计算机科学
COMPUTER SCIENCE

基于数据增强和两阶段训练的摘要忠实度评估

赵金爽，黄德根

引用本文

赵金爽，黄德根. 基于数据增强和两阶段训练的摘要忠实度评估[J]. 计算机科学，2025，52(10)：266-274.

ZHAO Jinshuang, HUANG Degen. Summary Faithfulness Evaluation Based on Data Augmentation and Two-stage Training [J]. Computer Science, 2025, 52(10): 266-274.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250100023

基于数据增强和两阶段训练的摘要忠实度评估

赵金爽　黄德根
大连理工大学计算机科学与技术学院　辽宁 大连 116024
(jinshuangz@mail.dlut.edu.cn)

摘　要　文本摘要的忠实度,即其与原文在事实层面的一致性,对于自动文本摘要的实际应用具有重要意义。现有的摘要忠实度评估方法在利用文本摘要数据集方面存在不足,且构建的不忠实摘要与原文差异显著,这限制了评估方法的有效性。针对此问题,提出一种基于数据增强和两阶段训练的摘要忠实度评估模型——FaithEval。首先,定义两种数据增强方法,即同主题相似检索和外插掩码填充,用于生成与原文内容相关联的不忠实摘要,应用这些方法从文本摘要数据集中提取训练数据;然后,充分利用数据集的信息,基于原文和参考摘要构建的训练数据,分两个阶段对模型进行训练,逐步强化模型的忠实度评估能力;最后,人工构建摘要忠实度评估测试集 SFETS,为检验模型性能提供基准。实验结果表明,在 SFETS 和 Rank19 数据集上,FaithEval 均表现出色,尤其在 SFETS 数据集上,达到了当前最优的效果。

关键词:文本摘要;忠实度评估;数据增强;两阶段训练;基准测试集
中图分类号　TP391

Summary Faithfulness Evaluation Based on Data Augmentation and Two-stage Training

ZHAO Jinshuang and HUANG Degen
School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China

Abstract　The faithfulness of text summaries, which refers to their factual consistency with the original texts, is very important for the practical application of automatic text summarization. Current methods for evaluating the faithfulness of summaries have shortcomings in utilizing text summarization datasets, and the constructed unfaithful summaries differ significantly from the original texts, which limit the effectiveness of these evaluation methods. To solve this problem, this paper proposes a summary faithfulness evaluation model, FaithEval, based on data augmentation and two-stage training. Firstly, two data augmentation methods are defined: Similarity Search with Same Topic and Insert and Fill External Mask, which are used to generate summaries that are related but not faithful to the original texts. These methods are used to extract training data from the text summarization dataset. Secondly, to fully utilize the dataset information, the model is trained in two stages based on the training data constructed from the original texts and the reference summaries, progressively strengthening the faithfulness evaluation ability of the model. Finally, the test set for summary faithfulness evaluation SFETS, is constructed manually to provide a benchmark for testing model performance. Experiments show that FaithEval performs well on both SFETS and Rank19 datasets, and achieves the current state-of-the-art performance on the SFETS dataset.

Keywords　Text summarization, Faithfulness evaluation, Data augmentation, Two-stage training, Benchmark test set

1　引言

文本摘要旨在对长文本进行归纳和总结,以形成具有概括性含义的短文本[1]。根据摘要生成的方式,文本摘要可分为抽取式和生成式。由于生成式方法更贴近人工生成摘要的过程,且在语法准确度和语义连贯性方面比抽取式方法更有优势,因此受到越来越多的关注[2]。近年来,随着大规模数据集的可用性和预训练模型的日益成熟,生成式文本摘要研究取得了重大进展[3-4]。然而,研究表明,生成式摘要模型生成的摘要中有 30%存在与原文事实不一致的情况[5],导致自动文本摘要在实际场景中难以落地应用。因此,评估文本摘要相对于原文的忠实度,已成为一个亟需解决的关键问题。

