DOI：10.11830/ISSN.1000-5013.202003035

多模型以动态权重相融合的词相似性分析

王仲昊[1]，万相奎[1,2,3]，李风从[1]，危竞[1]，刘俊杰[1]

（1. 湖北工业大学 太阳能高效利用及储能运行控制湖北省重点实验室，湖北 武汉 430068；
2. 湖北工业大学 太阳能高效利用湖北省协同创新中心，湖北 武汉 430068；
3. 湖北工业大学 湖北省电网智能控制与装备工程技术研究中心，湖北 武汉 430068）

摘要： 以 NLPCC-ICCPOL 2016 中文词语相似度比赛中的 PKU-500 数据集作为评价的参考标准，采用动态权重多模型融合的词相似性进行分析. 将得到的斯皮尔曼等级相关系数 0.568 与 NLPCC 2016 第 1 名的结果相比，效果提高了 9.6%. 结果表明：基于动态权重改进的多模型融合方法，提高计算词相似性的准确率.

关键词： 词相似性；统计模型；字典模型；改进的多模型融合；动态权重

中图分类号： TP 391.1 **文献标志码：** A **文章编号：** 1000-5013(2021)01-0121-07

Word Similarity Analysis by Multi-Model Fusion With Dynamic Weight

WANG Zhonghao[1], WAN Xiangkui[1,2,3], LI Fengcong[1], WEI Jing[1], LIU Junjie[1]

(1. Key Laboratory of Solar Energy Efficient Utilization and Energy Storage Operation Control in Hubei Province, Hubei University of Technology, Wuhan 430068, China;
2. Solar Energy Efficient Use of Hubei Province Collaborative Innovation Center, Hubei University of Technology, Wuhan 430068, China;
3. Hubei Province Power Grid Intelligent Control and Equipment Engineering Technology Research Center, Hubei University of Technology, Wuhan 430068, China)

Abstract: Taked the PKU-500 dataset in the NLPCC-ICCPOL 2016 Chinese word similarity competition as the reference standard for evaluation, used dynamic weights multi-model fusion to analyze word similarity, obtained a Spearman rank correlation coefficient of 0.568, which is 9.6% higher than the first place in the NLPCC-ICCPOL 2016. The results show that the improved multi-model fusion method based on dynamic weight improves the accuracy of calculating word similarity.

Keywords: word similarity; statistical model; dictionary model; improved multi-model fusion; dynamic weights

自然语言处理(natural language processing，NLP)在很多领域都有着广泛的应用，而词相似性是 NLP 应用(机器翻译、搜索引擎、人机交互、舆论分析等)的一个基础内容，所以研究提高词相似性的准

收稿日期： 2020-03-29

通信作者： 万相奎(1976-)，教授，博士，主要从事嵌入式系统设计与信号处理的研究. E-mail：wanxiangkui@163.com.

基金项目： 国家自然科学基金资助项目(61571182)

确率是非常重要的.目前,主流的分析词语相似性的模型有基于统计模型的词袋(bag of word,BOW)模型[1]、属于Word2Vec[2]的连续词袋(continuous bag of word,CBOW)模型[3]、跳字(skip-gram)模型[3]及Google公司最新提出的BERT(bidirectional encoder representations from transformers)训练模型[4].当前使用较多的是Word2Vec模型和Glove,BERT等预训练模型,若要求计算效率更高、语料特性变化性强且需在线学习,选择Word2Vec模型;若要求计算结果准确、语料特性稳定且无需在线学习,选择Glove,BERT等预训练模型.基于上、下文的统计模型对语料库的要求特别高,足够数量的语料库才能得到较准确的词相似度.这是由于统计模型的相似度计算并不是判断真正的词的意义,它是对每个词出现在文本中某一位置可能性的判断,这是统计模型的不足.基于人工分类的词典,如《同义词词林扩展版》[5]和《HowNet》[6]更能体现词之间意义的关联性且更贴近人的使用习惯;但是当词汇数量不足、覆盖率不够、更新率不高、颗粒度比较粗糙时,只能按照相同或不相同两种定义区分词的相似度.

