入，通过 softmax 函数得到目标词的预测 w_{t}。模型中只有投影矩阵 $\boldsymbol{M}_{\text{project}}$ 参与训练。

通常的词向量训练模型的初始词向量随机设定，没有任何语言学知识的约束。我们的方法是通过 Hownet 中的义原标注体系，经过由语言学知识构建的数学模型计算，得到每个词的词向量，所以初始词向量具有语言学意义，从而含义相近的词语的词向量也是相近的。通过这种方式，我们将义原这种符号主义离散形式的知识转化为连续稠密的数学模型，而且这种指定义原向量，通过计算得到词向量的方法是自然高效的，不需要进行训练的初始词向量，也可以取得与现有词向量可比的效果。经过在大规模语料上训练后的词向量则结合了语言学规则和上下文假设两个方面的特点，从而可以得到更好的词向量。

4 实验

4.1 基于 Hownet 的词向量表示

4.1.1 词向量训练数据集

我们词向量训练所用数据集包括 1993—2003 年共 11 年的人民日报新闻语料和搜狗实验室[2]提供的 2012 年 6—7 月 18 个频道的新闻语料，分词后共约 3.5 G，包括约 5.6 亿个词。

4.1.2 词向量训练模型配置

由于 Hownet 词表中的词是义原标注最准确的部分，所以目前我们的词向量模型词表是以 Hownet 词表为主体。首先根据 Hownet 词表对语料

[2] http://www.sogou.com/labs/resource/ca.php

进行分词, 对语料统计后发现, Hownet 词表中词的总频次对总词数的覆盖度为近 91%, 而 Hownet 词表外的高频词总频次占总词频的近 7%, 对于未登录高频词部分, 我们所用的分词系统[12]会给这些词一个类别标签: 时间、日期、数字、序数词、人名、地名、商标名、组织机构名和后缀类型名词), 我们将每个类别当作一个义原加入现有义原体系中, 属于这 9 类的词可视为单义原的词语。对于 Howent 词表外的低频词, 我们统一设定为'UNK'类别, 对应于一个'UNK'义原。这样我们在未改变 Howent 原架构的基础上做了适当的调整, 从而可以更好地结合本文提出的词向量表示模型。

词向量模型的词表大小为 118354, 包括 Hownet 原词表大小 118343、新增加的 9 个类别标签、1 个'UNK'标识和 1 个'@ZERO'标识(用于神经网络模型的零向量填充)。采用的 Hownet 版本的原始义原数量为 2176 个, 增加 9 个类别义原和 1 个'UNK'义原后, 新的义原总数共 2186 个。

模型的训练方式与 Skip-gram 模型相似, 指定窗口大小为 5, 即考虑目标词左右各两个词的大小。义原正交单位矩阵大小为[2186, 2185], 经过投影矩阵得到的词向量维度为 300。模型实现平台为 Tensorflow1.2.1, 初始化学习率为 1.0, 优化算法为 AdagradOptimizer, 模型训练中的负采样大小为 64。

4.1.3 词向量表示对比实验

我们选取语料中 3 个高频词和 3 个低频词, 分别通过 word2vec 和基于 Hownet 的词向量表示(H-WRL)计算这些词语的最近邻词, 结果如表 1 所示。从表 1 可以得到以下结论。

