关慧, 刘启华. 一种基于词嵌入和多重语义关系的词语相似度计算方法[J]. 智能计算机与应用,2026,16(4):180-186.
DOI:10.20169/j.issn.2095-2163.24060501

一种基于词嵌入和多重语义关系的词语相似度计算方法

关 慧[1,2], 刘启华[1]
(1 沈阳化工大学 计算机科学与技术学院, 沈阳 110142; 2 辽宁省化工过程工业智能化技术重点实验室, 沈阳 110142)

摘 要: 已有基于 WordNet 的词语相似度计算方法中,语义关系的研究多集中于"is-a"关系,导致其结果与人类判断偏差较大。而词嵌入的方法虽然计算效果好,但难以区分复杂的语义。因此,本文提出一种基于词嵌入和多重语义关系的模型,该模型引入词嵌入方法和概念间的多重语义,在 WordNet 中提取包括反义关系在内的 14 种语义关系,利用词嵌入将这些语义关系编码成向量,然后将反义关系单独作为一个影响因子,与其他 13 种语义特征共同构成的影响因子进行线性组合,从而利用反义关系的相异性对过高的相似度进行修正。实验结果表明,该模型在 WS203(WordSim-203)、SimLex666 数据集上的相关度比现有的基于 WordNet 和词嵌入的方法分别提高约 1.2%和 28%,有效地提高了词语相似度计算的准确性。

关键词: 词语相似度; WordNet; 反义关系; 词嵌入
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2026)04-0180-07

A method for calculating word similarity based on word embeddings and multiple semantic relations

GUAN Hui[1,2], LIU Qihua[1]
(1 School of Computer Science and Technology, Shenyang University of Chemical Technology, Shenyang 110142, China;
2 Liaoning Key Laboratory of Industrial Intelligence Technology on Chemical Process, Shenyang 110142, China)

Abstract: In the existing Wordnet-based methods of word similarity calculation, the semantic relationship is mostly focused on the "IS-A" relationship, which results in a large deviation from human judgment. Although the word embedding method has good computational effect, it is difficult to distinguish complex semantics. Therefore, this paper proposes a model based on word embeddings and multiple semantic relations. The model introduces multiple semantics between word embedment methods and concepts, extracts 14 semantic relations including antisense relations from WordNet, encodes these semantic relations into vectors using word embedment, and then takes antisense relations as a single influence factor. The similarity of the influence factors composed with 13 other semantic features is linearly combined, and the antisense relation is used to correct the excessive similarity. The experimental results show that the correlation degree of the proposed model on WS203 and SimLex666 data sets is about 1.2% and 28% higher than that of the existing WordNet and word embedding methods, which effectively improves the accuracy of word similarity calculation.

Key words: word similarity; WordNet; antisense relation; word embedding

0 引 言

在自然语言处理领域中,词语相似度计算一直是一个重要的研究方向。其可应用于机器翻译、检索系统、文本分类等多个领域[1]。目前基于 WordNet 计算方法主要有 3 大类:路径法[2-4]、信息内容法[5-7]、特征法[8-10],后续又发展出基于分布式来表示的词嵌入方法[11-13]。

路径法简单直接,其利用语义网络中的最短路径长度来测量词语间的相似度。Rada 等[2]、关慧等[3]、Cai 等[4]都在这方面进行了研究。信息内容法的主要思想是如果两个概念共享更多的信息量,那么这两个概念就会更加相似或相关。Resnik 等[5]、Batet 等[6]、Almousa 等[7]均提出了信息内容

基金项目: 辽宁省教育厅 2021 年度科学研究经费项目(LJKZ0434)。
作者简介: 刘启华(2000—),女,硕士研究生,主要研究方向:自然语言处理,语义处理,特征选择。
通信作者: 关 慧(1976—),女,博士,副教授,主要研究方向:软件演化,软件安全性,语义处理等。Email:575581567@qq.com。
收稿日期: 2024-06-05

法来计算词语相似度。特征法利用了概念之间的共同特征来计算相似度，避免了概念间边缘长度的一致性。Tversky 等[8]、Ezzikouri 等[9]、Wasti 等[10]曾在这方面进行研究。词嵌入是一种将词汇表中的词语映射为固定长度向量的技术。词嵌入模型虽然在计算相似度的效果上超过了 WordNet 方法，但其在区分一词多义、语义关系等方面不敏感。WordNet 方法由于其特有的结构，可以区分词语的语义关系。因此，将 WordNet 与词嵌入相结合是近年来研究的热点。赵福强等[1]、Lee 等[11]、Zhao 等[12]、陈丹华等[13]在结合 WordNet 与词嵌入的模型方面做了大量的研究。尽管基于 WordNet 的方法简单高效，但其研究大多集中于“is-a”关系，导致其结果与人类判断偏差加大。虽然有许多研究表明，词嵌入方法取得了巨大的成功，但这种方法对不同词产生差别向量，因此词嵌入方法不能表示一个词的不同含义。

