Computer Science and Application 计算机科学与应用, 2025, 15(9), 93-104
Published Online September 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.159227

Hans汉斯

基于对比学习的中英混杂文本相似度方法

廖红虹*，赵文博，廖海明，郭昊淞，刘剑波

联通(广东)产业互联网有限公司，广东 广州

收稿日期：2025年7月26日；录用日期：2025年8月26日；发布日期：2025年9月5日

摘　要

中英混杂(Code-Switching, CS)文本的语义相似度计算是自然语言处理中的一项重要挑战，其主要难点在于复杂的语言结构和缺乏标注数据。本文提出了一种针对中英混杂文本的对比学习框架CSCL，并设计了代码点迁移和语境感知回译两种数据增强策略，以生成高质量的正负样本对，帮助模型学习对语言切换不敏感且鲁棒的语义表示。在双塔孪生网络中应用该方法，使用Albert作为共享编码器。实验结果表明，CSCL方法在中英混杂文本相似度计算上表现优于多个基线模型，Spearman等级相关系数显著提升，相比对比方法提升了4个百分点，验证了该方法的有效性。

关键词

中英混杂，语义相似度，对比学习，数据增强，孪生网络

A Contrastive Learning Based Method for Chinese-English Code-Switching Text Similarity

Honghong Liao*, Wenbo Zhao, Haiming Liao, Haosong Guo, Jianbo Liu

China Unicom (Guangdong) Industrial Internet Co., Ltd., Guangzhou Guangdong

Received: Jul. 26th, 2025; accepted: Aug. 26th, 2025; published: Sep. 5th, 2025

Abstract

The semantic similarity computation for Chinese-English code-switching (CS) texts is a significant challenge in natural language processing, mainly due to the complex language structures and the scarcity of annotated data. This paper proposes a contrastive learning framework for code-switching

*通讯作者。

文章引用: 廖红虹, 赵文博, 廖海明, 郭昊淞, 刘剑波. 基于对比学习的中英混杂文本相似度方法[J]. 计算机科学与应用, 2025, 15(9): 93-104. DOI: 10.12677/csa.2025.159227

texts (Code-Switching Contrastive Learning, CSCL) and designs two data augmentation strategies: Code-Switching Point Shifting (CSPS) and Context-Aware Back-Translation (CABT), to generate high-quality positive and negative sample pairs that help the model learn robust semantic representations insensitive to language switching. The method is applied in a Siamese network structure with Albert as the shared encoder. Experimental results show that the CSCL method outperforms several baseline models in Chinese-English mixed-text similarity computation, compared with the comparison method, it has increased by 4 percentage points in Spearman's rank correlation, demonstrating the effectiveness of the proposed approach.

Keywords

Chinese-English Code-Switching, Semantic Similarity, Contrastive Learning, Data Augmentation, Siamese Network

1. 引言

随着全球化进程的加速和互联网的普及，中英混杂(Code-Switching, CS)已成为数字通信中的一种普遍且自然的语言现象。在微博、Twitter、小红书等社交媒体平台，以及在线论坛和即时通讯工具中，用户为了表达便捷、彰显身份或填补词汇空缺，频繁地在单一话语中嵌入不同语言的元素。例如，“这个 design 的 a-line 版型很显瘦”或“你的 proposal 需要 re-evaluate 一下”等表达方式屡见不鲜。这种现象的普遍性对自然语言处理(NLP)技术提出了新的要求。

准确地计算中英混杂文本之间的语义相似度，对于诸多下游任务至关重要，包括信息检索(例如，在混杂查询与纯中文或纯英文文档之间建立联系)、智能问答系统中的重复问题检测、以及文本聚类与推荐系统等。然而，由于其独特的语言结构，衡量混杂文本的语义相似度远比处理单语文本复杂，至今仍是一个未被有效解决的挑战。现有的文本相似度方法主要针对单语或跨语言场景设计，当直接应用于中英混杂文本时，其性能往往会显著下降。

将深度学习模型有效应用于中英混杂文本相似度计算，主要面临以下三个核心挑战：

1) 词汇与语法结构的混合导致的语义理解困难。中英混杂文本在句法和词法层面表现出高度的复杂性。句内混杂(intra-sentential switching)将不同语言的词汇(如英文名词、动词)嵌入到另一种语言的语法框架中(如中文的“SVO”结构[1])。这种“借用”打破了单一语言的词法和句法约束，可能导致标准 NLP 模型在词性标注、依存句法分析和最终的语义合成(semantic composition)上出现偏差。模型不仅需要理解单个中英文单词的含义，更要准确捕捉它们在混杂语境下组合生成的新语义，这是一个巨大的挑战。

