基于网络嵌入和预训练模型的义原预测

白 宇[1,2] 王之光[2] 刘懿萱[2] 蔡东风[1,2]

[1](南京航空航天大学计算机科学与技术学院 江苏 南京 211106)

[2](沈阳航空航天大学人机智能研究中心 辽宁 沈阳 110136)

摘 要 义原是构成《知网》概念描述的核心部件,义原预测是 HowNet 自动或半自动扩展中涉及的关键问题之一。提出一种基于网络嵌入和预训练模型的义原预测方法,通过对《知网》中的字-词-义项-义原及其关系的表示学习,融合预训练语言模型动态构建局部“义项-义原”关系网络,实现新概念与候选义原的动态匹配。实验结果中的义原预测 F1 值达到 0.623 7,表明该方法能够更有效地解决《知网》中未登录词的义原预测问题。

关键词 义原 预训练语言模型 网络嵌入

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.07.007

SEMEME PREDICTION BASED ON NETWORK EMBEDDING AND PRE-TRAINING MODEL

Bai Yu[1,2] Wang Zhiguang[2] Liu Yixuan[2] Cai Dongfeng[1,2]

[1](*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, Jiangsu, China*)

[2](*Human-Computer Intelligence Research Center, Shenyang Aerospace University, Shenyang 110136, Liaoning, China*)

Abstract Sememe is the core component of concept description in HowNet, and the predication of sememe description for new concepts is the key issue involved in automatic or semi-automatic expansion of HowNet. This paper proposes a sememe prediction method based on network embedding and the pre-training models. It realized the dynamic matching between the new concept and the candidate sememe by learning representation of the character-word-concept-sememe and their relationships in HowNet, and combining the pre-training language models to construct the partial "concept-sememe" relationship network. The predicted F1 value of the experimental results was 0.6237, which indicated that this method could solve the problem of semantic prediction of OOV words in HowNet more effectively.

Keywords Sememe Pre-training language model Network embedding

0 引 言

HowNet(知网)[1]自问世以来,受到自然语言处理领域的广泛关注,国内外学者在词汇语义消歧[2-3]、相似度计算[4-6]、文本分类[7]和信息检索等方面探索了 HowNet 的重要应用价值。研究发现,HowNet 通过统一的义原标注体系直接刻画语义信息,且每个义原含义明确固定,可被直接作为语义标签融入机器学习模型,使自然语言处理深度学习模型具有更好的鲁棒性和可解释性[8]。

义原(Sememe)是构成概念描述的核心部件。目前,《知网》构建了包含约 2 230 个义原的精细的语义描述体系,并为约 14.8 万概念标注了义原。

然而,与其他依靠人工构建的知识库系统一样,HowNet 存在着规模有限、更新扩展维护成本高的问题。相关研究[9]表明,没有 HowNet 背景知识和未经专门训练的人员难以较好地完成义原标注任务。这导致 HowNet 潜在的巨大应用价值与其自身规模有限、语义资源稀疏且难扩展的矛盾,解决这个矛盾的一个可行的途径就是开展 HowNet 的自动或半自动构建技术的研究,其核心问题之一就是为新概念的描述选择合适的义原。

新概念是指随时代发展而新出现或旧词新用的概念。随着互联网应用的普及,文本大数据中大量的新

收稿日期:2021-03-02。国家自然科学基金项目(U1908216)。**白宇**,讲师,主研领域:信息检索,语言与知识计算。**王之光**,硕士生。**刘懿萱**,硕士生。**蔡东风**,教授。

词不断出现,同时现有词语的含义被延伸和扩展,因此有必要对以义原为基础的语义知识库进行持续地修正和扩充。在词汇义原自动标注方面,Xie 等[23]提出了义原预测任务,该任务是在 HowNet 义原集合中选择出适合构建新概念的 Def 描述的义原子集。如图 1 所示,例如:在现有 HowNet 知识库中,“小米”的 Def 描述包含的义原集合为{material | 材料,edible | 食物,crop | 庄稼},但在目前实际语言环境中,“小米”除了具有一种农作物的概念外,还可以描述为一个公司名或电子产品的品牌。因此,其义原集合还可以包含{InstitutePlace | 场所}或{SpeBrand | 特定牌子}。

