基于义原表示学习的词向量表示方法

于宁[1,2], 王江萍[1,2], 石宇[1,2], 刘建毅[1†]
1.北京邮电大学/北京市海淀区
2.北京邮电大学可信分布式计算与服务教育部重点实验室/北京市海淀区
ningy@bupt.edu.cn
codingwjp@bupt.edu.cn
yus@bupt.edu.cn
liujy@bupt.edu.cn

摘要

本文利用知网（HowNet）中的知识，并将Word2vec模型的结构和思想迁移至义原表示学习过程中，提出了一个基于义原表示学习的词向量表示方法。首先，本文利用OpenHowNet获取义原知识库中的所有义原、所有中文词汇以及所有中文词汇和其对应的义原集合，作为实验的数据集。然后，基于Skip-gram模型，训练义原表示学习模型，进而获得词向量。最后，通过词相似度任务、词义消歧任务、词汇类比和观察最近邻义原，来评价本文提出的方法获取的词向量的效果。通过和基线模型比较，发现本文提出的方法既高效又准确，不依赖大规模语料也不需要复杂的网络结构和繁多的参数，也能提升各种自然语言处理任务的准确率。

关键词： 知网（HowNet）；义原表示学习；词向量

Word Representation based on Sememe Representation Learning

Ning Yu[1,2], Jiangping Wang[1,2], Yu Shi[1,2], Jianyi Liu[1†]
1.Beijing University of Posts and Telecommunications/Haidian, Beijing
2.Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education/Haidian, Beijing
ningy@bupt.edu.cn, codingwjp@bupt.edu.cn
yus@bupt.edu.cn, liujy@bupt.edu.cn

Abstract

This paper uses the knowledge in HowNet and transfers the ideas of the Word2vec to propose a word representation method based on sememes representation learning. Firstly, we use OpenHowNet to obtain all sememes, all Chinese words, all Chinese words and their corresponding sememe sets in the sememe knowledge base as the experimental data set. Then, based on the Skip-gram model, we train sememe representation learning model to obtain word vectors. Finally, through the word similarity task, word sense disambiguation task, word analogy and observation of the nearest neighbor sememe, to evaluate the effect of the word vector. Compared with the baseline model, it is found that the proposed method is efficient and accurate, does not use large-scale corpus, does not need complex network structure and numerous parameters, and can improve the accuracy of various natural language processing tasks.

Keywords: HowNet , Sememe Representation Learning , Word Representation

基金项目：国家自然科学基金（U1836108, U1936216）；北京邮电大学中央高校基本科研业务费行动计划项目

1 引言

词向量表示学习旨在将符号形式的自然语言表示成一个低维实数向量，是自然语言处理领域的基础工作和重要任务之一。从最简单的“one-hot”开始，研究者就开始研究怎么将抽象符号表示为计算机可以识别并处理的形式。“one-hot”虽然简单易得，但是在计算过程中通常会遇到数据稀疏性问题。近年来，随着语料库不断丰富、算力不断增强，词向量表示学习取得了巨大的进展。研究者开始利用神经网络进行词表示学习。Mikolov et al. (2013)提出了Word2vec模型，其中包含CBOW和Skip-gram两个模型，相比于“one-hot”，该词向量维度降低，利于计算，词向量可以表达词汇关系，具有相同上下文的词会具有相似的向量，并且无需人工标注，可以利用丰富的语料库自动学习。在此之后，一大批词表示学习方法涌现，如以GloVe(Pennington et al., 2014)为代表的语言模型。Matthew et al. (2018)提出了继承Word2vec所有优点的ELMo。Devlin et al. (2018)提出了BERT预训练模型，此模型不仅可以进行微调应用到下游任务，也提供了预训练好的词向量。

由此可见，借助已有大规模语料，即在数据指导下，训练神经网络得到词向量，是目前的主流方法。但是这种思路却存在以下两个问题：一是将“汉字”或者“词汇”视为了最小的语言单元。Bloomfield et al. (1926)在文章中指出，义原是人类语言最小的语义单元。也就是说虽然字或词是NLP领域中最小的语言使用单元，却不是最小的语义单元。二是忽视了现实世界中存在的大量语言现象、常识和知识。没有知识指导的自然语言处理过程必然存在一定的局限性。