采用多方法融合可以有效弥补各单一方法的不足之处.在NLPCC 2016词相似度比赛中,Guo等[7]采用类似的多方法融合策略,将得到的相似度进行一定的加权计算,以PKU-500[8]数据集为评价的参考标准,得到了0.518的斯皮尔曼系数,位列第一.实际上,这一方法所使用的融合策略还有提升空间.本文提出一种改进的融合方法,改变词典与统计模型相融合后结果的权重分配,并按语料类型分别训练不同的统计模型,有效提高计算词相似性的准确性[9].

1 多方法融合的词相似性计算

1.1 统计模型的选择

统计模型实际反映的是在上、下文中某个词出现的可能性,因此,一个词的意义是通过对其上、下文的建模而计算得到的.skip-gram模型主要根据某一词预测上、下文[10],与文中的主要需求不符,所以不采用.BOW模型忽略文本的语序和语法等要素,仅仅将其看作若干单词的集合,文本中每个单词的出现都是独立的,因此,该模型的计算结果准确度有限,也不采用.CBOW模型针对BOW模型的缺点进行改进,将文本的语序作为一个要素加入计算中,大幅提高计算的准确性和训练速度[2].CBOW模型将词向量化,使统计模型可以运用深度学习进行计算,进一步提高计算效率[1].

BERT是一种预训练模型,相对于Word2Vec模型,将全部上、下文纳入模型计算中,进一步提高计算的准确性.但这些预训练模型的结构十分庞大,且训练过程需要大量的数据和设备资源[11],所以在词相似性分析的过程中,必须针对语料数据的特点选取合适的模型.如稳定性强的词义规范性语料可采用BERT训练模型,因为预训练模型只需对数据集进行计算后就可以得到准确度很高的结果.但对于稳定性差的词义,非规范语料比较适合采用Word2Vec模型.因为这些语料主要来源于互联网,其特点是更新不断且变化快.若采用BERT训练模型,可能无法及时完成对非规范语料的处理.文中的数据集主要来源于互联网中的语料数据,对计算效率和更新速度有一定的要求.因此,采用Word2Vec[12]模型中的CBOW模型对语料库进行计算,以提升词相似度的准确性.

1.2 词典模型的选择

1.2.1 《HowNet》词典模型 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,义原有1 500个;每一个词语一般有一个或多个概念[13],例如"北京"一词的《HowNet》的结构描述,如图1所示.

NO.=006271
W_C=北京
G_C=N[bei3 jing1]
S_C=
E_C=
W_E=Beijing
G_E=N
S_E=
E_E=
DEF={place|地方: PlacSect={capital|国都,
belong="China|中国",modifier{ProperName|专}}

图1 《HowNet》中词的结构

Fig. 1 Structure of words in *HowNet*

由图1可知:义原是以多层结构体系分布[14].通过计算义原之间的相对距离,可以得到词语间的相似度[15],并将相似度作为多模型融合方法中的一部分.

1.2.2 《同义词词林扩展版》词典模型 《同义词词林》是由梅家驹等[5]编写的,随后又由哈尔滨工业大学信息检索实验室进一步扩充、更新,并发布了《同义词词林扩展版》.这一词典模型中包含了77 000多条词汇,这些词汇以树状结构进行组合.整个词典中的词汇分为12组大类、97组中类、1 400余组小类、

1 400 余组词群及 17 000 余组原子词群[16]. 树状结构的叶节点由一个个原子词群组成，原子词群由一个或几个词组成，在同一末端的词都是语义相同或有很强相关性的词. 词林中："="表示该原子词群中的词相等或同义；"@"代表独立，表示词林中该词语没有同义的或者相关的词；"#"表示该原子词群的词相关但不是同义的.《同义词词林扩展版》的组成形式，如图 2 所示.

通过《同义词词林扩展版》的这种结构模式，可以反映词的相似性，所以《同义词词林扩展版》在文中作为词相似性评价的一个部分.

