基于降噪排序学习的无监督语句高效嵌入方法

刘成龙 朱奕萌 何理扬 刘 淇
(中国科学技术大学认知智能全国重点实验室 合肥 230088)

摘 要 近年来,无监督的语句嵌入作为自然语言处理的关键任务,因其在处理海量无标注数据方面巨大的应用潜力而备受关注。对比学习方法虽然在该任务上取得了显著成效,却只考虑了语句对之间局部关联,忽略了全局排序结构;排序学习方法虽然提升了全局语义感知能力,但会因不恰当的批次采样和不一致的排序信号而引入显著噪声。此外,现有的列表排序损失函数在捕捉细粒度语义差异方面效率低下。为了解决这些问题,本文提出了基于排序学习的多维度降噪的数据预处理与模型训练架构 RankRefine。具体而言,本文方法采用离线共识采样和数据重组构建可靠的有序序列从而缓解噪声。同时,通过理论分析,本文揭示了列表排序损失函数梯度范数随批次规模增大的问题,受此启发并借鉴 LambdaRank 的思想,本文引入了一种成对比较策略和加权方案,以克服列表排序损失在捕捉细粒度语义上的低效问题。该策略实现了对排序位置敏感的损失加权,增强了模型对局部语义细微差别的感知能力,同时提高了训练效率。在此基础上,本文创新性地为文本语义任务设计了一种更稳定的采样方案 RankRefine+,通过自适应容错列表排序采样策略在一定程度上容忍排序不一致情况,增强难负例(hard negative)采样效果,显著提升共识序列质量,实现了模型性能的综合优化。实验结果表明,RankRefine 方法在语义文本相似度任务(semantic textual similarity,STS)、迁移学习任务(transfer learning,TR)中准确率较同尺寸最优模型分别平均提升 1.38%、1.14%。在 STS 任务中,RankRefine+在 RankRefine 的基础上取得了 0.11%的提升。在重排序任务(reranking)中,本文方法在多项指标上表现领先。消融实验进一步证明了本文方法的有效性和鲁棒性。

关键词 自然语言处理;语句嵌入;文本检索;无监督学习;对比学习;数据增强;排序学习
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2026.01527

An Efficient Denoising Learning-to-Rank Framework for Unsupervised Sentence Embedding

LIU Cheng-Long ZHU Yi-Meng HE Li-Yang LIU Qi
(State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei 230088)

Abstract As one of the most fundamental tasks in natural language processing (NLP), sentence embedding, which maps variable-length sentences into fixed-dimensional dense vectors such that semantically similar sentences are proximate in the representation space, underpins a wide spectrum of applications, including semantic retrieval, text clustering, text matching, and question answering. In particular, unsupervised sentence embedding has recently garnered considerable attention due to its tremendous potential in harnessing the ever-growing volume of large-scale unlabeled text data available on the web. Traditional supervised methods have achieved solid performance on a wide range of tasks, yet rely heavily on expensive and labor-intensive human-annotated datasets. In contrast, unsupervised methods learn general-purpose sentence representations directly from raw text corpora, offering better scalability and generalizability across a wide range of

收稿日期:2025-06-04;在线发布日期:2026-03-11。本课题得到国家自然科学基金重点项目(No. 62337001)和国家自然科学基金青年科学基金项目(A 类)(No. 62525606)资助。**刘成龙**,硕士研究生,中国计算机学会(CCF)学生会员,主要研究领域为数据挖掘与智能教育。E-mail:sepcnt@mail.ustc.edu.cn。**朱奕萌**,硕士研究生,中国计算机学会(CCF)学生会员,主要研究领域为数据挖掘与智能教育。**何理扬**,博士,中国计算机学会(CCF)学生会员,主要研究领域为信息检索、大语言模型与教育数据挖掘。**刘 淇(通信作者)**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为认知智能与数据挖掘。E-mail:qiliuql@ustc.edu.cn。

downstream tasks. Among unsupervised approaches, contrastive learning has achieved remarkable results in learning sentence representations by clustering similar pairs while preserving the gap among irrelevant items. However, these methods focus solely on pairwise relationships between sentences, while neglecting the global ranking structure across multiple sentences, which makes it tough to capture fine-grained and nuanced semantic information. Learning-to-rank methods enhance global semantic perceptual abilities, but they induce significant noise due to inappropriate batch sampling strategies and inconsistent ranking signals. Furthermore, existing listwise ranking loss functions demonstrate inefficiency in capturing fine-grained semantic distinctions. To tackle these challenges, we present RankRefine, a novel multidimensional denoising framework for data preprocessing and model optimization based on the learning-to-rank method. Specifically, RankRefine leverages offline consensus sampling and data reconstruction, which build sentence lists from distinct index models and retains only those ordinal relationships on which the index models reach agreement, to construct reliable and high-quality sentence ranking lists as model training data in order to mitigate gradient noise. Meanwhile, through theoretical analysis, we reveal that the gradient norm of listwise ranking loss functions increases with the batch size due to its mathematical form, heavily distorting the stability and convergence of training. Inspired by our findings and LambdaRank, we introduce a new pairwise loss with a positionally aware weighting scheme, overcoming the lack of competence of listwise ranking loss functions in capturing fine-grained semantics while enhancing training efficiency. Building on this, we innovatively propose RankRefine+, a more stable sampling scheme for text semantic tasks that significantly enhances hard negative sampling and improves consensus sequence quality through an adaptive fault-tolerant listwise sampling strategy, achieving comprehensive performance optimization. Extensive experiments on established benchmarks show that RankRefine outperforms the best baseline model with an average increase in Spearman correlation coefficient of 1.38% in semantic textual similarity (STS) tasks. An average accuracy gain of 1.14% in transfer learning (TR) tasks demonstrates that our proposed method is capable of being applied to various downstream tasks. RankRefine+ achieves further improvements of 0.11% over RankRefine in STS tasks. In reranking tasks, our proposed methods also achieve excellent performance across multiple metrics. Ablation studies verify the contribution of each component, and hyperparameter sensitivity analysis further confirms consistent and stable gains across diverse configurations, together demonstrating the practical robustness and effectiveness of our framework.

Keywords natural language processing; sentence embedding; text retrieval; unsupervised learning; contrastive learning; data augmentation; learning to rank

1 引 言

语句嵌入(sentence embedding)作为自然语言处理领域的基础性任务之一[1],通过将文本映射到高维稠密向量空间,为语义信息的数学表征和计算提供了有效途径。高质量的语句嵌入能够准确捕捉语句间的深层语义关联,在语义检索[2-4]、文本分类[5-8]、问答系统[9-12]等下游任务中发挥着关键作用,对大数据场景下的文本处理具有重要价值。

主流有监督方法[13-14]严重依赖大规模高质量标注数据,而在实际应用中,此类数据的获取成本高昂且在特定领域常难以实现。因此,如何通过无监督学习挖掘文本数据的内在语义结构,成为该领域的核心问题。对比学习[15-18]作为无监督语句嵌入的主流范式,通过在向量空间中拉近语义相似样本、推远无关样本来学习表征,代表性工作包括 SimCSE[19]、DiffCSE[20]、ConSERT[21]、ArcCSE[22]等。然而,这些方法仅考虑二元文本对关系,难以捕捉细粒度语义差异。RankCSE[23]、RankEncoder[24]等方法引入

列表级排序损失[25-26],将句间关系从文本对扩展至有序列表,以增强模型对语义层次的感知能力。

然而,现有基于排序的方法仍面临数据噪声与训练低效的双重挑战。一方面,现有方法采用随机采样构建训练批次,导致批次内文本语义相关性低、主题分散。这种语义鸿沟使得文本间的排序关系缺乏实际意义,产生大量噪声标签。尽管 RankCSE[23] 通过多模型预测结果的加权平均提升了标签可靠性,但对于语义无关文本的排序仍然具有高度不确定性,产生的误导性梯度严重影响模型的语义学习能力。另一方面,从优化角度分析,在大批次训练场景下,负对数似然损失函数表现出损失值和梯度范数增长过快的特性。现有方法对所有排序关系赋予相同权重,忽视了不同文本对在语义学习中的重要性差异。这种均等化处理不仅引入了大量冗余约束,还导致模型难以聚焦于关键的语义信息,严重影响了训练效率和收敛速度。

为系统性地解决上述问题,本文提出了一种高效的多维度降噪排序学习无监督语句嵌入框架:RankRefine,致力于从算法设计和数据处理两个维度协同优化,在捕捉细粒度语义的同时,克服损失噪声,保证训练效率。与现有降噪方法仅关注数据合成或逐语句条目处理不同,本文构建了针对训练批次内噪声的系统性降噪框架,无需生成额外文本数据即可显著提升数据质量。针对随机采样批次内语义鸿沟带来的噪声问题,本文设计了基于动态规划的共识排序采样策略和数据重组方案,通过利用索引模型间的一致排序结果,在提高批次内文本相关性的同时,主动识别并过滤模型间存在分歧噪声的排序关系。针对损失函数缺陷带来的训练低效问题,本文基于严格的理论分析揭示了 ListMLE 损失的梯度发散问题,并创新性地将共识序列信息直接融入损失函数设计,设计了基于文本相似度的加权列表排序函数,聚焦于重要可靠的排序对进行学习优化,实现了语义的细粒度捕捉和训练效率的双重提升。此外,针对严格共识采样可能导致的细粒度语义信息丢失问题,本文提出了 RankRefine+变体,一种自适应的容错列表排序采样策略,通过位置信息动态容忍细微差异,从而在保证语义一致性的同时增强难负例采样效果,提升数据利用率和学习效率。

实验结果①表明,基于 BERT[27]/RoBERTa[28] 预训练模型在维基百科语料上训练,采用 RankRefine 方法在语义文本相似度(semantic textual similarity, STS)基准测试上,嵌入准确率相较于现有先进方法,平均提升 1.38%,展现了显著的性能优势。消融实验进一步证明了本文方法的有效性。本文方法还在迁移学习任务(transfer learning, TR)上平均提升 1.14%,在重排序任务(reranking)上多项指标领先。

本文的贡献主要体现在以下三个方面:

(1)构建系统性批次降噪框架:区别于现有方法侧重数据合成或逐语句对处理的降噪策略,本文在数据预处理阶段引入动态规划的共识采样策略,构建高质量的数据批次,利用索引模型间的一致降序判断,以过滤引发梯度冲突的数据,提高排序数据的可靠性。

(2)理论驱动的损失函数创新:基于严格理论证明,在损失设计上引入基于成对方法的加权排序,创新性地将数据预处理阶段的共识信息融入损失函数,有效避免了语义鸿沟带来的无意义噪声,实现了细粒度语义信息捕捉,并提高了模型训练效率。

(3)全面的模型能力测评:设置了大量实验进行了全方位的模型能力评估。在不同基准测试上,本文方法均取得了超过现有最优方法的效果。

2　相关工作

语句嵌入(sentence embedding)任务源于自然语言处理领域对文本表示学习的持续探索。早期文本表示方法主要采用独热编码(one-hot encoding)或词袋模型(Bag-of-Words, BoW)[29],后者通常结合词频-逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)[30]进行词汇权重调整。然而,这些方法不仅存在维度灾难问题,而且难以有效捕捉词汇间的深层语义关系。

Word2Vec[31] 和 GloVe[32] 通过优化词向量表示实现了重要突破。前者基于局部上下文窗口学习多词模式,后者则利用全局共现矩阵构建语义结构。然而,静态词嵌入方法为每个词元生成固定向量,缺乏上下文感知能力,难以处理一词多义等问题。

为解决上述局限性,研究者在深度学习框架下探索了更为强大的文本表示方法。特别是以 Transformer 架构[33-34] 为基础的预训练语言模型(Pre-trained Language Models, PLMs),如 BERT[27] 及其改进版本 RoBERTa[28] 等,推动语句嵌入技术进

① 本文实验代码可在 https://github.com/RankRefine/repo 获取。

入动态、上下文感知的新阶段。这类深度模型通过在大规模未标注语料库上采用自监督学习策略，能够有效提取丰富的层次化语义和句法特征，并根据词元(token)所处具体语境生成动态、上下文相关的表征。预训练语言模型的出现标志着文本表征范式的重大转变，即从传统静态的离散符号表示转向基于深度网络、能够捕捉复杂上下文依赖的动态表征。

然而，预训练语言模型直接输出的上下文相关的词元序列嵌入，难以建立向量索引，进行高效的检索查询。因此，如何有效地聚合、提炼这些动态且变长的词元级信息，以生成能够准确表征文本核心语义的固定维度嵌入，成为亟待解决的关键问题。针对这一挑战，研究者提出了多种解决方案，其中 Sentence-BERT[14](SBERT)是典型代表。该方法在预训练模型基础上引入池化(pooling)机制，并结合孪生网络(Siamese Network)或三元组网络(Triplet Network)架构，在自然语言推断(Natural Language Inference，NLI)[35]等包含句对关系的任务上进行微调，从而生成语义信息更为凝练、适于相似度计算的嵌入向量。这一工作为后续工作，尤其是无监督语句嵌入方法的研究奠定了重要基础。

在实际应用中，大规模且高质量的标注数据往往难以获取，因此无监督语句嵌入(unsupervised sentence embedding)成为近年来的研究热点。该任务旨在摆脱对人工标注的依赖，利用自监督学习机制从文本数据中挖掘深层语义结构，实现语句到高维向量的有效映射。目前，对比学习已成为该领域的主流技术范式，其核心思想是通过构建正负样本对，使模型学习能够区分语义差异的向量表征[19-23,36-38]。该方法在语句嵌入领域衍生出了两条并行的技术发展路径。第一条路径为对比学习核心机制的优化，即致力于在不依赖模型、训练数据规模扩张的前提下，通过对算法本身的设计优化来提升嵌入质量。本文主要沿此路径展开，集中体现在数据增强策略与损失函数设计两个方面。