2) 缺少大规模、高质量的标注数据集。监督学习方法在标准的文本相似度任务(如 STSB [2])中取得了巨大成功，但这严重依赖于大规模的人工标注数据。对于中英混杂文本，构建一个类似规模的、由专家标注其语义相似度得分的数据集，成本极其高昂且耗费人力。标注过程不仅需要标注者具备双语能力，还需要其对混杂语言的使用习惯有深刻理解，以做出一致且可靠的判断。这种标注数据的稀缺性，极大地限制了监督学习方法的应用，使得研究重心不得不转向无监督或自监督的学习范式。

3) 现有预训练模型对混杂模式的适应性不足。尽管多语言预训练模型(如 mBERT [3]，XLM-R [4])在海量的多语种语料上进行了训练，具备了一定的跨语言理解能力，但它们并非为处理代码混杂现象而专

门设计。这些模型的预训练目标(如掩码语言模型)主要在各自独立的单语语料上进行，旨在学习语言的通用表示，而未能充分学习在语言边界(code-switching points)上两种语言如何进行语义交互和融合的细粒度知识。因此，当直接使用这些模型来编码中英混杂句子时，所生成的句子向量往往是次优的，难以精确地捕捉混杂文本间的细微语义差异。

鉴于此，现有主流的文本相似度计算方法在处理中英混杂文本时均表现出明显局限。一种直接的方法是利用多语言预训练模型(如 mBERT)提取词向量并进行平均池化(mBERT-avg)，但这种操作会严重损失句法结构和词序信息，导致生成的句子向量语义表达能力不足，难以区分细微的语义差异，即陷入“各向异性”问题。

为解决此问题而设计的句子表示模型，如 Sentence-BERT (SBERT) [5]，虽然通过在自然语言推断(NLI)和语义相似度(STS)任务上进行微调来生成高质量的句子向量，但其训练数据主要由单语文本构成。因此，这些模型并未显式地学习过句内代码混杂的特定语言模式，其对于如何融合中英文词汇形成连贯语义的能力是未经优化的，直接应用时效果欠佳。

另一方面，以 SimCSE [6]为代表的无监督对比学习方法，通过引入“Dropout”作为最小化的数据增强策略，在单语文本表示上取得了巨大成功。然而，对于中英混杂文本，其核心挑战在于语言的结构性切换。仅仅依赖 Dropout 这种随机且非结构化的噪声来构造正样本，可能不足以让模型学习到对语言切换这一结构化现象的语义不变性。例如，该方法无法显式地引导模型去理解“这个 design 很不错”与其近义表达“这个设计 is very good”之间的高度相似性。总而言之，现有方法缺乏一个专门为捕捉中英混杂文本内在结构和语义特性而设计的学习范式。

为应对上述局限性，本文提出了一种新颖的、专为中英混杂文本设计的对比学习框架，将其命名为中英混杂对比学习(Code-Switching Contrastive Learning, CSCL)。该框架旨在从未标注的混杂文本中学习高质量的句子级别语义表示。

本文的主要贡献可以概括为以下三点：

1) 提出一个专为中英混杂文本优化的对比学习框架。本文系统性地将对比学习范式应用于中英混杂文本相似度计算的研究。通过构建一个端到端的孪生网络，提出的 CSCL 框架能够从未标注数据中有效学习对语言切换不敏感的、语义上更鲁棒的句子向量。

2) 设计并验证了一套新颖的、针对混杂文本的数据增强策略。有效的正负样本构建是文本相似度方法成功的关键。因此，本文提出并实现了一系列针对混杂文本结构特性的数据增强方法，如代码点迁移(Code-Switching Point Shifting)和语境感知回译(Context-aware Back-translation)。这些策略超越了简单的随机噪声，通过模拟真实的语言使用变体来生成语义一致但形式多样的正负样本对，从而引导模型深入理解混杂语境下的语义内涵。

3) 在多个基准上取得了当前最佳(State-of-the-Art)性能。本文在自建的中英混杂文本相似度基准上进行了全面而深入的实验验证。结果表明，与直接应用 mBERT、标准 SBERT 以及通用 SimCSE 等一系列强基线模型相比，本文提出的 CSCL 方法在所有测试集上均取得了显著且一致的性能提升，为该任务树立了新的技术标杆。

2. 相关工作

2.1. 文本相似度计算

文本语义相似度计算作为自然语言处理领域的一项基础且关键的任务，旨在衡量文本间的语义关联性，其研究已从传统的统计与规则方法，演进至以深度学习为核心的表示学习新范式。早期研究主要依赖于编辑距离、N-Gram 等字面匹配方法，以及向量空间模型(VSM)、主题模型(如 LDA)等统计方法[7]。

这些方法虽易于实现，但在捕捉深层语义信息方面存在天然缺陷。为了弥补这一不足，部分研究尝试将命名实体等语义信息融入 N-Gram 图模型中[8]，或利用本体知识库进行语义扩展[9]，在一定程度上提升了传统方法的性能。