A105A06# 排头兵 三好生

A105B01= 先锋 急先锋 开路先锋 先行官 先遣 先行者

A105B02# 旗手 红旗手 持旗人 持旗者 突击手 弄潮儿

A105B03@ 创新者

A105C01= 功臣 元勋

A105C02@ 有功者

A105D01@ 偶像

A106A01= 爱好者 发烧友

图 2　《同义词词林扩展版》的组成形式

Fig. 2　Forms of *Synonym cilin extended edition*

1.3　多模型融合的方法

将多个模型进行融合，在一定条件下，各模型相互弥补缺点以提高计算词相似性的准确性. 在 NLPCC-ICCPOL 2016 的比赛中，Zou 等[17]取得第 2 名，采用《HowNet》单一词典模型，与 PKU-500 数据集进行相关性分析后，得到 0.457 的斯皮尔曼系数. Guo 等[7]获得第 1 名，采用简单多模型融合的方法，获得了 0.518 的斯皮尔曼系数，相比于第 2 名，效果提高了 13.3%. 由此可见，多模型的融合可以有效提高词相似性的准确率. 文中根据文本不同的特点采取不同的权重，从而有效利用不同模型的优势，以到达更好的效果. 进一步扩展数据集，将数据集分为规范性语料数据集、非规范性语料数据集、通用数据集[8].

1.4　整体系统的结构

对 3 个不同语料库计算的统计模型和两个词典模型采用动态权重，系统整体结构图，如图 3 所示.

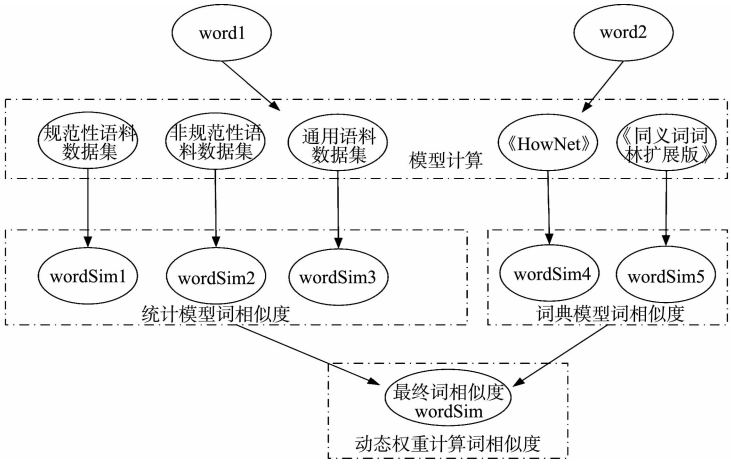

图 3　系统整体结构图

Fig. 3　System overall structure diagram

图 3 中：将第 1 层两个需要计算相似度的词 word1，word2 输入 5 个模型中，分别对第 2 层 5 个模型进行计算，得出两个词在该模型下的相似度. 分别根据 3 个统计模型和两个词典模型相似度的标准差，动态确定第 3 层的每个模型权重，最终可以得到两个词的相似度.

在系统整体结构第 2 层中，统计模型中两个词语相似度是通过两次词语所对应的词向量之间的余弦夹角定义的，即余弦相似性，计算公式为

$$\mathrm{sim}_1(W_1,W_2)=\cos\theta=\frac{\boldsymbol{v}_1\times\boldsymbol{v}_2}{|\boldsymbol{v}_1|\times|\boldsymbol{v}_2|}. \tag{1}$$

式(1)中：W_1，W_2 分别为 word1，word2；$\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 为 W_1 和 W_2 的词向量；$\cos\theta$ 取值范围为[0,1]，当两个词为同义词或完全相同的词时，两词向量的夹角为 0，词相似性为 1，而当两个词为完全不相同的词时，两词向量的夹角为 90°，词相似性接近于 0.

对于《HowNet》，刘群等[15]提出一种量化计算的方式. 所有的义原按照上、下文结构组成了一个层次体系，词语间的远近距离反映了词语间语义的相似程度，其数学关系为

$$\begin{cases}\lim\limits_{d\to\infty}\mathrm{sim}(W_1,W_2)=0,\\ \lim\limits_{d\to 0}\mathrm{sim}(W_1,W_2)=1.\end{cases}$$

上式中:词间的距离越短,则词义越相似;词间的距离越长,则词义越不相似.通过计算语义距离就可以量化词相似性,两个义原之间的语义距离定义为

$$\mathrm{sim}_2(W_1,W_2)=\frac{\alpha}{d(W_1,W_2)+\alpha}. \tag{2}$$

式(2)中:$d(W_1,W_2)$表示 W_1 和 W_2 在义原层次体系中的最短路径的长度;α 表示一个可调节的参数.