在无监督学习框架下，数据增强是提升训练数据质量与多样性的关键。无监督对比学习通常将同一批次中的其他样本视为负例，而正例则经由数据增强生成。不同的数据增强策略对模型性能有显著影响。例如，SimCSE[19]通过随机掩码(dropout)构造语义相近但表面形式不同的困难正例；ConSERT[21]在此基础上引入词序重排、序列裁剪及对抗扰动等方式，获得更具挑战性的正例集。近年来，大语言模型(Large Language Models，LLMs)为数据增强提供了新机遇。例如，SynCSE[37]、MultiCSR[38]与 SynCSE-r[39]利用大语言模型生成语义多样且语法合理的增强样本，显著提升了嵌入质量。然而，现有方法多从逐句或逐样本角度降低噪声，侧重数据合成或局部增强，忽视了训练批次内部因语义分散产生的结构性噪声问题。

在损失函数设计方面，信息噪声对比估计(Information Noise Contrastive Estimation，InfoNCE)[40-41]已成为无监督语句嵌入任务中的基础损失函数。其核心机制是最大化锚点样本与其正样本之间的互信息(Mutual Information，MI)，在嵌入空间中缩小二者距离，同时拉远与负样本的距离。从信息论视角来看，该目标可视为互信息最大化的变分下界估计[41]，从而提升模型对语义相似与不相似样本的区分能力。在此基础上，研究者提出了多种改进方法。如 DiffCSE[20]借鉴 ELECTRA[42]的思想，引入词元替换检测(Replaced Token Detection，RTD)损失以增强模型对细微语义变化的敏感性；ArcCSE[22]加入角度边界(angular margin)直接优化嵌入向量间夹角；RankCSE[23]与 RankEncoder[24]则采用列表级排序损失(如 ListNet[25]与 ListMLE[26])提升细粒度语义表达。然而，这些方法默认训练批次内所有排序关系同等重要，未考虑语义分散可能导致的排序噪声，也缺乏梯度稳定性的理论分析。

第二条技术路径则聚焦于模型规模与架构的演进，即通过更大规模数据、更高参数量和更先进基座模型提升性能上限。E5[43]与 GTE[44]是连接早期方法与此路线的代表：E5 将指令微调(Instruction Tuning)引入嵌入任务，并结合多阶段训练增强模型适应性；GTE 则优化编码器架构以提升通用表示能力。这一趋势在基于大语言模型的嵌入方法中更为显著，例如闭源模型 gemini-embedding-001[45]和开源模型 gte-Qwen2-instruct[46]、Qwen3-Embedding[47]均借助预训练大语言模型作为基座进行后训练，并取得先进性能。然而，这类方法高度依赖算力与部分有监督数据，训练与部署成本高昂。这凸显了本文所研究的价值，即在不显著增加计算与数据成本的情况下，通过高效算法提升模型性能。

3 预备知识

本节依次介绍语句嵌入的问题定义、无监督对比损失以及两类排序学习损失，为后续方法设计奠定理论基础。

3.1 问题定义

语句嵌入的核心任务是将原始文本集合 $T=\{x_i\}_{i=1}^N$ 的每段文本通过嵌入模型 f 投影为高维空间中的向量表示 $f(x_i)\in\mathbb{R}^d$。定义 $\phi(\cdot,\cdot)$为向量间相似度的度量函数(如余弦相似度),则语句嵌入的目标是使嵌入向量间的相似度 $\phi(f(x_i),f(x_j))$尽可能接近原始文本之间的语义关联程度。形式化地,该目标可表述为

$$\min_f \sum_{i,j} \mid \phi(f(x_i),f(x_j)) - \mathrm{sem}(x_i,x_j) \mid \tag{1}$$

其中,$\mathrm{sem}(x_i,x_j)$表示文本 x_i 和 x_j 之间的实际语义关联程度。

在实际应用中,通常采用三元组损失(Triplet Loss)作为替代性的优化目标。对于给定的锚点句(anchor sentence,s_a)、正样本句(positive sentence,s_p)和负样本句(negative sentence,s_n),三元组损失要求锚点句与正样本句之间的距离小于其与负样本句之间的距离,且二者之差至少为预设的边界值(margin,m)。记二者之差为

$$d(s_a,s_p,s_n) = \phi(f(s_a),f(s_n)) - \phi(f(s_a),f(s_p)) \tag{2}$$

则三元组损失函数定义为

$$L(s_a,s_p,s_n) = \max\{0, m - d(s_a,s_p,s_n)\} \tag{3}$$

3.2 无监督对比损失

对比损失(contrastive loss)是无监督语句嵌入任务中广泛采用的优化目标,SimCSE 提出了基于 InfoNCE 损失的对比学习框架。

考虑包含 N 个样本的数据批次$\mathcal{D}=\{x_k\}_{k=1}^N$。SimCSE 通过两次不同的随机掩码操作分别生成锚点样本 x_i 及其正样本 x_i',批次内其他样本作为负样本。InfoNCE 损失函数定义为:

$$\mathcal{L}_{\mathrm{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\phi(f(x_i),f(x_i'))/\tau}}{\sum_{j=1}^{N} e^{\phi(f(x_i),f(x_j'))/\tau}} \tag{4}$$

其中,$\phi(u,v)=\dfrac{u^{\mathrm{T}}v}{\|u\|\cdot\|v\|}$为余弦相似度,$\tau$ 为温度系数。当 $j=i$ 时,$f(x_j')$为正样本;否则为负样本。

3.3 列表排序损失

为了克服对比损失在捕捉细粒度语义信息时的局限性,RankCSE[23] 和 RankEncoder[24] 引入了列表排序损失(listwise ranking loss),用于优化模型预测的语义相似度排序与参考排序之间的一致性。

对于给定查询 q 下的一个文本列表 $X=\{x_1,x_2,\cdots,x_N\}$,模型 f 为每个文本 x_j 预测一个排序得分为

$$s_q(x_j)=\phi(f(q),f(x_j)) \tag{5}$$

ListNet[25] 通过 Softmax 函数计算每个文本 x_j 排在列表首位的概率,或视为其相关性的概率表示:

$$P_f(x_j \mid X)=\frac{\exp(s_q(x_j))}{\sum_{k=1}^{N}\exp(s_q(x_k))} \tag{6}$$

类似地,根据真实语义关联程度计算一个目标概率分布 $P_{sem}(X_j\mid X)$。最终,通过交叉熵作为损失函数,衡量这两个概率分布 $P_f(X)=\{P_f(x_j\mid X)\}_{j=1}^N$ 和 $P_{sem}(X)=\{P_{sem}(x_j\mid X)\}_{j=1}^N$ 之间的差异:

$$\mathcal{L}_{\mathrm{ListNet}} = -\sum_{j=1}^{N} P_{sem}(x_j \mid X)\log(P_f(x_j \mid X)) \tag{7}$$

ListMLE[26] 基于 Plackett-Luce 模型[48-49] 对排序概率进行建模。该模型为每个项目 x_i 赋予效用参数 $v_i>0$,排序 $\pi=(\pi_1,\pi_2,\cdots,\pi_N)$的概率表示为顺序选择概率的乘积:

$$P(\pi \mid V)=\prod_{j=1}^{N}\frac{v_{\pi_j}}{\sum_{k=j}^{N} v_{\pi_k}} \tag{8}$$

观察到参考排序 π^* 的概率为 $P(\pi^*\mid f)$,记 $s_q(i)=\phi(f(q),f(x_i))$,代入 $v_i=\exp(s_q(x_{\pi_i}))$,ListMLE 损失函数即为参考排序 π^* 的负对数似然:

$$\begin{aligned}\mathcal{L}_{\mathrm{ListMLE}} &= -\log P(\pi^* \mid f)\\ &= -\log\prod_{j=1}^{N}\frac{\exp(s_q(x_{\pi_j^*}))}{\sum_{k=j}^{N}\exp(s_q(x_{\pi_k^*}))}\\ &= \sum_{j=1}^{N}\left(\log\sum_{k=j}^{N}\exp(s_q(x_{\pi_k^*})) - s_q(x_{\pi_j^*})\right)\end{aligned} \tag{9}$$

pListMLE[50](Position-aware ListMLE)是对 ListMLE 的进一步推广与改进,为式(9)中每一求和项添加了重要性系数 $\alpha(i)$,其中 $\alpha(i)$单调递减。

3.4 成对排序损失

RankNet[51] 是一种开创性的排序学习算法,它将排序问题转化为一系列成对的分类问题。本文创新性地将其引入到无监督语句嵌入学习中,通过优化任意两个文本的相对顺序,以提升整体的排序列表质量。

对于任意一对文本(x_i,x_j),基于它们的语义关联程度,定义偏好关系 $S_{ij}\in\{-1,0,1\}$:

$$S_{ij}=\begin{cases}1, & \text{如果 } \mathrm{sem}(q,x_i) > \mathrm{sem}(q,x_j)\\ -1, & \text{如果 } \mathrm{sem}(q,x_j) > \mathrm{sem}(q,x_i)\\ 0, & \text{其他情况}\end{cases} \tag{10}$$

RankNet 对文本对的偏序关系进行概率化建模。首先，定义目标概率 $\bar{P}_{ij}$：

$$\bar{P}_{ij}=\frac{1}{2}(1+S_{ij}) \tag{11}$$

接着，通过 Sigmoid 函数 $\sigma(\cdot)$来预测 x_i 比 x_j 更相关的概率 $\hat{P}_{ij}$：

$$\begin{aligned}\hat{P}_{ij}(s_q(x_i),s_q(x_j))&=\sigma(s_q(x_i)-s_q(x_j))\\&=\frac{1}{1+e^{-\alpha(s_q(x_i)-s_q(x_j))}}\end{aligned} \tag{12}$$

RankNet 采用二元交叉熵(Binary Cross-Entropy)作为其损失函数 C_{ij}，用以衡量目标概率与模型预测概率之间的差异：

$$C_{ij}=-\bar{P}_{ij}\log\hat{P}_{ij}-(1-\bar{P}_{ij})\log(1-\hat{P}_{ij}) \tag{13}$$

对于一个查询中所有具有明确偏序关系的文本对，总的损失函数是所有单个文本对损失之和：

$$C=\sum_{(i,j):S_{ij}\neq 0}C_{ij} \tag{14}$$

为了通过梯度下降法优化嵌入模型 f 的参数，需要计算损失函数 C 关于模型得分 $s_q(x_k)$的梯度。对于一个特定的文本对(x_i,x_j)，损失 C_{ij} 关于相似度得分的偏导数为

$$\frac{\partial C_{ij}}{\partial s_q(x_i)}=-\frac{\partial C_{ij}}{\partial s_q(x_j)}=\alpha(\hat{P}_{ij}-\bar{P}_{ij}) \tag{15}$$

在排序列表较长时，排序靠后的文本对相比于靠前的文本对重要性显著更低。LambdaRank[52] 是对 RankNet 的一种改进，旨在更直接地优化那些不可导或难以直接优化的排序评价指标，其梯度定义为

$$\begin{aligned}\lambda_{ij}^{\text{LambdaRank}}&=\frac{\partial C^{*}}{\partial(s_q(x_i)-s_q(x_j))}\\&\approx\lambda_{ij}^{\text{RankNet}}\cdot|\Delta Z_{ij}|\\&=\frac{-\alpha}{1+e^{\alpha(s_q(x_i)-s_q(x_j))}}\cdot|\Delta Z_{ij}|\end{aligned} \tag{16}$$

其中，C^{*} 表示任意损失指标，$|\Delta Z_{ij}|$为交换 x_i 和 x_j 的位置对该损失指标造成的改变量。

4 RankRefine：高效排序学习降噪方案

在列表排序学习中，噪声与效率是两项亟待解决的核心挑战。噪声主要源于两个方面：一是数据构造不当，导致排序结果缺乏实际意义，二是在不同语义粒度下，损失的重要性难以量化。设计不合理的损失函数不仅削弱了模型区分有效梯度的能力，还显著降低训练效率。

为应对上述问题，本文提出基于语义的共识采样与数据重组策略，以及基于成对比较的加权排序损失。两者分别从数据与算法两个层面协同提升排序鲁棒性与训练效率。

图 1 展示了本文的数据预处理与模型训练架构，由以共识采样为核心的数据预处理模块(见 4.1 节)和融合对比学习与加权排序损失的训练模块(见 4.2 节)两部分构成。此外，本文引入自适应容错策略以增强难负例采样效果(详见 4.3 节)。

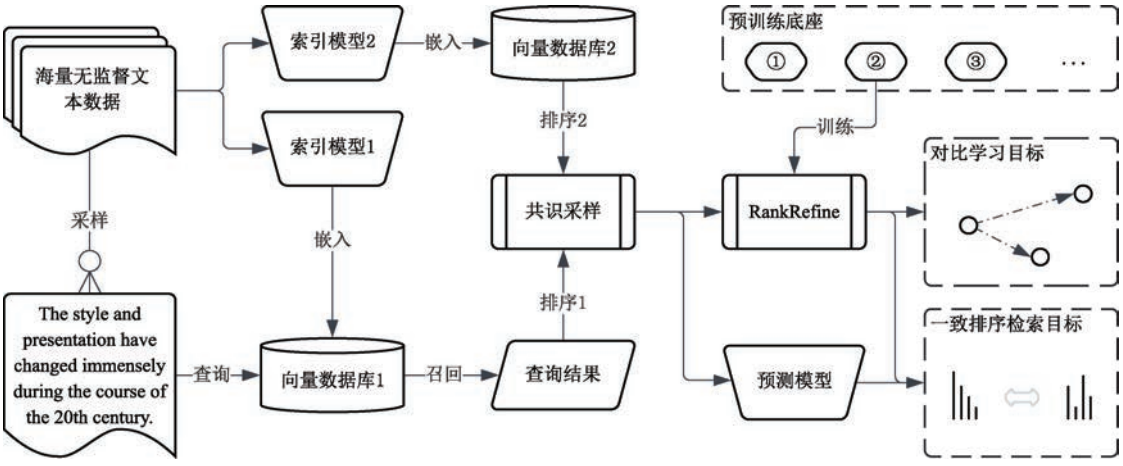

图 1 RankRefine 算法的整体架构图线

4.1 基于语义的共识采样与数据重组

现有的无监督语句嵌入学习方法在构造数据批次时通常采用随机采样策略[19-24,37-38]，即在整个训练数据集上进行无放回的随机遍历。由于语义空间的稀疏性，这种策略对于仅使用 InfoNCE 及其变体的方法而言是充分的。