为了解决上述存在的问题，本文将词嵌入与 WordNet 知识库有效的结合，从概念语义关系的角度考虑，引入概念间的多重语义关系，并提出一种基于多重语义关系和词嵌入的词语相似度计算模型。通过 WordNet 的语义网络结构，提取出词语对应概念的同义关系、上位关系、实例上位关系、下位关系、实例下位关系、成员整体关系、成员部件关系、物质整体关系、部分整体关系、物质部件关系、部分部件关系、定义、例子、反义关系等 14 种语义关系，其中将反义关系单独作为一种影响因子，其余 13 种语义关系共同作为另一个影响因子，然后利用词嵌入的方法结合两种影响因子，从而提高相似度计算结果的准确性。

1　基于词嵌入和多重语义关系的词语相似度计算模型

在本节中，针对现有的计算词语相似度方法存在高度依赖人工标注的知识库以及基于词嵌入模型难以区分复杂语义等不足，本文将词嵌入引入到基于 WordNet 的方法中，使用词嵌入对在 WordNet 中提取出的多重语义关系进行预处理。此外，针对基于 WordNet 的词语相似度与人类判断结果相差较大的情况，本文引入反义关系来进行修正，并提出了一种基于词嵌入和多重语义关系的词语相似度计算模型。图 1 展示了基于词嵌入和多重语义关系的词语相似度计算模型的流程图。

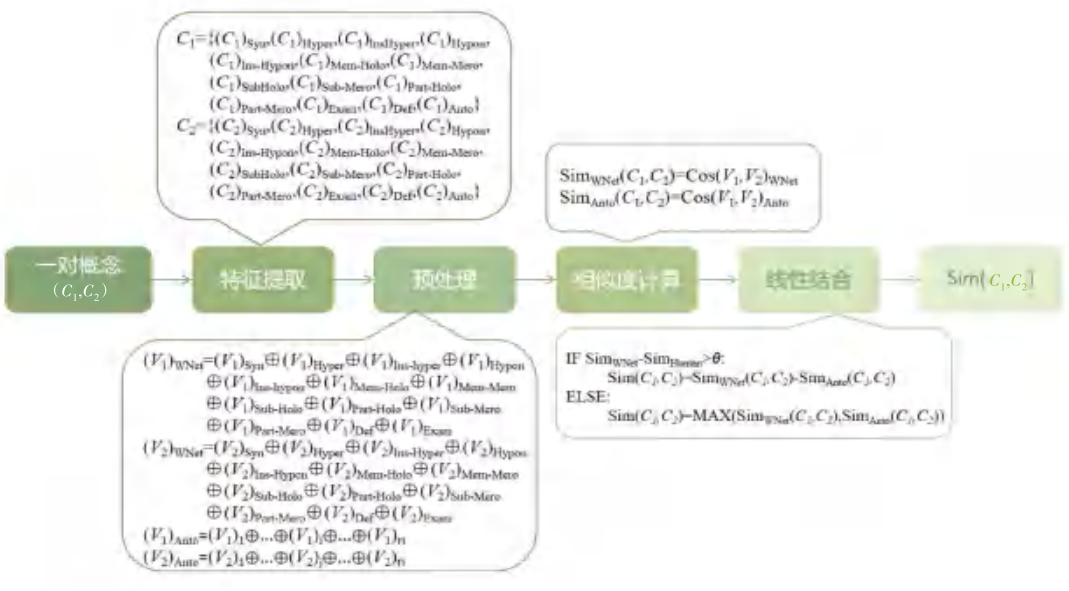

图 1　基于词嵌入和多重语义关系的词语相似度计算模型的流程图

Fig. 1　Flow chart of word similarity calculation model based on word embeddedness and mupltiple semantic relations

1.1　数据的预处理

本文的数据是从 WordNet 中提取的语义关系，需要将语义关系转化成模型可以识别和处理的信息，本文使用词嵌入的方法对数据进行预处理。

在已有的研究中，利用 WordNet 中的语义关系计算词语相似度的文章较少，并且在这些研究中用到的语义关系较少或仅使用信息内容的方法计算词语相似度。例如，Batet[6]的方法利用概念的上位词、下位词、同义词和多义词来改进基于信息内容的相似度计算；Hussain[14]等提出了一种向量空间方法，在该方法中，将 WordNet 中 gloss，lemmas，example，sister-term，derivations，holonyms，meronyms，hypernyms，hyponyms

等 9 种语义特征与维基百科的 page gloss,hyperlink 的特征相结合,并使用信息内容的方法来计算词语的相似度。Hussain 的方法虽然将 WordNet 与维基百科中的特征相结合,扩充了语义关系,但是基于信息内容的方法对语义关系的表示能力不强,无法有效提高词语的相似度。Almousa[7] 提出了两种方法:一种是基于语义关系的信息内容法,另一种是基于非分类关系的路径加权重法来计算词语相似度,该方法中使用了所有类型的非分类关系并将其分为三大类:部分关系,成员关系和物质关系,并在不同粒度级别上部署非分类关系。虽然 Almousa 的方法利用到了 antonymy 关系,但是在该方法中,antonymy 关系仅属于三大类中的一个分类,这弱化了 antonymy 关系对词语相似度的影响。