由式(2)可知:两义原的相似度的取值范围为$[0,1]$,当两个词为同义词或完全相同词时,$d(W_1,W_2)$为 0,词相似性为 1;当两个词为完全不相同词时,$d(W_1,W_2)$为一个很大的值,词相似性接近于 0.

由于《同义词词林》采用 5 层的树状结构,所以给每一层赋予一个权重,然后,根据两个词所在的最低层次计算两词的相似性.一般两词的相似性就是该层的权重,所有的权重在$[0,1]$取值,两词 W_1 和 W_2 的相似性定义为

$$\mathrm{sim}_3(W_1,W_2)=\begin{cases}a, & \text{在同第 5 级分支下且标志位为“=”时},\\ b, & \text{在同第 4 级分支下},\\ c, & \text{在同第 5 级分支下且标志位为“\#”时},\\ d, & \text{在同第 3 级分支下},\\ e, & \text{在同第 2 级分支下},\\ f, & \text{在同第 1 级分支下},\\ g, & \text{标志位为“@”时}.\end{cases} \tag{3}$$

将第 2 层计算的结果分别按照统计模型和词典模型进行计算,由于整体系统采用动态权重,而标准差可以反映一个数据集的离散程度,所以根据不同统计模型或词典模型间的标准差分配权重.若不同模型间的离散程度大,说明部分模型的结果误差较大,需要给单一模型情况下准确率高的模型分配更大的权重,从而减小误差;若不同模型间的离散程度小,说明计算结果大体一致,给各模型均匀分配权重即可.通过标准差的引入,可以进一步提高整体计算效果.

对于一个集合 S,定义指示函数 $l_S(\cdot)$为

$$l_S(x)=\begin{cases}1, & x\in S,\\ 0, & \text{其他}.\end{cases}$$

令 σ_S^2 为各统计模型计算结果的方差,将方差的分布范围划分为若干个区域,即

$$\sigma_S^2\in[b_0,b_1)\cup[b_1,b_2)\cup\cdots\cup[b_{i-1},b_i).$$

为每个区间分配一个对应的权重矢量 $\boldsymbol{h}_{S,i}$,即

$$[b_{i-1},b_i)\to\boldsymbol{h}_{S,i}.$$

由于提出的方法使用了 3 个统计模型,因此,权重矢量包含 3 个元素,即 $\boldsymbol{h}_{S,i}\in\mathbf{R}^3$.权重矢量为

$$\widetilde{\boldsymbol{h}_S}(\sigma_S^2)=\sum_i 1_{[b_{i-1},b_i)}(\sigma_S^2)\boldsymbol{h}_{S,i}.$$

同理,两个字典模型所使用的权重计算公式为

$$\widetilde{\boldsymbol{h}_D}(\sigma_D^2)=\sum_j 1_{[b_{j-1},b_j)}(\sigma_D^2)\boldsymbol{h}_{D,j}.$$

上式中:σ_D^2 为两个字典模型计算结果的方差;$[b_{j-1},b_j)$代表方差分布的第 j 个区间;$\boldsymbol{h}_{D,j}\in\mathbf{R}^3$ 代表第 j 个区间对应的权重矢量.

2 计算过程与结果分析

2.1 计算使用的数据集

在统计模型方面,采用的数据集分为 3 个:1) 通过网络爬虫获取 4 GB 的贴吧、微博语料库(4 GB 非规范),将其作为非规范性文本的数据集进行计算;2) 通过网络爬虫获取共 4 GB 的百度百科和维基百科语料库(4 GB 规范),将其作为规范性文本的数据集进行计算;3) 将腾讯 AILab 的开源 NLP 数据

集(Tencent-NLP)作为通用数据集进行计算，分别得到不同语境下的 3 个统计模型. 在词典模型方面，采用《HowNet》和《同义词词林扩展版》作为词典模型.

对中文词相似度的计算结果的评价使用 NLPCC-ICCPOL 2016 中文词相似度评测比赛的 PKU-500 数据集[8]，数据集总共有 500 组词语，每一组的两个词都由人工在 1～10 的范围给定一个分数.