表 1 最近邻词计算实验结果

Table 1 Nearest word computation result

词	百万词频	word2vec	H-WRL
学校	0.16	中小学, 中学, 幼儿园, 小学, 学生, 院校, 高校, 这所, 该校, 高等, 师范, 班级, 校, 6．44 万, 校长	进修学校, 夜校, 寄宿学校, 母校, 教育机构, 学堂, 校规, 校旗, 建校, 兴学, 本校, 高小, 初小, 小学, 大学
发表	0.14	讲话, 声明, 里斯本机场, 反腐倡廉论, 林登·拉罗什, 干部作风论, 谈话, 内塔尼亚胡办公室, 创新思维论, 乌拉伊, 演说, 社论, 菲利普·鲍林, 科威特机场, 塞尔索·阿莫里姆	登载, 发行, 出刊, 登出, 报载, 刊载, 连载, 刊印, 刊行, 刊登, 刊出, 印发, 印行, 宣称, 宣讲
图片	0.08	附, 压题, (, 黄耀高, 柳晓力, 1 , 何春凤, 徐烨, 关威, 图, 余锦华, 刘妍, 胡学迅, 题图, 王绿霞	图解, 左视图, 图表, 剖视图, 剖面图, 附图, 平面图, 卡通画, 卡通, 同心圆, 平截头体, 缩尺图, 矢量图, 轴线, 座标
出丑	0.0002	跑神, 露馅儿, 意志不坚定者, 说重, 宁伤, 越规, 死要面子, 小报告, 咬手, 丰城供电公司, 出乖露丑, 怕失, 费唇舌, 酒疯, 吹牛拍马	坍台, 丢份, 蒙羞, 丢丑, 蒙垢, 丢脸, 丢面子, 丢人, 丢人现眼, 蒙尘, 露丑, 当场出彩, 出洋相, 现丑
匡正	0.0002	祛邪, 止恶, 编译者们, 兴廉, 时弊, 正本清源, 概念系统, 助善, 黄钟大吕式, 镜鉴, 颓风, 丁临一, 鞭挞, 言为心声, 良知, 此则	浪子回头金不换, 浪子回头, 雠, 矫治, 批改作业, 重新做人, 弃邪归正, 革面洗心, 改邪归正, 改过自新, 改过向善, 改过迁善, 迁善改过, 改恶向善, 改恶从善
答理	0.00002	示于, 演鬼, 肯叫, 割袍断义, 且听下回分解, 闲饭, 偷车者, 装穷, 斥之, 刘团长, 一点灵犀, 关心费, 贫嘴薄舌, 瞧得起, 下轿	应, 对, 答应, 卷, 报, 答案, 答记者问, 回执, 答词, 回条, 回帖, 回单, 答复, 应对如流, 对答如流

（1）word2vec 的稳定性不强。以“匡正”和“出丑”为例，虽然 word2vec 可以在其最近邻的词中发现意思相近的词，但是如果增大其近邻词的范围，其近义程度降低。对比 H-WRL 在“匡正”这一词的近邻词结果可以发现，得益于知识库义原的约束，即使增大近邻词的范围，其近邻词在语义上相似性也很强。

（2）word2vec 的低频词表示质量不高。根据齐夫定律，语料中必存在一个庞大的低频词集合，word2vec 基于数据的训练方式使得其对高频词的训练相对较多，低频词出现次数少，所以得到的训练较少，如果词频低到一定程度，质量就难以保证。通过对比低频词“答理”(百万词频仅为 0.00002)在 word2vec 和 H-WRL 的近邻词，可以发现 word2vec 的近邻词结果是不好的。H-WRL 是基于 Hownet 义原的，这些低频词不仅受文本中出现次数的影响，还受到知识库中已有规则的约束，故可以保证低频词的质量。

（3）word2vec 的基于数据驱动的训练方式得到的词向量表示更侧重于语义关联性即当前词和近邻词共同出现在一个上下文的情况，这一现象在“图片”这个词最为明显。本文模型因为结合了 Hownet 中的语言学知识，得到的近邻词更具有语义相似性，即近邻词和当前词是同义词。

4.2 任务一：词相似度计算

词相似度计算任务用于评价词向量的质量，方法是根据词向量计算给定词对的相似度。

4.2.1 数据集

评价词相似度所用的是 wordsim-297 标准数据集，此数据集每行的格式都是(w_1, w_2, score)，其中 w_1 和 w_2 为一对词语，score 为人工评分，评分区间是 0~5。通过式(8)计算给定两个词语的相似度:

$$Model_{score} = d\left(Embedding_{w1}, Embedding_{w2}\right), \quad (8)$$

其中 d(**Embedding**$_{w1}$, **Embedding**$_{w2}$)为余弦相似度计算，然后将 Model$_{score}$ 正规化到与数据集的人工评分区间相同，计算 Model$_{score}$ 和人工评分之间的 Spearman 相关系数。

4.2.2 实验结果

词相似度计算的实验结果如表 2 中所示，其中 word2vec 是通过 Skip-gram 模型训练得到的词向量，H-WRL 为基于 Hownet 义原的词向量表示，H-WRL-w2v 是将 H-WRL 词向量和 word2vec 词向量进行拼接得到的词向量。我们将 word2vec 作为 baseline，与另外两种模型进行比较。通过表 2 可以得到以下结论。