随着深度学习技术的兴起，研究重心全面转向基于神经网络的模型[10]。该范式下的方法可大致分为两类：表示型和交互型。表示型方法，特别是基于孪生网络(Siamese Network)的架构，通过共享权重的编码器(如 CNN、LSTM 及其混合体 RCNN)将两个文本独立映射到向量空间，再计算其相似度[11]。这种结构在问答系统和智能客服等场景中取得了良好应用，但其将交互延迟到最后的匹配阶段，可能损失细粒度的对齐信息[12]。预训练语言模型，尤其是 BERT 及其变体的出现，标志着一个新的里程碑。这类模型通过在大规模无标注语料上进行预训练，能够生成深度上下文化的文本表示，显著提升了各项任务的基准[10]。学者们发现，不同的预训练模型(如 BERT、RoBERTa、ERNIE)因其训练目标和语料的差异而各具优势，因此，通过模型集成策略，如均值、元学习或改进的 Adaboost 算法，融合多个模型的判断，能够进一步提升相似度计算的准确性和鲁棒性[13]。此外，还有研究探索将结构相似性(如编辑距离)与语义相似性(如 Sentence-BERT)通过层次分析法(AHP)等方式进行加权融合，证明了多维度信息融合的有效性[14] [15]。

尽管深度学习方法成果斐然，但其对大规模标注数据的依赖以及在小数据集场景下的训练难题，催生了新的研究方向。其中，对比学习(Contrastive Learning)为解决小样本和无监督场景下的相似度计算问题提供了有效途径[16]。近期研究深入到对比学习的目标函数(如 InfoNCE [17] [18])层面，通过分析并解耦反向传播中的正负样本耦合算子，有效抑制了梯度衰减，从而大幅提升了模型在小数据集上的收敛速度和最终性能，这对于将相似度分析技术应用于法律、医疗等专业领域的实际问题具有重要意义。

总体来看，当前文本相似度计算的研究展现出一条清晰的演进脉络：从浅层统计到深度表示，再到预训练模型的范式迁移，最终开始关注学习过程本身的优化。这些研究在模型架构(如混合模型、集成模型)和学习策略(如对比学习、多维度信息融合)上均取得了显著进展。然而，研究领域仍存在一些挑战：针对中英文混杂文本的相似度计算仍需进一步探索，同时，高质量、大规模的中英文混杂的相似度数据集依然稀缺，多数模型验证仍依赖有限的公开数据集或自建小规模语料。

2.2. 自然语言处理中的对比学习

近年来，对比学习(Contrastive Learning)已成为自监督表示学习领域最具影响力的范式之一。其核心思想并非直接预测标签，而是在一个度量空间中，通过拉近“正样本对”的表示，同时推远“负样本对”的表示来学习高质量的特征。这种方法使得模型能够从未标注数据中捕捉到丰富的语义信息，并生成一个结构良好的表示空间，其中语义相近的样本在空间上彼此靠近。

这一范式在学习句子级别语义表示的任务中取得了突破性进展。其中，SimCSE [6]是该领域的里程碑式工作。SimCSE 的巧妙之处在于其提出了一种极其简洁而有效的正样本构建方法：在无监督设置下，它将同一句子两次输入到同一个带有标准 Dropout 的预训练语言模型编码器(如 BERT)中。由于 Dropout 层的随机性，两次前向传播会产生两个略有差异但语义上完全一致的向量表示，这两个向量便构成了一个高质量的正样本对。批次内的所有其他句子则自然地成为负样本。通过在这种简单的“噪声”下进行对比学习，SimCSE 极大地提升了单语种句子的表示质量，在多个标准语义相似度(STS)基准上达到了当时的最佳水平。

然而，尽管 SimCSE 在单语种文本上表现卓越，其核心的数据增强策略在直接应用于结构更复杂的中英混杂文本时，暴露了其潜在的不足。SimCSE 所依赖的 Dropout 本质上是一种与结构无关的随机噪声。但中英混杂现象的核心挑战并非随机的单词缺失，而是一种高度结构化的语言切换。简单地通

过 Dropout 来扰动模型内部的激活状态，并不能显式地引导模型去理解语言切换点两侧的语义等价关系。例如，它无法直接教会模型“这个 design 很出色”与“这个设计 is excellent”这两句在语义上是高度相似的，因为这两者在词汇层面差异巨大。因此，如何设计出能够感知并利用混杂文本结构特性的数据增强策略，从而引导模型学习到对语言切换更具鲁棒性的语义表示，便构成了本文研究的核心动机。

3. 本文方法

本文提出了基于对比学习的中英文混杂的文本相似度计算方法，旨在满足当前流行的中英文混杂使用的趋势。在 3.1 节重点介绍中英文混杂文本结构特性的数据增强策略，3.2 节介绍共享编码器的双塔孪生网络结构及其损失函数，3.3 节介绍模型训练方式。