文本摘要的忠实度评估是一项关键且复杂的任务,其目标在于衡量文本摘要是否准确反映了原文的事实信息[6]。目前,尚缺乏专门用于忠实度评估的监督训练数据集。人工创建大规模、高质量的数据集非常昂贵且耗时,因此,部分研究者采用弱监督方法,通过数据增强构建合成数据,以此训练摘

到稿日期:2025-01-06　返修日期:2025-04-29
基金项目:云南省重点研发计划(202203AA080004);国家自然科学基金(U1936109)
This work was supported by the Key R&D Program of Yunnan Province(202203AA080004) and National Natural Science Foundation of China (U1936109).
通信作者:黄德根(huangdg@dlut.edu.cn)

要忠实度评估模型。在文本摘要数据集中，参考摘要通常被视为忠实于原文，因此，当前面临的主要挑战是构建有效的不忠实摘要。Kryscinski 等[7]从原文中抽取单句，并替换其中的实体、数字、代词等文本单元构建不忠实摘要。Cao 等[8]在前者的基础上，选择破坏参考摘要而非原文单句，创建了 4 种类型的不忠实摘要：实体错误、数字错误、日期错误和指代错误。这些工作通过替换原文单句或参考摘要中的关键词生成不忠实摘要，会导致不忠实摘要与原文存在明显差异，这并不利于训练忠实度分类模型。为了进一步提高不忠实摘要的质量，研究人员尝试基于掩码语言模型[9]（Masked Language Model，MLM）任务构建数据。例如，Lee 等[10]通过遮蔽原文和摘要的部分内容，并利用模型填充遮蔽部分，以生成与原文内容更相关的不忠实摘要。然而，该方法面临模型过度依赖常见的上下文对应关系，导致遮蔽部分被正确填充的问题。如表 1 所列，在掩码语言模型的填充结果中，“各级”后被填充为“法院”，“刑事”后被填充为“案件”，“冤假错案”前被填充为“重大”，均被正确填充。填充句与原始句事实一致，不能作为不忠实摘要。

表 1 掩码填充结果示例

Table 1 Examples of mask filling result

文本类别	例句
原始句	周强表示，去年各级**法院**再审改判刑事**案件** 1 317 件，其中**纠正**一批**重大**冤假错案。
掩码句	周强表示，去年各级[**MASK**]再审改判刑事[**MASK**]1 317 件，其中[**MASK**]一批[**MASK**]冤假错案。
填充句	周强表示，去年各级**法院**再审改判刑事**案件** 1 317 件，其中**包括**一批**重大**冤假错案。

综合分析上述方法，可以发现存在两个局限性：1）所构建的不忠实摘要的质量尚未达到理想水平，这在一定程度上影响了模型的训练效果；2）在利用文本摘要数据集的信息方面存在不足，大多数方法仅基于原文或参考摘要来构建不忠实摘要，而未尝试全面整合并利用这两种信息以优化模型的训练效果。

针对上述问题，本文提出一种基于数据增强和两阶段训练的摘要忠实度评估模型 FaithEval。首先，定义两种数据增强方法：同主题相似检索（Similarity Search with Same Topic，S3T）和外插掩码填充（Insert and Fill External Mask，IFEM）。S3T 方法根据主题将数据集中的文章进行分组，并在同主题文章中筛选出语义最相近的单句作为不忠实摘要。IFEM 方法在单句中插入额外的掩码（“[MASK]”）并进行填充，以生成与原文内容相关的不忠实摘要。应用这些方法，从文本摘要数据集中提取训练数据。然后，充分利用文本摘要数据集的信息，对模型进行两个阶段的训练。第一阶段使用基于原文提取的训练数据，训练模型掌握对基础事实一致性的判断能力；第二阶段使用基于参考摘要提取的训练数据，提升模型在复杂语境下的忠实度评估能力。本文的主要贡献如下：