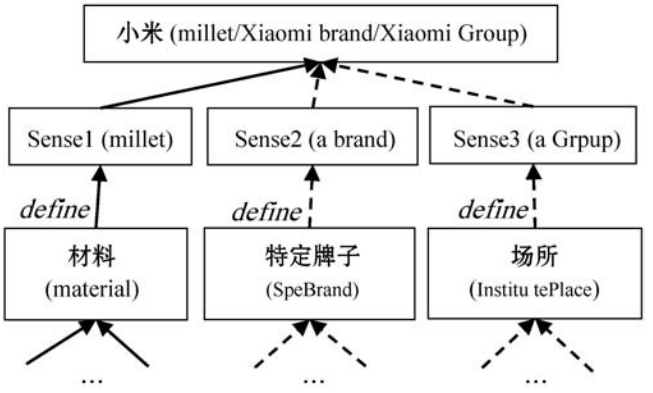


图 1　新概念发现及义原推荐示例

Xie 等[23]基于词嵌入(Word Embeddings)和义原嵌入(Sememe Embeddings)提出了多种义原预测模型,借助协同过滤和矩阵分解的方法,从已有的人工标注数据集学习词汇与义原的通用的关系,从而自动构造出新词的义原。Jin 等[21]提出了基于词语内部字信息和外部上下文信息的义原预测框架,通过将内部模型和外部模型融合,提升了低频词义原预测的效果。Li 等[25]提出了基于字和多标记分布序列到序列(Label Distributed Seq2seq Model)模型,利用词的定义和描述信息进行义原预测。张磊等[22]基于多标签分类模型架构,通过将句子中的词作为模型输入,减小了用字作为最小单位的歧义性。杜家驹等[24]提出了义原相关池化模型,利用局部语义相关性来预测义原,该方法依赖于定义文本的获取质量。上述方法均以字或词的分布式表示为基础,忽略了基于词向量与基于 HowNet 义原信息的词语相似关系不同构问题。此外,部分方法实现义原推荐的过程中依赖词语的定义句,这对低资源的情况提出了挑战。

通常,语义相似的词语或概念之间会共享相同的义原。因此,解决新概念义原选择问题,可以借鉴协同过滤(Collaborative Filtering,CF)的思想,利用 HowNet 已有概念的 Def 描述中的义原集合来预测当前新概念的 Def 描述最可能使用的义原集合,其关键在于度量新旧词语或概念之间的语义相似性。

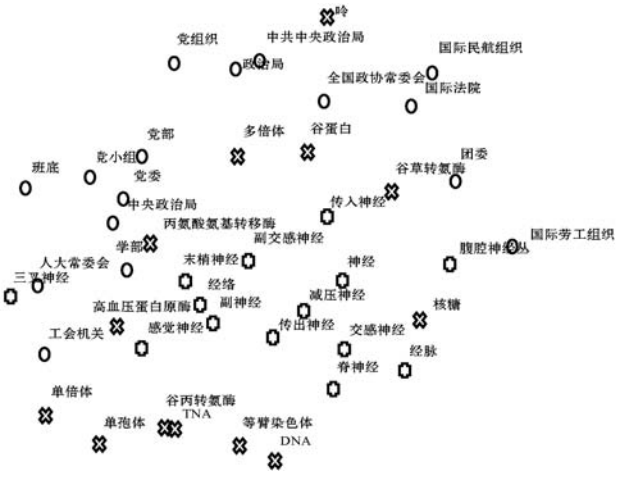


图 2　基于词向量与基于 HowNet 的词语相似关系比较
(HowNet 中语义相近的词语使用相同符号表示)

近年来,基于词的分布式表示方法已成为计算词语相似度的主流方法,但是,词语的表示学习过程与 HowNet 对词语或概念的描述的形成过程存在着本质区别。如图 2 所示,对随机抽取的词语(本例中,选取的 3 个中心词为神经束膜、团委、原生质)进行对比,结果可见,基于词的分布式表示方法得到的相似度计算结果与基于义原的相似度度量方法之间往往存在差异,表现为词语之间语义距离的度量不一致,这里称为“相似性异构”问题。为了更好地为新概念选择合适的义原,需要建模一种新的相似度计算方法,使其计算得到的度量结果能够逼近基于知网义原的相似度计算结果,即达到“相似性同构”。因此,本文提出一种基于网络嵌入和预训练模型的新概念义原预测方法。通过对《知网》中的字-词语-义项-义原及其关系的表示学习,融合预训练语言模型,实现词语与候选义原的动态匹配。主要贡献包括:(1) 提出了一种词语分布式表示与义原预测联合学习方法;(2) 解决了低资源情况下义原预测的定义句依赖问题,基于与训练语言模型,捕获更丰富的语义信息,有助于实现新概念义原推荐的多样化;(3) 以字为基础的词语表示学习可将本文义原预测方法推广到多粒度义原标注场景。

