基于 BERT 的术语使用规范度自动检测研究

张庆国[1] 薛德军[1]
ZHANG Qingguo XUE Dejun

摘 要 术语使用规范度人工检测方法存在领域知识障碍，其效率和准确率低。利用 BERT（bidirectional encoder representation from transformers）模型捕捉文本语义和结构特征，将自然语言表述的文字组合映射到高维向量空间，可使用向量间的相似性衡量文字组合间的相似性。通过与原始术语向量在一定阈值下的相似性比对，实现了实时变化的文本流中术语使用规范度的自动检测。在包含 5 万个术语的数据集上进行测试，准确率为 0.912 4，F_1 值为 0.913 5。所提出的方法达到了工程化应用的程度，且与领域知识无关。

关键词 BERT；术语使用规范度；相似性计算；双指针滑动窗口；池化策略

doi：10.3969/j.issn.1672-9528.2024.06.007

0 引言

术语是指一个专业领域内具有特定含义和用途的必须规范使用的标准词汇。由于领域知识欠缺、常识误导、打字手误、OCR 识别错误等原因，易造成标准术语的不规范使用。

国家有关部门对图书等出版物质量有专门规定，对报纸、期刊的质量要求甚至细化到具体的编校差错率。传统的“三审三校”制度主要依靠人工完成，尽管专业编审的知识素养较高，但面对大量的待审校内容，工作负担重、效率低下。

利用 BERT 模型的文本语义和结构特征的表征能力，本文提出了一种兼顾效率和准确率并与领域无关的方法。首先将术语逐一输入 BERT 模型编码形成输出向量，存入向量数据库，并做向量索引；然后将当前文本采用双指针滑动窗口策略得到某符串，输入 BERT 模型得到编码后的输出向量；最后，基于输出向量检索向量数据库，获得相似度大于一定阈值的最相似的术语，比较和标记差异，并移动指针，继续文本流中其他字符串的检测。向量数据库专为向量查询与检索设计，能够为万亿级向量数据建立索引，高效的向量索引保证了计算效率；BERT 类模型的应用保证了准确率和领域无关性，无需额外的监督或者无监督学习，适合工程化应用。

1 相关研究工作概述

术语使用规范度检测技术在某种程度上是一种介于短语和句子之间的相似性计算技术，借助短语或句子相似性，查找候选术语，对比当前词语与候选术语的差异，在一定差异容错范围内的词语被认为是不规范使用的术语。

1.1 短语相似性计算技术

短语通常字数较少，无上下文信息，语义信息较少，可以利用的信息不多。短语相似性计算技术主要分为基于统计的平均互信息[1]、相关熵[2]、词共现[3]、LDA[4]等方法和基于 Word Net[5]、How Net[6]、同义词词林[7]等的语义计算方法以及综合考虑相似元的字面、语义及统计关联等多层特征的字符串相似度计算方法[8]。除此之外，编辑距离算法是一种比较两个字串之间由一个字串转换成另一个字串所需的最少编辑操作次数的算法，编辑操作主要是替换、插入、删除三种操作[9]。编辑距离算法也可以作为短语相似性的计算方法。

1.2 句子相似性计算技术

句子相似性计算技术总体上可以分为基于字符序列的相似计算技术、基于语义的相似性计算技术和基于深度学习的相似性计算技术。

基于字符序列的句子相似性计算技术包括最大边缘相关的方法[10]、基于向量空间模型的方法[11]、基于编辑距离的算法[9]等。这类句子相似性计算方法只考虑句子出现的字符或词组合，未考虑其他信息，准确率较低，文献 [11] 在 0.1 阈值下的查全率、查准率和 F 值分别仅为 77.42%、74.65% 和 76.01%。