在这种有缺陷的思路下训练得到的词向量会出现以下问题：首先，具有多种含义的词汇，也就是常见的一词多义现象，只能得到一个向量、一种表示；其次，由于语料中低频词在训练过程中不能得到充分训练，基于此训练得到的词向量在任务中的表现较差，明显不如高频词。

基于以上思考，本文利用Dong et al. (2003)标注的大型语言常识知识库知网（HowNet）和Tomas Mikolov提出的Word2vec中的Skip-gram模型结构，研究如何获取更好的词向量。本文提出了一种新的思路——通过训练模型，获得义原向量，进而求得词向量。通过义原组合获得词向量，首先显然从根本上解决了一词多义的问题，其次低频词也不再囿于自身词频。基于此，本文提出了一个模型——W2VBS（**W**ord**2**vec **B**ased on **S**ememe），其优点是：一是改变了以往模型的输入粒度，即处理的最小单元，且引入外部语言知识，形成知识指导的词向量模型。二是模型属于轻量型，且训练数据集极小，训练速度快，资源消耗少。三是本文在多个下游任务上进行实验，达到了类似或好于其它方法的效果，验证了本文提出的模型的有效性。

本文安排如下：第二节简要介绍近年来研究者利用知网(HowNet)进行的各种工作；第三节介绍知网（HowNet）中可以利用的信息以及本文提出的基于Skip-gram的义原表示学习模型和基于义原的词向量表示方法；第四节介绍本文提出的方法计算得到的词向量在词相似度任务、词义消歧任务、词汇类比和观察最近邻义原上的表现；第五节总结本文的整体思路并提出存在的缺陷和未来的研究方向。

2 相关工作

知网（HowNet）是董振东、董强父子毕三十年之功标注的大型语言常识知识库。从2002年开始，研究者就开始利用知网（HowNet）进行自然语言处理。Liu et al. (2002)首次利用知网（HowNet）进行词汇相似度计算，在没有足够计算资源和复杂模型的情况下，充分利用其中每个概念的丰富的语义信息，得到的结果与人的直觉比较符合，词汇相似度值刻画也比较细致。Dang et al. (2010)提出一种基于知网（HowNet）的中文句子情感倾向判别方法，取得了良好的效果。Qi et al. (2019)首次考虑外部知识，将义原信息引入到语义组合问题中。Li et al. (2019)利用义原知识进行关系抽取，解决了分词不准确的问题和一词多义问题。Qin et al. (2020)将义原引入循环神经网络中，提升了模型的性能。Zhang et al. (2020)将义原与Transformer模型结合，提出三种Transformer变体，具有极高的鲁棒性。

近年来，研究者意识到静态的知识库存在人工标注耗时耗力、多人标注存在标准不一致性问题、人工标注的词汇存在噪声问题和不完整性问题、词汇的义原是动态变化的且存在时效性问题以及新词的义原缺失问题，所以开始探索知网（HowNet）中的内容自动扩充问题。Xie et al. (2017)首次尝试进行词汇的自动义原获取，提出SPWE、SPSE、SPASE三种模型，利用Word Embedding 和Sememe Embedding 进行义原预测。Jin et al. (2018)尝试利用词汇的内部字符信息和外部上下文信息，解决了低频词的义原预测问题。Du et al. (2020)尝试使用局部语义匹配的思想进行新词的义原标注。Qi et al. (2020)提出建立多语种义原知识库。

模型	不依赖 大规模语料库	不依赖 预训练词向量	利用 外部知识库	自下而上
Sun et al. (2016)	✗	✗	✔	✗
Niu et al. (2017)	✗	✔	✔	✗
Chen et al. (2019)	✗	✔	✔	✔
W2VBS	✔	✔	✔	✔

表 1: 与其它模型的比较及不同之处

在融合义原进行词向量表示方面，Sun et al. (2016)将知网（HowNet）引入词表示学习模型中，验证了义原能帮助获得语义更丰富的词向量。Niu et al. (2017)基于Skip-gram模型提出SE-WRL方法，其中包括SSA、SAC、SAT三个模型，旨在利用义原信息完善并提高词向量的表示，解决一词多义问题。Chen et al. (2019)提出一种正交化义原向量的方法，改变以往通过预训练词向量反过来指导义原向量表示的思路，从自上而下改为自下而上，通过训练义原向量，进而获得词向量，更清晰地诠释了义原向量和词向量之间的关系。本文集合以上模型的优点，不依赖大规模语料库也不依赖预训练词向量，利用外部知识库，自下而上，训练义原向量，进而获取词向量。表1展示了本文提出的模型与以往模型的不同之处。