1　相关概念与模型

1.1　HowNet 的义原标注体系

《知网》(HowNet)是一个以汉语和英语的词语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[10],它通过事件、万物、属性、属性值、部件、空间和时间等 7 个维度进行世界的描述,如表 1 所示。HowNet 的构建秉承还原论思想,即所有词语的含义可以由最基本的、不再分割的最小语义单位——“义原”构成,

而《知网》期望利用有限的义原描述无限的概念集合。

表 1 知网哲学中构成世界的 7 个维度

概念类型	概念示例	DEF 描述
事件	打鼾	{MakeSound | 发声:cause = {ill | 病态}, time = {sleep | 睡}}
	赖床	{arise | 起身:TimeAfter = {sleep | 睡}, time = {late | 迟}}
万物	床	{furniture | 家具:{sleep | 睡:location = {~}}}
	睡意	{aspiration | 意愿:CoEvent = {expect | 期望:content = {sleep | 睡}}}
属性	睡相	{Posture | 姿势:host = {human | 人}, scope = {sleep | 睡}}
属性值	香	{BehaviorValue | 举止值:scope = {joyful | 喜悦:scope = {sleep | 睡}}}
部件	床架	{part | 部件:whole = {furniture | 家具:{sleep | 睡:location = {~}}}}
空间	铺位	{location | 位置:{sleep | 睡:location = {~}}}
时间	就寝时间	{time | 时间:{sleep | 睡:time = {~}}}
	睡眠时间	{time | 时间:{sleep | 睡:duration = {~}}}

如图 3 所示,在 HowNet 中,义原被划分为 Entity | 实体、Event | 事件、Attribute | 属性、Value | 值、Secondary-Feature | 第二特征等 5 个大类。

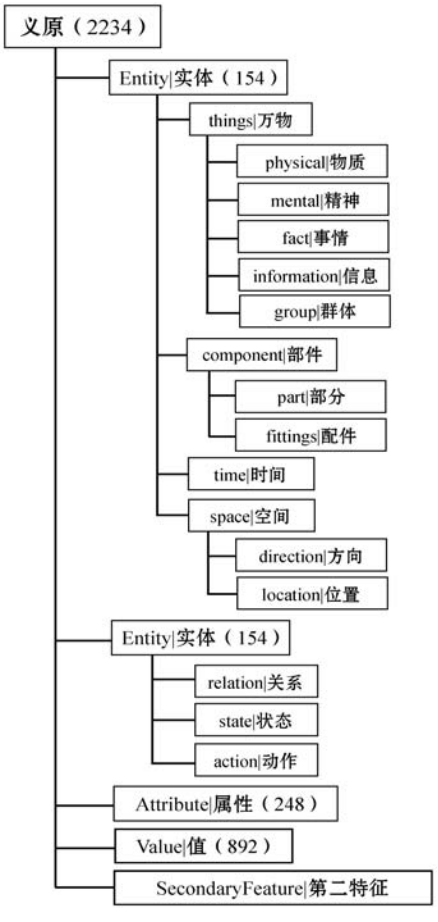


图 3 HowNet 中义原的分类树(数量)

1.2 “词语-义项-义原”关系网络

在 HowNet 中,由义原向上构建概念,由概念向上定义词,词语、义项(概念,Def)、义原的关系如图 4 所示。这里以词语“小米”为例,由“material | 材料”、“edible | 食物”和“crop | 庄稼”等义原及其关系构成了的概念义项的描述 Def = {material | 材料:MaterialOf = {edible | 食物}, material = {crop | 庄稼}}(senseID:177381),再由这个概念义项定义了词语“小米(millet)”。