基于语义的句子相似性计算主要使用语法、句法结构和语义词典等。李彬等人[12]在 2003 年提出了基于语义依存的汉语句子相似度计算技术，准确率达到了 81.4%。该方法对依存树进行完全匹配的计算量巨大，仅使用了关键的有效搭配对相似程度进行计算，未能利用所有语义信息。李茹等人[13]采用多框架语义全面分析句子的语义，在无噪声数据集上的准确率为 91.11%。在噪声数据逐渐加大的情况下，多框架语义方法准确率逐渐下降，基于 VSM 的方法准确率逐渐提高并变为最优方法。由此可见，多框架语义方法的鲁棒性较差。田堃等人[14]提出基于语义角色标注的汉语句子相似度算法，平均召回率、准确率和 F 值分别为 75.49%、70%

1. 同方知网数字出版技术股份有限公司 北京 100192

和 72.28%，其语义角色标注以动词为核心，忽略了其他词性。周艳平等人 [15] 提出了一种基于同义词词林的句子语义相似度方法，召回率、准确率和 F 值分别 69.17%、92.63% 和 79.20%，相似度阈值为 0.7，该阈值相对宽泛。袁绍正等人 [16] 提出基于句子的多属性融合相似度计算方法，通过提取句子的词频属性、词序属性、词性属性及句长属性综合考量，召回率、准确率和 F 值分别 85.24%、89.09% 和 87.12%，对语义信息的利用较少。雷歆等人 [17] 通过分辨时态和定位定语，融入语言特征，能有效提升句子相似度计算的准确率，在汉老双语的句子相似度任务上 F_1 值达到了 77.67%。英文的句子语义相似度的计算也同样使用词法、句法、同义反义等，同义反义信息通常使用 Word Net。Vasile Rus 等人 [18] 使用词汇 - 句法图识别句子释义是否相同，并基于 MSPC 数据集验证了其有效性。

以神经网络为代表的深度学习近年来在许多 NLP 任务中取得了不错的成绩。Tai 等人 [19] 提出一种树形 LSTM 算法计算句子相似度，充分利用了语义信息。Jonas Mueller 等人 [20] 提出了一个基于 LSTM 的孪生网络结构处理句子相似性，该模型在 SemEval 2014 数据集上取得了 State-of-the-Art 的结果。Chi Ziming 等人 [21] 将注意力机制引入 LSTM 网络，在建模句子时给予不同的单词不同的注意力。胡艳霞等人 [22] 提出了一种基于多头注意力机制 Tree-LSTM 的句子语义相似度计算方法。近年来，基于预训练语言模型的方法在 NLP 各个领域成为研究热点。Sentence-BERT 使用孪生网络和三胞胎网络生成具有语义的固定维数的句子向量，以大幅度减少向量相似度比较的计算量 [23]。BERT-flow 模型将 BERT 的输出空间由一个锥形可逆地映射为标准的高斯分布空间，在一定程度上解决了使用 BERT 获取的语义向量在相似度计算方面效果不佳的问题，减少了词嵌入空间分布各向异性带来的影响 [24]。Su 等人 [25] 在分析 BERT-flow 模型的基础上，使用一个线性变换达到了与 BERT-flow 相近的效果。SimCSE 采用自监督来提升模型的句子表示能力，通过防过拟合技术和基于 NLI 数据集的数据监督学习，在多项 NLP 任务中刷榜，在 STS 任务上，提升了接近十个点 [26]。Wang 等人 [27] 复现了各类模型和方法并进行了评测，证明了 SimCSE 的卓越性能，但是仍需要领域内的无标签语料来训练。

1.3 术语使用规范度检测技术

在术语使用规范度的检测过程中，某一术语出现缺字、添字、替换字、字序错乱等情况，往往使得该术语失去实际含义，成为无意义的字符串。因此，并非所有的短语 / 句子相似度算法都适用于术语使用规范度检测任务。编辑距离算法是纯粹的字符串运算方法，可以检测术语的不规范使用，但是字序的调换对编辑距离算法的影响很大，识别字序调换的术语召回率较低，且计算量大，计算规模取决于术语集的规模。鉴于汉字是象形文字，文字的字形带有一定的含义，可以借助图像处理技术处理文字，通过图像的相似度来衡量文字的相似度，但该过程需要大量的 GPU 计算资源，其准确率较低。预训练语言模型，尤其是 BERT 类模型，采用典型 MLM 算法训练，与术语使用规范度检测任务有相通之处，基于 BERT 类模型进行术语使用规范度检测理论上是可行的。