其中值得一提是，清华大学自然语言处理实验室(Qi et al., 2019)发布了一个基于HowNet的工具——OpenHowNet。该工具的核心数据文件包含223767个中英文词和词组所代表的概念。该工具还提供多种功能接口，包括检索词汇对应的概念标注的完整信息、检索词汇的义原集合、计算基于义原的词相似度、检索词的最近邻词、获取两个义原之间的关系，还包括检索两个义原之间是否存在关系、存在什么关系等。

3 方法

本节中将介绍本文如何充分利用知网（HowNet）中的信息，并介绍本文提出的模型和方法：基于Skip-gram的义原表示学习模型和基于义原的词向量表示方法。

3.1 HowNet中的义原信息

知网（HowNet)是一个以汉语和英语的词汇所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。知网（HowNet)的全部主要文件包括知识词典构成了一个有机结合的知识系统。其中的哲学的根本点是：世界上一切事物（物质的和精神的）都在特定的时间和空间内不停地运动和变化。它们通常是从一种状态变化到另一种状态，并通常由其属性值的改变来体现。知网（HowNet）中存在义原、义项、词汇三种粒度的信息。知网中还标注了各种粒度的信息之间的关系。

图1和图2展示了知网（HowNet）中的词汇的标注信息和结构。

```
Display 0 sememe tree
sense苹果
 └────[None]computer—电脑
 ├────[modifier]PatternValue—样式值
 │  └────[CoEvent]able—能
 │  └────[scope]bring—携带
 │  └────[patient]$
 └────[patient]SpeBrand—特定牌子
Display 1 sememe tree
sense苹果
 └────[None]fruit—水果
```

图 1: 词汇“苹果”在知网（HowNet）中的标注

图1是词汇“苹果”在知网（HowNet）中的标注信息和结构。从表中可知，“苹果”有两个义项。第一个义项下包含“电脑”、“样式值”、“能”、“携带”和“牌子”五个义原；第二个义项下包含“水果”一个义原。

Display 0 sememe tree
sense水分
└———[None]Dampness—湿度
└———[host]physical—物质
Display 1 sememe tree
sense水分
└———[None]information—信息
└———[content]boast—夸大

图 2: 词汇“水分”在知网（HowNet）中的标注

图2是词汇“水分”在知网（HowNet）中的标注信息和结构。从表中可知，“水分”也有两个义项。第一个义项下包含“湿度”和“物质”两个义原；第二个义项下包含“信息”和“夸大”两个义原。

3.2 基于Skip-gram的义原表示学习模型

本文提出了一个将Word2vec迁移至义原向量表示学习过程中的思路。沿用Skip-gram模型的结构和思想，仅改变模型的输入和输出。同时，改变自上而下的思路，自下而上地训练义原向量，进而获得词向量。

Skip-gram模型的思想是：有相同上下文的词，应该有相似的词向量。本文的思想是：被包含在相似义原集合中的义原，应该有相似的义原向量。基于以上，本文的策略是：遍历知网（HowNet）中所有标注的中文词汇的义原集合。通过给定某一义原，预测可能与之搭配使用的义原们（这些义原组合起来可以表示一个词的概念），来训练义原向量表示学习模型。