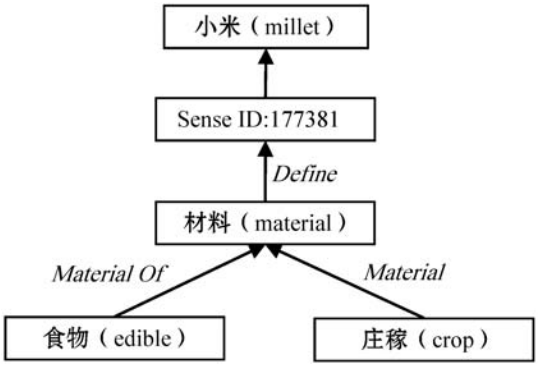


图 4 词语、义项(概念)、义原的关系示例

由于存在一词多义现象,因此,在 HowNet 中每个词语对应一个或多个概念义项的定义(即,概念描述 Def),每个 Def 包含一个或多个义原。任一概念描述中的义原之间通过体现概念与概念和概念的属性与属性之间的相互关系连接,通常一个义原可能存在于多个概念义项的定义当中。在对多个词语的语义关系进行建模时,就形成了“词语-义项-义原”关系网络。

为了发现新词语的不同义项,将“词语-义项-义原”关系网络中的词语节点删除,形成“义项-义原”关系网络。“义项-义原”关系网络的每个极大联通子网络潜在表示了词语的不同义项。为体现同义词语的义原共享性,可以建立仅由义项与构成该义项 Def 的全部义原之间的连接关系构成的“义项-义原”残差关系网络。

1.3 网络嵌入模型

网络嵌入(Network Embedding)旨在学习网络中节点的低维度潜在表示,其基本思想是找到一种映射函数 $f: V \rightarrow \mathbf{R}^d$,该函数将网络中的每个节点转换为低维度的潜在表示 $f(x) \in \mathbf{R}^d$,其中 $x \in V$ 是网络中的一个节点。目前,同质网络嵌入方法主要有 DeepWalk[11]、Node2vec[12]、LINE[13] 和 SDNE[14] 等。其中,DeepWalk 和 Node2vec 都是基于网络上的随机游走并使用 Skip-gram 模型进行节点嵌入。本文在网络表示学习层采用 Node2vec 方法对字及义原节点进行向量化表示。

1.4 预训练语言模型

随着深度学习的发展,卷积神经网络(CNNs)、递归神经网络(RNNs)、图神经网络(GNNs)和注意机制等神经网络模型被广泛应用于解决自然语言处理

(NLP)任务。相比于非神经网络模型方法严重依赖于离散的手工特征,神经网络方法通常使用低维稠密向量隐式表示语言的语法或语义特征[15]。近年来,大量的研究表明,使用大规模文本语料库进行训练得到的预训练模型(PTMs)可以学习近似通用语言表示,在对特定任务的小数据集微调后,可在显著降低单个自然语言处理任务的难度的同时提升系统性能。

在词语相似度计算方面,由于分布式表示方法可以通过将单词表示为低维稠密实数向量,捕捉词语间的关联信息。因此该方法可在低维空间中高效计算单词间的语义关联,有效解决数据稀疏问题[16]。相关研究表明[17],以 Word2vec 为代表的词语表示学习模型已经在词语相似度计算任务中取得了较好的效果。然而,Word2vec 中每一个词语被映射到一个唯一的稠密向量,它无法处理一词多义问题。此外,现有的表示学习模型根据词语的上下文分布来学习词语的表示向量,对于出现次数较少或未登录的词语,将无法学习出一个好的表示。上述问题成为影响词语分布式表示方法在相似度计算任务上发挥作用的主要障碍。

相比以 Word2vec 为代表的分布式表示方法,BERT 的一个比较突出的优势就是词语表示的动态性,能建模一词多义的现象。在新概念的义原推荐任务中,对于未登录词(OOV)的分布式表示需求普遍存在,为了减缓 OOV 的影响,一种方法可以通过扩大词典,以提升模型训练过程中词语的覆盖度,但该方法不能从根本上解决 OOV 问题。另一种方法,可采用基于字的 BERT 模型,利用 BERT 编码器最上层的字的隐层向量得到当前词语的向量。然而,简单地利用[CLS]的输出作为句向量的方法被证明效果并不理想。在 HowNet 中,概念相似性的度量关键是计算概念对应的义原序列之间的相似度。

Sentence-Transformers[18]模型基于 Transformer 模型(使用 BERT/XLNet 进行句子嵌入)并针对语义相似性进行了微调,因此在序列语义相似度(如,句子相似度)计算方面表现出了良好的性能。本文将待预测义原的 Token(词或句)看作字的序列,其中,每个字的向量表示由其所在 Token 的 Sentence-Transformers 编码和网络嵌入表示拼接而成。同时,Sentence-Transformers 被用来在《知网》中选择与输入 Token 相似的词语,形成候选词语集合。