2 基于 BERT 的术语使用规范度自动检测

Hinton 教授在 2006 年发表的论文是深度学习里程碑式的新起点 [28]。深度学习被引入到自然语言处理领域，将语言模型的研究从传统的统计语言模型提升到预训练语言模型 Word2Vec[29]，随后，ELMo[30]、GPT[31]、BERT[32]、ENRIE[33]、XLNet[34]、RoBERTa[35] 等各类预训练语言模型被提出。除早期 ELMo、GPT 等模型采用了单向上下文表征外，其他模型都是双向上下文表征，能更好地表达上下文信息。2017 年，Ashish Vaswani 等人 [36] 提出了神经网络结构 Transformer，成为自然语言处理技术加速发展的催化剂。基于 Transformer 的 GPT、BERT、GPT2、GPT3、GPT4 先后在多项 NLP 任务中刷榜。更大规模的训练数据和更大规模的参数，给预训练语言模型带来了更好的效果。

BERT 模型的结构分为嵌入层和编码层，前者处理输入，后者为对输入进行编码，由自注意力层、前馈层等构成。在自注意力层，输入向量经过三个全连接层被转化成 $\boldsymbol{Q}$（query）、$\boldsymbol{K}$（key）、$\boldsymbol{V}$（value）三个向量。经典的自注意力权重为：

$$Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

式中：d_k 为 $\boldsymbol{K}$ 的维度。BERT 预训练阶段采用了掩码语言模型（MLM）和下一句预测（NSP）两个非监督机器学习任务。MLM 随机掩盖每一个句子中 15% 的词，用其上下文来做预测。被选中随机掩盖的 15% 的词，分别以 80%、10%、10% 的概率用 [MASK] 标记替换，用固定词替换，保持不变不替换。MLM 和 NSP 让 BERT 模型拥有强大的语言表征能力和特征提取能力。

预训练的 BERT 模型可以仅用一个额外的输出层进行微调，可以为很多任务比如分类、问答和语言推理等创建当前最优模型，而无需对任务特定架构做出大量修改。这种设计提升了 BERT 的普适性。

对于连续的文本流，术语的不规范使用可能会出现在文本的任何位置，因此，在文本中自动定位并且发现术语的不规范使用还需要一定的策略，采用双指针滑动窗口策略，具体步骤如下。

步骤 1：输入术语样本集，将每一个术语输入 BERT 模型，输出多维向量。

步骤 2：将每一个术语生成的词向量，存入向量数据库，并做向量索引。

步骤 3：对传入的文本流进行句子的划分，并用句子数组记录句子的起始和结束位置，设定最大词汇长度 MaxLen。

步骤 4：从句子数组中顺序取出一个句子，对句子进行切分词语处理，设定句子指针 p 指向句子起始位置，句子指针 e 指向句子结束位置，转步骤 5；若所有句子处理完毕，则执行步骤 8，结束处理过程。

步骤 5：若从指针 p 开始到指针 e 的长度大于 MaxLen，则从指针 p 指向位置开始选择个 MaxLen 字符，标记为 S，令指针 e 指向从指针 p 开始的第 MaxLen+1 个字符，转步骤 6；否则，如果从指针 p 开始到指针 e 之间的字符数小于 4 个汉字且指针 e 并未指向句子结束位置，则指针 p 前进当前词语长度，指向下一个词的开始位置，指针 e 指向句子结束位置，转步骤 5 的开始步骤；否则，如果从指针 p 开始到句子结束位置之间的字符数小于 4 个汉字且指针 e 指向句子结束位置，则结束当前句子处理并转步骤 4；否则，选择从指针 p 开始到指针 e 的所有字符，标记为 S。