为解决上述存在的问题,本文首先找出 WordNet 中关于词语的 13 种语义关系,例如:同义关系(Synonymy)、上位关系(Hypernymy)、实例上位关系(Instance Hypernymy)、下位关系(Hyponymy)、实例下位关系(Instance Hyponymy)、成员整体关系(Member Holonymy)、成员部件关系(Member Meronymy)、物质整体关系(Substance Holonymy)、部分整体关系(Part Holonymy)、物质部件关系(Substance Meronymy)、部分部件关系(Part Meronymy)、例子(Example)和定义(Definition),并将这些语义关系共同作为影响因子,然后引入词嵌入的方法对在 WordNet 中提取出的多重语义关系进行预处理。步骤如下:

首先,在 WordNet 中,找出概念 C 的上述除反义关系的 13 种语义特征,将其分别组成集合 C_{Syn}、C_{Hyper}、$C_{\text{Ins-Hyper}}$、C_{Hypon}、$C_{\text{Ins-Hypon}}$、$C_{\text{Mem-Holo}}$、$C_{\text{Mem-Mero}}$、$C_{\text{Sub-Holo}}$、$C_{\text{Sub-Mero}}$、$C_{\text{Part-Holo}}$、$C_{\text{Part-Mero}}$、C_{Exam} 和 C_{Def},以同义关系特征集合 C_{Syn} 为例,C_{Syn} 表示为:

$$C_{\text{syn}} = (c_1, c_2, \cdots, c_n) \tag{1}$$

其中,$c_1, c_2, \cdots, c_n$ 表示概念 C 的同义关系。

然后,使用词嵌入获取集合中概念对应的向量,这些向量的集合分别称为 $\boldsymbol{V}_{\text{Syn}}$、$\boldsymbol{V}_{\text{Hyper}}$、$\boldsymbol{V}_{\text{Ins-Hyper}}$、$\boldsymbol{V}_{\text{Hypon}}$、$\boldsymbol{V}_{\text{Ins-Hypon}}$、$\boldsymbol{V}_{\text{Mem-Holo}}$、$\boldsymbol{V}_{\text{Mem-Mero}}$、$\boldsymbol{V}_{\text{Sub-Holo}}$、$\boldsymbol{V}_{\text{Sub-Mero}}$、$\boldsymbol{V}_{\text{Part-Holo}}$、$\boldsymbol{V}_{\text{Part-Mero}}$、$\boldsymbol{V}_{\text{Exam}}$ 和 $\boldsymbol{V}_{\text{Def}}$,以同义关系特征向量集合 $\boldsymbol{V}_{\text{Syn}}$ 为例,$\boldsymbol{V}_{\text{Syn}}$ 表示为:

$$\boldsymbol{V}_{\text{Syn}} = (\boldsymbol{V}_{c_1}, \boldsymbol{V}_{c_2}, \cdots, \boldsymbol{V}_{c_n}) \tag{2}$$

其中,$\boldsymbol{V}$ 表示特征 $c_1, c_2, \cdots, c_n$ 的向量集合。

接下来,将概念 C 上述的 13 种语义特征的向量集合合并为 $\boldsymbol{V}_{\text{WNet}}$,表示为:

$$\begin{aligned}\boldsymbol{V}_{\text{WNet}} = {} & \boldsymbol{V}_{\text{Syn}} \oplus \boldsymbol{V}_{\text{Hyper}} \oplus \boldsymbol{V}_{\text{Ins-Hyper}} \oplus \boldsymbol{V}_{\text{Hypon}} \oplus \\ & \boldsymbol{V}_{\text{Ins-Hypon}} \oplus \boldsymbol{V}_{\text{Mem-Holo}} \oplus \boldsymbol{V}_{\text{Mem-Mero}} \oplus \\ & \boldsymbol{V}_{\text{sub-Holo}} \oplus \boldsymbol{V}_{\text{Sub-Mero}} \oplus \boldsymbol{V}_{\text{Part-Holo}} \oplus \\ & \boldsymbol{V}_{\text{Part-Mero}} \oplus \boldsymbol{V}_{\text{Exam}} \oplus \boldsymbol{V}_{\text{Def}}\end{aligned} \tag{3}$$

其中,操作符⊕表示两个向量的连接操作。

最后,按 300 维度取概念 C 的特征向量集合 $\boldsymbol{V}_{\text{WNet}}$ 的平均向量 $\overline{\boldsymbol{V}_{\text{WNet}}}$,定义如下:

$$\overline{\boldsymbol{V}_{\text{WNet}}} = \frac{1}{R}\sum_{n=1}^{R}\overline{\boldsymbol{V}_{\text{WNet}}}[n] \tag{4}$$

其中,R 为平均向量 $\overline{\boldsymbol{V}_{\text{WNet}}}$ 的长度。

1.2 基于反义关系的模型

在大多数情况下,人们计算词语相似度时,主要利用路径和信息量的方法来计算相似度,然而这些方法计算的相似度结果与人类判断有较大偏差。随着相似度的计算过程逐渐模拟人类思维,对反义关系的研究变得尤为重要。