1）定义了两种数据增强方法（S3T 和 IFEM），旨在构建与原文内容高度相关的不忠实摘要。

2）提出了两阶段训练策略，利用从原文和参考摘要提取的训练数据，分阶段逐步强化模型的摘要忠实度评估能力。

3）人工构建了中文摘要忠实度评估测试集 SFETS，填补了该领域的数据空白，并为检验模型提供了标准化的测试平台。

4）实验结果显示，在 SFETS 和 Rank19[11]数据集上，FaithEval 均展现出优越的性能，尤其是在 SFETS 数据集上，达到了当前最优水平。

2 相关工作

现有的摘要忠实度评估方法主要分为无监督和弱监督两类。无监督方法依赖现有工具评估摘要的忠实度；弱监督方法则致力于设计专门用于评估忠实度的模型，这些模型通常在自动生成的数据集上进行训练[12]。

无监督方法包括基于三元祖、基于问答以及其他类型的评估方式[13]。其中，通过比较原文和摘要中事实三元组的重叠率来衡量忠实度是简单可行的评价方式。为提取三元组，Goodrich 等[14]训练了一个基于固定模式的关系提取模型。其中，固定模式是指预定义了一个关系集合，只有关系在该集合中的三元组才会被抽取出来。该方法显著提高了三元组的可比较性。Scialom 等[15]提出了基于问答的评估指标 QuestEval。QuestEval 通过生成关于原文和摘要的问题，并评估问答模型回答的相似度，结合精确分数和召回分数计算忠实度分数。问答模型的阅读理解能力使这类方法取得了良好的表现。然而，上述方法的计算成本较为昂贵。除专门设计的方法外，还有几种简单有效的评估方法。Durmus 等[16]提出事实一致性的一个直接指标是，摘要单句和原文之间的单词重叠或语义相似性。基于单词重叠的度量，如 ROUGE[17]，首先计算摘要单句和每个原文单句之间的重叠，然后取所有原文单句的平均分或最高分。基于语义相似性的度量，如 BERTScore[18]，通过计算两个单句标记嵌入的余弦相似度之和来评估相似性。

弱监督方法利用来自文本摘要的合成数据训练忠实度分类模型，旨在检验摘要与原文表达的事实是否一致。例如，Kryscinski 等[7]提出忠实度分类模型 FactCC。该模型的训练数据是通过从文本摘要数据集中抽取原文句子，并应用规则转换（如句子否定和实体交换）创建的。类似地，Cao 等[8]转换参考摘要而非原文句子，引入不忠实摘要的常见错误，并训练模型进行修正。这种基于规则的数据构建方法取得了一定的性能提升，同时也带来了性能瓶颈。一方面，其难以模拟所有类别的不忠实错误；另一方面，通过简单替换原文单句或参考摘要中的关键词生成不忠实摘要，会导致不忠实摘要与原文和忠实摘要之间差异显著，这不利于忠实度分类模型的训练。为解决这一问题，Lee 等[10]提出一种方法来构建不忠实摘要。他们将原文和参考摘要的部分内容遮蔽，然后训练模型进行填充。实验表明，使用这种不忠实摘要训练的忠实度分类模型在多数情况下优于现有模型。

近年来，大规模语言模型（Large Language Model，LLM）凭借其强大的文本理解和生成能力，在自然语言处理领域

备受瞩目。研究表明，LLM 在评估生成任务方面具有有效性，包括摘要的事实一致性评估[19-20]。例如，G-EVAL[21]利用 GPT-4[22]模型，结合思维链推理和结构化表格填写范式，能够对自然语言生成的文本进行细致的评价，实现更贴近人类判断的评价标准。然而，LLM 对文本指令和输入的敏感性，及其高昂的计算成本，限制了其在实践中的广泛应用[23]。对此，Gekhman 等[24]提出了 TrueTeacher 方法，该方法通过使用 LLM 注释不同模型生成的摘要来生成合成数据。实验结果表明，使用这种合成数据训练的摘要事实一致性评估模型，在性能上明显优于 LLM 教师模型。