步骤 6：使用选择的字符串 S 生成向量并检索术语向量数据库，如果存在大于一定相似度阈值的检索结果，则选择相似度最大的检索结果，转步骤 7；若没有大于一定阈值的检索结果，则句子的指针 p 不变，指针 e 向指针 p 的方向前进一个词汇的长度（亦即指针 e 后退字符串 S 的最后一个词的长度，指针 p 和 e 之间的字符数减少，长度减少一个词的长度），转步骤 5。

步骤 7：针对检索结果，计算当前字符串与检索结果对应的术语的差异并标记差异，指针 p 前进字符串 S 的长度，指针 e 指向句子结束位置，转步骤 5。

步骤 8：结束处理。

其中，MaxLen 为术语集术语最大长度，也即双指针滑动窗口的最大长度。双指针滑动过程示意图如图 1 所示。

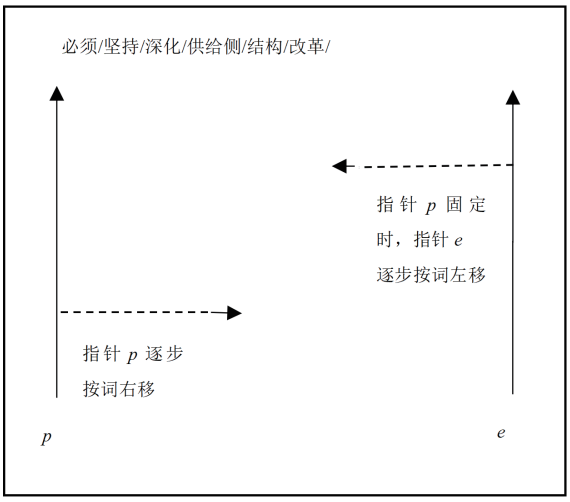

图 1 双指针滑动窗口示意图

3 实验与结果分析

术语规范度自动检测过程是一套复杂的流程，实验对术语不规范使用的核心匹配算法进行验证。实验的基础样本集是 5 万个不同领域的术语，结果的评价指标采用准确率（accuracy）、召回率（recall）、精准率（precision）和 F-measure（简称 F_1）。F_1 定义为：

$$F_1=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{2}$$

3.1 不同检测方法的实验与分析

BERT 微调方法以 5 万个正样本术语为基础，通过编写程序，分别构造了随机减少 1 个字、随机增加 1 个字、随机替换其中某个字、随机调换某几个字的位置 4 种操作，获得了格式如“1\t 真空紫外光 CVD\t 真空紫外盯光 CVD”的正样本训练集；通过相似度计算，选择与当前正样本术语最相似的 4 条术语，构成格式如“0\t 真空紫外光 CVD\t 真空紫外滤光片”的 4 组负样本训练集。训练集和测试均包含 40 万个词对。BERT 微调采用中文 chinese_L-12_H-768_A-12 模型，学习率设为 2e-5，batch size 设为 8，训练 2 个 epochs。其中，20 万条负样本测试集均识别正确，20 万条正样本测试集识别正确 159 595 条。微调后的模型对完全不相关的词的判断较为准确，但对正样本测试集的识别效果稍差。

原生 BERT 方法以 5 万个正样本术语为基础，通过编写程序，分别构造了随机减少 1 个字、随机增加 1 个字、随机替换其中某个字、随机调换某几个字的位置 4 种操作，获得了 4 组各 12 500 个不重复的词条的测试集。另外，增加 1 组 50 000 个与术语集不重复的词条作为第 5 组测试集。5 组测试集词条逐一与术语样本集进行比较。实验选用 BERT Base 中文预训练语言模型（chinese_L-12_H-768_A-12）。选取 0.93 作为相似度阈值，选取 BERT 第 12 层作为向量输出层，选取 REDUCE_MEAN 作为输出层池化策略，综合性能最优。