3.1. 数据增强策略

为了克服 SimCSE 等方法中结构无关数据增强策略的局限性，我们提出了一套专为中英混杂文本的结构特性而设计的增强策略组合。这些策略旨在生成语义一致但形式多样的高质量正样本对，从而引导模型学习对语言切换更具鲁棒性的语义表示。本文提出以下三种正样本构建策略来生成中英文混杂的高质量训练数据，它们可以组合使用以生成丰富的训练数据。为准确获取文本数据的结构特征，我们对所有的文本语句使用结巴分词[19]工具进行了处理，以便能够精确获取每个包含结构特性的词语。

3.1.1. 代码点迁移(Code-Switching Point Shifting, CSPS)

该策略的核心思想是：在保持句子核心语义不变的前提下，通过翻译句中的部分语块来改变中英文切换点(code-switching point)的位置。这能直接教会模型，一个概念的语义不应因其表达语言或切换模式的改变而改变。

由于混杂文本的关键结构特征在于语言边界，因此，CSPS 通过直接操纵这一结构，迫使模型学习跨越语言边界的语义等价关系，这是随机 Dropout 无法实现的。对于一个给定的混杂句子，我们首先识别出其中的中、英文语块。然后，随机选择一个语块(通常是较短的那个)并使用一个高质量的翻译引擎将其翻译成另一种语言，然后替换原文的对应部分。

代码点迁移的具体示例如表 1 所示。

Table 1. Examples for CSPS
表 1. 代码点迁移示例

原始句子	这个 design 真的很有创意，让人 impressed。	
迁移方式一	这个设计真的很有创意，让人印象深刻。	语言切换点消失
迁移方式二	这个 design is truly creative，让人 impressed。	语言切换点发生迁移和变化

3.1.2. 语境感知回译(Context-Aware Back-Translation, CABT)

回译是生成释义的经典方法，但我们对其进行了改造以适应混杂语境。CABT 仅对句子中的一种语言成分进行回译，而保持另一种语言的语境不变。

完整句子的回译(如中英混杂→英→中英混杂)可能会破坏原有的混杂结构。CABT 通过固定一种语言作为“锚点”，可以更精细地生成保留了混杂特性且语义一致的样本。首先，分离出句子中所有的中文词汇和英文词汇。然后，随机选择一种语言(例如，中文)，将其所有词汇片段拼接后进行回译(中→英→中)，最后将得到的新中文词汇放回它们在原句中的位置。

语境感知回译的具体示例如表 2 所示。

Table 2. Examples for CABT
表 2. 语境感知回译示例

原始句子	我 feel **这个** solution 有点**复杂**。
直接翻译	我 feel **this** solution 有点 **complicated**。
回译	我 feel **这** solution 有点**繁杂**。

3.1.3. 同义词替换(Synonym Replacement, SR)

这是一种旨在增加词汇多样性的增强策略，我们将其同时应用于句中的中英文部分。同义词替换的目的是增加模型对近义词的认知，避免模型过分依赖于特定的词汇表达。

我们利用大规模的中英文同义词词典，本文中文同义词基于哈工大同义词词林[20]，英文同义词基于WordNet [21]。在句子中随机选择一定比例的非停用词，并从词典中随机选择一个同义词进行替换。

同义词替换的具体示例如表 3 所示。

Table 3. Examples for SR
表 3. 同义词替换示例

原始句子	这个任务的截止日期非常**紧迫**。
同义词替换	这个任务的截止日期非常**紧急**。
伴随翻译	这个 project 的 deadline 非常紧急。

3.1.4. 负样本构建

与正样本构建的精心设计不同，在负样本方面，我们采用了高效且广泛使用的批内负样本(in-batch negatives)策略。具体而言，对于一个大小为 N 的训练批次，每个原始句子会生成一个对应的正样本，构成 N 个正样本对。对于其中任意一个句子(锚点，anchor)，其配对的增强句子是其唯一的正样本，而该批次内所有其他的 $2(N-1)$个句子(包括其他句子的原始版本和增强版本)都被视为负样本。这种方法无需额外的计算和存储开销，已被证明在对比学习中非常有效。

3.2. 模型架构及损失函数

3.2.1. 孪生网络编码器

本文的模型架构采用孪生网络结构，两个编码器塔共享同一套 Albert 权重。对于输入的任意一个中英混杂句子，编码器会将其处理后，输出其对应的句子级别向量表示。本文对最后一隐藏层所有词向量采用平均池化(average-pooling)的方式来获得最终的句子向量，实验证明这种方式在生成句子表示时通常优于直接使用[CLS]标记。详细模型架构如图 1 所示。

本文基础网络架构选择 Albert [22]作为句子编码器。与 mBERT 和 XLM-R 等模型相比，Albert 通过两项关键的参数削减技术，即因式分解的嵌入参数化和跨层参数共享，在大幅减少模型参数量的同时，依然保持了强大的语言表示能力。选用 Albert 的主要优势在于：