然而，随机采样策略导致数据批次中语义相似度较高的文本极为稀少，批次内文本涵盖截然不同

的话题与结构,缺乏可比性。这种情况导致排序结果包含严重噪声,降低了参考价值。

在此背景下,为提高批次内排序结果的参考价值,通过保证批次内文本内容一定程度相关性进行困难样本的构造是一个很好的选择。在无监督前提下,借助已有的模型进行相似文本的召回和排序有助于实现这一目标,但是该方案依然面临着用于召回和排序的模型本身能力限制带来的噪声问题。

针对此问题,本文提出基于共识序列提取的采样与数据重组策略,通过综合多个索引模型的召回和排序结果,提取共识排序信息,以提升批次内文本相关性并增强排序可信度。

具体而言,本文首先使用训练好的索引模型对无监督数据进行语句嵌入,并建立向量索引[53]。基于多个索引模型构建的文本索引,采用共识采样策略构造数据批次。

如图 2 所示,本文将不同索引模型的检索结果经动态规划对齐和共识采样后,生成高置信度的排序序列,并重组得到训练批次。

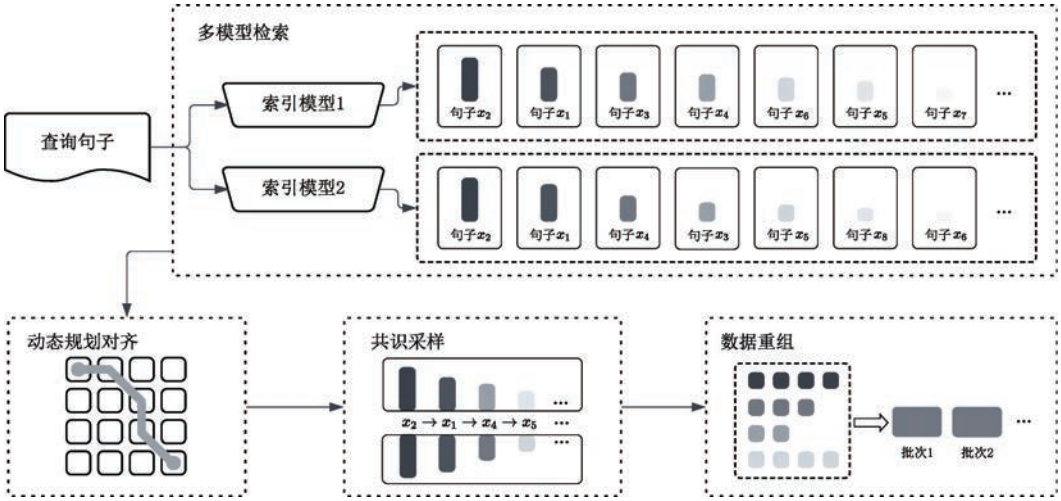

图 2　不同数据比例下 STS 平均表现

本文首先在一个索引模型 T_1 视角下,检索距离目标语句 x_q 最近的 k 段文本,满足和目标语句间语义关联性的递减关系:

$$\phi(f^{T_1}(x_{q,1}^{T_1}),f^{T_1}(x_q)) \geqslant \cdots \geqslant \phi(f^{T_1}(x_{q,k}^{T_1}),f^{T_1}(x_q)) \tag{17}$$

而后由另一个不同的索引模型 T_2 对检索得到的顺序序列进行嵌入,得到语义相似度计算结果:

$$\{\phi(f^{T_2}(x_{q,1}^{T_1}),f^{T_2}(x_q)),\cdots,\phi(f^{T_2}(x_{q,k}^{T_1}),f^{T_2}(x_q))\} \tag{18}$$

算法 1. 最长一致降序列算法

输入:由模型 T_2 嵌入得到的相关性序列 arr,长度 n

输出:序列中的最长一致降序列 seq

1. 初始化长度记录和前驱索引:
2. 创建数组 $len[0\cdots n-1]$和 $prevIdx[0\cdots n-1]$。
3. $maxLen \leftarrow 0, maxIdx \leftarrow -1$。
4. FOR $i \leftarrow 0$ TO $n-1$ DO
5. $len[i] \leftarrow 1$
6. $prevIdx[i] \leftarrow -1$
7. END FOR
8. 动态规划计算最长降序子序列长度:
9. FOR $i \leftarrow 1$ TO $n-1$ DO
10. FOR $j \leftarrow 0$ TO $i-1$ DO
11. IF $arr[j] \geqslant arr[i]$ AND $len[j]+1 > len[i]$ THEN
12. $len[i] \leftarrow len[j]+1$
13. $prevIdx[i] \leftarrow j$
14. END IF
15. END FOR
16. END FOR
17. 找到最长子序列的长度及结束索引:
18. FOR $i \leftarrow 0$ TO $n-1$ DO
19. IF $len[i] > maxLen$ THEN
20. $maxLen \leftarrow len[i]$
21. $maxIdx \leftarrow i$
22. END IF
23. END FOR
24. 根据前驱索引数组重建降序子序列:
25. 创建数组 $seq[0\cdots maxLen-1]$。
26. $curIdx \leftarrow maxIdx$
27. $curPos \leftarrow maxLen-1$
28. WHILE $curIdx \neq -1$ DO
29. $seq[curPos] \leftarrow arr[curIdx]$

30. $curIdx \leftarrow prevIdx[curIdx]$
31. $curPos \leftarrow curPos-1$
32. END WHILE
33. RETURN *seq*

如算法 1 所示，通过动态规划寻找跨索引模型的最长一致降序子序列，确保排序结果在多个语义视角下均成立：

$$\begin{aligned}&\{x_{q,d_1}^{T_1}, x_{q,d_2}^{T_1}, \cdots, x_{q,d_t}^{T_1}\},\\&\text{s.t.}\\&\phi(f^{T_1}(x_{q,d_1}^{T_1}), f^{T_1}(x_q))\\&\quad \geqslant \cdots \geqslant \phi(f^{T_1}(x_{q,d_t}^{T_1}), f^{T_1}(x_q)),\\&\phi(f^{T_2}(x_{q,d_1}^{T_1}), f^{T_2}(x_q))\\&\quad \geqslant \cdots \geqslant \phi(f^{T_2}(x_{q,d_t}^{T_1}), f^{T_2}(x_q))\end{aligned} \tag{19}$$

算法 1 的时间复杂度为 $O(n^2)$，辅助空间复杂度为 $O(n)$。对于本文实验的数据规模，单次运行耗时仅为微秒级，影响甚微。

上述方法具有良好的拓展性，提升索引的总文本量，可以进一步提升采样的批次质量。记上述过程采样得到的一致降序文本组成的集合为 $R_q=\{x_q, x_{q,d_1}^{T_1}, \cdots, x_{q,d_t}^{T_1}\}$，通过不放回采样，可以得到若干这样的数据集合 $R_{q_1}, R_{q_2}, \cdots, R_{q_p}$，最终将这些集合作为训练数据。

4.2 基于成对方法的加权排序损失设计

现有无监督语句嵌入方法通常采用基于负对数似然的列表排序损失（如 ListMLE）进行训练。然而，这类方法在实践中存在两个主要问题：一是其梯度中往往夹杂大量噪声，难以有效捕捉细粒度的语义差异；二是随着批次规模的增大，梯度范数迅速放大，导致模型更新不稳定，严重时甚至破坏模型结构。正如 RankCSE[23] 的消融实验所示，ListMLE 的这一缺陷严重限制了 InfoNCE 的作用。其实验表明，移除对比损失对性能几乎无影响，这说明模型无法通过对比损失有效学习新特征。

为更深入地分析这种不稳定性，首先考察 ListMLE 损失函数的梯度。令 $y_i=s_q(x_{\pi_i^*})$，其梯度可表示为

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{ListMLE}}}{\partial y_i}&=\frac{\partial}{\partial y_i}\sum_{k=1}^{N}\left(\log\sum_{j=k}^{N}\exp(y_j)-y_k\right)\\&=-1+\sum_{k=1}^{i}\frac{\exp(y_i)}{\sum_{j=k}^{N}\exp(y_j)}\end{aligned} \tag{20}$$

从式(20)可直观地看出，当 N 显著增大时，其梯度未必收敛。在无监督学习的常见场景中，批次内往往包含大量与查询语义无关的文本，此时梯度的不收敛意味着模型很难在这些文本上学习到有效的区分性特征。为了严谨说明这一问题，将在下文给出式(20)对应不收敛性的严格数学证明，从而论证改进排序损失的必要性与合理性。

定理 1. 对于采用 ListMLE 损失函数训练的排序模型，在满足批次中排序靠后的 M 个文本的预测得分 $y_j, \forall j\in[N-M+1, N]$ 独立同分布于任意分布 $\mathcal{D}_f$ 的条件下，当训练批次大小 $N\to\infty$ 且无关文本数量 $M=\alpha N$（其中 $0<\alpha<1$ 为常数）时，若 $N-R$ 为固定的小常数（即 R 为排序接近列表末尾的无关文本），则损失函数对 y_R 的偏导数的期望下界以 $\Theta(\ln N)$ 的速率随 N 增加而发散。

证明. 重点分析 ListMLE 损失函数 $\mathcal{L}_{\text{ListMLE}}$ 对 y_R（$R\in[N-M+1, N]$）的期望梯度，有

$$\frac{\partial \mathcal{L}_{\text{ListMLE}}}{\partial y_R}=-1+\sum_{k=1}^{R}\frac{\exp(y_R)}{\sum_{j=k}^{N}\exp(y_j)} \tag{21}$$

定义 $S_{k,N}=\sum_{j=k}^{N}\exp(y_j)$，则上述梯度表达式可进一步拆分为

$$\sum_{k=1}^{R}\frac{\exp(y_R)}{S_{k,N}}=\underbrace{\sum_{k=1}^{N-M}\frac{\exp(y_R)}{S_{k,N}}}_{\mathcal{P}_R^1}+\underbrace{\sum_{k=N-M+1}^{R}\frac{\exp(y_R)}{S_{k,N}}}_{\mathcal{P}_R^2} \tag{22}$$

对于第二部分和式 $\mathcal{P}_R^2$，由于区间 $k\in[N-M+1, R]$ 中所有文本的得分 y_j 均独立同分布于 $\mathcal{D}_f$：

$$\begin{aligned}\mathbb{E}\left[\frac{\exp(y_j)}{S_{k,N}}\right]&=\frac{1}{N-k+1}\sum_{i=k}^{N}\mathbb{E}\left[\frac{\exp(y_i)}{S_{k,N}}\right]\\&=\frac{1}{N-k+1}\mathbb{E}\left[\frac{\sum_{i=k}^{N}\exp(y_i)}{S_{k,N}}\right]\\&=\frac{1}{N-k+1}\end{aligned} \tag{23}$$

令 $p=N-k+1$，则当 $k=N-M+1$ 时，$p=M$；当 $k=R$ 时，$p=N-R+1$，于是可推导得到 $\mathcal{P}_R^2$ 的期望：

$$\begin{aligned}\mathbb{E}[\mathcal{P}_R^2]&=\sum_{k=N-M+1}^{R}\frac{1}{N-k+1}\\&=\sum_{p=N-R+1}^{M}\frac{1}{p}\\&=H_M-H_{N-R}\end{aligned} \tag{24}$$

其中，$H_M=\sum_{p=1}^{M}\frac{1}{p}$ 表示第 M 个调和数，约定 $H_0=0$。注意到 $\mathcal{P}_R^1$ 显然非负，因此梯度期望下界为

$$\mathbb{E}\left[\frac{\partial \mathcal{L}_{\text{ListMLE}}}{\partial y_R}\right]=-1+\mathbb{E}[\mathcal{P}_R^1]+H_M-H_{N-R}\geqslant -1+H_M-H_{N-R} \tag{25}$$

当批次大小 $N\to\infty$ 且 $M=\alpha N$(其中 $0<\alpha<1$ 为常数)时,有调和数近似关系 $H_M\approx \ln M+\gamma$(γ 为欧拉-马斯克若尼常数)。在此情形下,若 $N-R$ 为固定的小常数(即 R 靠近列表末尾),则梯度将以 $\Theta(\ln N)$的速率随 N 增加而发散。

证毕.

定理 1 表明,在使用 ListMLE 作为损失函数时,当训练批次较大且包含大量得分相近的无关文本时,位于排序列表末尾的文本梯度范数可能显著增大。这些末尾文本由于难以区分,排序结果通常带有较强的随机性。梯度的这种发散性会放大噪声水平,从而削弱模型训练的稳定性与效率。

除梯度问题外,批次内数据分布也是损失设计的关键因素。语义结构具有多向发散特性,对不同发散方向的样本强制全局排序缺乏实际意义。因此,训练应优先选择模型间一致性高、排序明确的文本对。然而 ListMLE 受 Plackett-Luce 模型严格假设限制,无论采用截断还是 p-ListMLE 的加权形式,损失梯度都会剧烈震荡导致训练不稳定。

针对上述问题,受 LambdaRank[52] 启发,本文提出加权成对排序损失。与 RankCSE 等方法随机构建批次并等权处理所有排序关系不同,RankRefine 将多模型交叉验证得到的高置信度共识序列融入损失设计,仅关注排序信息可靠的语句对,从算法层面同时提升数据有效性与训练效率。

定义排序损失,对数据批次中的两个样本 x_i、x_j,以及给定查询 q,相似度得分形式见式(5),则可定义梯度:

$$\lambda_{ij}=\frac{\partial \mathcal{L}_{ij}}{\partial s_q(x_i)}=-\frac{\partial \mathcal{L}_{ij}}{\partial s_q(x_j)} \tag{26}$$

损失函数整体形式如下:

$$\mathcal{L}_{ij,\text{Rank}}=\log[1+\exp(-\alpha(s_q(x_i)-s_q(x_j)))]\cdot\mu_{ij} \tag{27}$$

$$\mathcal{L}_{i,\text{Rank}}=\sum_{j;i\neq j}\mathcal{L}_{ij,\text{Rank}} \tag{28}$$

其中,α 为温度系数,μ_{ij} 为基于共识序列信息调整的截断系数。满足

$$\mu_{ij}=\begin{cases}1, & \text{如果 } j\leqslant \alpha_q\cdot|R_{g(x_i)}|\\ 0, & \text{其他情况}\end{cases} \tag{29}$$