2　义原预测模型

2.1　义原协同推荐框架

我们期望利用"词语-义项-义原"关系网络的结构特征和基于协同过滤机制为词语推荐相关义原。

协同过滤机制是推荐系统所采用的最为重要的技术之一[19]。其基本原理是假设两个用户如果具有相类似的购买行为,则他们对同一类商品感兴趣的程度也就会比较接近,那么当前用户很有可能会对另一个相似用户所喜欢的商品感兴趣。在语义相似性度量方面,结合 HowNet 中对概念描述方式的规定,可以认为语义相似的概念应具有相似的义原关联集合。因此,义原推荐的任务中,采用协同过滤机制是一种可行的途径。基本原理是根据相似的概念义项所包含的义原来推荐当前词语尚未被标记,但在实际语言环境中很可能表达其某方面语义信息的义原。

本文提出的基于协同推荐框架的义原预测模型如图 5 所示。该模型包括表示学习模块、候选义原选择模块和义原序列排序模块。在表示学习模块,通过网络表示学习方法获得 HowNet 中的字和义原的向量表示,通过 Sentence-Transformers 模型得到输入 Token 的表示向量,并将其拼接到 Token 包含的字的向量表示上。拼接后的向量表示经过 BiLstm 和线性层得到输入 Token 的新的向量表示,该向量表示被用来与候选义原序列向量进行相似性判别,进而实现义原序列排序。

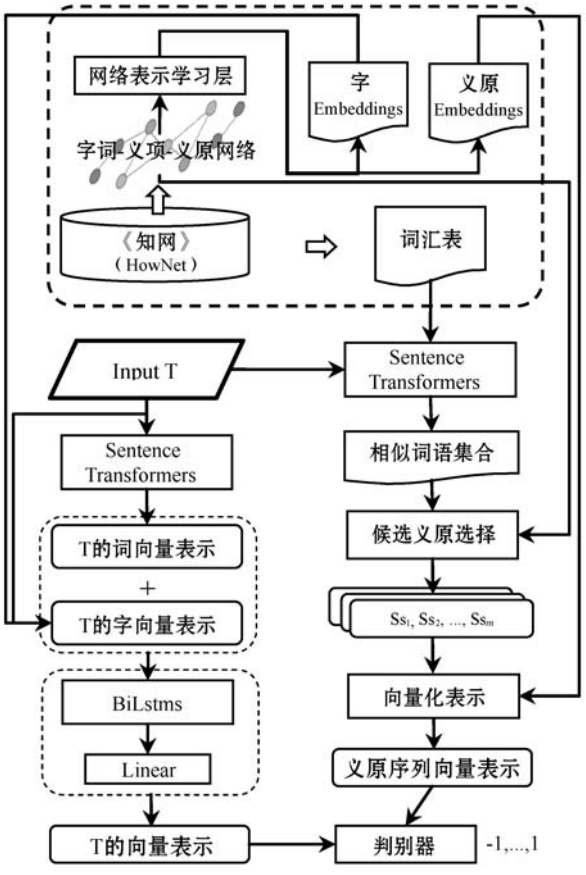


图 5　基于协同推荐框架的义原预测模型

2.2 候选义原选择

新概念义原推荐任务的目标是为 HowNet 中未登录的概念推荐合适的义原,因此,需要找到一种义原未知概念与义原已知概念之间的相似度计算模型。由于概念以词的形式表现,上述问题可转换为未登录词语与词表词的相似度计算问题。本文采用 Sentence-Transformers 在《知网》中选择与未登录词语相似的词表词作为候选义原。

利用相似词语对应的概念获得待推荐义原的集合。首先,通过在 Hownet 中进行相似词检索来获得查询词对应的相似词集合;其次,基于上述全部词语、词语对应的概念义项及其义原,构建“词语-义项-义原”关系子网络并基于网络节点重要性排序方法进行候选义原节点的选择。

在网络中,中心性(Centrality)表示了边和点的重要度。这里使用两种中心性的度量方法评估义原节点的重要度。

度中心性(Degree Centrality)是在网络分析中刻画节点中心性(Centrality)的最直接度量指标。一个节点的节点度越大就意味着这个节点的度中心性越高,该节点在网络中就越重要。标准化度中心性测量公式:

$$C_{\mathrm{d}}(v_i) = \sum_j x_{ij} / \max(C_{\mathrm{d}}(v_j)) \quad i \neq j \in \mathbf{N}$$