2.2　计算结果的评价标准

采用斯皮尔曼等级相关系数(ρ)评价多模型融合计算词相似性的效果. ρ 越大，说明两组数的相关性越高. 通过算法计算的两词相似度与 PKU-500 数据集中人工给定的相似度计算 ρ，ρ 越高，说明算法计算的结果与人工标定的结果有更高的相关性，也就是说明算法计算的结果更符合人的实际使用场景. 斯皮尔曼等级相关系数的计算公式为

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(R_{X,i} - R_{Y,i})^2}{n(n^2-1)}. \tag{4}$$

2.3　相似度计算对比试验的方案

分别对统计模型、词典模型、简单权重的多模型融合、动态权重的多模型融合进行计算，并得到相应的 ρ，评价各方案性能.

1) 统计模型词相似度性能评价实验. 分别对 4 GB 非规范语料库、4 GB 规范语料库、Tencent-NLP 计算词相似度，并将其与 PKU-500 数据集进行计算.

2) 词典模型词相似度性能评价实验. 分别对《HowNet》和《同义词词林扩展版》计算词相似度，并在 PKU-500 数据集中，计算 ρ.

3) 简单权重的多模型融合的词相似度性能评价实验. 将容量为 4 GB 非规范语料库、4 GB 规范语料库、Tencent-NLP 数据集、《HowNet》和《同义词词林扩展版》计算词相似度，权重设置为固定权重，并在 PKU-500 数据集中，计算 ρ.

4) 动态权重的多模型融合的词相似度性能评价实验. 将统计模型和词典模型中各数据集的权重分开计算，各个模型内部的计算，以及各个模型之间的计算都采用相同的权重公式(根据节 1.4)，设置相应的权重后得到最终的词相似度，并在 PKU-500 数据集中，计算 ρ.

2.4　计算结果的分析

2.4.1　统计模型的词相似度计算　统计模型的词相似度计算中，基于 Word2Vec 和 Glove 训练模型对各数据集进行计算. 其中，Tencent-NLP 数据集已计算得到词向量，所以不用再次计算；4 GB 非规范语料库和 4 GB 规范语料库需要计算后得到结果. Tencent-NLP 数据集是基于 Word2Vec 计算出的数据集，所以在 Glove 训练模型的计算中只使用 4 GB 非规范语料库和 4 GB 规范语料库. 不同数据集对 PKU-500 数据集的词汇覆盖率，如表 1 所示.

表 1　不同数据集对 PKU-500 数据集的词汇覆盖率

Tab. 1　Vocabulary coverage of PKU-500 dataset by different datasets

数据集	未包含词语个数	词语覆盖率/%
4 GB非规范	84	91.6
4 GB规范	16	98.4
Tencent-NLP	5	99.5

由表 1 可知：同样大小的数据集下，规范语料库比非规范语料库的覆盖率更高；Tencent-NLP 数据集由于是通用数据，所以覆盖率最高，但该数据集中只含有中文词汇，对于 PKU-500 数据集中的英文缩写如 WTO，GDP 则无法覆盖.

根据各数据集训练所得的 Word2Vec 模型，分别计算 PKU-500 中词汇组的词相似度，并进一步计算了在 PKU-500 数据集中人工打分的 ρ. 结果如下：4 GB 非规范语料库的 ρ 为 0.384；4 GB 规范语料库的 ρ 为 0.396；Tencent-NLP 数据集的 ρ 为 0.497.

根据各数据集训练所得的 Glove 训练模型，分别计算了 PKU-500 中词汇组的词相似度，并进一步计算了与 PKU-500 中人工打分的斯皮尔曼等级相关系数. 结果如下：4 GB 非规范语料库的 ρ 为 0.427；4 GB 规范语料库的 ρ 为 0.445.

综上可知：数据集越大，最终的计算结果越准确，Glove 训练模型的计算效果优于 Word2Vec. 但由于 Glove 训练模型的窗口范围为全局，计算消耗的时间大幅增加，对计算性能要求较高，并且其为预训

练模型,无法实现 Word2Vec 的在线学习,所以统计模型的选择必须根据语料的特点进行选择.非规范语料库由于其词义稳定性差且更新快,采取 Word2Vec 更合适.词义稳定,基本没有更新的规范语料则采用 Glove 训练模型更合适.规范语料库比非规范语料库拥有更好的效果.这是由于 PKU-500 数据集的人工打分是由专家、学者完成的,词的意义应该更接近于规范语境的使用情况.