其中,$|R_{g(x_i)}|$ 为 x_i 所在降序列的长度,截断阈值 α_q 较敏感,过小可能限制学习充分的排序信息,过大则可能引入噪声并增加计算量。

式(26)的计算复杂度低于式(20),且充分利用共识采样中的启发信息,不再关心批次内相关性较低样本的排序关系。本文方法利用共识采样,在成对排序与列表排序损失间进行有效折中:避免了列表排序损失的复杂性和计算成本,同时比成对排序损失更关注整体列表优化。如 5.6 节所述,该方法保证了损失和梯度范数在不同批次大小下的稳定性。

最终,本文采用加权混合排序损失和对比损失两部分,定义优化损失函数:

$$\mathcal{L}=\mathcal{L}_{\text{InfoNCE}}+\gamma\cdot\mathcal{L}_{\text{Rank}} \tag{30}$$

4.3　RankRefine+:自适应容错列表排序采样策略

本文提出的共识采样方法能有效提升排序信息的可靠性。然而,索引模型存在固有误差,过于严格的采样策略可能导致细微差异文本被错误剔除、采样序列长度缩短等问题。针对上述问题,本文设计了 RankRefine+方法。该方法结合文本数据特性,采用自适应容错列表排序采样策略。这一策略对排序的细微差异具有一定容忍度,能够在保留有效信息的同时提升数据质量。

容错阈值的设计可类比为独立排序过程。理想情况下,具有共同话题的排序应保持一致,但实际中需允许合理差异存在。自适应容错机制通过动态设定容忍边界处理这种差异,其中 *base_delta* 是边界设定的关键参数。

算法 2. 动态规划实现的容错最长降序列算法

输入:序列分数 *scores*,基础容忍范围 *base_delta*,最大上升次数 *max_up*。

输出:布尔掩码数组 *keep_mask*,指示保留的元素。

1. $n\leftarrow$length(*scores*)。
2. 初始化动态规划表:$dp[i][u]$表示 i 结尾、消耗 u 次上升的最长子序列长度
3. 创建二维数组 $dp[0\cdots n-1][0\cdots max_up]$并初始化为 1,4. 创建二维数组 $prev[0\cdots n-1][0\cdots max_up]$并初始化为$-1$。
5. 填充动态规划表:
6. FOR $i\leftarrow 0$ TO $n-1$ DO
7. $delta_i\leftarrow base\_delta\times\left(1-\frac{i}{n}\right)$
8. FOR $j\leftarrow 0$ TO $i-1$ DO
9. $sc_j\leftarrow scores[j]$, $sc_i\leftarrow scores[i]$
10. $up\_needed\leftarrow 0$
11. IF $sc_i>sc_j$ AND $(sc_i-sc_j)>delta_i$ THEN
12. $up\_needed\leftarrow 1$
13. END IF

```
        FOR u_prev←0 TO max_up DO
          IF dp[j][u_prev]>0 AND (u_prev + up_needed)≤max_up AND dp[j][u_prev] + 1>dp[i][u_prev + up_needed] THEN
            dp[i][u_prev+up_needed]=p[j][u_prev]+1
            prev[i][u_prev + up_needed]=j
          END IF
        END FOR
      END FOR
    END FOR
    寻找最优解：
    best_len←0,best_i←0,best_u←0
    FOR i←0 TO n−1 DO
      FOR u←0 TO max_up DO
        IF dp[i][u]>best_len THEN
          best_len←dp[i][u]
          best_i←i
          best_u=u
        END IF
      END FOR
    END FOR
    回溯构建保留掩码：
    创建布尔数组 keep_mask[0…n−1]并初始化为 False。
    设 cur_i←best_i,cur_u←best_u。
    WHILE cur_i≠−1 DO
      keep_mask[cur_i]←True
      prev_i←prev[cur_i][cur_u]
      IF prev_i≠−1 THEN
        delta_cur←base_delta×(1−cur_i/n)
        up_needed←0
        IF scores[cur_i]>scores[prev_i] AND (scores[cur_i] −scores[prev_i])>delta_cur THEN
          up_needed←1
        END IF
        cur_u←cur_u − up_needed
      END IF
      cur_i = prev_i
    END WHILE
    RETURN keep_mask
```

算法 2 展示了自适应容错列表排序采样具体实现。考虑到高分文本样本间差异较小，RankRefine+为高分段设定了更宽的容忍区间。区间内的文本不被视为破坏排序单调性。为控制噪声，算法引入最大上升次数参数 max_up 来限制容错范围。当该参数为 0 时，算法退化为严格最长降序列方法。

算法 2 的时间复杂度为 $O(n^2 \cdot max\_up)$，空间复杂度为 $O(n \cdot max\_up)$。在 $max\_up \ll n$ 时，空间占用接近 $O(n)$。

为使容错机制能够自适应地适配不同数据的分布特性，本文从跨模型排序不一致性的结构化视角出发，系统分析相邻逆序扰动的幅度分布特征，并据此给出基础容忍尺度的自适应归一化定义。

设候选项按索引模型 T_1 的相似度得分递减排序为 $\{x_t\}_{t=1}^{n}$，并记 $r_t=r(x_t)$ 为模型 T_2 在该固定顺序下给出的观测得分。定义相邻差分为

$$y_t:=r_{t+1}-r_t,t=1,\cdots,n-1 \tag{31}$$

为保证差分幅度对阈值位置具有加权敏感性，对正向相邻差分引入位置归一化：

$$\widetilde{y}_t=\frac{y_t}{1-\frac{t}{n}},t=1,\cdots,n-1 \tag{32}$$

注意到，若对所有 t 均有 $y_t\leqslant 0$，则序列 $\{r_t\}$ 关于索引 t 单调不增，即模型 T_2 在模型 T_1 所诱导的排序下不存在任何局部逆序。此时两个模型在该候选序列上的排序结果完全一致，无需引入任何容错机制，原始序列可被完整保留。

反之，若存在至少一个 t 使得 $y_t>0$，则正向差分集合非空，定义为

$$\mathcal{D}=\{\widetilde{y}_t \mid \widetilde{y}_t>0,t=1,\cdots,n-1\} \tag{33}$$

基于上述正向差分集合，定义经验生存函数：

$$\hat{S}(\delta)=\frac{1}{|\mathcal{D}|}\sum_{\widetilde{y}\in\mathcal{D}}1\{\widetilde{y}>\delta\},\delta>0 \tag{34}$$

在实践中，$\hat{S}(\delta)$ 在对数尺度下通常呈现由缓慢衰减向快速衰减过渡的形态，该转折位置反映了由随机噪声主导的微小逆序扰动与显著逆序之间的分界。为识别该转折点，定义如下归一化操作。约定若下述归一化分母为零，则将归一化结果定义为 0。

$$\widehat{\log\delta}=\frac{\log\delta-\min_{\delta'\in\mathcal{D}}\log\delta'}{\max_{\delta'\in\mathcal{D}}\log\delta'-\min_{\delta'\in\mathcal{D}}\log\delta'} \tag{35}$$

$$\hat{S}^{\mathrm{norm}}(\delta)=\frac{\hat{S}(\delta)-\min_{\delta'\in\mathcal{D}}\hat{S}(\delta')}{\max_{\delta'\in\mathcal{D}}\hat{S}(\delta')-\min_{\delta'\in\mathcal{D}}\hat{S}(\delta')} \tag{36}$$

据此，定义基础容忍尺度为生存函数与对数尺度差值最大化处对应的阈值。若最优解不唯一，则取其中最小的 δ：

$$base\_delta=\mathrm{argmax}_{\delta\in\mathcal{D}}(\hat{S}^{\mathrm{norm}}(\delta)-(1-\widehat{\log\delta})) \tag{37}$$

容错列表排序采样后，每个查询通常对应两个不同索引模型生成的排序序列。因此，同一备选项

可能在多个序列中重复出现。为充分利用有限训练资源并减少冗余,RankRefine+采用贪心最大覆盖(Greedy Maximum Coverage)[54]策略从中选择最优子集。该问题可形式化为:给定集合族$\mathcal{S}=\{S_1,S_2,\cdots,S_m\}$,其中每个$S_i$包含对应排序序列中的文本元素,目标是给定整数$k$,在约束$|\mathcal{C}|\leqslant k$下,选择子集$\mathcal{C}\subseteq\mathcal{S}$,使得并集$\bigcup_{S\in\mathcal{C}}S$的基数最大。

虽然最大覆盖问题在一般情况下是NP困难的,但贪心算法能够提供$(1-1/e)$近似保证[55]。本文采用基于堆的高效贪心实现,每轮迭代选择覆盖增益最大的序列并动态更新增益值。

该策略的时间复杂度为$O(m\log m\cdot n)$,其中m为候选序列数,n为平均序列长度。由于本文框架下m和n都相对较小,计算开销可忽略不计。

5　实验结果与分析

为全面验证所提算法的有效性,本节依次介绍实验环境、数据集、实验设置、训练效率、损失函数分析与案例分析;性能评估涵盖STS任务、迁移学习任务与重排序任务;最后通过消融实验检验各组件及低资源条件下的表现,并在基于大语言模型的嵌入模型微调实验中验证方法的可扩展性。

5.1　实验环境介绍

本文实验环境配备了32核中央处理器(CPU),基准频率2.7 GHz,512 GB DDR4内存。图形计算单元(GPU)使用了一块NVIDIA A100显卡(40 GB显存,PCIE接口)。

5.2　实验数据集设置

语义文本相似度(Semantic Textual Similarity,STS)任务旨在评估模型捕捉文本对之间语义相关性的能力。相关数据集通常由成对的句子构成,并配有人工标注的相似度分数,标签范围为0-5的整数。本文选取了以下主流评测数据集,STS12-16[56-60]和STS Benchmark[61]源自历年SemEval语义评测任务,是英文语义相似度评测的标准基准。这些数据集的文本来源广泛,涵盖新闻标题、图片描述和论坛帖子等多种文本类型。SICK-Relatedness[62](Sentences Involving Compositional Knowledge)数据集则专注于评估模型对句子结构和组合语义的理解能力。该数据集的句子对蕴含丰富的词法、句法和语义现象,为模型的深层语义理解能力提供了更具挑战性的测试场景。在评测过程中,本文遵循与SimCSE、RankCSE等主流方法一致的配置:使用SentEval[63]工具库比较模型嵌入结果间的余弦相似度与人工标注标签的一致性,使用Spearman相关系数[64]作为最终的评价指标。

Spearman相关系数ρ是一种基于等级的非参数统计方法,用以衡量模型预测得分与人工标注得分间单调关系的强弱,取值范围为$[-1,1]$。记等级差为d_i,样本对数量为n,其计算公式为

$$\rho=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}\tag{38}$$

ρ越接近+1表示排序一致性越高。

5.3　实验设置

为确保实验的公平性和可比性,本文采用与SimCSE[19]相同的训练数据集——维基百科数据集(Wiki 1M)。实验在BERT[27]和RoBERTa[28]两种主流预训练语言模型的各个尺寸版本上进行微调,以验证所提方法的通用性和适应性。

在数据预处理阶段,本文参考RankCSE[23]的配置策略,选用DiffCSE-BERT$_{\text{base}}$和SimCSE-BERT$_{\text{large}}$作为索引模型。这两个模型相对简单且性能适中,适合用于初步筛选。为确定最优的初始召回数量k,本文开展预实验进行分析。图3展示了不同k值下的召回采用率曲线。横坐标表示召回文本的排序位置,纵坐标表示该位置文本通过共识过滤后的最终采纳概率。

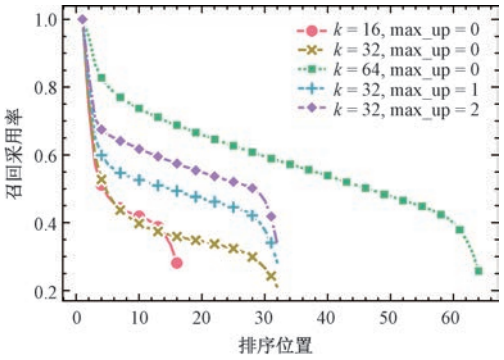

图3　不同共识采样参数下的召回采用率曲线

实验结果揭示了排序位置与采用率之间的显著负相关性。排序靠前的文本始终表现出较高的采用率,这验证了共识采样在高质量样本识别方面的有效性。当k从16增至32时,候选序列长度明显增加,但采用率曲线形态保持稳定。然而,当k增至64时,曲线斜率显著降低,表明低相关性样本的采纳概率上升,可能引入噪声并增加计算开销。综合考虑样本质量、噪声控制和计算效率,本文选定$k=32$作为后续实验的标准参数。该设置能够确保充

足的高质量候选样本，同时有效控制噪声影响并维持合理的计算成本。

数据预处理是一次性的离线操作。在本文的实验配置下，整个过程耗时约 30 分钟，其中主要时间消耗在使用索引模型计算数据嵌入。生成的排序数据被存储并用于所有后续的模型训练实验。

本文在排序学习阶段所采用的预测模型及权重选取方案见表 1。部分超参数沿用相关工作的默认设置，包括：对比损失的温度系数 $\tau=0.05$，损失加权参数 $\gamma=1.0$，平均批次大小为 128。RankRefine+方法的基础参数 $base\_delta$ 设为 0.004，该值由数据特征计算获得（参见节 4.3），max_up 取 1，学习率与截断阈值 α_q 的取值见表 2。

表 1 标签预测模型设置

模型底座	预测模型
$\text{BERT}_{\text{base}}$	$\frac{1}{3}\text{SimCSE-BERT}_{\text{large}}+\frac{2}{3}\text{SynCSE-BERT}_{\text{base}}$
$\text{BERT}_{\text{large}}$	$\frac{1}{3}\text{DiffCSE-BERT}_{\text{base}}+\frac{2}{3}\text{SynCSE-BERT}_{\text{large}}$
$\text{RoBERTa}_{\text{base}}$	$\frac{1}{3}\text{SimCSE-BERT}_{\text{large}}+\frac{2}{3}\text{SynCSE-RoBERTa}_{\text{base}}$
$\text{RoBERTa}_{\text{large}}$	$\frac{1}{3}\text{SimCSE-BERT}_{\text{large}}+\frac{2}{3}\text{SynCSE-RoBERTa}_{\text{large}}$