已有的研究表明[3],反义关系会对相似度的结果产生一定程度的负面影响,从而修正相似度的偏差。然而,在利用反义关系计算词语相似度时,现有的研究主要集中在基于路径和信息内容的方法,而忽略了词嵌入的方法以及反义关系作为语义关系的影响,这可能导致在计算相似度时缺少语义信息,使结果与人类判断偏差较大。因此,本文基于之前的工作,利用 WordNet 的固有结构,找出概念的多重语义特征,其中包括反义关系。将反义关系作为一个单独的影响因子,对过高的词语相似度计算结果进行修正。除反义关系外,还考虑了其他 13 种语义关系,将其共同作为一个影响因子,并提出了一种基于反义关系的词语相似度计算模型,以更好地发挥反义关系在语义计算中的作用。

为了研究概念的反义关系,首先需要查询反义关系的情况。本文选择 NLTK 自然语言处理工具处理 WordNet 中各种函数。例如,想要查询概念 C 的反义关系,首先要查询概念 C 的词条 lemma,再通过函数 antonymy( ) 查询 lemma 的反义关系,anto(C) 表示反义关系的集合,公式如下:

$$\text{anto}(C) = \{\text{antonymys}(\text{lemma}) \mid \text{lemma} \in C\} \tag{5}$$

其中,$a_i \in \text{anto}(C)$,函数 synsets( ) 用于查询对应词条所在的概念,设 anti(C) 代表概念 C 的反义关系集合。则公式如下:

$$\text{anti}(C) = \{\text{synsets}(a_i) \mid a_i \in \text{anto}(C)\} \tag{6}$$

在引入反义关系到相似度计算中时,考虑到概念直接反义关系较少的情况,将概念节点到公共祖先结点路径上节点的反义关系纳入考量是一个较优

的做法。假设概念 C 与其 LCA 间的最长路径表示为 path(C,LCA)，将路径上结点的反义关系集合起来，称此集合为最近公共祖先路径，定义如下：

$$\mathrm{NLAP}(C)=\{\mathrm{anti}(n)\mid n\in \mathrm{path}(C,\mathrm{LCA})\} \quad (7)$$

在 NLAP 里，路径 path(C,LCA) 中的概念可以当成是概念 C 的扩展，扩展出的结点对应的反义关系词集称为 NLAP(C)。设 $\boldsymbol{V}(C)$ 代表概念 C 的反义词集对应的向量集合，其公式如下：

$$\boldsymbol{V}(C)=\{v(n)\mid n\in \mathrm{NLAP}(C)\} \quad (8)$$

接下来，按 300 维度取向量集合 $\boldsymbol{V}(C)$ 的平均向量，得到平均向量$\overline{\boldsymbol{V}(C)}$定义如下：

$$\overline{\boldsymbol{V}(C)}=\frac{1}{R}\sum_{i=1}^{R}\boldsymbol{V}(C)[i] \quad (9)$$

在这里本文将概念 C 的反义关系集合 NLAP(C) 表示为 C_{Anto}，平均向量$\overline{\boldsymbol{V}(C)}$表示为$\overline{\boldsymbol{V}_{\mathrm{Anto}}}$，以便于与概念 C 的其他 13 种语义关系的集合 $\boldsymbol{V}_{\mathrm{WNet}}$ 和平均向量$\overline{\boldsymbol{V}_{\mathrm{WNet}}}$区分。在将上述两种影响因子分别转变成平均向量后，本文使用余弦相似度的距离计算方法分别计算概念对的两种影响因子的相似度。余弦相似度可表示为：

$$\mathrm{Sim}(C_i,C_j)=\mathrm{Cos}(\bar{V}_i,\bar{V}_j) \quad (10)$$

$$\mathrm{Cos}(\bar{V}_i,\bar{V}_j)=\frac{\bar{V}_i\cdot\bar{V}_j}{\|\bar{V}_i\|\times\|\bar{V}_j\|}=\frac{\sum_{i,j=1}^{n}\bar{V}_i\cdot\bar{V}_j}{\sqrt{\sum_{i=1}^{n}(\bar{V}_i)^2}\sqrt{\sum_{j=1}^{n}(\bar{V}_j)^2}} \quad (11)$$

其中，C_i,C_j 表示为一对概念；$\bar{V}_i$,$\bar{V}_j$ 表示概念 C_i,C_j 的平均向量；函数 $\mathrm{Cos}(V_i,V_j)_{\mathrm{WNet}}$ 和 $\mathrm{Cos}(V_i,V_j)_{\mathrm{Anto}}$ 分别表示计算两种影响因子的余弦相似度。

在基于 WordNet 的词语相似度计算中，将计算得到的实验结果与人类评分标准在相同环境下进行比较。发现在数据集中，大约有一部分词语的相似度高于人类评分标准，而另一部分则低于人类评分标准。举例来说，在 Miller and Charles 数据集中，与人类评分标准相比，有 17 对词语相似度的值高于人类评分标准，另外有 13 对词语相似度的值低于人类评分标准。因此，本文提出的基于反义关系的词语相似度模型主要针对高于人类评分标准的概念对进行修正，具体如图 2 所示。