1) 参数高效性：更少的参数量意味着更快的训练速度和更低的显存消耗，这使得我们能够在使用大规模未标注混杂语料进行自监督训练时更加高效。

句子关系建模：Albert 在预训练阶段引入了句子顺序预测(Sentence Order Prediction, SOP)任务，这使其本身就具备了较强的辨别句子对关系的能力，与本文研究的语义相似度任务目标天然契合。

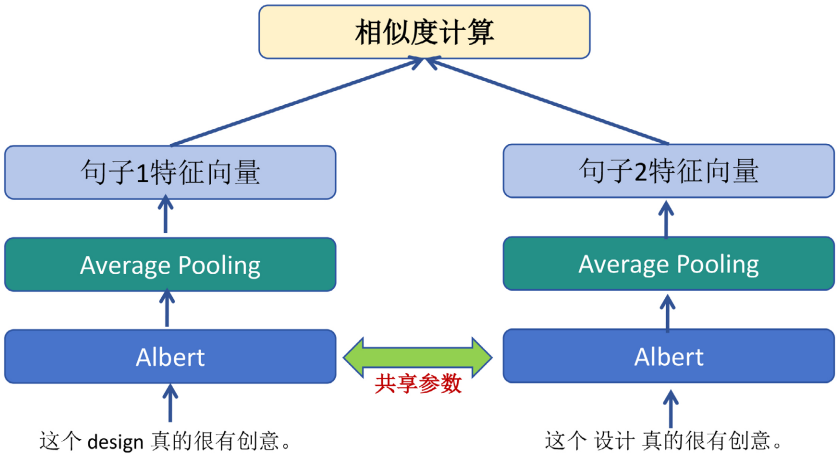

Figure 1. Siamese-network for the proposed methods
图 1. 本文提出方法的孪生网络结构

3.2.2. 损失函数：困难负样本对比损失

标准的 InfoNCE 损失函数平等地对待所有负样本，但这可能导致训练效率不高，因为模型在训练后期会花费大量计算资源去推开那些已经“离得很远”的简单负样本。为了让模型聚焦于学习更具挑战性的、细粒度的语义差异，本文采用了困难负样本对比损失(Hard Negative Contrastive Loss) [23]。

一个“困难负样本”指的是，在当前的表示空间中，与锚点(anchor)句子语义不同，但向量距离却非常近的样本。这些样本是模型最容易混淆的对象。困难负样本对比损的目标是，在每个训练步中，集中“火力”去推开这些最困难的负样本。

因此，本文采用基于三元组(Triplet)的损失形式来实例化这一思想。对于一个由锚点句子 x_i，其增强后的正样本 x_i^+，其在批次内挖掘出的最困难的负样本 x_i^- 构成的三元组，其损失函数定义如下：

$$L\left(h_i,h_i^+,h_i^-\right)=\max\left(0,M+\mathrm{sim}\left(h_i,h_i^+\right)-\mathrm{sim}\left(h_i,h_i^-\right)\right) \tag{1}$$

其中，h_i,h_i^+,h_i^- 分别为 x_i,x_i^+,x_i^- 经过 Albert 编码器得到的特征向量，$\mathrm{sim}(\cdot,\cdot)$代表相似度计算函数，M 是一个预设的边界(Margin)超参数，它要求正样本对的相似度至少要比困难负样本对的相似度高出 M。困难负样本 x_i^- 是在同一批次内的所有其他句子表示中，除了证样本增强得到的证样本 x_i^+ 之外，与锚点 x_i 相似度最高的那个。通过优化该损失，模型被迫将注意力集中在那些最容易混淆的样本对上，从而学习到一个边界更清晰、判别能力更强的表示空间。

3.3. 模型训练

本文使用的孪生网络模型采用端到端的方式进行训练，具体流程如下：

1) 数据批次构建：从大规模未标注中英混杂语料库中随机抽取一批次(mini-batch)的句子。

2) 正样本生成：对批次中的每一个句子，应用我们在 3.1 节中提出的增强策略，使用一个随机组合的方式组合 1 个到 3 个数据增强策略来生成其对应的正样本，构成正样本对。