注：表格中的“预测模型”部分由模型名称与模型权重组成，如“$\frac{1}{3}\text{SimCSE-BERT}_{\text{large}}+\frac{2}{3}\text{SynCSE-BERT}_{\text{base}}$”表示在预测模型中，模型 $\text{SimCSE-BERT}_{\text{large}}$ 的预测结果权重占比为 $\frac{1}{3}$，而模型 $\text{SynCSE-BERT}_{\text{base}}$ 的预测结果权重占比则为 $\frac{2}{3}$”表示在预测模型中，模型 $\text{SimCSE-BERT}_{\text{large}}$ 的预测结果权重占比为 $\frac{1}{3}$，而模型 $\text{SynCSE-BERT}_{\text{base}}$ 的预测结果权重占比则为 $\frac{2}{3}$。

表 2 超参数设置

模型底座	$\text{BERT}_{\text{base}}$	$\text{BERT}_{\text{large}}$	$\text{RoBERTa}_{\text{base}}$	$\text{RoBERTa}_{\text{large}}$
RankRefine 学习率	3×10^{-5}	2×10^{-5}	3×10^{-5}	2×10^{-5}
RankRefine+ 学习率	3×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
α_q	2.0	4.0	2.0	4.0

在模型训练过程中，本文采用了分阶段保存与评估策略：每间隔 500 步对模型进行一次保存，并在验证集上进行性能测评。最终，选取在语义文本相似度任务和迁移学习任务验证集上综合表现最优的模型和超参数配置作为最终实验结果。

本文将提出的模型与若干经典的无监督语句嵌入方法进行了对比，包括且不限于 SimCSE[19]、DiffCSE[20]、PCL[65]、PromptBERT[66]等经典对比学习方法，使用排序学习的 RankCSE，使用合成数据的 DebCSE[36]、SynCSE[37]、MultiCSR[38]。本文实验基线方法的选择确保了与该领域最先进方法进行公平比较，其中使用的所有数据集、索引模型、预测模型均为无监督的，以充分确保对比的公平性。

本文在不同任务上基线方法的选择有所差异。在 STS 任务上，本文选择了涵盖对比学习、排序学习和针对数据层面进行无监督强化等各种技术路线的代表性方法进行全面对比；在迁移学习任务和重排序任务上，由于部分方法闭源且在原始论文中缺少该模块，本文主要选择在此任务上开源先进方法进行复现与对比，以验证方法的有效性。

5.4 训练效率分析

为评估训练效率，本文选取在损失函数设计上同样采用排序学习的 RankCSE 作为主要对比基线，以保证比较的公平性。本节将重点分析本文引入的一次性数据预处理步骤对整体训练效率的增益。

根据表 3 的数据，本文所提出的方法在训练前需进行约 30 分钟的离线预处理。其时间消耗集中在文本嵌入与召回阶段。通过高效的 Faiss[53]向量检索库，索引构建过程较为高效且高度可拓展。而本文提出的核心共识采样算法复杂度极低，无论是 RankRefine 还是 RankRefine+在该规模的数据上都仅需数十秒，相比于其他部分复杂度可以忽略不计。这种设计确保了整个预处理流程的可扩展性，使其能无缝应用于更大规模的语料库。

表 3 模型训练效率对比

	RankCSE		RankRefine		RankRefine+	
	base	large	base	large	base	large
轮次数目	4	4	1	1	1	1
每轮次耗时	30 min	55 min	25 min	48 min	18 min	40 min
预处理时长	—	—	30 min	30 min	30 min	30 min
总耗时	120 min	220 min	55 min	78 min	48 min	70 min

这一过程仅需执行一次，即可在后续的多模型训练或超参数搜索中被完全复用，从而摊销其开销。在显存需求与训练数据规模保持一致的条件下，RankRefine 仅用 1 个轮次即可超过 RankCSE 在 4 个轮次下的效果，总训练时长缩短约 50%～65%。

在此基础上，RankRefine+进一步优化了合成批次的覆盖率与数据利用率：其批次构造策略能够在同等批次规模下覆盖更多互斥样本对，并显著提高有效梯度更新的样本比例。这种优化使得模型在

更少的参数更新步数中即可充分利用训练数据,将单轮训练时间进一步降低。因此本文提出的方法尤其适用于需要高频迭代、快速验证及大规模超参搜索的实验场景,可显著提升整体训练效率。

5.5　案例分析

为直观展示所提出框架的优越性,本节选取一个典型案例,分析单一模型在处理复杂语义时易出现的"主题漂移"风险,并说明本文提出的共识采样机制如何利用交叉验证对结果进行纠偏与降噪,从而获得高质量训练数据。

案例选取索引 868284 的样本作为分析对象,其查询句(Query)为:"Bay Village and The Village of Indian Hill are cities despite the word "Village" in their names, and 16 villages have "City" in their names."。该查询的核心意图在于探讨地理行政区划的官方名称与法律地位之间的差异,这对模型语义理解能力提出了较高要求。

首先考察索引模型 SimCSE 的初步召回结果(表 4),该模型在该例中未能准确捕捉核心语义,出现了明显的主题漂移现象。从表中可以看出,SimCSE 虽然识别了地名概念,但将相关性错误地关联到汽车领域。例如,序号 2、3、7、8 等语句均涉及"日产(Nissan)""GT-R 赛车"等内容。这类无关信息的引入不仅无法有效支撑任务目标,还会造成训练数据的语义污染,削弱模型学习真实任务模式的能力。同时,与主题高度相关的语句(如序号 4 和 19)反而被排在部分不相关样本之后,进一步暴露了依赖单一模型构造数据的固有风险。

接下来,展示同一查询在另一个索引模型 DiffCSE 下的召回结果,如表 5 所示。

表 4　单一模型召回出现的主题漂移现象

序号	SimCSE 召回文本	DiffCSE 打分
0	Bay Village and The Village of Indian Hill are cities despite the word "Village" in their names…	1.0000
…	…	…
2	Following a disappointing performance in the 2015 24 Hours of Le Mans, the program's remaining schedule…	0.9566
3	Nismo has also developed production class Nissan GT-R cars for the 24 Hours of Nürburgring.	0.9430
4	List of cities in Ohio	0.9458
…	…	…
7	The engine is an RB28DETT Z2 (a normal GT-R engine with a stroked displacement of 2.8 liters…	0.9480
8	Nissan announced in June 2014, that Nismo will enter the LMP1 category to fight for the FIA World Endurance Championship…	0.9308
…	…	…
19	Smaller municipalities are villages.	0.9477

注:灰色高亮行表示与查询句核心语义无关,出现了向"赛车运动"的主题漂移。

表 5　共识采样策略对数据的过滤情况

序号	DiffCSE 召回	SimCSE 打分	RankRefine	RankRefine+
0	Bay Village and The Village of Indian Hill are cities despite the word "Village" in their names…	1.0000	保留	保留
…	…	…	…	…
4	Smaller municipalities are villages.	0.7885	丢弃↑	保留
5	She commenced operations there in January 1943 with bombardments of …	0.7669	丢弃	丢弃
6	List of cities in Ohio	0.8496	丢弃↑	保留
…	…	…	…	…
9	Shortly after midnight on 4-5 July, she participated in the bombardment of Vila and Bairoko Harbor…	0.8113	保留	保留

注:灰色高亮行是与查询句话题相关但语义相关性弱的样本;↑表示 SimCSE 打分的非递减项。

对比之下,DiffCSE 的召回结果(表 5)在本例中相对优于 SimCSE。其中的多个语句(如序号 4 和 6)与"行政区划"主题高度契合,初步召回样本的相关性明显提升。这表明,不同预训练模型由于训练方法与数据分布差异,捕捉的语义特征具有差异性与互补性,这种差异为多模型协同提供了基础,但该增益效应难以在随机批次中完全体现。

然而,即便是在 DiffCSE 的召回结果中,仍可观察到一定的排序分值波动,如序号 4 和 6 的得分高于部分前序样本,同时仍夹杂少量无关语句(如序号 5 和 9)。针对这种情况,本文提出的共识采样策略在初步召回基础上进行了二次降噪:通过寻找最长一致递减子序列,过滤与主题连贯性不符的样本,例如剔除序号 5;同时识别出排序位置存在波动但与

主题高度相关的语句，如序号 4 和 6，并借助容错机制予以保留，从而在确保语义精准性的同时，保持了训练集的信息丰富性。RankRefine＋相比之下保留了更多话题相关但实际语义缺乏关联的样本，即许多有监督方法致力构造的“难负例”，这进一步提升了采样得到的批次质量。

综上所述，该案例表明单一模型存在固有的语义理解偏差，容易导致主题漂移并引入噪声数据，最终影响训练效果。虽然模型间的系统性偏差相似时可能限制降噪效果，但多模型共识框架通过整合不同模型的互补信息，仍能有效抑制偏差并提升数据质量。相比单模型或随机采样方法，该框架在监督信号质量上具有显著优势，同时将训练数据规模控制在合理范围内，为性能提升奠定了基础。

5.6 损失函数分析

实验结果（图 4）进一步印证了定理 1 基于梯度公式（20）所揭示的梯度发散规律。

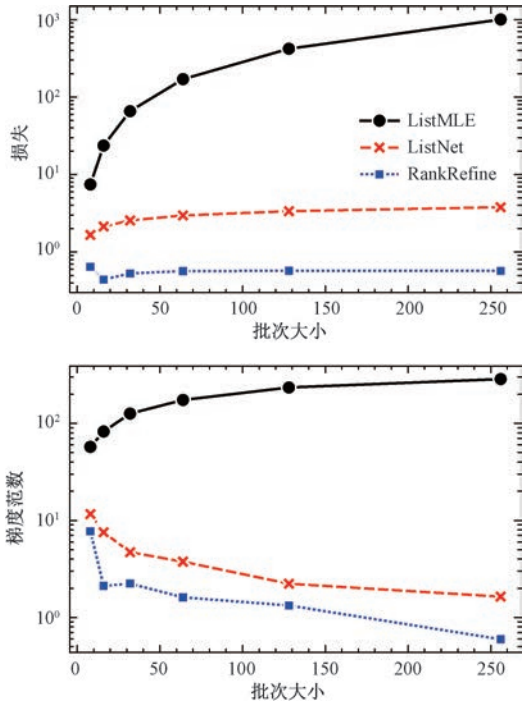

图 4 不同方法的损失值与梯度范数随批次大小变化对比

随着批次大小 N 的增加，ListMLE 与 ListNet 的损失值均持续上升，其中 ListMLE 的损失几乎呈指数增长，梯度范数也随批次规模显著放大。大批次训练中，这种梯度异常放大会导致模型在局部标签模式上产生过拟合。

相比之下，基于式（30）定义的 RankRefine 损失在不同批次规模下的损失值与梯度范数均保持近乎平稳，不会因批次扩容而引发数值激增。这一结果直观验证了本文在 4.2 节中提出的加权成对损失策略能有效抑制梯度发散，并与理论分析保持一致。此外，由于式（26）中的截断权重 μ_{ij} 能过滤语义相关性较低的排名对，RankRefine 在提升优化稳定性的同时，也增强了模型的鲁棒性与泛化能力。

5.7 语义文本相似度实验结果

表 6 展示了各方法在语义文本相似度任务上的性能对比。RankRefine 和 RankRefine＋在全部四种模型配置下均取得了最优或次优性能。为验证这种性能提升的统计显著性，本文采用成对 t 检验（paired t-test），对 RankRefine 与 RankRefine＋相对于各基线方法在所有 STS 任务上的性能差异进行检验。结果表明，在显著性水平 $p<0.05$ 下，性能提升均具有统计学意义，验证了方法改进的可靠性。在基准测试中，RankRefine 展现出稳定的性能优势。以 $\text{BERT}_{\text{base}}$ 为例，该方法达到 81.55％的平均 Spearman 相关系数，相比 SimCSE 基线提升 5.30 个百分点。这一提升幅度在四种模型配置中保持一致，表明本文方法具有良好的泛化能力。与采用列表排序学习的 RankCSE 相比，RankRefine 的平均性能提升达 2.72 个百分点。该结果表明，优化批次采样与损失截断能进一步捕捉细粒度的文本语义差异。RankRefine＋进一步提升了框架性能。通过引入算法 2 中的自适应容错机制，在无监督的前提下，该方法在所有模型配置上超越了有监督的 SimCSE 方法。容错机制的核心在于动态调整容忍阈值，如公式（37）所示。该设计允许算法根据数据分布特征自适应地平衡噪声过滤与信息保留。实验结果表明，RankRefine＋在保持高性能的同时，训练效率同样取得了显著提升。

容错机制的引入产生了复杂的性能影响。在多数情况下，RankRefine＋超越或持平 RankRefine 的性能。然而，在特定任务上出现了轻微性能下降，如 $\text{BERT}_{\text{large}}$ 在 STS12 上从 77.33％降至 77.19％。这种现象可通过算法 2 的工作原理解释：参数 max_up 允许序列包含有限的非单调元素，可能引入少量噪声。尽管如此，整体性能与训练效率的提升证明了这一设计权衡的合理性。

值得注意的是，RankRefine＋在 $\text{RoBERTa}_{\text{large}}$ 配置下，RankRefine＋相比 RankRefine 提升 0.23 个百分点，高于小规模模型的提升幅度。这表明容错机制与模型自身高质量表征的结合能够产生协同效应。

表 6　STS 任务上的语句表征表现(Spearman 相关系数)