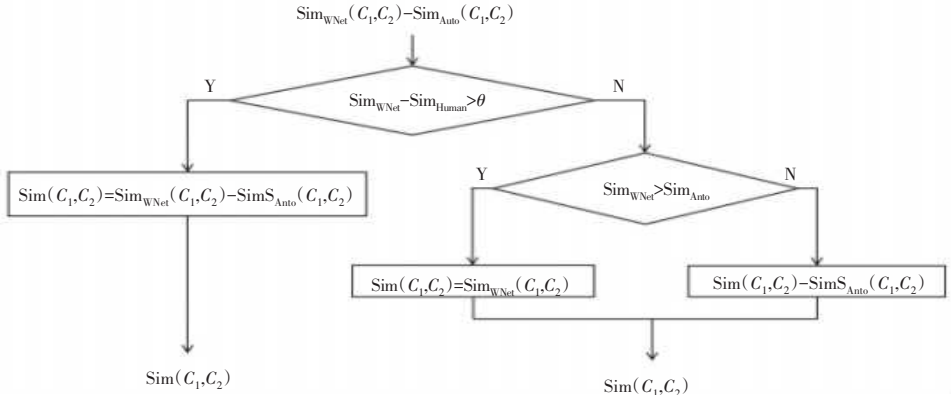

图 2　基于多重语义关系的词语相似度计算模型

Fig. 2　Word similarity calculation model based on multiple semantic relations

对于高于人类评分标准的概念对，本文将其两种影响因子的相似度进行线性结合；而对于低于人类评分标准的概念对，则是选取两种影响因子相似度的最大值作为其最终的相似度结果。因此，本文提出一种基于反义关系的模型，将词语相似度计算方法定义为线性组合，公式如下式：

$$\mathrm{value}=\mathrm{Sim}_{\mathrm{WNet}}(C_1,C_2)-\mathrm{Sim}_{\mathrm{Human}}(C_1,C_2) \quad (12)$$

其中，$\mathrm{Sim}_{\mathrm{WNet}}$、$\mathrm{Sim}_{\mathrm{Anto}}$、$\mathrm{Sim}_{\mathrm{Human}}$ 分别表示为概念 C_1、C_2 的两种影响因子的相似度值及人类评分标准。

2　实验部分

2.1　数据集

在此研究中，本文统一使用 WordNet3.0 版本计算相似度，并选择 4 个标准数据集来评估所提出的模型，分别为 Miller and Charles(MC30)[15]、Rubenstein and Goodenough (RG65)[16]、Agirre (WS203)[17] 和 Hill(SimLex)[18] 数据集。MC30 是从 65 对 RG65 数据集中提取的 30 对英语单词组成的数据集，WS203 数据集包含 203 对单词，包括名词和动词。SimLex 是最大的词语相似度数据集，包括 666 对名词、222 对动

词和 111 对形容词。本文在实验中只选择 SimLex 的名词对。

2.2 评价指标

本文使用 Pearson 相关系数 r 表示测量结果与人类评分标准的相似程度,计算公式如下:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times (y_i - \bar{y})^2}} \tag{13}$$

其中,x_i 表示人类判断集合 x 中第 i 个元素;y_i 表示计算结果集合 y 中的第 i 个元素;$\bar{x}$ 和 $\bar{y}$ 分别代表集合 x 和 y 的平均值;n 是集合 x 或者 y 的元素的总数。

2.3 实验结果与讨论

为了科学评估所提出的模型在不同权重参数下对相似度的贡献,本文使用了 3 种词嵌入模型:分别是基于 GoogleNews 语料库训练的 word2vec 模型、在 Common Crawl 语料库上训练的 GloVe 模型和在 Common Crawl 语料库上使用 CBOW 算法训练的 FastText 模型,3 种模型的向量维度都是 300 维。

2.3.1 多重语义关系的聚合消融实验

本文中,结合概念的多重语义特征可以提高概念的向量表示。然而,词语相似度模型的性能主要由两个步骤决定:(1)语义特征选择标准;(2)融合不同语义特征的方法。因此,本文进行了详细的消融实验,以显示各种语义特征组合对词语相似度的影响。最后,基于消融实验的结果,本文整合了最适合的语义特征,建立了一个更准确的相似性模型。