2.4.2 词典模型的词相似度计算 在词典模型的词相似度计算中,《同义词词林扩展版》需要根据词在词典中位置与词典的结构,定义其相似度的参数.《同义词词林扩展版》相似度参数,如表 2 所示.

表 2 《同义词词林扩展版》相似度参数

Tab. 2 Similarity parameters of *Synonym cilin extended edition*

参数	数值	参数所使用的条件
a	0.95	在同第 5 级分支下且标志位为"="时
b	0.70	在同第 4 级分支下
c	0.50	在同第 5 级分支下且标志位为"#"时
d	0.40	在同第 3 级分支下
e	0.20	在同第 2 级分支下
f	0.10	在同第 1 级分支下
g	0	标志位为"@"时

《HowNet》词典的词相似性可以根据刘群等[15]的方法计算得到,不用设置参数.

不同词典对 PKU-500 数据集的词汇覆盖率,如表 3 所示.根据两词典模型和两词典模型加权分别计算 PKU-500 数据集中词汇组的词相似度,并计算人工打分的 ρ,结果如下:《HowNet》的 ρ 为 0.373;《同义词词林扩展版》的 ρ 为 0.460;《HowNet》+《同义词词林扩展版》的 ρ 为 0.476.《同义词词林扩展版》在合适的权重下计算词相似度的效果好于《HowNet》.这主要是因为《同义词词林扩展版》的词覆盖率更高、分布结构更为合理.由于两词典无法覆盖的词并不相同,词典模型加权可以使得词覆盖率进一步提高,ρ 进一步增大,计算效果也更好.

表 3 不同词典对 PKU-500 数据集的词汇覆盖率

Tab. 3 Vocabulary coverage of different dictionaries on PKU-500 dataset

数据集	未包含词语	词语覆盖率/%
《HowNet》	170	83.0
《同义词词林扩展版》	86	91.4

2.4.3 简单权重多模型融合的词相似度计算 在简单权重的多模型融合的词相似度计算中,将 4 GB 非规范语料库、4 GB 规范语料库、Tencent-NLP 数据集、《HowNet》和《同义词词林扩展版》5 个模型分两种不同的权重对比.权重组合 1 为 0.2,0.2,0.2,0.2,0.2,这种权重分配方式反映数据的集中趋势,可以一定程度消除极端数据的影响,提高结果的准确率;权重组合 2 为 0.10,0.10,0.30,0.25,0.25,这种权重分配方式给予单一模型下计算效果更好的模型以更高的权重,可以提高结果的准确率.权重组合 1 的 ρ 为 0.503;权重组合 2 的 ρ 为 0.516.因此,当多模型融合时,可以提高词相似度计算的效果,比单一模型的效果好,而且权重不同融合后的效果也不同.权重组合 2 的效果优于权重组合 1,这是由于覆盖率更大,词相似度效果更好的模型所占的权重更大.

2.4.4 动态权重多模型融合的词相似度计算 在动态权重的多模型融合的词相似度计算中,分两步确定权重.

步骤 1 对统计模型的计算结果的方差分布 σ_S^2 划分为两个区间,即

$$\sigma_S^2 \in [b_0, b_1) \cup [b_1, b_2).$$

上式中:$b_0 = 0$,$b_1 = 0.12^2$,$b_2 = \infty$,两个区间对应的权重分别 $\boldsymbol{h}_{S1} = [0.15 \quad 0.15 \quad 0.20]^{\mathrm{T}}$,$\boldsymbol{h}_{S2} = [0.05 \quad 0.05 \quad 0.40]^{\mathrm{T}}$.

步骤 2 对词典模型的计算结果的方差分布 σ_D^2 划分为两个区间,即

$$\sigma_D^2 \in [b_0, b_1) \cup [b_1, b_2).$$

上式中:$b_0 = 0$,$b_1 = 0.15^2$,$b_2 = \infty$,两个区间对应的权重分别为 $\boldsymbol{h}_{D1} = [0.2 \quad 0.3]^{\mathrm{T}}$,$\boldsymbol{h}_{D2} = [0.1 \quad 0.4]^{\mathrm{T}}$.