模型底座	方法	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	平均值
$BERT_{base}$	SimCSE†	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
	DiffCSE†	72.28	84.43	76.47	83.90	80.54	80.59	71.23	78.49
	PCL♠	72.84	83.81	76.52	83.06	79.32	80.01	73.38	78.42
	DebCSE†	76.15	84.67	78.91	85.41	80.55	82.99	73.60	80.33
	$RankCSE_{ListMLE}$	75.66	**86.27**	77.81	84.74	81.10	81.80	75.13	80.36
	SynCSE	74.53	82.14	78.22	83.46	80.66	81.42	**80.51**	80.13
	MultiCSR	75.88	82.39	78.80	84.42	80.54	82.23	80.03	80.61
	* RankRefine	76.47	83.41	79.29	85.88	82.17	**83.82**	79.78	81.55
	* RankRefine+	**76.51**	83.69	**79.36**	**85.93**	**82.24**	83.74	79.82	**81.61**
$BERT_{large}$	SimCSE†	70.88	84.16	76.43	84.50	79.76	79.26	73.88	78.41
	PCL♠	74.87	86.11	78.29	85.65	80.52	81.62	73.94	80.14
	DebCSE†	76.82	86.36	79.81	85.80	80.83	83.45	74.67	81.11
	$RankCSE_{ListMLE}$	75.48	**86.50**	78.60	85.45	81.09	81.58	75.53	80.60
	SynCSE	75.23	84.28	79.41	84.89	82.09	83.48	**81.79**	81.60
	MultiCSR	75.56	85.19	80.14	85.91	82.40	84.19	81.65	82.15
	* RankRefine	**77.33**	85.81	80.52	86.49	**83.18**	**85.13**	81.53	82.86
	* RankRefine+	77.19	85.87	**80.68**	**86.62**	83.16	85.07	81.64	**82.89**
$RoBERTa_{base}$	SimCSE†	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
	DiffCSE†	70.05	83.43	75.49	82.81	82.12	82.38	71.19	78.21
	PCL♠	71.13	82.38	75.40	83.07	81.98	81.63	69.72	77.90
	DebCSE†	74.29	85.54	79.46	85.68	81.20	83.96	74.04	80.60
	PromptRoBERTa♣	73.94	84.74	77.28	84.99	81.74	81.88	69.50	79.15
	$RankCSE_{ListMLE}$	73.20	85.95	77.17	84.82	82.58	83.08	71.88	79.81
	SynCSE	76.15	84.41	79.23	84.85	82.87	83.95	**81.41**	81.81
	MultiCSR	77.03	84.72	79.71	85.80	82.68	84.24	80.64	82.12
	* RankRefine	**77.57**	86.04	81.12	86.67	83.41	**85.73**	80.94	83.07
	* RankRefine+	77.51	**86.32**	**81.28**	**86.76**	**83.64**	**85.73**	80.53	**83.11**
$RoBERTa_{large}$	SimCSE†	72.86	83.99	75.62	84.77	81.80	81.98	71.26	78.90
	PCL♠	74.08	84.36	76.42	83.49	81.76	82.79	71.51	79.49
	DebCSE†	**77.68**	87.17	80.53	85.90	83.57	85.36	73.89	82.01
	$RankCSE_{ListMLE}$	73.47	85.77	78.07	85.65	82.51	84.12	73.73	80.47
	SynCSE	75.92	85.01	80.43	85.83	84.40	85.05	**81.99**	82.66
	MultiCSR	74.42	84.46	79.17	84.76	83.67	84.23	81.50	81.74
	* RankRefine	76.70	86.55	81.58	**87.92**	85.24	86.58	81.40	83.71
	* RankRefine+	76.77	**87.24**	**81.95**	87.88	**85.39**	**87.01**	81.37	**83.94**

注:所有结果均基于无监督学习设置。加粗数字表示该任务中的最高性能,下划线标注次高性能。后续表格沿用此标注规范。部分基线闭源故引用了原始论文中的数据,"†":数据来源于文献[36],"♠":数据来源于文献[23],"♣":数据来源于文献[38],"*":本文方法,其余方法的结果均基于其官方开源代码库在相同实验条件下复现所得;其他表同理。

5.8　迁移学习任务

迁移学习任务(TR)旨在通过跨领域的小样本微调,评估句子嵌入在下游分类任务中的泛化能力。本文基于 SentEval[63]工具包,在七个标准分类数据集上进行实验。这些任务覆盖多个领域与任务类型,均依赖语义相似度建模,能够有效检验语义表示的质量:

(1)MR(Movie Review Dataset)[67]:包含电影评论片段,用于二元情感分类;

(2)CR(Customer Review Dataset)[68]:包含各类产品的顾客评论文本,用于二元情感分类;

(3)SUBJ(Subjectivity Dataset)[69]:包含从电影评论和剧情摘要中抽取的文本,用于判断其主观性或客观性;

(4)MPQA(Multi-Perspective Question Answering Opinion Corpus)[70]:包含新闻文章中标注的观点短语,用于观点极性分类;

(5)SST-2(Stanford Sentiment Treebank)[71]:基于电影评论的语句情感分类数据集,任务为二元情感分类;

(6)TREC(Text Retrieval Conference Question Dataset)[72]:包含问题文本,任务是将问题分类到预定义的 6 种类型中;

(7)MRPC(Microsoft Research Paraphrase

Corpus)[73]：包含从新闻源中抽取的文本对，任务是判断两个文本是否在语义上等价。

表 7 的结果显示，RankRefine 在 $RoBERTa_{base}$ 和 $RoBERTa_{large}$ 两种底座模型下均取得了最高平均准确率，分别达到 88.29%和 89.73%，相较 SimCSE 基线模型平均提升+3.45%和+3.62%。在超过一半的任务上（如 CR、SUBJ、SST-2、MRPC），RankRefine 明显优于其他方法，表明其对多领域、多任务具有一致的泛化能力。这种显著提升与第 4.2 节提出的加权成对排序损失设计高度相关。本文方法在捕捉了细粒度语义差异的同时，抑制噪声排序关系的影响，从而增强对源域外样本的鲁棒性。结果表明，该方法显著提升了语义表示的判别性与稳定性。

表 7 迁移任务上的语句表征表现（准确率）

模型底座	方法	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	平均值
$RoBERTa_{base}$	SimCSE	81.04	87.74	93.28	86.94	86.60	84.60	73.68	84.84
	DiffCSE	82.42	88.34	93.51	87.28	87.70	86.60	76.35	86.03
	PCL	81.83	87.55	92.92	87.21	87.26	85.20	76.46	85.49
	$RankCSE_{ListMLE}$	83.53	89.22	94.07	88.97	89.95	**89.20**	76.52	87.35
	SynCSE	85.41	91.44	93.39	89.91	91.21	84.40	76.87	87.52
	* RankRefine	85.42	**91.60**	**94.21**	89.96	**92.09**	87.20	**77.57**	**88.29**
	* RankRefine+	**85.46**	91.44	94.20	**90.25**	90.61	88.60	77.28	88.26
$RoBERTa_{large}$	SimCSE	82.74	87.87	93.66	88.22	88.58	92.00	69.68	86.11
	PCL	84.47	89.06	94.60	89.26	89.02	**94.20**	74.96	87.94
	$RankCSE_{ListMLE}$	84.47	89.51	94.65	89.87	89.46	93.00	75.88	88.12
	SynCSE	87.18	92.02	94.16	90.76	91.65	86.80	76.87	88.49
	* RankRefine	**88.04**	92.10	95.10	**90.92**	**92.86**	90.80	**78.26**	**89.73**
	* RankRefine+	87.84	**92.53**	**95.22**	90.88	92.64	90.60	77.22	89.56

5.9 重排序任务

为进一步验证所提方法在排序能力上的提升效果，本文在标准语义相似度任务之外，引入了重排序任务（reranking）作为补充评测。这类任务关注多文本间的相对语义关系，是语义相似度建模的自然延伸，同时对模型在跨领域及分布外数据上的泛化能力提出更高要求。

实验设置遵循 SynCSE[19] 的方案，在四个公开的重排序基准上进行评测，并与多个代表性方法比较：

（1）AskUbuntuDupQuestions（AskUbuntu Duplicate Questions）[74]：收录 Ubuntu 技术问答社区中大量重复或相近问题。任务是在给定问题候选列表中，将语义重复的问题排在前列；

（2）MindSmallReranking（MIND-Microsoft News Dataset，small version for reranking）[75]：基于 MIND 新闻推荐数据集的精简版本。任务是根据用户历史或查询新闻，对候选新闻按相关性重排；

（3）SciDocsRR（Scientific Documents Reranking）[76]：面向科学文献检索，根据给定查询论文，对候选文献清单进行相关性排序；

（4）StackOverflowDupQuestions（StackOverflow Duplicate Questions）[77]：来自 StackOverflow 编程问答社区，涵盖海量编程技术问题，需识别并优先排列语义重复的问题。

实验结果（表 8）显示，该任务覆盖多个异构领域，要求模型既能准确捕捉文本间细微语义差异，也能在分布外数据上保持稳定表现。

如表 8 所示，这四个任务覆盖技术问答、新闻检索、科学文献和编程社区等领域，要求模型既能捕捉微小的语义差异，又能在分布差异显著的数据集上保持稳定表现。在对比结果中，虽然 SynCSE 与 MultiCSR 等方法在 STS 任务中表现优异，但在重排序任务上的平均准确率通常低于 SimCSE，其主要原因可能在于这些方法依赖大语言模型生成的合成训练语料。尽管生成文本质量不低，但其分布与真实语料存在差异，且可能包含特定模式或“人工痕迹”。为了适应这些特定模式，模型的语义表示空间容易发生分布偏移，从而削弱了在真实且复杂排序任务中的泛化能力。

相比之下，本文方法在所有模型底座上均展现出良好且稳健的性能。在 $BERT_{base}$ 和 $RoBERTa_{base}$ 上，RankRefine 分别以 48.58%和 47.78%的平均正确率取得了最优结果，在大型模型上也表现出强大的竞争力。其核心优势在于，RankRefine 规避了引入外部合成数据所带来的分布漂移风险。它直接作用于模型在原始、真实数据上产生的排序列表，通过一种细粒度的、面向排序的优化目标来调整表示，从而增强模型对语义相对次序的辨别力。此过程在

提升排序精度的同时,最大程度地保留了原始数据分布的完整性,确保了模型学到的语义知识具有更强的鲁棒性和泛化能力,因此能够在这些异构的重排序基准上取得更优异的成绩。

表 8　重排序任务上的全类平均正确率比较

模型底座	方法	AskU.	MindSmall	SciDocsRR	StackO.	平均值
$BERT_{base}$	SimCSE	51.81	29.25	69.52	40.56	47.78
	PCL	52.23	28.21	69.18	**40.88**	47.62
	SynCSE	51.79	28.96	69.49	39.88	47.53
	MultiCSR	51.04	29.04	69.32	39.50	47.22
	$RankCSE_{ListNet}$	51.77	**29.90**	**72.04**	40.49	48.55
	* RankRefine	**52.94**	29.49	71.24	40.65	**48.58**
	* RankRefine+	51.48	29.51	70.99	40.56	48.55
$BERT_{large}$	SimCSE	**53.69**	30.60	73.02	41.18	**49.62**
	PCL	51.96	29.20	70.58	**41.48**	48.30
	SynCSE	51.36	30.56	71.33	40.06	48.33
	MultiCSR	51.62	29.47	71.31	39.76	48.04
	* RankRefine	52.58	**30.69**	73.15	40.35	49.19
	* RankRefine+	52.75	30.64	**73.23**	40.04	49.16
$RoBERTa_{base}$	SimCSE	52.54	**29.63**	66.68	39.54	47.10
	PCL	51.02	28.40	65.10	40.72	46.31
	SynCSE	52.59	27.58	63.39	38.81	45.59
	MultiCSR	51.91	27.97	62.83	39.35	45.51
	* RankRefine	**53.92**	29.22	67.38	**40.62**	**47.78**
	* RankRefine+	53.61	29.39	**67.39**	40.17	47.64
$RoBERTa_{large}$	SimCSE	54.95	29.30	68.82	**42.54**	48.90
	PCL	53.08	28.73	66.20	41.57	47.40
	SynCSE	55.22	**29.88**	69.33	39.00	48.36
	MultiCSR	54.01	29.16	**69.73**	40.50	48.75
	* RankRefine	54.78	29.67	69.60	40.94	48.75
	* RankRefine+	**55.44**	29.63	69.61	41.05	**48.93**

与 SynCSE、MultiCSR 等方法相比,本文方法的优势在于有效避免了由外部合成数据引入的分布漂移风险。它直接在模型基于原始真实语料生成的排序列表上施加捕捉细粒度语义差异的优化目标,以调整语义表示结构,增强模型对相对语义次序的辨别能力。这样的设计不仅提升了排序精度,还最大限度保留了原始数据分布特征,显著增强了模型在跨领域与分布外任务中的鲁棒性与泛化能力。因此,本文提出的方法在多种异构的重排序基准任务中均取得了优异成绩。

5.10　消融实验

为评估各模块的独立贡献,本文基于 $RoBERTa_{large}$ 进行消融实验,重点分析损失函数设计与共识序列数据构造的影响。具体而言,RankRefine+ 采用自适应容错列表排序采样策略,设置最大上升次数参数 $max\_up=1$,允许排序过程中的有限容错;而 RankRefine 对应 $max\_up=0$ 的情形,此时算法退化为严格最长降序列方法,不允许任何排序上升。本文还测试了 $max\_up=2,3$ 的设置以评估容错阈值的敏感性。

损失函数消融中,分别移除式(30)中对比损失项或排序损失项。数据构造消融中,使用随机采样替代共识采样,或使用单一索引模型替代多索引组合,所有实验保持相同超参数配置。表 9 展示了各变体在 STS 任务上的平均 Spearman 相关系数。

表 9　消融实验结果(Spearman 相关系数)

方法	平均值	差值
RankRefine+	83.94	0
$max\_up=2$	83.79	−0.15
$max\_up=3$	83.76	−0.18
RankRefine	83.73	−0.21
w.o. 对比损失	15.02	−68.92
w.o. 排序损失	59.69	−24.25
w.o. 共识降序采样	83.15	−0.79
w.o. 多索引召回[1]	83.51	−0.43
w.o. 多索引召回[2]	83.53	−0.41

注:上标[1] 表示 $SimCSE\text{-}BERT_{large}$,[2] 表示 $DiffCSE\text{-}BERT_{base}$。

实验结果揭示三点发现。首先,损失函数两部分均不可或缺。移除对比损失导致性能骤降至 15.02,移除排序损失则降至 59.69,验证了 4.2 节的设计理念:加权成对排序损失捕捉同主题内细微语义差异,对比损失提供全局语义聚合信号,二者缺一不可。

其次，关于自适应容错机制的影响，RankRefine+($max\_up=1$)相比严格最长降序列的 RankRefine($max\_up=0$)提升了 0.21 个百分点，表明适度的容错阈值有助于保留细微差异文本的有效信息，避免过于严格的采样策略导致的信息损失。然而，进一步放宽容错范围至 $max\_up=2$ 反而导致轻微性能下降(−0.15)，揭示过度容错可能引入噪声，破坏排序单调性。