表 1 展示了不同语义关系对相似度的影响。组合一考虑了 WordNet 的语义特征,包括同义、上位、实例上位、下位、实例下位、成员整体、成员部件、物质整体、物质部件、部分整体、部分部件、例子和定义的语义特征。组合二、三、四分别除同义关系、上位关系和和实例上位关系、下位关系和实例下位关系外,结合了组合一中其他的语义特征。观察表 1 发现,相比于组合一,组合二、三、四分别除同义关系、上位关系和实例上位关系、下位关系和实例下位关系外,结合了组合一中其他的语义特征。观察表 1 发现,相比于组合一,组合二、三、四都获得了较低的 Pearson 相关值,这表明忽略同义关系、上位关系和实例上位关系、下位关系和实例下位关系会对词语相似度产生负面影响。组合五和组合六分别除物质整体关系和物质部件关系、成员整体关系和成员部件关系外,结合了组合一中其他的语义特征。观察表 1 可知,这两个组合所有 Pearson 相关值与组合一完全相同,说明物质整体关系和物质部件关系、成员整体关系和成员部件关系对词语的相似度没有明显的影响。组合六和组合七分别除部分整体关系和部分部件关系、例子、定义外,结合了组合一中其他的语义特征。由表 1 可知,相较于组合一,组合七和组合八中有些数据集的 Pearson 相关值高于组合一,造成这种现象的原因可能是在大型数据集上概念的部分整体关系和部分部件关系、例子和定义的信息较多,会产生冗余,会对词语相似度产生消极影响。

表 1 不同的语义关系在 MC30、RG65 和 WS203 数据集上的 Pearson 相关系数结果

Table 1 Pearson correlation coefficient results of different semantic relationships on MC30, RG65 and WS203 datasets

模型	Word2Vec				GloVe				FastText			
	MC30	RG65	WS203	SimLex	MC30	RG65	WS203	SimLex	MC30	RG65	WS203	SimLex
组合一	0.89	0.88	0.75	0.54	0.87	0.84	0.60	0.48	0.88	0.88	0.71	0.54
组合二	0.86	0.83	0.69	0.48	0.84	0.81	0.48	0.41	0.85	0.82	0.61	0.46
组合三	0.88	0.86	0.74	0.53	0.86	0.84	0.59	0.47	0.87	0.86	0.70	0.53
组合四	0.85	0.85	0.73	0.48	0.83	0.78	0.58	0.44	0.83	0.83	0.68	0.49
组合五	0.89	0.88	0.75	0.54	0.87	0.84	0.60	0.48	0.88	0.88	0.71	0.54
组合六	0.89	0.88	0.75	0.54	0.87	0.84	0.60	0.48	0.88	0.88	0.71	0.54
组合七	0.89	0.88	0.74	0.56	0.87	0.84	0.61	0.48	0.88	0.88	0.71	0.56
组合八	0.88	0.87	0.74	0.55	0.85	0.83	0.67	0.54	0.88	0.90	0.75	0.58

上述详细的消融实验考察了本文在 WordNet 中提取的每个语义关系的意义,并对各种语义关系在词语相似度上的贡献有了更好的理解。消融实验表明,忽略一些语义关系,如定义、例子等,可以提高 Pearson 的相关值。然而,没有一个单一的语义关系被消除后,可以提高所有基准的词语相似度。反而,组合一结合了所有的语义关系,实现了一致的结果,并在所有基准测试中更加稳定,因此,在本文所提出的模型中,将 WordNet 的所有语义关系与反义关系结合起来计算相似度,公式如下:

$$\mathrm{Sim}(C_1, C_2)_{\mathrm{WNet-Anto}} = \begin{cases} \mathrm{Sim}_{\mathrm{WNet}}(C_1, C_2) - \mathrm{Sim}_{\mathrm{Anto}}(C_1, C_2), & \text{if value} > \theta \\ \max(\mathrm{Sim}_{\mathrm{WNet}}(C_1, C_2), \mathrm{Sim}_{\mathrm{Anto}}(C_1, C_2)), & \text{else} \end{cases} \tag{14}$$

其中,θ 的取值范围为[0,1],$\mathrm{Sim}_{\mathrm{WNet}}$、$\mathrm{Sim}_{\mathrm{Anto}}$、

Sim_{Human}分别表示为概念 C_1、C_2 的两种影响因子的相似度值及人类评分标准。

2.3.2　对参数 θ 的讨论

本文在基于反义关系模型的实验中，对于参数 θ 的取值做了进一步的讨论。其中，参数 θ 取值范围为[0~1]，然后使用上述 3 种词嵌入模型分别对 MC30、RG65、WS203、SimLex666 这 4 个数据集进行相似度的计算，如图 3~6 所示。从图 3~6 中可以看出，θ 的取值在[0.2,0.3]之间时，相似度的准确率达到最高点。在 MC30 数据集中，word2vec 模型、GloVe 模型和 FastText 模型均在 θ 取值为 0.2 时，准确率最高；在 RG65 数据集中，word2vec 模型、GloVe 模型在 θ 取值为 0.2 时，准确率最高，而 FastText 模型则是在 θ 取值为 0.3 时，准确率达到高峰；在 WS203 数据集中，word2vec 模型和 FastText 模型在 θ 取值为 0.2 时，表现最好，GloVe 模型则是在 θ 取值为 0.3 时，表现最好；在 SimLex666 数据集中，word2vec 模型当 θ 取值为 0.2 时，表现出色，而 GloVe 模型和 FastText 模型在 θ 取值为 0.3 时，表现最为出色。所以在本文后续的实验中，参数 θ 的取值为 0.2。