3 摘要忠实度评估模型 FaithEval

3.1 模型框架

摘要忠实度评估模型框架如图 1 所示。

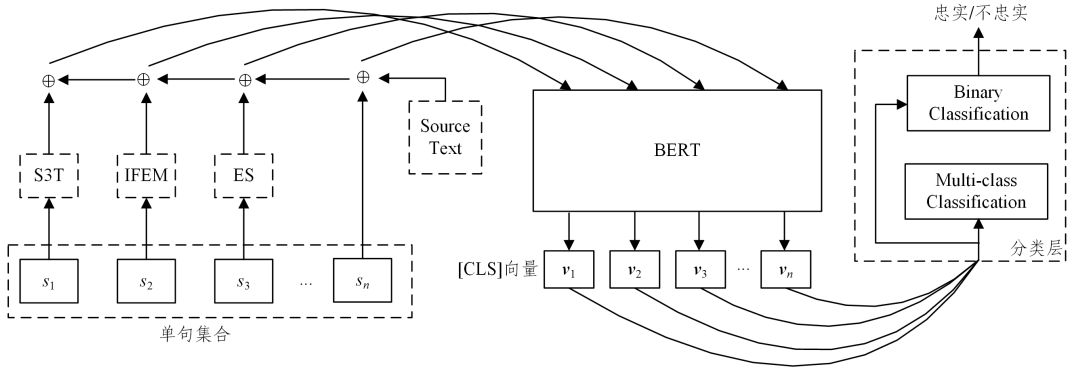

图 1 摘要忠实度评估模型框架

Fig. 1 Framework of summary faithfulness evaluation model

首先，从文本摘要数据集的原文和参考摘要出发，应用 3 种数据增强方法构建训练数据。具体而言，从数据集的原文中抽取单句，获得单句集合 $S_o=\{s_1,s_2,\cdots,s_n\}$，其中 n 表示单句总数。对于单句 s_i，$i\in[1,n/2]$，分别应用实体替换(Entity Swap，ES)、同主题相似检索(S3T)和外插掩码填充(IFEM)3 种数据增强方法生成不忠实摘要；对于 s_i，$i\in[n/2+1,n]$，直接将其作为忠实摘要。数据集中参考摘要的处理方式与原文相同。生成的摘要 x_i 如式(1)所示：

$$x_i=\begin{cases}ES(s_i), & i\%3=0 \text{ 且 } i\in[1,n/2]\\ S3T(s_i), & i\%3=1 \text{ 且 } i\in[1,n/2]\\ IFEM(s_i), & i\%3=2 \text{ 且 } i\in[1,n/2]\\ s_i, & i\in[n/2+1,n]\end{cases} \tag{1}$$

然后，对模型实施两阶段和多任务的联合训练。将摘要 x_i 和相应原文 o_i 进行拼接，得到输入序列 I_i，如式(2)所示：

$$I_i=[\text{CLS}]+x_i+[\text{SEP}]+o_i+[\text{SEP}] \tag{2}$$

将 I_i 作为 BERT[9]模型的输入。BERT 模型能够有效捕捉输入序列的上下文信息，其输出中“[CLS]”标记对应的向量 $\boldsymbol{v}_i$（$\boldsymbol{v}_i\in\mathbb{R}^d$，$d$ 表示向量的维度）可作为整个输入序列的表示。将 $\boldsymbol{v}_i$ 同时输入至二分类和多分类的分类器中进行预测分类，实现多任务学习。第一阶段训练使用基于原文构建的训练数据，第二阶段训练使用基于参考摘要构建的训练数据。两阶段训练由浅入深，逐步深化模型理解能力。