式中:$x_{ij}=1$ 表示节点 i 与节点 j 之间存在直接联系,否则,$x_{ij}=0$;N 为网络中全部节点的集合。由于节点的度(Degree)的计算过程没有考虑图中邻接节点的重要性,不能很好地体现词语之间对义原的共享特征,因此,需要从路径这个维度来度量节点的中心性,这里引入基于介性中心度(Betweenness Centrality)的中心性度量方法。

计算网络中任意两个节点的所有最短路径,如果这些最短路径中有很多条都经过了某个节点,那么就认为这个节点的介性中心度高。介性中心度测算公式为:

$$C_{\mathrm{b}}(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

式中:σ_{st}表示从节点 s 到 t 的最短路径数;$\sigma_{st}(v_i)$ 表示从节点 s 到 t 的,且经过 v_i 的最短路径数。

节点的介性中心度是以经过某个节点的最短路径数目来刻画节点重要性的指标。一个义原节点的介性中心度较高,说明其被相似词语共享的可能性越大。

结合义原节点的标准化度中心性和介性中心度,计算义原节点的推荐指数:

$$R(v_i) = C_{\mathrm{d}}(v_i) \times \log(C_{\mathrm{b}}(v_i) + 1)$$

2.3 义原序列排序

对候选义原进行组合,构成不同长度的义原序列,该序列将作为新概念义原标注的候选结果。义原序列得分:

$$R(\boldsymbol{t},\boldsymbol{s}) = 1 - f_{\mathrm{d}}(\boldsymbol{t},\boldsymbol{s})$$

$$f_{\mathrm{d}}(\boldsymbol{t},\boldsymbol{s}) = \left\| \boldsymbol{t} - \sum_{\boldsymbol{s}_i \in \boldsymbol{s}} \boldsymbol{s}_i \right\|_2^2$$

式中:$\boldsymbol{t}$ 为输入 Token 的向量表示,由 BiLSTM 模型训练得到;$\boldsymbol{s}$ 为义原序列的向量表示,由序列中包括的义原对应的向量相加得到。

3 实验与结果分析

3.1 实验设置

3.1.1 数据集

本文进行义原预测实验使用的数据集从 HowNet 知识库中抽取产生。选择《知网》中长度大于 1 的登录词构成数据集,共计 106 384 个词。按 8∶1∶1划分为相似词选择模型的训练数据集、验证集和测试集。

3.1.2 对比模型

选择目前义原推荐效果较好的两阶段协同过滤方法作为义原推荐实验的比较模型,其中第一阶段选择的词语相似计算模型如表 2 所示。

表 2 词语相似计算模型

序号	模型	说明
1	Word2vec	基于 Word2vec,输入为词对
2	BERT	基于 BERT,输入为词对
3	BERT^2 + CNN	基于孪生 BERT + CNN,输入为词对

候选义原选择阶段,采用的对比模型是 4 种(M1 - M4)基本模型及其组合模型,如表 3 所示。

表 3 候选义原选择方法

模型		编号	说明
CSNet	Glo	M1	“义项-义原”网络全局中心性推荐指标
	Loc	M2	“义项-义原”网络局部中心性推荐指标
ResCSNet	Glo	M3	“义项-义原”残差网络全局中心性推荐指标
	Loc	M4	“义项-义原”残差网络局部中心性推荐指标

在模型训练阶段,本文模型采用的主要超参数的设置如表 4 所示。

表 4　模型超参数设置

序号	参数名	值
1	优化器(optimizer)	Adam
2	学习速率(learning rate)	0.1
3	损失函数(loss func.)	MSELoss
4	激活函数(activation function)	ReLU
5	训练轮数(epochs)	30
6	batch_size	100

3.2　实验结果分析

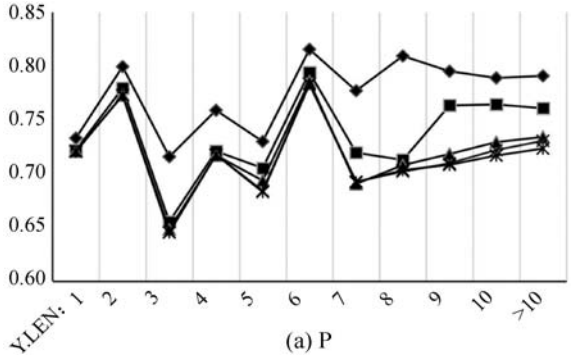
本文评估方法采用 F1 值来作为评价指标。表 5 展示了各模型的义原预测 F1 值。可知,结合知网网络嵌入和与训练语言模型信息的模型(SemRec)在实现词语分布式表示与义原预测联合学习的情况下,取得了最优的结果,F1 值达到 0.623 7。