最后，数据构造策略影响同样显著。共识降序采样相比随机采样提升 0.79 个百分点，验证了基于语义共识采样在降低索引模型排序噪声方面的有效性；共识采样相比单一模型提升约 0.4 个百分点，表明在主题复杂或语义歧义较高的情境下，共识采样能够缓解单一索引模型的固有误差。这与节 5.5 的案例分析一致，共同为模型训练提供高质量数据基础。

进一步分析表明，当索引模型数量≥3 时，构建跨所有模型一致的长降序序列变得困难。模型间的偏好差异被放大，导致共识序列过短或频繁断裂，无法为训练提供充足的高质量排序信号，形成“采样困难”问题。虽然同时增大召回样本数量 k 和容错区间可以部分缓解这一问题，但会带来显著的计算开销。因此，选取两个具有适度多样性的索引模型是在共识序列可靠性与可获得性之间的最优权衡。

本文进一步分析了截断阈值 α_q 对模型在 STS 任务中的平均性能影响(见图 5)。实验结果表明，无论 α_q 取值如何变化，本文方法均优于现有最优方法，表现出较高的鲁棒性。

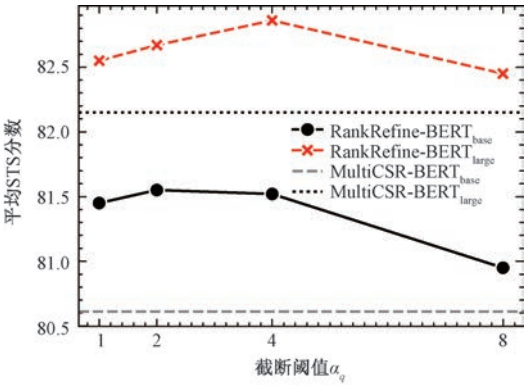

图 5 超参数 α_q 对 STS 测试平均表现的影响

结合图 5 与表 10，可观察到最优截断阈值与模型嵌入维度呈正相关关系，这意味着模型捕捉排序信息的效率与其自身能力密切相关。过小或过大的取值都会削弱有效排序信息的获取，从而导致性能下降，这一现象与理论分析一致。在推广至其他架构时，可依据嵌入维度对 α_q 取值进行按比例缩放，以提升适配效果。

表 10 模型架构信息与超参数 α_q

模型底座	最优 α_q	参数量	嵌入维度
BERT$_{base}$	2.0	110M	768
BERT$_{large}$	4.0	340M	1024
RoBERTa$_{base}$	2.0	125M	768
RoBERTa$_{large}$	4.0	355M	1024

为评估低资源情境对所提方法的影响，本文在语义文本相似度(STS)任务中，对不同训练数据比例进行了消融实验，其结果如图 6 所示。实验表明，随着训练数据比例增加，各模型性能总体稳定提升；在仅使用 20%数据的低资源条件下，基线模型已接近全量数据训练的性能，并超过对比方法。进一步增加数据量时，性能虽继续提升，但增幅逐渐缩小，呈现典型的边际收益递减现象。这说明模型在早期阶段即可利用有限数据有效捕捉关键语义信息，而在全部数据下达到最优表现。该结果凸显了本文方法在低资源场景下的数据利用效率和泛化潜力，对训练样本受限的实际应用尤具价值。

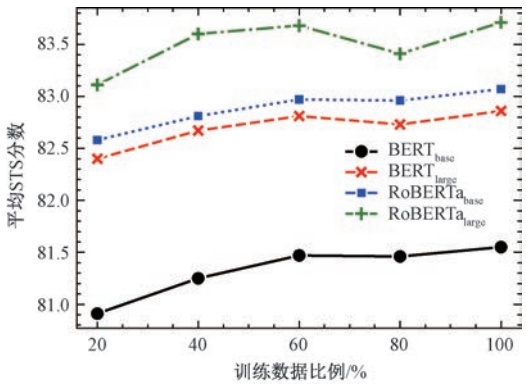

图 6 不同数据比例下 STS 平均表现

5.11 基于大语言模型的嵌入模型微调实验

为进一步验证所提出方法在更大规模语句嵌入模型上的适用性，本文在 Qwen3-Embedding[47] 系列模型上开展了补充实验，旨在考察方法在模型规模显著增大时的稳定性与可行性。

具体而言，本文选取 Qwen3-Embedding-0.6B、4B 与 8B 作为基座模型，在保持 RankRefine+框架整体结构不变的前提下进行参数高效微调[78]。共识采样阶段采用 Qwen3-Embedding-4B 与 Qwen3-Embedding-8B 作为索引模型，通过同一模型家族的不同规模视角构建排序共识序列，以降低模型架构差异引入的干扰。

实验仅对学习率与 LoRA 秩等关键超参数进

行了必要调整,以适配大语言模型的数值稳定性需求。考虑到模型参数量显著增加带来的显存与算力开销,实验在多卡环境下结合参数高效微调与分布式训练框架[79]完成,以确保训练过程的稳定性。相关实现遵循主流大模型训练实践。

实验在语义文本相似度(STS)任务上评测模型性能,结果如表 11 所示。经 RankRefine+微调后模型均取得了性能提升,Qwen3-Embedding-0.6B、4B 和 8B 模型的平均 Spearman 相关系数分别获得约 1.7%、0.5%和 0.4%的相对提升。

表 11　Qwen3-Embedding 实验结果

方法	平均值	差值
0.6B(28 层,1024 维)	81.42	0
$r=24$	**82.83**	+1.41
$r=32$	82.80	+1.38
4B(36 层,2560 维)	85.76	0
$r=24$	86.17	+0.41
$r=32$	**86.20**	+0.44
8B(36 层,4096 维)	86.21	0
$r=24$	86.53	+0.32
$r=32$	**86.54**	+0.33

随模型规模增大,性能提升幅度呈边际递减趋势,这与高参数量模型本身已具备较强语义表征能力的特点相一致。上述结果表明,本文方法无需调整核心算法结构即可稳定应用于更大参数规模的嵌入模型,并在算力受限条件下取得有效收益。

6　结论与未来工作

本文针对语句嵌入领域中无监督学习的关键挑战,提出了一套系统性的优化方案,旨在提升模型的语义表征能力与训练效率。通过算法改进和数据处理优化两个维度的创新设计,本文在语义文本相似度(STS)、迁移学习(TR)和重排序任务上取得了显著的性能提升。在数据处理方面,本文提出了基于动态规划的一次性离线共识排序采样策略与数据重组方法。该策略能够有效规避预测模型排序冲突所带来的噪声,显著提高参照排序结果的可靠性。由于不依赖特定数据特征,该方法可适应不同领域任务,展现出较强的通用性;在算法层面,本文基于理论分析,提出了融合启发信息的加权排序学习损失函数,可细粒度捕捉语义差异,并显著提升训练效果。该设计与具体模型架构解耦,因而可应用于多种基础模型。实验结果表明,本方法不仅在多项基准测试的性能上超过现有方法,还在训练效率上表现出显著优势。对不同架构的预训练模型进行评估时,均获得一致的性能提升,验证了方法的鲁棒性与适用性。

然而,当前方案仍需借助外部模型完成部分过程。未来可探索构建训练优化与数据质量优化的闭环框架,以进一步提升方法的完整性与整体性能。未来研究可进一步发展损失加权策略的自适应机制,并在更丰富的应用场景中检验其泛化能力。RankRefine 方法具有良好的扩展潜力,包括在更为充足的算力条件下,应用于更大规模数据集、与大语言模型嵌入技术结合以实现长文本匹配与跨语言检索,以及迁移至视觉、多模态等嵌入学习任务。凭借模型无关性与数据特征无关性,RankRefine 可以为多领域无监督嵌入研究提供新的思路。

致　谢　本文得到了国家自然科学基金重点项目(No.62337001)和国家自然科学基金青年基金(A)类项目(No.62525606)的支持,在此表示诚挚的感谢!

参　考　文　献

[1] Pang Liang, Lan Yan-Yan, Xu Jun, et al. A survey on deep text matching. Chinese Journal of Computers, 2017, 40(4): 985-1003 (in Chinese)
(庞亮,兰艳艳,徐君等.深度文本匹配综述.计算机学报,2017,40(4):985-1003)

[2] Wang Jin, Chen En-Hong, Shi De-Ming, et al. Semantic similarity-based information retrieval method. Pattern Recognition and Artificial Intelligence, 2006, 19(6): 696-701 (in Chinese)
(王进,陈恩红,施德明等.一种基于语义相似度的信息检索方法.模式识别与人工智能,2006,19(6):696-701)

[3] He Li-Yang, Huang Zhen-Ya, Chen En-Hong, et al. An efficient and robust semantic hashing framework for similar text search. ACM Transactions on Information Systems, 2023, 41(4): 1-91

[4] Tong Wei, He Li-Yang, Li Rui, et al. Efficient similar exercise retrieval model based on unsupervised semantic hashing. Journal of Computer Applications, 2024, 44(1): 206-216 (in Chinese)
(佟威,何理扬,李锐等.基于无监督语义哈希的高效相似题检索模型.计算机应用,2024,44(1):206-216)

[5] Guo Xian-Wei, Lai Hua, Yu Zheng-tao, et al. Emotion classification of case-related microblog comments integrating emotional knowledge. Chinese Journal of Computers, 2021, 44(3): 564-578 (in Chinese)
(郭贤伟,赖华,余正涛等.融合情绪知识的案件微博评论情绪分类.计算机学报,2021,44(3):564-578)

[6] Liu Xiao-Ming, Li Cheng-Zheng-Xu, Wu Shao-Cong, et al. A survey of text classification algorithms and application scenarios. Chinese Journal of Computers, 2024,47(6):1244-1287 (in Chinese)
(刘晓明,李丞正旭,吴少聪等.文本分类算法及其应用场景研究综述.计算机学报,2024,47(6):1244-1287)

[7] Huang Zhen-Ya, Liu Qi, Chen En-Hong, et al. A survey on text analysis methods and applications for educational questions. Journal of Chinese Information Processing, 2022, 36(10):1-16 (in Chinese)
(黄振亚,刘淇,陈恩红等.面向学科题目的文本分析方法与应用研究综述.中文信息学报,2022,36(10):1-16)

[8] Huang Wei, Chen En-Hong, Liu Qi, et al. Hierarchical multi-label text classification: An attention-based recurrent network approach//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York, USA, 2019:1051-1060

[9] Yu Kai, Chen Lu, Chen Bo, et al. Cognitive technology in task-oriented dialogue systems: concepts, advances and future. Chinese Journal of Computers, 2015,38(12):2333-2348 (in Chinese)
(俞凯,陈露,陈博等.任务型人机对话系统中的认知技术——概念、进展及其未来.计算机学报,2015,38(12):2333-2348)

[10] Zhang Wei-Nan, Zhang Yu, Liu Ting. A topic inference based translation model for question retrieval in community-based question answering services, Chinese Journal of Computers, 2015,38(2):313-321 (in Chinese)
(张伟男,张宇,刘挺.一种面向社区型问句检索的主题翻译模型.计算机学报,2015,38(2):313-321)

[11] Zhu Jia-Jun, Bao Mei-Kai, Zhang Kai, et al. A commonsense question answering method based on multi-source knowledge infusion. Computer Engineering & Science, 2025,47(2):349-360 (in Chinese)
(朱嘉骏,包美凯,张凯等.基于多源知识注入的常识问答方法研究.计算机工程与科学,2025,47(2):349-360)

[12] Chen En-Hong, Liu Qi, Wang Shi-Jin, et al. Key techniques and application of intelligent education oriented adaptive learning. CAAI Transactions on Intelligent Systems, 2021, 16(5):886-898 (in Chinese)
(陈恩红,刘淇,王士进等.面向智能教育的自适应学习关键技术与应用.智能系统学报,2021,16(5):886-898)

[13] Conneau A, Kiela D, Schwenk H, et al. Supervised learning of universal sentence representations from natural language inference data//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017:670-680

[14] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019:3982-3992

[15] He Guo-Cai, Liu Xia-Bi. Unsupervised visual representation learning with image triplets mining. Chinese Journal of Computers, 2018,41(12):2787-2803 (in Chinese)
(何果财,刘峡壁.基于图像三元组挖掘的无监督视觉表示学习.计算机学报,2018,41(12):2787-2803)

[16] Zhang Bei-Chen, Li Liang, Zha Zheng-Jun, et al. Contrastive cross-modal representation learning based active learning for visual question answer. Chinese Journal of Computers, 2022,45(8):1730-1745 (in Chinese)
(张北辰,李亮,查正军等.基于跨模态对比学习的视觉问答主动学习方法.计算机学报,2022,45(8):1730-1745)

[17] Zhang Bing-Bing, Zhang Jian-Xin, Li Pei-Hua. Contrastive language-video learning model based on spatio-temporal information auxiliary supervision. Chinese Journal of Computers, 2024,47(8):1769-1785 (in Chinese)
(张冰冰,张建新,李培华.基于时空信息辅助监督的语言-视频对比学习模型.计算机学报,2024,47(8):1769-1785)

[18] Chen Gong-Guan, Liu Hui, Li Heng-Tai, et al. Research on subgraph matching-based contrastive learning for multi-modal pretraining. Chinese Journal of Computers, 2025,48(4):893-909 (in Chinese)
(陈公冠,刘慧,李恒泰等.面向多模态预训练的子图匹配式对比学习方法研究.计算机学报,2025,48(4):893-909)

[19] Gao Tian-Yu, Yao Xing-Cheng, Chen Dan-Qi. SimCSE: Simple contrastive learning of sentence embeddings//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online, Punta Cana, Dominican Republic, 2021:6894-6910

[20] Chuang Y S, Dangovski R, Luo Hong-Yin, et al. DiffCSE: Difference-based contrastive learning for sentence embeddings//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. Seattle, USA, 2022:4207-4218

[21] Yan Yuan-Meng, Li Ru-Mei, Wang Si-Rui, et al. ConSERT: A contrastive framework for self-supervised sentence representation transfer//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021:5065-5075

[22] Zhang Yu-Hao, Zhu Hong-Ji, Wang Yong-Liang, et al. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022:4892-4903

[23] Liu Ji-Duan, Liu Jia-Hao, Wang Qi-Fan, et al. RankCSE: Unsupervised sentence representations learning via learning to rank//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 2023:13785-13802