3.2 数据增强方法

3.2.1 实体替换

对于单句 s，按照实体出现的顺序，其包含的实体集合为 $E_s=\{e_1,e_2,\cdots,e_l\}$，其中，$l$ 表示实体总数，e_i 表示第 i 个实体。每个实体 e_i 具有名称 b_i 和类别 c_i。例如，若 e_1 为“[北京，LOC]”，则 b_1 为“北京”，c_1 为地点类别“LOC”。ES 方法旨在将单句中的实体替换为相同类别、不同名称的其他实体，且优先在单句内部进行替换，以最大限度保持与原句内容的相关性。具体而言，如果在单句 s 中类别为 c_i 的实体数量 $l_{c_i}>1$，则该类别的实体互相交换位置；如果在单句 s 中类别为 c_i 的实体数量 $l_{c_i}=1$，则从原文中寻找相同类别的实体进行替换。

3.2.2 同主题相似检索

S3T 方法的核心目标是为每个单句筛选出一个在语义上高度相似但并非完全一致的单句，以此作为不忠实摘要。图 2 展示了 S3T 方法的示例。

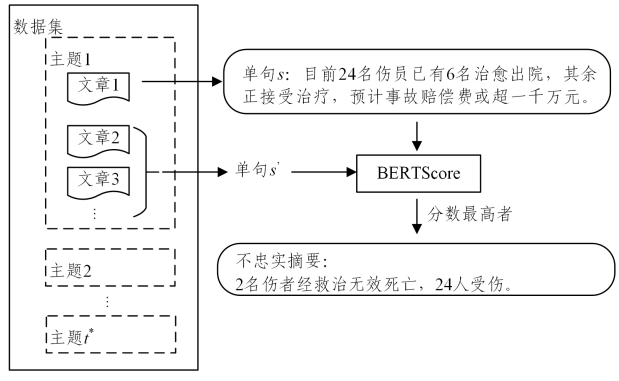

图 2 S3T 方法示例

Fig. 2 Example of S3T method

首先，将数据集中的文章根据主题进行分组。本文使用 LDA[25](Latent Dirichlet Allocation)主题模型挖掘数据集中的潜在主题。LDA 是一种无监督学习算法，通过分析数据集中单词的分布情况，可以推断出潜在的主题分布，具体计算如式(3)所示：

$$LDA(A,t)=\{z_1,z_2,\cdots,z_t\} \tag{3}$$

其中，A 表示文本摘要数据集，z_i 表示第 i 个主题的概率分布，t 表示主题数目。LDA 能够将数据集 A 中每篇文章的主题表示为一种概率分布，并根据这些主题进行主题聚类或文本分类。确定最优的主题数目 t^* 对于提高主题聚类的准确性至关重要。本文在数据集 A 上训练多个 LDA 模型，主题数

目 t 各不相同，使用主题一致性（Coherence）指标比较不同模型的性能，以确定最优的主题数目 t^*，并据此将文章按照主题进行分组。t^* 的计算式如式(4)所示：

$$t^* = \arg\max_{t} Coherence(LDA(A,t)) \tag{4}$$

然后，对于单句 s，利用 BERTScore[18] 在相同主题的文章中检索语义最相似的单句 s^*，具体计算式如式(5)所示：

$$s^* = \arg\max_{s' \in A_s} BERTScore(s,s') \tag{5}$$

其中，A_s 表示与单句 s 同主题的文章集合（排除单句 s 所属文章），s' 表示 A_s 中包含的单句，$BERTScore(s,s')$ 表示 s 和 s' 的语义相似度。

3.2.3 外插掩码填充

IFEM 方法旨在保留单句原始内容的基础上，插入额外的掩码并引导模型填充，从而生成与原始内容高度相关的不忠实摘要。遮蔽原始内容时，信息缺失导致模型有一定概率能够正确填充掩码，尤其是对于常见的事实对应关系。相较之下，IFEM 方法要求模型在完整句子信息的基础上，生成额外的内容填充掩码，这增加了引入与原文不一致的事实错误或不相关内容的概率。