实验测试了 RoBERTa 和 SimCSE 预训练语言模型。RoBERTa 模型选择 3 层的精简模型 RoBERTa-tiny3L768-clue，测试其对术语检测任务的表现，选择相似度阈值 0.95，选择 REDUCE_MEAN 池化策略，选择第一层输出向量，综合性能最好，F_1 值达到 0.886 1。SimCSE 模型选择从 Hugging Face 下载的预训练模型 SimCSE-bert-base。SimCSE 的表现一般，选择相似度阈值 0.89 时，F_1 值为 0.694 1。

基于图像相似度方法的测试过程中将汉字转为不同规格的图像，综合比较下选择将单一汉字转为 244 像素 ×244 像素，采用 VGGNet 卷积神经网络抽取图像特征计算图像相似性以判断文字的相似性，结果相对较优。

不同模型 / 方法的测试情况汇总为表 1。

表 1 不同模型（方法）的比较

模型 / 方法	阈值	Accuracy	Recall	Precision	F_1
SimCSE	0.89	0.740 3	0.627 8	0.776 0	0.694 1
RoBERTa tiny	0.95	0.885 9	0.898 4	0.874 0	0.886 1
编辑距离	0.75	0.790 9	0.790 9	1.0	0.883 2
基于图像相似度	0.965	0.811 0	0.783 3	0.876 5	0.827 3
BERT 微调	无	0.899 0	0.798	1.0	0.845 5
BERT	0.93	**0.912 4**	**0.925 3**	0.902 0	**0.913 5**

从实验结果看，SimCSE 方法在术语使用规范度检测任务中表现一般，SimCSE 效果的提升与训练和测试的领域数据集相关度较高，不适合领域分布较广的数据集，鲁棒性稍差。采用 BERT 微调的方法，本质上是 0、1 为标签的二分类，其准确率非常高，所有负样本均分类正确，但对正样本的分类错误率稍高，召回率偏低。编辑距离和基于图像的相似度的方法总体上表现尚可，鲁棒性高，是较为通用的方法。BERT 训练时采用的 MLM 方法与实验任务有相通之处，因此，无需微调、无需修改网络结构和参数，原生 BERT 即可在实验任务中取得了最优效果。

3.2 基于 BERT 的检测方法的相关影响因素分析

通常取 BERT 倒数第二层输出向量，在多数情况下是最优选择，实验数据表明，针对术语使用规范度检测这项任务，如图 2 所示，最后一层效果最优。

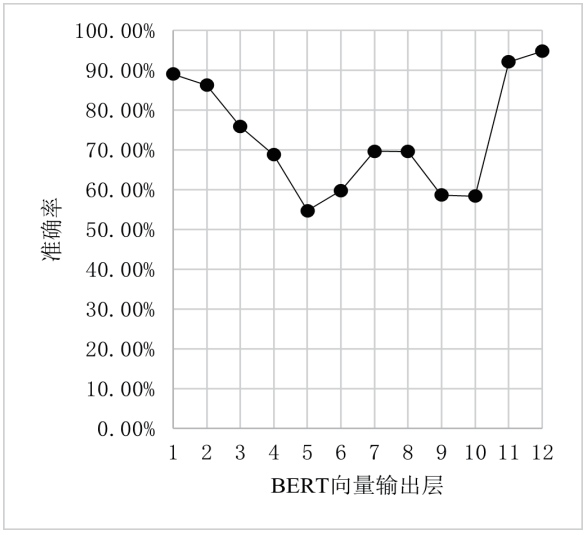

图 2 BERT 向量输出层与准确率的关系

综合各组的测试数据，取 BERT 第 12 层（最后一层）作为向量输出层，在不同的相似度阈值下计算精准率、召回率和 F_1，如图 3 所示。从图 3 中可以看出，取最低相似度阈值 0.92 和 0.93，F_1 达到最优。综合考虑，取 0.93 作为最低相似度阈值。