(1)在 wordsim-297 数据集上，我们的词向量模型得到的词向量在词相似度计算任务的结果好于 word2vec 的结果，说明我们的词向量表示能够更好地计算词向量表示之间的语义关联。

(2)将基于 Hownet 的词向量表示与 word2vec 的词向量表示进行拼接，可以提升整体词向量的质量。表明将人工知识和基于训练的词向量进行结合是合理有效的。

4.3 任务二：词义消歧

在这部分，通过词义消歧任务对基于 Hownet 的词向量表示进行评测。

4.3.1 数据集

评测采用 Senseval-3 的 Task5-Chinese Lexical sample 数据集，选取数据集中的“把握”、“材料”等 6 组词(共 198 个训练例和 96 个测试例)进行评测，这些词的义项数从 2~6 不等。

4.3.2 实验设置

我们主要对 3 种词向量表示进行测评：word2vec 词向量、本文提出的基于 Hownet 义原词向量表示(H-WRL)以及将 word2vec 词向量和基于 Hownet 的词向量进行拼接得到的词向量表示(H-WRL-w2v)。

词义消歧任务的目的是评测不同的词向量表示作为输入时的效果。我们参考目前广泛使用的循环神经网络模型 GRU，模型架构如图 2 所示。

句子 S={w_{t-4},...,w_t,...,w_{t+4}}作为模型的输入，基于目标消歧词得到两个上下文：$L_{context}$={w_{t-4},...,w_t}和 $R_{context}$={w_t,...,w_{t+4}}。通过词向量索引 look-up 操作，将词向量作为 GRU 的输入，得到两个输出 L_{state} 和 R_{state}，作为上下文的特征拼接，经过两个全连接网络层和 Softmax 函数进行分类，得到预测输出。

模型通过 keras 实现，算法采用十折交叉验证，保证了结果的准确性。模型中词向量的维度为 300(将两种词向量表示进行拼接，维度为 600)，GRU 中的节点维度为 100，全连接网络的隐层节点为 100 维，在两层全连接网络层之间用到 dropout，参数为 0.1。

我们的词义消歧任务随机选择义项作为 baseline，并将实验结果与相关工作，如 Li[13]的朴素贝叶斯方法、Wang[14]的 PageRank 方法以及孙茂松[6]的义项敏感模型等进行对比，结果如表 3 所示。

表 2 任务一词相似度计算实验结果

Table 2　Experimental result of word similarity computation

词向量表示模型	Spearman 相关系数
word2vec	66.27
H-WRL	67.33
H-WRL-w2v	**68.78**

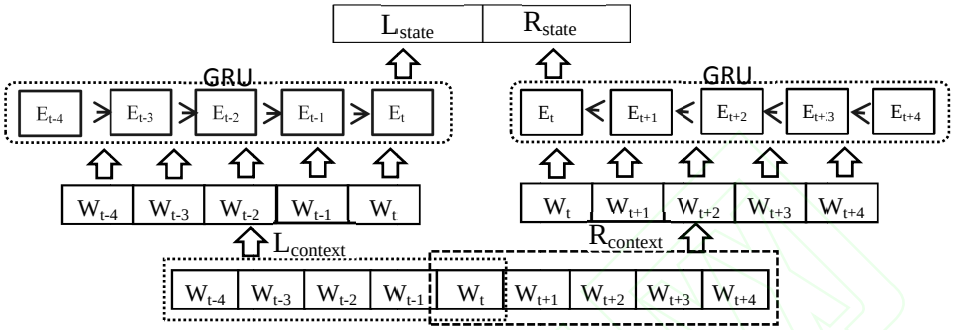

图 2 词义消歧模型架构

Fig. 2　Word sense disambiguation model

表 3 任务二词义消歧实验结果

Table 3　Experimental result of word sense disambiguation result

编号	歧义词	baseline	Li[13]	Page Rank[14]	义项敏感模型	H-WRL	word2vec	H-WRL-w2v
1	把握	0.25	0.32	0.53	–	**0.60**	0.47	0.53
2	材料	0.33	0.74	0.6	–	**0.80**	**0.80**	0.70
3	老	0.1	0.26	0.42	–	**0.46**	0.42	0.38
4	没有	0.25	0.39	**0.73**	–	0.67	0.47	**0.73**
5	研究	0.17	0.67	0.5	–	**0.86**	0.67	0.80
6	突出	0.33	0.27	0.47	–	0.73	0.60	**0.87**
平均准确率		0.24	0.44	0.54	0.58	**0.69**	0.57	0.67