[24] Seonwoo Y, Wang Guo-Yin, Seo C, et al. Ranking-enhanced

unsupervised sentence representation learning//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 2023:15783-15798

[25] Cao Zhe, Qin Tao, Liu Tie-Yan, et al. Learning to rank: from pairwise approach to listwise approach//Proceedings of the 24th International Conference on Machine Learning. New York, USA, 2007:129-136

[26] XIA F, LIU T Y, WANG J, et al. Listwise approach to learning to rank: theory and algorithm//Proceedings of the 25th International Conference on Machine Learning. New York, USA, 2008:1192-1199

[27] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers). 2019:4171-4186

[28] Liu Yin-Han, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv: 1907.11692, 2019

[29] Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11):613-620

[30] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 1972,28(1):11-21

[31] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781, 2013

[32] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014:1532-1543

[33] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. California, USA, 2017,30:6000-6010

[34] Liu Jian-Wei, Wang Yuan-Fang, Luo Xiong-Lin. Research and development on deep memory network. Chinese Journal of Computers, 2021,44(8):1549-1589 (in Chinese)
(刘建伟,王园方,罗雄麟.深度记忆网络研究进展.计算机学报,2021,44(8):1549-1589)

[35] Bowman S R, Angeli G, Potts C, et al. A large annotated corpus for learning natural language inference//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015:632-642

[36] Miao Pu, Du Ze-Yao, Zhang Jun-Lin. DebCSE: Rethinking unsupervised contrastive sentence embedding learning in the debiasing perspective//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. New York, USA, 2023:1847-1856

[37] Zhang Jun-Lei, Lan Zhen-Zhong, He Jun-Xian. Contrastive learning of sentence embeddings from scratch//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023:3916-3932

[38] Wang Hui-Ming, Li Zhao-Dong-Hui, Cheng Li-Yin, et al. Large language models can contrastively refine their generation for better sentence representation learning//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico, 2024:7874-7891

[39] He Li-Yang, Liu Cheng-Long, Li Rui, et al. Refining sentence embedding model through ranking sentences generation with large language models//Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria, 2025:10627-10643

[40] Rusak E, Reizinger P, Juhos A, et al. InfoNCE: Identifying the gap between theory and practice//Proceedings of The 28th International Conference on Artificial Intelligence and Statistics. Mai Khao, Thailand, 2024:4159-4167

[41] Van Den Oord A, Li Ya-Zhe, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv: 1807.03748, 2018

[42] Clark K, Luong M T, Le Q V, et al. ELECTRA: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv: 2003.10555, 2020

[43] Wang Liang, Yang Nan, Huang Xiao-Long, et al. Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv: 2212.03533, 2022

[44] Zhang Xin, Zhang Yan-Zhao, Long Ding-Kun, et al. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. Miami, USA, 2024:1393-1412

[45] Lee J, Chen Fei-Yang, Dua S, et al. Gemini embedding: Generalizable embeddings from Gemini. arXiv preprint arXiv: 2503.07891, 2025

[46] Li Ze-Han, Zhang Xin, Zhang Yan-Zhao, et al. Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv: 2308.03281, 2023

[47] Zhang Yan-Zhao, Li Ming-Xin, Long Ding-Kun, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv: 2506.05176, 2025

[48] Luce R D, et al. Individual choice behavior: A theoretical analysis Vol. 4. New York, USA, 1959

[49] Luce R D. The choice axiom after twenty years. Journal of Mathematical Psychology, 1977,15(3):215-233

[50] Lan Yan-Yan, Zhu Ya-Dong, Guo Jia-Feng, et al. Position-aware ListMLE: A sequential learning process for ranking//Proceedings of the Thirtieth Conference on Uncertainty in

Artificial Intelligence. Arlington, USA, 2014:449-458

[51] Burges C, Shaked T, Reneshaw E, et al. Learning to rank using gradient descent//Proceedings of the 22nd International Conference on Machine Learning. New York, USA, 2005: 89-96

[52] Burges C, Ragno R, Le Q. Learning to rank with nonsmooth cost functions//Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver, Canada, 2006:193-200

[53] Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 2021,7(3): 535-547

[54] Nemhauser G L, Wolsey L A, Fisher M L. An analysis of approximations for maximizing submodular set functions—I. Mathematical Programming, 1978,14(1):265-294

[55] Feige U. A threshold of ln n for approximating set cover. Journal of the ACM, 1998,45(4):634-652

[56] Agirre E, Cer D, Diab M, et al. SemEval-2012 task 6: A pilot on semantic textual similarity//SEM 2012: The First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). Montréal, Canada, 2012:385-393

[57] Agirre E, Cer D, Diab M, et al. * SEM 2013 shared task: Semantic textual similarity//Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. Atlanta, USA, 2013:32-43

[58] Agirre E, Banea C, Cardie C, et al. SemEval-2014 task 10: Multilingual semantic textual similarity//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland, 2014:81-91

[59] Agirre E, Banea C, Cardie C, et al. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and Pilot on interpretability//Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado, 2015:252-263

[60] Agirre E, Banea C, Cer D, et al. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation//Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego, USA, 2016:497-511

[61] Cer D, Diab M, Agirre E, et al. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation//Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Vancouver, Canada, 2017:1-14

[62] Marelli M, Menini S, Baroni M, et al. A SICK cure for the evaluation of compositional distributional semantic models//Proceedings of the Ninth International Conference on Language Resources and Evaluation. Reykjavik, Iceland, 2014: 216-223

[63] Conneau A, Kiela D. SentEval: An evaluation toolkit for universal sentence representations//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan, 2018,1699-1704

[64] Spearman C. The proof and measurement of association between two things. The American Journal of Psychology, 1904,15(1):72-101

[65] Wu Qi-Yu, Tao Chong-Yan, Shen Tao, et al. PCL: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022:12052-12066

[66] Jiang Ting, Jiao Jian, Huang Shao-Han, et al. PromptBERT: Improving BERT sentence embeddings with prompts//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022:8826-8837

[67] Pang Bo, Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). Ann Arbor, USA, 2005:115-124

[68] Hu Min-Qing, Liu Bing. Mining and summarizing customer reviews//Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2004:168-177

[69] Pang Bo, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain, 2004: 271-278

[70] Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 2005,39:165-210

[71] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013:1631-1642

[72] Voorhees E M, Tice D M. Building a question answering test collection//Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000:200-207

[73] Dolan B, Brockett C. Automatically constructing a corpus of sentential paraphrases//Proceedings of the Third International Workshop on Paraphrasing (IWP2005). Jeju Island, Republic of Korea, 2005:9-16

[74] Lei T, Joshi H, Barzilay R, et al. Semi-supervised question retrieval with gated convolutions//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics. San Diego, USA, 2016:1279-1289

[75] Wu Fang-Zhao, Qiao Ying, Chen J H, et al. MIND: A large-scale dataset for news recommendation//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020:3597-3606

[76] Cohan A, Feldman S, Beltagy I, et al. SPECTER: Document-level representation learning using citation-informed transformers//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 2270-2282

[77] Liu Xue-Qing, Wang Chi, Leng Yue, et al. Linkso: a dataset for learning to retrieve similar question answer pairs on software development forums//Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering. New York, USA, 2018:2-5

[78] Hu E J, Shen Ye-Long, Wallis P, et al. LoRA: Low-rank adaptation of large language models//Proceedings of the International Conference on Learning Representations. Online,2022

[79] Rasley J, Rajbhandari S, Ruwase O, et al. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA, 2020:3505-3506

LIU Cheng-Long, M. S. candidate. His research interests include data mining and intelligent education.

ZHU Yi-Meng, M. S. candidate. Her research interests include data mining and intelligent education.

HE Li-Yang, Ph. D. His research interests include information retrieval, educational data mining, and large language models.

LIU Qi, Ph. D., professor. His research interests include cognitive intelligence and data mining.

Background

Unsupervised sentence embedding has become the cornerstone of natural language processing (NLP), making it possible to implement semantic representation without extensive annotated datasets. These embedding methods have delivered significant performance improvements in downstream tasks including semantic search, text clustering, and paraphrase identification. Despite the significant progress of contrastive learning paradigm (like SimCSE) in recent years, and the existing preliminary exploration of learning-to-rank methods (like RankCSE), unsupervised methods are still faced with challenges in capturing heuristic semantic information robustly and efficiently.

The current mainstream contrastive learning approaches mainly focus on semantic correlation between local pairwise sentences, restricting the ability to represent global semantic structure, which plays a significant role in deeply fine-grained text understanding. Meanwhile, emerging LTR methods are designed to capture global structures, yet they are frequently affected by noise. Such noise primarily comes from suboptimal batch sampling tactics, where unreliable ranking signals are generated by sentences with huge semantic gaps, and inconsistent guidance from distinct teacher models. Additionally, listwise ranking loss functions adopted by traditional LTR frameworks often lack efficiency in distinguishing fine-grained semantic differences and stability when dealing with large batch sizes and noisy data. These challenges make it difficult to balance accurate semantic modeling, and optimizing stability, greatly hindering optimal convergence and performance.

To address these limitations, we introduce RankRefine, an innovative and efficient multidimensional denoising structure for data preprocessing and model optimization based on learning-to-rank method. RankRefine significantly enhanced data quality and training effectiveness through two key stages. Firstly, the data preprocessing phase adopts offline consensus sampling, combined with longest consistent decreasing subsequence algorithm, to pick out reliable orderly sentence lists from consensus of several indexing models. This approach effectively filters out conflicting and noisy sorting signals to acquire training data with higher quality. Secondly, during the training stage, RankRefine puts forward a new optimizing target inspired by LambdaRank, which combines pairwise comparison in consistent sequence with location sensitive weighting mechanism, to prioritize the use of more informational sorting pairs. The final optimizing target integrates ranking loss as well as standard InfoNCE contrastive loss, effectively captures global structure information and local pairwise information.

Experiments show that RankRefine achieves average accuracy improvements of 1.38%, and 1.14% over the best-performing models of the same size on semantic textual similarity (STS), and transfer learning (TR), respectively. In the reranking task, RankRefine demonstrates superior performance across multiple metrics. This research is an important approach towards robust and efficient NLP methods. By addressing key problems including denoising and optimizing stability in unsupervised sentence embedding, our work pro-

vides a higher-fidelity semantic representation, reduces reliance on expensive manual annotations, and lays the foundation for more complex NLP applications.

This work was supported by the National Natural Science Foundation of China under Grant No. 62337001 and Grant No. 62525606.

附录 A. 数据集信息

表 12 和表 13 分别统计了 STS 数据集和 TR 数据集的训练集、验证集、测试集的数据信息。在 STS 任务中,本文采用各 STS 数据集的测试集做为评测数据集,使用 STS-B 的验证集寻找最佳超参数和最佳检查点。在 TR 任务中,本文采用 SentEval 工具库的默认设置在除了 SST 以外的数据集上进行十折交叉验证(采用 SST 的默认划分)。

表 12 STS 数据集的训练/验证/测试数据统计信息

数据集	训练集	验证集	测试集
STS12	—	—	3108
STS13	—	—	1500
STS14	—	—	3750
STS15	—	—	3000
STS16	—	—	1186
STS-B	5749	1500	1379
SICK-R	4500	500	4927

表 13 TR 数据集的训练/验证/测试数据统计信息

数据集	训练集	验证集	测试集
MR	10662	—	—
CR	3775	—	—
SUBJ	10000	—	—
MPQA	10606	—	—
SST	67349	872	1821
TREC	5452	—	500
MRPC	4076	—	1725

附录 B. 实验数据补充

本附录记录了未直接出现在正文中的实验真实数据记录。表 14 和表 15 分别记录了与图 4 所对应的不同方法的损失值与梯度范数随批次大小变化的情况。表 16 展示了图 5 中超参数 α_q 对于 STS 测试平均表现影响的相关实验数据。表 17 则对应消融实验部分对于低资源情景下的讨论,为图 6 数据来源。

表 14 不同批次大小下各方法的 Loss 对比

Batch Size	ListMLE	RankRefine	ListNet
8	7.4453	0.6432	1.6620
16	23.6346	0.4411	2.1304
32	65.7142	0.5296	2.5566
64	170.3015	0.5682	2.9623
128	420.6910	0.5723	3.3687
256	1004.2458	0.5715	3.7892

表 15 不同批次大小下各方法的梯度范数对比

Batch Size	ListMLE	RankRefine	ListNet
8	57.2968	7.7169	11.6215
16	82.3188	2.1134	7.5487
32	126.1111	2.2348	4.7103
64	174.2454	1.6097	3.7545
128	233.7330	1.3299	2.2244
256	285.6163	0.5970	1.6371

表 16 对 α_q 的敏感性分析:STS 平均得分结果

模型底座	$\alpha_q=1$	$\alpha_q=2$	$\alpha_q=4$	$\alpha_q=8$
$\text{BERT}_{\text{base}}$	81.45	81.55	81.52	80.95
$\text{BERT}_{\text{large}}$	82.55	82.67	82.86	82.45

表 17 不同数据比例下 STS 任务平均表现

模型底座	20%	40%	60%	80%	100%
$\text{BERT}_{\text{base}}$	80.91	81.25	81.47	81.46	81.55
$\text{BERT}_{\text{large}}$	82.40	82.67	82.81	82.73	82.86
$\text{RoBERTa}_{\text{base}}$	82.58	82.81	82.97	82.96	83.07
$\text{RoBERTa}_{\text{large}}$	83.11	83.60	83.68	83.41	83.71

附录 C. 大语言模型的嵌入模型微调实验参数设置

表 18 列出了 5.11 小节中 Qwen3 Embedding 模型微调实验的详细超参数配置。

表 18 Qwen3 Embedding 微调实验的超参数设置

类别	超参数	取值
模型参数	截断阈值 α_q	2.0
LoRA 配置	缩放因子(lora_alpha)	8
	丢弃率(lora_dropout)	0.1
优化器配置	学习率	6×10^{-6}
	梯度裁剪阈值	2.0
训练配置	批次大小(0.6B/4B)	256
	批次大小(8B)	128
	梯度累积步数(0.6B/4B)	1
	梯度累积步数(8B)	4
	混合精度格式	BF16
	DeepSpeed 优化阶段	Stage 2