表 5　不同模型组合 F1 值

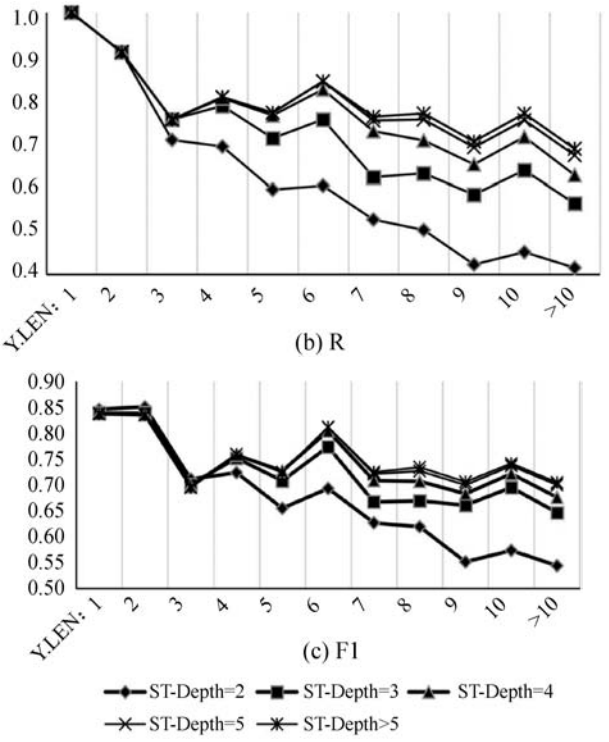
模型	最优义原选择模型	F1
MCA[20]	—	0.454 5
CSP[21]	—	0.563 3
Word2vec	M2 + M4	0.442 4
BERT	M1 + M2 + M4	0.250 4
$BERT^2$ + CNN	M2	0.268 3
SemRec	—	**0.623 7**

为考察候选义原集合构建环节对系统性能的影响,将 Sentence Transformers 替换成《知网》相似度计算模型[4]模拟在知网相似度完全拟合情况下的模型性能。随机选取约相同数量的由不同义原长度表示的词语构成观察数据,对不同深度的义原树和义原数量情况下的模型表现进行分析。

从图 6 的结果看,在构建候选义原集合时,遍历义原树的深度越深,越有利于召回相关义原,反之则有利于义原选择的准确率。从总体看,选择较深的遍历义原树深度,在义原数量较多的情况(义原数量 >3)能表现出更好的性能。在"义项-义原"网络中选择义项邻居和邻居的邻居节点作为候选义原则在义原数量较少的情况表现出更好的性能。



(a) P



(b) R

(c) F1

ST-Depth=2　ST-Depth=3　ST-Depth=4
ST-Depth=5　ST-Depth>5

图 6　选义原集合构建环节截取不同深度的义原树在各义原数量情况下对模型性能的影响

4　结　语

本文提出了一种基于网络嵌入和预训练模型的新概念义原预测方法,通过对《知网》中的词语-义项-义原及其关系的表示学习,融合预训练语言模型,动态构建"义项-义原"关系网络,实现词语与候选义原的动态匹配,通过义原标注提升了表示学习的可解释性。在未来的研究中,可针对不同粒度的语言单位尝试义原预测,并根据预测结果实现可解释的语义相似度计算。

知网相似度完全拟合情况下的模型性能分析表明,与《知网》相似度同构的计算模型有助于提升义原预测的性能,因此下一步工作可以尝试学习《知网》词语语义相似度训练学习候选义原集合选择模型,在整体上提高对义原预测的质量。

参　考　文　献

[1] Dong Z D, Dong Q. HowNet and the computation of meaning [M]. London: World Scientific Publishing, 2006.

[2] Duan X Y, Zhao J, Xu B. Word sense disambiguation through sememe labeling[C]//20th International Joint Conference on Artificial Intelligence, 2007: 1594 – 1599.

[3] 刘鹏远,赵铁军. 利用语义词典 Web 挖掘语言模型的无指导译文消歧[J]. 软件学报, 2009, 20(5): 1292 – 1300.