图 3 展示了 IFEM 方法的示例。首先，对单句进行词法分析，确定插入掩码的位置。相较于在随机位置插入掩码，选择性地在名词、动词、形容词等特定词类前插入掩码，可以生成更稳定和可控的不忠实摘要。单句 s 包含的属于特定词类的词语集合 $W_s=\{w_1,w_2,\cdots,w_q\}$，其中，$q$ 表示词语总数，w_i 表示单句 s 中的第 i 个词语。基于单句整体的长度分布，从 W_s 中随机选择 m 个词语，如图 3 所示，掩码句中的加粗部分表示被选中的词语，在这些词语之前进行掩码插入。

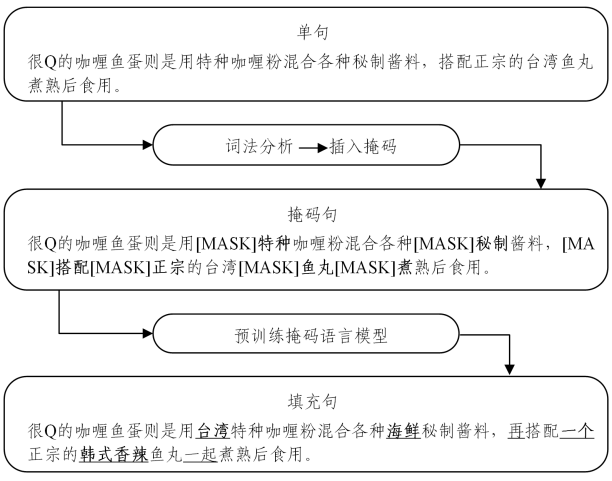

图 3 IFEM 方法示例

Fig. 3 Example of IFEM method

然后，使用预训练的掩码语言模型[26]填充掩码。采样算法使用 Top-K，即在每一步保留 k 个最高概率的词作为候选。对于每条输入，模型返回 k 条输出，计算每条输出 g_i 和单句 s 的 ROUGE-1[17] 分数，分数最低者为最终结果 g^*。设 R 为 ROUGE-1 分数，计算式如式(6)所示：

$$g^* = \arg\min_{g_i, i\in[1,k]} R(s,g_i) \tag{6}$$

图 3 的填充句中，下划线部分表示模型填充的内容，加粗部分表示不忠实的内容。

3.3 两阶段训练策略

1）第一阶段训练

基于原文单句，采用 ES，S3T 和 IFEM 这 3 种数据增强方法构建训练数据。这些方法直接作用于原文单句，生成的训练数据抽象程度较低，且事实错误较为明显。因此，该阶段的目标是训练模型准确识别直接且明确的事实对应关系，实现对基础一致性和差异性的敏锐捕捉。

2）第二阶段训练

基于参考摘要单句，采用与第一阶段相同的数据增强方法构建训练数据。这些方法作用于参考摘要单句，生成的训练数据抽象程度较高，且与原文的事实对应关系较为复杂。因此，在第一阶段训练的基础上，第二阶段旨在引导模型掌握更深层次的一致性判断逻辑，要求模型在概括性语境下保持高水平的忠实度评估能力。

为进一步提升模型的训练效果，在每个训练阶段均引入多任务学习机制。对于输入序列 I_i，从 BERT[9] 模型的输出中提取“[CLS]”标记对应的向量 $\boldsymbol{v}_i$，将其作为特征输入至由全连接层构成的二分类和多分类的分类器中，实现联合训练。具体如式(7)和式(8)所示：

$$\hat{\boldsymbol{y}}_i^{\text{bin}} = \sigma(\boldsymbol{W}_1 \boldsymbol{v}_i + \boldsymbol{b}_1) \tag{7}$$

$$\hat{\boldsymbol{y}}_i^{multi} = \sigma(\boldsymbol{W}_2 \boldsymbol{v}_i + \boldsymbol{b}_2) \tag{8}$$