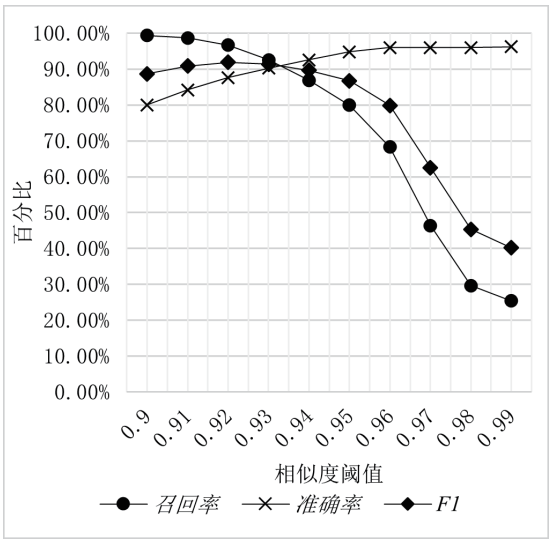

图 3 不同相似度阈值下召回率、精确率和 F_1 值的变化

选定相似度阈值 0.93，五组测试集准确率变化不大，但是带有错字的术语测试集的准确率较其他组偏低，如图 4 所示。

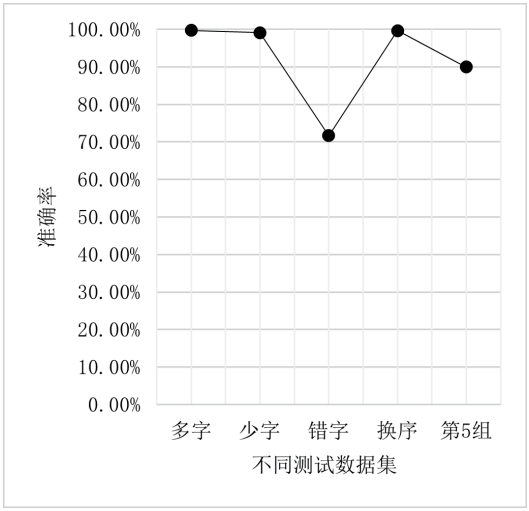

图 4 不同测试集数据测试准确率

以上的测试数据，均采用默认的 REDUCE_MEAN 池化策略。不同的池化策略如表 2 所示。

表 2 不同池化策略说明

池化策略	描述
REDUCE_MEAN	在时间轴上取编码层隐藏状态的平均值
REDUCE_MAX	在时间轴上取编码层隐藏状态的最大值
REDUCE_MEAN_MAX	分别做 REDUCE_MEAN 和 REDUCE_MAX 然后在最后一个轴上将它们连接在一起，产生 1536 维向量
CLS_TOKEN 或者 FIRST_TOKEN	得到对应的隐藏状态 [CLS]，即第一个 token
SEP_TOKEN 或者 LAST_TOKEN	得到对应的隐藏状态 [SEP]，即最后一个 token

不同池化策略下测试数据集的表现如图5所示，REDUCE_MEAN池化策略在准确率上性能最优。

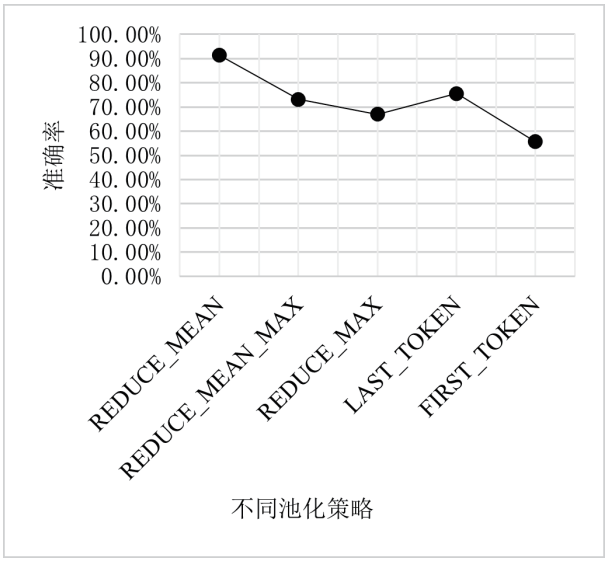

图5 不同池化策略下测试数据集的表现

综上图2至图5，在术语使用规范化检测任务中，选取0.93作为相似度阈值，选取BERT第12层作为向量输出层，选取REDUCE_MEAN作为输出层池化策略，综合性能最优。