3) 数据编码：将批次内所有的原始句子和增强后的句子一同送入共享权重的 Albert 编码器，得到它们各自的句子向量。

4) 困难负样本挖掘：对于批次中的每一个锚点，遍历批内所有其他样本，根据余弦相似度找到其最困难的负样本。

5) 损失计算与优化：使用上述定义的困难负样本对比损失函数计算损失，并通过反向传播算法更新

Albert 编码器的参数。

模型训练完成后，其核心价值在于能够高效地为任意中英混杂文本生成高质量的语义向量。推理阶段非常高效，无需孪生结构，其详细执行过程如下：

1) 单向编码：对于任意给定的两个待比较相似度的中英混杂句子 S_A 和 S_B，将它们分别、独立地输入到单个已经训练完毕的 Albert 编码器中。

2) 获取向量：经过一次前向传播，我们即可获得它们各自的句子特征向量 v_A 和 v_B。

3) 计算相似度：两个句子的语义相似度最终由这两个向量的余弦相似度来度量：

$$\text{Similarity} = \frac{v_A \times v_B}{v_A v_B} \tag{2}$$

这种推理方式极大地提高了计算效率，使其非常适合需要对海量文本对进行比较的实际应用场景。

4. 实验及实验结果分析

4.1. 实验数据集介绍

4.1.1. 训练数据集构建

为有效地进行无监督对比学习，我们构建了一个大规模的中英混杂文本语料库，命名为 Uni-Corpus。该语料库主要由从两个主流社交媒体平台——新浪微博和小红书——上采集的公开帖子和评论组成，时间跨度为 2023 年至 2024 年。选择这两个平台是因为它们的用户群体广泛，且中英混杂现象非常普遍和自然。

由于爬取下来的数据存在很多噪声和格式问题，因此对数据清洗与预处理步骤如下：

1) 去重：移除了完全重复的文本。

2) 过滤：删除了长度小于 5 个单词或超过 200 个单词的句子，以及纯中文或纯英文的句子，以确保语料库聚焦于混杂现象。

3) 数据标准化：将所有文本转换为小写，移除了特殊字符、URL 链接和@提及，并统一了标点符号。

经过处理后，我们的 Uni-Corpus 包含了约 150 万条高质量的中英混杂句子，为模型学习提供了丰富且真实的语境。在模型训练过程中，我们使用经过中文预训练的 Albert 模型[24]作为初始化模型。由于该初始模型使用的字典并未对英文做特殊处理，仅以 26 个字母作为 token，对英文的表达能力欠缺，因此，本文采用 BPE (byte-pair encoding) [25]算法对初始化模型的词典进行了扩展，并在 Uni-Corpus 数据上对模型进行微调训练，以使得基础 Albert 模型兼顾中英文双语特性，微调参数为参考文献[24]的默认配置参数。

4.1.2. 评测基准数据集构建

为了全面评估模型的性能，我们采用了两个中英混杂语义相似度数据集，分别命名为 Uni-STS-B 和 Uni-STS-H。

1) Uni-STS-B：这是一个改编自公开代码混杂数据集的基准 LinCE 基准[26]中的部分数据。具体来说，我们选取了其自然语言推断(NLI)任务中的句子对，并将其标签(蕴含、矛盾、中立)映射为语义相似度得分。我们将“蕴含”关系对视为高相似度(得分 4~5)，“中立”关系对视为中等相似度(得分 2~3)，“矛盾”关系对视为低相似度(得分 0~1)。

2) Uni-STS-H (Human-annotated)：考虑到现有基准的局限性，我们构建了一个全新的、由人工精细标注的中英混杂文本相似度数据集，命名为 Uni-STS-H。该数据集包含 1500 对精心挑选的中英混杂句子对，涵盖了日常对话、产品评论、科技讨论等多个领域。

我们招募了 10 位精通中英双语且熟悉网络用语的标注者。每对句子由三位标注者独立进行打分，评分范围为 0 到 5 (0 代表完全不相关，5 代表语义完全等价)。最终的得分为三位标注者评分的平均值。为了确保标注质量，我们计算了标注者之间的皮尔逊相关系数，平均值达到了 0.86，表明标注结果具有高度的一致性和可靠性。

在模型训练时，由于最终计算相似度的取值为区间[0, 1]，因此，在模型训练时，所有的相似度得分均除以数值 5 之后，再输入到模型计算损失。

4.2. 实验设置

4.2.1. 评价指标

本文采用斯皮尔曼等级相关系数(Spearman's rank correlation coefficient, ρ)作为主要的评价指标。与皮尔逊相关系数不同，斯皮尔曼系数衡量的是模型预测的相似度得分排序与人类标注的排序之间的一致性，它对得分的具体数值不敏感，更适合评估 STS 任务。本文将计算出的系数乘以 100 以便于报告和比较。

4.2.2. 对比模型

本文提出的 CSCL 方法将与以下四个强有力的基线模型进行比较：

1) TF-IDF + Cosine：这是一种经典的、非深度学习的基线模型，通过计算句子的 TF-IDF 向量的余弦相似度来判断。

2) mBERT-avg [3]：使用预训练的多语言 BERT 对句子进行编码，然后对输出的词向量进行平均池化，最后计算两个句子向量的余弦相似度。

3) SBERT [5]：使用一个强大的、预训练好的多语言句子表示模型 paraphrase-multilingual-mpnet-base-v2，该模型在大量的单语和跨语言释义对上训练过。