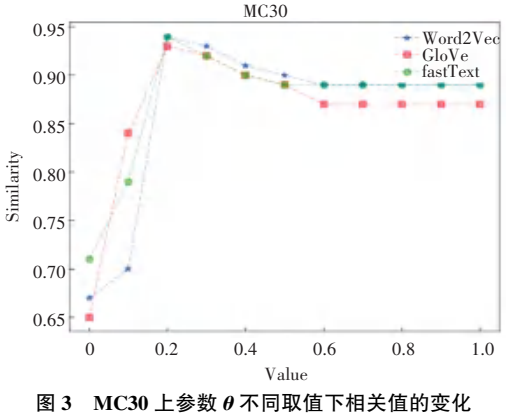

图 3　MC30 上参数 θ 不同取值下相关值的变化

Fig. 3　Changes of correlatiion values with different values of parameters θ on MC30

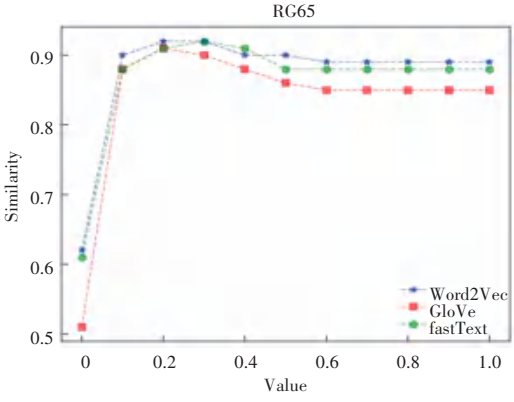

图 4　RG65 上参数 θ 不同取值下相关值的变化

Fig. 4　Changes of correlatiion values with different values of parameters θ on RG65

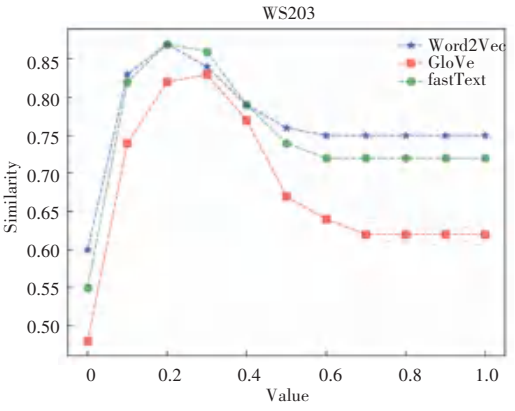

图 5　WS203 上参数 θ 不同取值下相关值的变化

Fig. 5　Changes of correlatiion values with different values of parameters θ on WS203

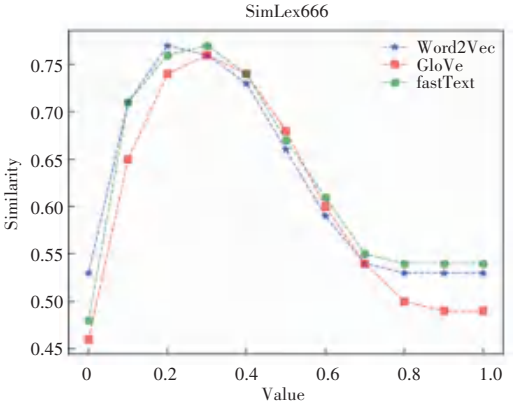

图 6　SimLex 上参数 θ 不同取值下相关值的变化

Fig. 6　Changes of correlatiion values with different values of Parameters θ on SimLex

2.3.3　与其他方法的比较

本节将所提出的模型的结果与目前已有的词语相似度算法进行比较，以评估模型的有效性。表 2 展示了本文模型和其他方法的 Pearson 相关值。观察表 2 可以发现，概念的多重语义关系与词嵌入结合的模型得到的词语相似度计算结果比传统方法的 Pearson 相关值更高。在本文的实验中，仅考虑单一的语义或仅使用词嵌入方法得到的 Pearson 相关值较低，这表明在计算词语相似度时，组合多重语义和词嵌入方法对相似度结果有积极的作用。实验结果证实了以下观点：

(1)在词语相似度计算中考虑更多有效的语义特征，可以使得相似度结果会更接近于人类判断。

(2)在考虑相同的语义特征的前提下，结合其他方法，如词嵌入，可以得到更接近人类判断的相似度结果。

由此可知，在计算相似度时考虑更多的语义特征并有效结合多种方法可以显著提高模型的相似度计算结果。在文本提出的模型中，反义关系作为一

个单独的影响因子引入到相似度计算过程中,用于对计算结果进行修正。

表 2 本文结果与现有方法结果对比

Table 2 Comparison of experimental results and existing methods

算法	MC30	RG65	WS203	SimLex
Rada[2]	0.75	0.80	0.64	0.55
Guan[3]	0.89	0.86	0.77	0.59
Cai[4]	0.92	0.89	0.81	
Resnik[5]	0.81	0.84	0.74	0.53
Batet[6]	0.84	0.84	0.67	
Wasti[10]	0.84	0.85	0.67	0.59
Lee[11]	0.88	0.83	0.80	0.48
Zhao[12]	0.94	0.92	0.86	0.60
Hussain[14]	0.89	0.89	0.80	0.62
WordNet(Word2Vec)	0.89	0.88	0.75	0.55
WordNet(GloVe)	0.87	0.84	0.60	0.48
WordNet(FastText)	0.88	0.88	0.71	0.54
Sim_{Anto}	0.74	0.78	0.76	0.53
Sim_{WNet}	0.89	0.88	0.75	0.55
$Sim_{WNet-Anto}$	0.93	0.92	0.87	0.77