4.3.3 实验结果

我们通过对实验结果表 3 进行分析，可以得出如下结论。

（1）本文提出的基于 Hownet 的词向量表示的方法得到的结果是最好的，表明我们的词向量表示具有一定的词义消歧能力。

（2）将 H-WRL 和 word2vec 的词向量进行拼接，提高了 word2vec 词向量的性能，可见通过知识引入可以对 word2vec 词向量表示进行有效的补充。

（3）在其中的 4 组词中，Hownet 的准确率都取得较大的优势，只有“材料”和“没有”与之前的最好结果相同，其原因可总结为两点:

①“材料”一词只有两个义项，与其他多义项词相比较简单;

②“没有”在语料中出现高达 51 万次，属于高频词汇，在 word2vec 的训练机制中得到充分的训练故可以有很好的效果。

（4）从整体效果来看，基于 Hownet 的词向量比于 word2vec 词向量的性能更加稳定，而且基于义原对词向量进行表示的方法不会出现低频词欠缺训练的问题，故基于 Hownet 的低频词词向量质量更高。

5 结论与展望

本文提出一种基于 Hownet 的词向量表示方法，基本思想是通过对义原向量进行建模将符号主义离散的知识转化为可以连续表示的数学模型。将义原向量化，所有义原共同构成一个义原向量空间，每个词语可以表示为在这个义原子空间的一个投影，通过这种方式我们将人工知识与现有基于训练的词向量表示方法进行有效结合，并通过最近邻、词相似度计算和词义消歧等实验对基于 Hownet 的词向量表示方法进行了测评。测评结果显示，其结果均好于现有最好结果，证明了这种词向量表示的有效性。总体来说，基于 Hownet 义原指定从而计算得到的词向量表示自然、高效且具有重要的语言学意义，不仅可以单独作为一种词向量的表示方式，也可以与基于训练的词向量进行拼接使用，提升基于训练的词向量表示的性能。

目前我们的研究中每个词的表示为这个词管辖下的所有义原，而当这个词为多义词时，用到的义原应该是不同的，且这些义原的重要程度也不一样。在未来的工作中，我们将对义项进行细分，对这些义原的重要程度进行量化研究，使得基于 Hownet 的词向量表示更强大，可以用于更多的自然语言处理任务中。

参考文献

[1] Harris Z S. Distributional structure. Word, 1981, 10(2/3):146–162

[2] Mikolov T, Chen Kai, Corrado G, et al. Efficient estimation of word representations in vector space // ICLR. Scottsdale, 2013: 1–12

[3] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations // Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Atlanta, 2013: 746–751

[4] Pennington J, Socher R, Manning C. Glove: global vectors for word representation // Conference on Empirical Methods in Natural Language Processing. Doha, 2014:1532–1543

[5] Antoniak M, Mimno D. Evaluating the stability of embedding-based word similarities. Transactions of the Association for Computational Linguistics, 2018, 23(6):107–19

[6] 孙茂松，陈新雄. 借重于人工知识库的词和义项的向量表示：以 HowNet 为例. 中文信息学报，2016，30(6): 1–6

[7] 董振东，董强.《知网》[DB/OL]. (2000) [2018–04–01]. http://keenage.com

[8] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors // Neurocomputing: Foundations of Research. MIT Press, 1988:533–536

[9] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3: 1137–1155

[10] 唐共波，于东，荀恩东. 基于知网义原词向量表示的无监督词义消歧方法. 中文信息学报, 2015, 29(6): 23–29

[11] Niu Y, Xie R, Liu Z, et al. Improved word representation learning with sememes // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, 2017: 2049–2058

[12] 罗智勇. 现代汉语通用分词系统的技术与实现. 北京: 北京工业大学, 2002

[13] Li W, Mccallum A. Semi-supervised sequence modeling with syntactic topic models // National Conference on Artificial Intelligence.Pittsburgh, 2005:813–818

[14] Wang J, Liu J, Zhang P. Chinese word sense disambiguation with PageRank and HowNet // IJCNLP. Hyderabad, 2007: 39–44