其中，$\hat{\boldsymbol{y}}_i^{\text{bin}}$ 和 $\hat{\boldsymbol{y}}_i^{\text{multi}}$ 表示预测第 i 个样本属于各类别的概率分布，$\hat{\boldsymbol{y}}_i^{\text{bin}} \in \mathbb{R}^2$，$\hat{\boldsymbol{y}}_i^{\text{multi}} \in \mathbb{R}^4$，$\mathbf{W}_1$ 与 $\mathbf{b}_1$ 分别表示二分类器的权重和偏置，$\boldsymbol{W}_2$ 与 $\boldsymbol{b}_2$ 分别表示多分类器的权重和偏置，σ 表示激活函数。在多分类任务中，模型的目标是准确识别由 ES，S3T，IFEM 方法生成的 3 类不忠实摘要以及忠实摘要，共 4 个类别。如果模型能够有效区分这些类别，则表明其能够敏锐捕捉摘要间的细微差异，这对于二分类任务（即区分忠实摘要与不忠实摘要）十分关键。分类任务采用交叉熵作为损失函数，二分类任务和多分类任务的损失函数如式(9)和式(10)所示：

$$L_{\text{bin}} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{\text{bin}}\log(\hat{y}_i^{\text{bin}}) + (1-y_i^{\text{bin}})\log(1-\hat{y}_i^{\text{bin}})\right) \tag{9}$$

$$L_{\text{multi}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{4} y_{ic}^{\text{multi}}\log(\hat{y}_{ic}^{\text{multi}}) \tag{10}$$

其中，y_i^{bin} 表示第 i 个样本的忠实度标签，$\hat{y}_i^{\text{bin}}$ 表示第 i 个样本被预测为忠实摘要的概率，y_{ic}^{multi} 表示第 i 个样本多分类标签的独热编码（One-Hot Encoding）中第 c 个元素，$\hat{y}_{ic}^{\text{multi}}$ 表示第 i 个样本被预测为第 c 个类别的概率。

训练的总损失 L 由二分类损失 L_{bin} 和多分类损失 L_{multi} 组成：

$$L = \alpha L_{\text{bin}} + (1-\alpha) L_{\text{multi}} \tag{11}$$

其中，α 表示损失权重。通过合理的权重分配，模型能够在两个任务之间实现有效的学习平衡，进一步提升在摘要忠实度评估任务上的准确性。

4 实验与分析

4.1 摘要忠实度评测数据集 SFETS

SFETS 数据集包含 844 个样本，每个样本由忠实度标签、摘要和原文构成。其中，忠实摘要源于最先进摘要模型的

输出，不忠实摘要则源于3个部分。第一部分是最先进摘要模型输出的不忠实摘要，经人工筛选后纳入数据集。第二部分是基于模型输出人工构建的不忠实摘要。根据训练数据的转换规则，在摘要中人为引入事实错误。相较于自动转换，人为引入的错误能够保证与原文内容的相关性更高，从而更有效地检验模型的忠实度评估能力。第三部分来自 Kryscinski 等[7]发布的验证集和测试集中的不忠实摘要。这些摘要经过翻译、校正和相关性筛选后，被纳入数据集。对于未通过相关性筛选的摘要，则按照第二部分的方法进行处理。数据集中不忠实摘要的示例如表2所列。

表2 SFETS数据集不忠实摘要示例

Table 2 Examples of unfaithful summaries of SFETS dataset

序号	原文	不忠实摘要
1	韩媒称，国际油价暴跌让中国笑逐颜开。……**美银美林指出**，国际油价每下跌10%，中国GDP会增长0.15%。	**韩媒**：油价每跌10%中国gdp会增长0.15%
2	8月7日，**湖北省**十一家奥迪经销商高管集体赴京，与一汽-大众奥迪的代表共同**接受发改委约谈**。……	**湖北武汉**十一家奥迪品牌经销商集体赴京正式**约谈国家发改委**