4) SimCSE [6]：我们在自建的 Uni-Corpus 上，使用 SimCSE 的官方实现重新训练了一个模型。这代表了将最先进的通用对比学习方法直接应用于混杂语料的结果。

4.2.3. 实现细节

本文提出的模型(CSCL)及 SimCSE 基线均基于参考文献[24]提供的源代码和基础模型进行初始化。所有实验均使用 PyTorch 框架在 NVIDIA A100 GPU 上完成。关键超参数设置如表 4 所示，以确保实验的可复现性。

Table 4. Hyperparameters for model training
表 4. 模型训练超参数

超参数	取值
优化器	AdamW
批量大小(Batch Size)	128
最大序列长度	128
训练轮数(Epochs)	3
损失函数边界(Margin M)	0.3
池化方式	平均池化(Average Pooling)

4.3. 实验结果分析

4.3.1. 主要结果

如下表 5 所示，本文提出的 CSCL 方法在两个评测基准上均显著优于所有基线模型。从结果可以看出，基于深度学习的方法远优于传统的 TF-IDF，经过微调的 SBERT 和 SimCSE 比简单的 mBERT-avg 表

现更好，说明针对句子表示的优化是有效的。本文的 CSCL 方法在两个数据集上均取得了最佳性能，相较于同样在 Uni-Corpus 上训练的 SimCSE，在更具挑战性的人工标注数据集 CS-STS-H 上提升了超过 4 个百分点，这有力地证明了本文提出的面向混杂文本的数据增强策略和困难负样本挖掘机制的有效性。

Table 5. Experimental results
表 5. 实验结果

模型	Uni-STS-B	Uni-STS-H
TF-IDF + Cosine	45.32	48.15
mBERT-avg	62.78	65.54
SBERT	73.15	75.88
SimCSE	75.91	78.03
CSCL (本文方法)	**79.56**	**82.41**

4.3.2. 消融实验

为了验证我们提出的各个组件的贡献，我们进行了一系列消融实验，消融实验结果如表 6 所示。我们从完整的 CSCL 模型中逐一移除关键的数据增强策略，并在 Uni-STS-H 数据集上进行评测。实验结果清晰地表明：每一种我们提出的数据增强策略都对最终性能有正向贡献，其中，代码点迁移(CSPS)的贡献最大，移除它会导致性能大幅下降，这说明显式地建模语言切换结构是至关重要的。与仅使用 Dropout 作为增强策略的模型(即 SimCSE 基线)相比，本文提出的完整模型的巨大优势(+4.38 个点)再次证明了针对性、结构化的数据增强远比通用的随机噪声更适合中英混杂这一特定任务。

Table 6. Results of ablation study
表 6. 消融实验结果

模型配置	Uni-STS-H ($\rho \times 100$)
CSCL (完整模型)	82.41
w/o Code-Switching Point Shifting	80.19 (−2.22)
w/o Context-aware Back-translation	81.05 (−1.36)
w/o Synonym Replacement	81.76 (−0.65)
仅使用 Dropout (类似 SimCSE)	78.03 (−4.38)

4.3.3. 定性分析

为了更直观地理解 CSCL 的优势，本文分析了一些基线模型判断错误而 CSCL 判断正确的案例。

案例：

句子 A：“这个 new feature 的 UI 设计得太棒了！”

句子 B：“The UI of this new feature is excellent.”

模型预测相似度 (0-1)：

mBERT-avg：0.65，由于词汇重叠少(仅“UI”，“feature”)，模型给出了一个不确定的分数。

SimCSE：0.72，表现有所改善，但仍不够自信。

CSCL (本文方法)：0.91，准确捕捉到了两句话的语义等价性。

从模型预测相似度可以看出，mBERT-avg 和 SimCSE 在一定程度上受到了表面词汇差异的影响。而本文提出的 CSCL 模型，得益于代码点迁移等策略的训练，已经学习到“太棒了”和“excellent”在赞美 UI 设计的语境下是高度等价的。模型不再仅仅依赖词汇的重叠，而是真正理解了跨越语言边界的深层语义。

4.3.4. 参数敏感性分析

本文实验过程中，同步分析了损失函数中的关键超参数——边界值 M——对模型性能的影响。在实验过程中从[0.1, 0.2, 0.3, 0.4, 0.5]的范围内调整 M，并在 Uni-STS-H 开发集上观察性能变化。实验发现，当 M 设置在 0.2 到 0.4 之间时，模型性能较为稳定且表现最佳。过小的 M，如 0.1，会使得模型训练不足，难以拉开困难负样本的距离；而过大的 M，如 0.5，则可能导致模型难以收敛。最终我们选择 M = 0.3 作为最优配置。

5. 结论

本文提出的中英混杂对比学习框架 CSCL 有效地提升了中英混杂文本的语义相似度计算精度。通过设计专为混杂文本优化的数据增强策略，尤其是代码点迁移和语境感知回译，成功克服了现有方法无法有效捕捉语言切换点语义的限制。实验结果表明，CSCL 在多个基准数据集上均超越了强有力的基线模型，特别是在 Uni-STS-H 数据集上，取得了显著提升。未来的工作可以进一步探索在更大规模和更多语言的混杂文本上的应用，并优化模型的推理速度和训练效率。

基金项目

2024 年中国联通联通(广东)产业互联网有限公司云犀实时动态 AI 引擎研究项目(Y91R240EGH0003)。

参考文献

[1] 谷波, 王瑞波, 李济洪, 等. 基于 RNN 的中文二分结构句法分析[J]. 中文信息学报, 2019, 33(1): 35-45.