图3以“智能”一词为例，展示了本文提出的义原表示学习模型结构，是一个只有两层的神经网络。

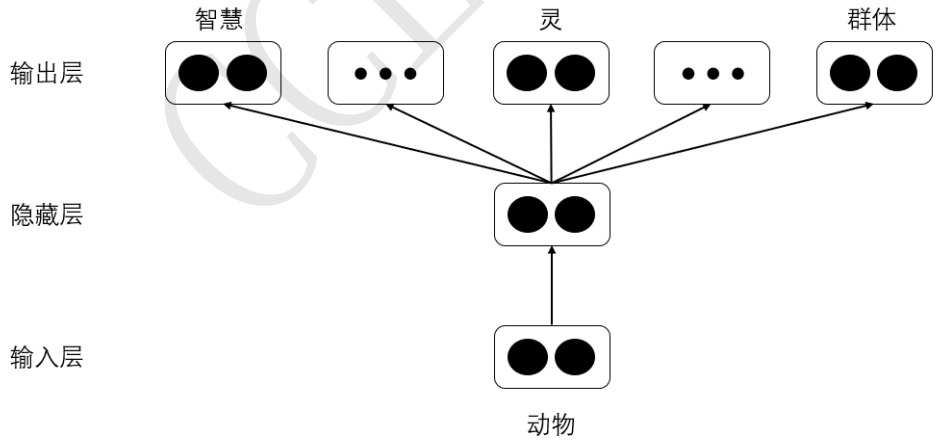

图 3: 基于Skip-gram的义原表示学习模型

图3可解释为：“智能”一词在知网（Hownet）的义原集合为{“动物”，“智慧”，“灵”，“群体”}。按照实验设置的窗口大小，在义原集合上进行滑动。首先将“动物”作为输入层，然后经过一个隐藏层，输出层为“智慧”、“灵”和”群体”。然后再次滑动窗口，依次进行训练。训练完成后取神经网络的第一层权重参数作为义原向量的Look-up Table。

模型的形式化解释为：输入某个义原集合$setS$中的义原 S_i ,预测义原集合中的其它义原S_0，S_1，S_{i-1}，S_{i+1}, ... , S_n。模型的目标函数为:

$$L = \sum_{S_i \in setS} logp(setS - S_i|S_i) \tag{1}$$

3.3 基于义原的词向量表示方法

本文考虑知网（HowNet)对词汇的标注信息和结构，探索基于义原的词向量表示方法。在直观上，一个词汇可以由一个有限的义原集合表示，那么词向量也可以由其对应的义原向量求得。3.2中已经讨论了如何进行义原向量表示学习，所以到目前为止，词向量也很容易求得:

1）当我们只考虑义原向量时，词向量可由其对应的义原向量相加或平均取得。

$$W_{vector} = \sum_{i=1}^{n} S_i \tag{2}$$

其中W_{vector}是该词汇的词向量，S_i是该词汇对应的义原集合中的第i个义原向量，n是义原集合的大小。

2）当我们考虑一词多义问题时，词向量可由该词汇某个义项下对应的义原向量相加或平均取得。

$$W_{vector} = \sum_{i=1}^{n} Se_i \tag{3}$$

其中W_{vector}是该词汇的词向量，Se_i是该词汇某个义项下对应的义原集合中的第i个义原向量，n是该义项下的义原的数量。

3）当我们考虑增强预训练词向量的语义丰富性时，词向量可由1）2）求得的词向量和预训练的词向量相加或连接取得。

$$W_{vector} = \sum_{i=1}^{n} S_i + Pretrained\ Word\ Embeddings \tag{4}$$

其中前半部分是1）2）中的结果，后半部分是预训练的词向量。

基于3.2和3.3，本文将提出的模型命名为W2VBS（**W**ord**2v**ec **B**ased on **S**ememe）。

4 实验

本节中将介绍本文的实验数据集和模型在词相似度任务、词义消歧任务、词汇类比和观察最近邻义原上的表现。

关于实验的原始训练数据集和处理后的训练数据集。首先，本文利用清华大学自然语言处理实验室发布的OpenHowNet工具，获取义原知识库中的所有义原、所有中文词汇以及所有中文词汇和其对应的义原集合，作为本文的原始训练数据集。其次，本文考虑到，每个词汇对应的义原是一个集合（元素之间无顺序），而且对应集合中的义原都彼此相关。为了应用适用于序列（元素之间有顺序）的Skip-gram模型，且使模型得到充分训练，本文将每个词汇的义原集合随机打乱10次，使集合中的每个义原都能与彼此出现在一个窗口内，形成处理后的训练数据集。在实验过程中，本文设置窗口大小为1和2，即窗口内为3个义原和5个义原，迭代次数为5，使用hierarchical softmax技巧，训练长度为300维的义原向量。表2展示了本文利用知网（HowNet)构造的训练数据集的项以及各项对应的数量。