动态权重的多模型相融合 ρ 为 0.568,比 NLPCC-ICCPOL 2016 评测比赛的第 1 名高出 9.6%.

3 结束语

提出多模型相融合的词相似性分析的方法,将 Word2Vec,《HowNet》和《同义词词林扩展版》3 个模型融合在一起,通过赋予动态权重的方法,提高了词相似性的准确率.对于 PKU-500 数据集,采用多模型相融合的相似性进行分析,获得 0.568 的斯皮尔曼等级相关系数,其与 NLPCC 2016 第 1 名的结果

相比，效果提高了 9.6%. 文中各模型结果的动态权重策略还有进一步优化的空间，从而进一步提高词相似性的准确率.

参考文献：

[1] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]// International Conference on Machine Learning. Pennsylvania: ICML, 2014: 1188-1196.

[2] HU Fei, LI Li, ZHANG Zili, *et al*. Emphasizing essential words for sentiment classification based on recurrent neural networks[J]. 计算机科学技术学报：英文版，2017，32(4)：785-795. DOI: 10.1007/s11390-017-1759-2.

[3] BOJANOWSKI P, GRAVE E, JOULIN A, *et al*. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146. DOI: 10.1162/tacl_a_00051.

[4] 李舟军，范宇，吴贤杰. 面向自然语言处理的预训练技术研究综述[J]. 计算机科学，2020，47(3)：162-173. DOI: 10.11896/jsjkx.191000167.

[5] 梅家驹，竺一鸣，高蕴琦，等. 同义词词林[M]. 上海：上海辞书出版社，1983.

[6] 马永起，韩德培，蒙立荣，等. 基于 How-net 的词语语义相似度算法[J]. 计算机工程，2018，44(6)：151-155. DOI: 10.19678/j. issn. 1000-3428. 0047061.

[7] GUO Shaoru, GUAN Yong, LI Ru. Chinese word similarity computing base in combination strategy[C]// Proceedings of NLPCC 2016, Lecture Notes in Artificial Intelligence. Beijing: [s. n.], 2016: 744-752. DOI: 10.1007/978-3-319-50496-4_67.

[8] WU Yunfang, LI Wei. Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word similarity measurement [J]. Lecture Notes in Artificial Intelligence, 2016, 10102: 828-839. DOI: 10.1007/978-3-319-50496-4_75.

[9] 吴军. 数学之美[M]. 北京：人民邮电出版社，2014.

[10] MIKOLOV T, SUTSKEVER I, CHEN K, *et al*. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the 27th Annual Conference on Neural Information processing Systems. New York: NIPS, 2013: 3111-3119.

[11] 廖胜兰，吉建民，俞畅，等. 基于 BERT 模型与知识蒸馏的意图分类方法[J/OL]. [2020-05-07]. https://doi.org/10.19678/j. issn. 1000-3428. 0057416.

[12] 陈慧，田大钢，冯成刚. 多种算法对不同中文文本分类效果比较研究[J]. 软件导刊，2019，18(5)：73-78. DOI: 10.3969/j. issn. 1003-0077. 2019. 03. 005

[13] 朱靖雯，杨玉基，许斌，等. 基于 HowNet 的语义表示学习[J]. 中文信息学报，2019，33(3)：33-41. DOI: 10.11907/rjdk. 182489.

[14] 赵倩倩. 词语相似度计算及其在语义选择限制知识获取中的应用研究[D]. 郑州：郑州大学，2018.

[15] 刘群，李素建. 基于《知网》的词汇语义相似度计算[C]// 第三届汉语词汇语义学研讨会论文集. 台北：[s. n.]，2002：59-76.

[16] 吕立辉，梁维薇，冉蜀阳. 基于词林的词语相似度的度量[J]. 现代计算机(专业版)，2013(1)：3-6，9. DOI: 10.3969/j. issn. 1007-1423(z). 2013. 01. 001.

[17] ZOU Yinfeng, OUYANG Chunping, LIU Yongbin, *et al*. A similarity algorithm based on the generality and individuality of words[C]// Proceedings of NLPCC 2016, Lecture Notes in Artificial Intelligence. Beijing: [s. n.], 2016: 549-588. DOI: 10.1007/978-3-319-50496-4_48.

（**责任编辑**：陈志贤　　**英文审校**：吴逢铁）