[4] 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.
[5] 李峰,李芳. 中文词语语义相似度计算—基于《知网》2000[J]. 中文信息学报,2007(3):99 - 105.
[6] 江敏,肖诗斌,王弘蔚,等. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报,2008(5):84 - 89.
[7] 程玉胜,梁辉,王一宾,等. 基于风险决策的文本语义分类算法[J]. 计算机应用,2016,36(11):2963 - 2968.
[8] Niu Y L, Xie R B, Liu Z Y, et al. Improved word representation learning with sememes[C]//55th Annual Meeting of the Association for Computational Linguistics,2017:2049 - 2058.
[9] Duan X Y, Zhao J, Xu B. Word sense disambiguation through sememe labeling[C]//20th International Joint Conference on Artificial Intelligence,2007:1594 - 1599.
[10] 董振东,董强,郝长伶. 知网的理论发现[J]. 中文信息学报,2007(4):3 - 9.
[11] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations[C]//20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2014:701 - 710.
[12] Grover A, Leskovec J. Node2vec: Scalable feature learning for networks[C]//22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2016:855 - 864.
[13] Tang J, Qu M, Wang M Z, et al. LINE: Large-scale information network embedding[C]//24th International Conference on World Wide Web,2015:1067 - 1077.
[14] Wang D X, Cui P, Zhu W. Structural deep network embedding[C]//22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2016:1225 - 1234.
[15] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]//26th International Conference on Neural Information Processing Systems,2013:2787 - 2795.
[16] 孙飞,郭嘉丰,兰艳艳,等. 分布式单词表示综述[J]. 计算机学报,2019,42(7):1605 - 1625.
[17] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[EB]. arXiv:1301.3781,2013.
[18] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks[C]//9th International Joint Conference on Natural Language Processing,2019:3982 - 3992.
[19] Su X Y, Khoshgoftaar T M. A survey of collaborative filtering techniques[J]. Advances in Artificial Intelligence, 2009,4(12):2.
[20] Fu X H, Liu G, Guo Y, et al. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon[J]. Knowledge-Based Systems, 2013,37:186 - 195.
[21] Jin H M, Zhu H, Liu Z Y, et al. Incorporating Chinese characters of words for lexical sememe prediction[C]//56th Annual Meeting of the Association for Computational Linguistics,2018:2439 - 2449.
[22] 张磊,李响,刘媛媛. 基于深度学习和词典定义的义原预测方法[J]. 信息工程大学学报,2019,99(5):96 - 101.
[23] Xie R B, Yuan X C, Liu Z Y, et al. Lexical sememe prediction via word embeddings and matrix factorization[C]//26th International Joint Conference on Artificial Intelligence, 2017:4200 - 4206.
[24] 杜家驹,岂凡超,孙茂松,等. 基于局部语义相关性的定义文本义原预测[J]. 中文信息学报,2020,34(5):1 - 9.
[25] Li W, Ren X C, Dai D M, et al. Sememe prediction: Learning semantic knowledge from unstructured textual wiki descriptions[EB]. arXiv:1808.05437,2018.

(上接第 12 页)

[21] Aafer Y, Du W L, Yin H. DroidAPIMiner: Mining API-level features for robust malware detection in android[C]//International Conference on Security and Privacy in Communication Systems,2013:86 - 103.
[22] Agrawal R, Stokes J W, Marinescu M, et al. Neural sequential malware detection with parameters[C]//IEEE International Conference on Acoustics, Speech and Signal Processing,2018:2656 - 2660.
[23] Arzt S, Rasthofer S, Fritz C, et al. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps[J]. ACM SIGPLAN Notices,2014, 49(6):259 - 269.
[24] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [EB]. arXiv:1810.04805v1,2018.
[25] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[EB]. arXiv:1301.3781,2013.
[26] Google. UI/Application exerciser monkey[EB/OL]. [2021 - 04 - 20]. https://developer.android.com/studio/test/monkey.
[27] 中国反网络病毒联盟-移动互联网恶意程序描述格式[EB/OL]. [2021 - 04 - 20]. https://white.anva.org.cn/rel/file/ydwj.pdf.
[28] 2020 上半年度中国手机安全状况报告[EB/OL]. [2021 - 04 - 20]. https://pdf.dfcfw.com/pdf/H3_AP202009181414354503_1.pdf.