4.1 词相似度任务

4.1.1 数据集

本文在标准数据集wordsim-240上，测试本文提出的模型获取的词向量是否可以提高任务性能。

wordsim-240数据集中，包含240个词对儿，其中每一对词都有一个人工打分，这个打分的含义是两个词在语义上的相似程度。打分的范围为1-10。结构如“李白 诗 9.2”。

项	大小
义原	2187
中文词汇	127262
训练数据集	347280

表 2: 基于知网(HowNet)构造的数据集

模型	与人工打分的相关性
CBOW	55.85
Skip-gram	53.42
GloVe	48.22
SSA	58.90
W2VBS	53.09
W2VBS + CBOW	59.32

表 3: 不同模型在词相似度任务的上的表现

在实验过程中，由于“华佗”、“本拉登”、“海外华人”和“社会调查”四个词在知网（HowNet）中没有标注信息，所以在利用W2VBS进行实验的过程中忽略了包含这四个词的四个词对。在W2VBS + CBOW中，正常进行实验。

4.1.2 实验结果

本文选取词向量表示学习的经典模型——CBOW、Skip-gram和GloVe，以及Niu et al. (2017)提出的**SSA**（**S**imple **S**ememe **A**ggregation **M**odel），作为比较的基线模型。表3展示了利用各基线模型和本文提出的模型得到的词向量，计算出的词间相似度与人工打分的Spearman相关系数。

词间相似度的计算公式为：

$$similarity(W_a, W_b) = \cos\theta = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} {a_i}^2} \times \sqrt{\sum_{i=1}^{n} {b_i}^2}} \tag{5}$$

其中，W_a和W_b是两个词汇，a_i和b_i分别是各自词向量的第i个分量，n是词向量的长度。

通过表3,本文得到以下结论：

1）在标准数据集wordsim-240上，本文提出的W2VBS模型得到的词向量与经典词向量基线模型差异不大。在不依赖大规模语料也不利用复杂的网络结构和繁多的参数来训练的情况下，能取得与基线模型差不多的效果。这说明利用知网（HowNet）获取词向量的方法，是值得关注的。

2）本文提出的将W2VBS模型得到的词向量和在大规模语料上预训练得到的词向量结合，作为完整的词向量，明显改善了任务效果。这说明将知网（HowNet）作为外部语言知识，引入当前的词向量表示学习中，是必要的。

4.2 词义消歧任务

4.2.1 数据集

本文在Semeval2007中文词义消歧任务上，测试本文提出的模型获取的词向量是否可以提高任务性能。

同Sun et al. (2016)的实验数据集一样，本文选取了具有一词多义性质的六个词——“把握”、“材料”、“老”、“没有”、“突出”和”研究”作为实验数据集。

4.2.2 实验结果

本文选取随机选择义项、Li et al. (2005)的朴素贝叶斯、Wang et al. (2008)的PageRank，以及Sun et al. (2016)提出的义项不敏感模型，作为比较的基线模型。表4展示了利用各基线模型和本文提出的模型获得的词向量，在词义消歧任务中的六个候选词上的平均准确率。

模型	平均准确率
随机选择义项	0.24
朴素贝叶斯	0.44
PageRank	0.54
义项不敏感	0.56
W2VBS	0.57
W2VBS + CBOW	0.61

表 4: 不同模型在词义消歧任务上的表现

词汇	背债	蠢笨	二赖子	匡谬	不宣而战
SogouT词频	99	95	51	10	1
CBOW 给出的最近邻词	替前夫 家欠 堕胎费	自大无比 愚蠢无知 懦弱无用	横眉瞪目 张五魁 花荣志	吉金录 曲话 笋谱	西方 伊朗人 冲突国
Sun et al. (2016) 给出的最近邻词	债款 债务 债项	木讷 呆头呆脑 愚蠢	恶棍 穷凶极恶 逞凶	错误 订正 讹误	杀敌 拔寨 整军
本文模型 给出的最近邻词	欠账 亏空 呆账 倒账 该欠 该账	不协调 粗拙 呆笨 粗手笨脚 呆痴 呆钝	浑球儿 混混儿 孽障 泼皮 地头蛇 小流氓	补偏救弊 改悔 匡正时弊 弃暗投明 纠偏 补过	开仗 兴兵 开战 动兵 开打 交兵

表 5: 词汇类比对比

通过表4,本文得到以下结论:

1）在Semeval2007中文词义消歧任务上，利用本文提出的W2VBS得到的词向量，提高了任务的准确率。这说明基于义原表示学习的词向量表示方法能够很好地获得句子的语义信息。

2）本文提出的将W2VBS模型得到的词向量和在大规模语料上预训练得到的词向量结合，作为完整的词向量，得到了最高的准确率。这说明基于义原向量的词向量弥补了预训练词向量在语义上的缺失。

4.3 词汇类比

一个好的词向量表示方法应该将具有相似语义的词映射到向量空间中的近邻位置。为直观展示本文提出的模型的性能，本文选取CBOW、Sun et al. (2016)的词向量结果作为基线模型。由于基线模型选取了SogouT语料中的5个低词频——“背债”、“蠢笨”、“二赖子”、“匡谬”、“不宣而战”，本文也在这五个词上进行实验，分别给出它们的6个最近邻词，与基线模型的结果进行比较。表5中展示了对比结果。

从表5中可以看出，本文模型给出的最近邻词明显好于CBOW和Sun et al. (2016)给出的最近邻词。如词频是51的“二赖子”，它明显是一个名词。本文模型可以很好的计算给出它的近义词，并且词汇的词性都相同，如“混混儿”，而且“混混儿”一词在SogouT中的词频也很小。这说明词频小的词汇在以往的模型中没有得到很好的训练，词向量的质量也就随之下降。再有词频是1的“不宣而战”，CBOW给出的最近邻词明显不应该属于它的近邻词，而我们的模型给出的“开仗”、“动兵”等词汇，完全符合其近义词的概念。

4.4 观察最近邻义原

本文通过训练义原向量，进而计算获得词向量，也就是说义原向量是最底层的表示。如果我们能发现底层的表示符合人们的感知和预期，那么该表示就可以被广泛推广到各种下游任务中。本文类比“具有相似语义的词之间的距离小”的原理，得出“同类义原之间的距离小”的结

义原	最近邻义原top20
女	男、直系事情、贞洁、同辈、 人、后裔、小辈、配偶、 已婚、长辈、家庭、正经、 成年、妖媚、社交性、直系、 旁系、不忠、不善交往、冷漠
生气	情绪、感叹、态度、激动、 可信、感情特性值、坏脾气、羞愧、 仇恨、失望、不合理、报复、 悲哀、感激、外移、表示情感、 生物人、不说、喜悦、无序
深度	高度、明暗、坡度、宽度、 高贵、体积、厚度、电阻、 示怒、下次、光洁度、水域、 辉煌、苏格兰、美洲、方位特性、 变外观、密度、薄思想、序数

表 6: 最近邻义原

论。为直观展示本文提出的模型是否将同类义原映射到了向量空间中的近邻位置，表6中分别列举了与“女”、“生气”和“深度”三个义原最近邻的20个义原。

从表6中可以看出，本文模型计算的义原向量非常符合预期。相似的义原被映射到了向量空间中的近邻位置。如表达性别和社会关系的“女”以及其近邻义原、表达情绪的“生气”以及其近邻义原和表达属性的“深度”以及其近邻义原，都能很好的被算法挖掘，显示了本文提出的模型有很大的潜力。

5 总结

本文尝试改变以往将“字符”或“词汇”当做最小的模型处理单元的思路，将大规模外部知识集——知网（HowNet）引入自然语言处理过程中，并尝试思考基于义原的词向量表示方法，旨在表示当前自然语言处理工作不仅需要大规模语料、设计包含大量参数的复杂模型，而且一定是需要知识的指导的。自然语言是人类独有的符号化语言，其更新快、复杂多样、歧义性高，要想让计算机看懂自然语言、分析自然语言，甚至能够和人进行交流，达到一定的智能程度，必然需要人类语言知识的指导。没有外部知识指导的自然语言处理过程必然受到一定的限制，各个环节都不能突破一定的瓶颈。