4 结论

本文探索了术语使用规范度检测任务的各类方法，并基于构造的测试数据进行测试，实验表明，基于BERT Base预训练语言模型的检测方法，选取0.93作为相似度阈值，选取最后一层作为向量输出层，选取REDUCE_MEAN作为输出层池化策略时，综合性能最优，可以有效检测术语的不规范应用。为加速向量的相似性计算过程，使用了向量数据库。在实际应用中，RoBERTa tiny较BERT Base综合性能近似，其运算速度更快，更适合工程化应用。

大模型（LLM）在最近一两年发展迅速。Freestone M等人研究发现，PaLM和ADA这两个基于LLM的模型在单词类比任务上能够捕获有意义的语义并保持高精度，它们也与SBERT在语义上保持一致，在资源有限时，SBERT可能是一种有效的替代方案[37]。截至2024年5月27日，huggingface MTEB榜单前10名几乎都使用了7 B及以上规模大模型。使用大模型进行短文本的语义表示，会是下一步研究的重点，但工程应用仍需要平衡资源投入与回报之间的关系，选择性价比高的方案。

参考文献：

[1]PETER F, STEPHEN A, VINCENT J, et al. Word-sense disambiguation using statistical methods[C]//Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL'91).New York:ACM,1991:264-270.

[2]LILLIAN L. Similarity-based approaches to natural language processing[EB/OL].(1997-08-19)[2024-4-20].https://arxiv.org/abs/cmp-lg/9708011.

[3] 张涛，杨尔弘．基于上下文词语同现向量的词语相似度计算[J]. 电脑开发与应用，2005(3):41-43.

[4] 吕亚伟，李芳，戴龙龙．基于LDA的中文词语相似度计算[J]. 北京化工大学学报（自然科学版），2016,43(5):79-83.

[5]ZHANG P Y. Word similarity computation based on WordNet and HowNet[C]//International Conference on Measurement, Instrumentation and Automation. Switzerland: Queensland University of Technology,2013:336-338.

[6] 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.

[7] 梅家驹．同义词词林[M]. 上海：上海辞书出版社，1983.

[8] 章成志．基于多层特征的字符串相似度计算模型[J]. 情报学报，2005,24(6):696-701.

[9]LEVENSHTEIN V I. Binary codes capable of correcting deletions, insertions, and reversals[J].Sov Phys Dokl, 1966(2): 707-710.

[10]JAIME C, JADE G. The use of MMR diversity-based reranking for reordering documents and producing summaries[J]. Proceedings of ACM SIGIR'98,2017,51(2):209-210.

[11] 苏小虎．基于改进VSM的句子相似度研究[J]. 计算机技术与发展，2009,19(8):113-116.

[12] 李彬，刘挺，秦兵，等．基于语义依存的汉语句子相似度计算[J]. 计算机应用研究，2003(12):15-17.

[13] 李茹，王智强，李双红，等．基于框架语义分析的汉语句子相似度计算[J]. 计算机研究与发展，2013,50(8):1728-1736.

[14] 田堃，柯永红，穗志方．基于语义角色标注的汉语句子相似度算法[J]. 中文信息学报，2016,30(6):126-132.

[15] 周艳平，李金鹏，蔡素．基于同义词词林的句子语义相似度方法及其在问答系统中的应用[J]. 计算机应用与软件，2019, 36(8):65-68+81.