3 结束语

为了解决现有的 WordNet[19-20] 词语相似度计算方法存在的问题,本文提出了一种基于词嵌入和多重语义关系的词语相似度计算模型,通过引入多重语义关系和词嵌入在一定程度上补充概念的语义信息。与现有的基于 WordNet 的方法相比,该模型提取出 WordNet 中多重语义关系,在计算相似度的过程中引入词嵌入的方法对提取的语义关系预处理,使得相似度结果更加准确。在计算概念对的相似度时,使用反义关系对概念之间的过高的相似度计算结果进行修正,使反义关系在语义计算中发挥更好的作用。实验结果表明,本文的方法在数据集 MC30、RG65、WS203、SimLex666 上具有较高的相关性,分别为 0.93,0.92,0.87 和 0.77。

参考文献

[1] 赵福强. 基于词嵌入和 WordNet 的词汇相似度计算模型[D]. 重庆:重庆大学,2022.

[2] RADA R, MILI H, BICKNELL E, et al. Development and application of a metric on semantic nets[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1989, 19(1): 17-30.

[3] 关慧,马天宇,王广伟. 相异性在语义相似度计算中的应用[J]. 沈阳化工大学学报,2022,36(2):167-179.

[4] CAI Y, PAN S, WANG X, et al. Measuring distance-based semantic similarity using meronymy and hyponymy relations[J]. Neural Computing and Applications, 2020, 32: 3521-3534.

[5] RESNIK P. Using information content to evaluate semantic similarity in a taxonomy[C]//Proceedings of the 14th International Joint Conference on Artificial Intelligence. IJCAI, 1995: 448-453.

[6] BATET M, SÁNCHEZ D. Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic information content[J]. Artificial Intelligence Review, 2020, 53(3): 2023-2041.

[7] ALMOUSA M, BENLAMRI R, KHOURY R. Exploiting non-taxonomic relations for measuring semantic similarity and relatedness in WordNet[J]. Knowledge-Based Systems, 2021, 212: 106565.

[8] TVERSKY A. Features of similarity[J]. Psychological Review, 1977, 84(4): 327.

[9] EZZIKOURI H, MADANI Y, ERRITALI M, et al. A new approach for calculating semantic similarity between words using WordNet and set theory[J]. Procedia Computer Science, 2019, 151: 1261-1265.

[10] WASTI S H, HUSSAIN M J, HUANG G, et al. Assessing semantic similarity between concepts: A weighted-feature-based approach [J]. Concurrency and Computation: Practice and Experience, 2020, 32(7): e5594.

[11] LEE Y Y, KE H, YEN T Y, et al. Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement[J]. Journal of the Association for Information Science and Technology, 2020, 71(6): 657-670.

[12] ZHAO F, ZHU Z, HAN P. A novel model for semantic similarity measurement based on wordnet and word embedding[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(5): 9831-9842.

[13] 陈丹华,王艳娜,周子力,等. 基于 Word2Vec 的 WordNet 词语相似度计算研究[J]. 计算机工程与应用, 2022,58(3):222-229.

[14] HUSSAIN M J, BAI H, WASTI S H, et al. Evaluating semantic similarity and relatedness between concepts by combining taxonomic and non-taxonomic semantic features of WordNet and Wikipedia[J]. Information Sciences, 2023, 625: 673-699.

[15] MILLER G A, CHARLES W G. Contextual correlates of semantic similarity[J]. Language and Cognitive Processes, 1991, 6(1): 1-28.

[16] RUBENSTEIN H, GOODENOUGH J B. Contextual correlates of synonymy[J]. Communications of the ACM, 1965, 8(10): 627-633.

[17] AGIRRE E, ALFONSECA E, HALL K, et al. A study on similarity and relatedness using distributional and wordnet-based approaches[C]//Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. ACL, 2009: 19-27.

[18] HILL F, REICHART R, KORHONEN A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation [J]. Computational Linguistics, 2015, 41(4): 665-695.

[19] GUAN H, JIA C, YANG H. Intelligent recognition of semantic relationships based on antonymy [J]. Multiagent and Grid Systems, 2020, 16(3): 263-290.

[20] GUAN H, MA T, WANG G. An improved dissimilarity based approach to semantic similarity calculation[C]//Proceedings of the 9th International Conference on Dependable Systems and Their Applications (DSA). Piscataway, NJ: IEEE, 2022: 1023-1028.