本文在实验过程中，也总结了未来工作可以提升的地方。一是本文将义原集合中的元素视为平等的，但是知网（HowNet）中对义项、义原以及彼此之间的关系都进行了标注，忽略这些信息使得我们虽然解决了一词多义问题的第一步——义原向量的表示，但是下一步我们应当着重利用这些层次信息，为词汇选择正确的义项，真正实现针对一词多义问题的完整流程；二是知网（HowNet）的覆盖率不足，这限制了模型只能在收录的词汇上发挥作用。在第二部分相关工作中，本文也调研了当前有一部分工作在做新词的义原预测问题，这也是我们今后研究的方向；三是本文只是将知网（HowNet）引入了词向量表示学习中，HowNet的知识体量很大，怎么将它运用到各项自然语言处理过程中，如文本分类、信息抽取、文本摘要、自动问答等，是我们今后研究的重点。

参考文献

CHEN Yang and LUO Zhiyong. 2019. A Word Representation Method Based on Hownet. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 55(1): 22-28.

Devlin J, Chang M W, Lee K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

F Qi, Huang J and Yang C 2019. Modeling Semantic Compositionality with Sememe Knowledge. *Meeting of the Association for Computational Linguistics. 2019.*

Huiming Jin, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin and Leyu Lin. 2018. Incorporating Chinese Characters of Words for Lexical Sememe Prediction. *ACL 2018.*

Jiaju Du, Fanchao Qi, Maosong Sun and Zhiyuan Liu. 2020. Lexical Sememe Prediction using Dictionary Definitions by Capturing Local Semantic Correspondence. *arXiv e-prints 2020.*

L.Bloomfield. 1926. A set of postulates for the science of language *Language* , vol.2 no.3, pp.153-164.

Lei Dang and Lei Zhang. 2010. Method of discriminant for Chinese sentence sentiment orientation based on HowNet. *Application Research of Computers 2010.*

Li, W. and A. McCallum. 2005. Semi-supervised Sequence Modeling with Syntactic Topic Models. *Proceedings of AAAI:*p.813.

Maosong Sun and Xinxiong Chen. 2016. Embedding for Words and Word Senses Based on Human Annotated Knowledge Base: A Case Study on HowNet. *JCIP.*

Matthew E. Peters and Mark Neumann and Mohit Iyyer and Matt Gardner and Christopher Clark and Kenton Lee and Luke Zettlemoyer. 2018. Deep contextualized word representations. *peters2018deep:81.*

Niu Y, Xie R and Liu Z. 2017. Improved Word Representation Learning with Sememe. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Volume 1: Long Papers.

Pennington,J., R. Socher and C.D. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of EMNLP.*

Qi F, Chang L and Sun M. 2020. Towards Building a Multilingual Sememe Knowledge Base: Predicting Sememes for BabelNet Synsets. *Proceedings of the AAAI Conference on Artificial Intelligence, 2020*,34(5):8624-8631.

Qi, Fanchao and Yang, Chenghao and Liu, Zhiyuan and Dong, Qiang and Sun, Maosong and Dong, Zhendong. 2019. OpenHowNet: An Open Sememe-based Lexical Knowledge Base. *arXiv preprint arXiv:*1901.09957.

Qun Liu and Sujian Li. 2002. Word Similarity Computing Based on HowNet. *International Journal of Computational Linguistics Chinese Language Processing 2002.*

Ruobing Xie, Xingchi Yuan, Zhiyuan Liu and Maosong Sun. 2017. Lexical Sememe Prediction via Word Embeddings and Matrix Factorization. *IJCAI 2017.*

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space.

Wang, J.,J. Liu and P. Zhang. 2008. Chinese Word Sense Disambiguation with PageRank and HowNet. *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing.*

Yuhui Zhang, Chenghao Yang, Zhengping Zhou, Zhiyuan Liu. 2020. Enhancing Transformer with Sememe Knowledge. *Rep4NLP 2020.*

Yujia Qin, Fanchao Qi, Sicong Ouyang, Zhiyuan Liu, Cheng Yang, Yasheng Wang, Qun Liu and Maosong Sun. 2020. improving Sequence Modeling Ability of Recurrent Neural Networks via Sememes. *TASLP 2020.*

Zhendong Dong and Qiang Dong. Introduction to HowNet.

Ziran Li, Ning Ding, Zhiyuan Liu, Haitao Zheng and Ying Shen. 2019. Chinese Relation Extraction with Multi-Grained Information and External Linguistic Knowledge. *Proceedings ofthe 57th Annual Meeting ofthe Association for Computational Linguistics:*4377–4386.