[2] Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I. and Specia, L. (2017) SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. *Proceedings of the* 11*th International Workshop on Semantic Evaluation* (*SemEval*-2017), Vancouver, August 2017, 1-14. https://doi.org/10.18653/v1/s17-2001

[3] Devlin, J., Chang, M.W., Lee, K., *et al*. (2019) 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*, **1**, 4171-4186.

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., *et al*. (2020) Unsupervised Cross-Lingual Representation Learning at Scale. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, 8440-8451. https://doi.org/10.18653/v1/2020.acl-main.747

[5] Reimers, N. and Gurevych, I. (2019) Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, November 2019, 3982-3992. https://doi.org/10.18653/v1/d19-1410

[6] Gao, T., Yao, X. and Chen, D. (2021) SimCSE: Simple Contrastive Learning of Sentence Embeddings. *Proceedings of the* 2021 *Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, November 2021, 6894-6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[7] 李莹, 伍胜, 徐聪, 等. 语义文本相似度计算方法研究综述[J]. 软件导刊, 2024, 23(11): 1-11.

[8] 于营, 周显春, 贾树文, 等. 基于命名实体 N-Gram 图的文本相似性度量[J]. 现代计算机, 2022, 28(2): 73-77.

[9] 张克亮, 李芊芊. 基于本体的语义相似度计算研究[J]. 郑州大学学报(理学版), 2019, 51(2): 52-59.

[10] 徐传丽, 周世杰, 吴春江. 深度学习中文本相似度计算研究综述[J]. 计算机应用与软件, 2024, 41(11): 1-14.

[11] 杨德志, 柯显信, 余其超, 等. 基于 RCNN 的问题相似度计算方法[J]. 计算机工程与科学, 2021, 43(6): 1076-1080.

[12] 纪明宇, 王晨龙, 安翔, 等. 面向智能客服的句子相似度计算方法[J]. 计算机工程与应用, 2019, 55(13): 123-128.

[13] 苏锦钿, 洪晓斌, 余珊珊. 基于多模型集成的语义文本相似性判断[J]. 华南理工大学学报(自然科学版), 2022, 50(4): 1-9.

[14] 左玉生, 张礼. 基于深度神经网络的文本语义相似性度量[J]. 南京理工大学学报, 2022, 46(1): 83-88.

[15] 温雨, 王琦, 严武军. 基于相似度融合的中文文本相似性度量方法研究[J]. 计算机应用, 2023(10): 36-39.

[16] 董勃, 罗森林. 小数据集文本语义相似性分析模型的优化与应用[J]. 信息安全研究, 2023, 9(10): 980-985.

[17] Oord, A., Li, Y. and Vinyals, O. (2018) Representation Learning with Contrastive Predictive Coding. https://arxiv.org/abs/1807.03748

[18] Rusak, E., Reizinger, P., Juhos, A., *et al*. (2024) InfoNCE: Identifying the Gap between Theory and Practice. https://arxiv.org/abs/2407.00143

[19] 结巴中文分词[EB/OL]. https://github.com/fxsjy/jieba, 2025-07-13.

[20] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版), 2010, 28(6): 602-608.

[21] Fellbaum, C. (2010) WordNet. In: *Theory and Applications of Ontology*: *Computer Applications*, Springer, 231-243. https://doi.org/10.1007/978-90-481-8847-5_10

[22] Lan, Z., Chen, M., Goodman, S., *et al*. (2019) ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. https://arxiv.org/abs/1909.11942

[23] Kalantidis, Y., Sariyildiz, M.B., Pion, N., *et al*. (2020) Hard Negative Mixing for Contrastive Learning. *Advances in Neural Information Processing Systems*, **33**, 21798-21809.

[24] GitHub Repository (2019) Albert_zh. https://github.com/brightmart/albert_zh

[25] Gage, P. (1994) A New Algorithm for Data Compression. *The C Users Journal*, **12**, 23-38.

[26] Aguilar, G., Kar, S. and Solorio, T. (2020) LinCE: A Centralized Benchmark for Linguistic Code-Switching Evaluation. *Proceedings of the* 12*th Language Resources and Evaluation Conference*, Marseille, May 2020, 1803-1813.