[16] 袁绍正，周艳平．基于句子的多属性融合相似度计算方法[J]. 计算机系统应用，2022,31(4):303-308.

[17] 雷歆，周蕾越，周兰江．融合语法及结构特征的汉老双语句子相似度计算方法[J]. 中文信息学报，2023,37(9):73-82.

[18]RUS V, MCCARTHY P M, LINTEAN M C, et al. Paraphrase identification with lexico-syntactic graph subsumption[C]// Twenty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS-21).Menlo

Park:AAAI Press, 2008: 201-206.

[19]TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics.Stroudsburg:Association for Computational Linguistics, 2015:1556-1566.

[20]MUELLER J, THYAGARAJAN A. Siamese recurrent architectures for learning sentence similarity[J].Thirtieth AAAI conference on artificial intelligence, 2016(16):2786-2792.

[21]CHI Z, ZHANG B. A Sentence similarity estimation method based on improved siamese network[J]. Journal of intelligent learning systems and applications,2018, 10(4):121-134.

[22] 胡艳霞 , 王成 , 李弼程 , 等 . 基于多头注意力机制 Tree-LSTM 的句子语义相似度计算 [J]. 中文信息学报 , 2020, 34(3):23-33.

[23]REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg:Association for Computational Linguistics, 2019: 3982-3992.

[24]LI B, ZHOU H, HE J, et al. On the sentence embeddings from pre-trained language models[C]//Empirical Methods in Natural Language Processing (EMNLP).Stroudsburg:Association for Computational Linguistics, 2020:9119-9130.

[25]SU J, CAO J, LIU W, et al. Whitening sentence representations for better semantics and faster retrieval[EB/OL].(2021-03-29)[2024-05-10].arXiv preprint arXiv:2103.15316 ,2021, https://arxiv.org/pdf/2103.15316.pdf.

[26]GAO T, YAO X, CHEN D. Simcse: simple contrastive learning of sentence embeddings[C]//SimCSE: Simple Contrastive Learning of Sentence Embeddings.Stroudsburg:Association for Computational Linguistics,2021:6894-6910.

[27]WANG B, KUO C C J, LI H. Just rank: rethinking evaluation with word and sentence similarities[C]//Annual meeting of the Association for Computational Linguistics. Stroudsburg:Association for Computational Linguistics, 2022: 6060-6077.

[28]HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554..

[29]MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL].(2013-09-17)[2024-05-09].https://arxiv.org/pdf/1301.3781.pdf.

[30]PETERS M E, NEUMANN, IYYER M, et al. Deep contextualized word representations[J].Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: human language technologies. association for computational linguistics, 2018(1):2227-2237.

[31]RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB/OL].(2018-09-15)[2024-05-01].https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

[32]DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Conference on the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg:ACL, 2019:4171-4186.

[33]ZHANG Z, HAN X, LIU Z, et al. ERNIE: Enhanced language representation with informative entities[C]//Annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 1441-1451.

[34]YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding[C]//Conference on Neural Information Processing Systems.Red Hook: Curran Associates, 2020:5730-5740.

[35]LIU Z, LIN W, SHI Y, et al. A robustly optimized BERT pre-training approach with post-training[C]//Proceedings of the 20th Chinese National Conference on Computational Linguistics. Cham:Springer,2021:1218-1227.

[36]VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL].(2017-06-12)[2024-05-02].https://arxiv.org/abs/1706.03762.

[37]FREESTONE M, SANTU S K K. Word embeddings revisited: do llms offer something new?[EB/OL].(2024-02-16)[2024-03-16]. https://arxiv.org/abs/2402.11094.

【作者简介】

张庆国（1976—），男，山东临沂人，硕士，高级工程师，研究方向：人工智能技术应用、自然语言处理、数据要素市场。

薛德军（1970—），男，湖南辰溪人，博士，副总经理兼总工程师，研究方向：数据要素市场、数字出版、人工智能。

（收稿日期：2024-05-27）