4.2 实验设置

实验使用 LCSTS[27]（Large Scale Chinese Short Text Summarization Dataset）新闻摘要数据集。从数据集的原文中抽取36万个单句，构建第一阶段训练集；从参考摘要中抽取72万个单句，构建第二阶段训练集。训练集中忠实与不忠实样本的比例保持1∶1，不忠实样本中通过3种数据增强方法构建的样本比例亦保持1∶1∶1。中文实验的基线系统采用12层的 BERT-base-Chinese[9]模型，使用 Adam 算法进行参数优化，初始学习率设置为 2×10^{-5}，批次大小设置为12，梯度积累设置为8，epoch 设置为10。使用 SFETS 数据集作为验证集和测试集进行参数调优和模型选择。若模型在连续3个 epoch 内未见性能提升，则训练自动结束。英文实验的基线系统采用12层的 BERT-base-uncased 模型，在 CNN/DM[28]新闻摘要数据集上采用相同的方法进行训练。实验在配备有1块 NVIDIA GeForce RTX 3090 GPU 和 24GB 内存的计算机环境下进行。

1）中文实验部分

基于 SFETS 数据集对模型性能进行评估。对于分类式基线方法，选用平衡准确率（Blanced Accuracy，BACC）作为评价指标；对于打分式基线方法，选用受视工作特征曲线下的面积（Area Under Curve，AUC）作为评价指标。受视工作特征曲线是一种图形化工具，用于展示不同分类阈值下真正例率与假正例率的关系。AUC 不受类别不平衡的影响，并且独立于预测阈值，因而被广泛用于评估不同模型的性能。AUC 值的取值为[0，1]，表示模型预测的忠实摘要得分高于不忠实摘要的概率。AUC 值越高，表明模型性能越好。

2）英文实验部分

基于 Rank19[11]数据集对模型性能进行评估。该数据集由 Falke 等提出，在先前研究中被广泛用作生成式摘要测试数据集。Rank19 数据集包含373个样本，每个样本包含来自 CNN/DM 数据集的一个原文单句和两个覆盖相同内容的摘要单句，其中一个摘要句忠实于原文，另一个则不忠实。通过衡量忠实摘要相较于不忠实摘要获得更高评分的频率（Accuracy，ACC）来评价各方法的性能。

4.3 对比实验

4.3.1 基线方法

本文选取以下方法进行对比：

1）ROUGE 系列方法。ROUGE[17]是评估摘要质量的常用指标，ROUGE-N 衡量候选摘要和参考摘要之间 n-gram（通常是单词或双词）的重叠，ROUGE-L 衡量两者间的最长公共子序列，该指标考虑了句子层面的结构相似性。

2）基于语义相似度的方法。BERTScore[18]通过计算候选摘要与参考摘要的 BERT[9]嵌入向量的余弦相似度来评估两者的语义相似度。

3）基于问答的方法。QuestEval[15]通过生成摘要和原文的相关问题，比较这些问题的答案与标准答案的一致性，进而计算摘要的事实分数，其中标准答案主要为实体或关键短语。

4）基于文本蕴含的方法。Kryscinski 等[7]通过替换原文单句中的关键词构建不忠实摘要，训练了忠实度分类模型 FactCC；Cao 等[8]通过替换参考摘要中的关键词构建不忠实摘要，训练了忠实度校正器（Factual Error Correction for Abstractive Summarization Model，FECASM）；Lee 等[10]通过遮蔽原文和参考摘要的部分内容，并利用模型填充遮蔽部分构建不忠实摘要，训练了 MFMA 模型。

为了更全面地评估本文方法的有效性和竞争力，特别设计了与当前领先的两个大型语言模型（LLaMA[29]和 ChatGLM[30]），以及基于 LLM 实现的一致性评估方法 TrueTeacher[24]的对比实验。通过这种比较，不仅可以验证本文方法的实际效果，还可以体现不同模型在处理复杂自然语言处理任务时的优势与局限。

LLaMA 是由 Meta 开发的一种大规模语言模型，其在多项语言任务中表现出色。Llama 3.1 系列模型包括 8B，70B 和 405B 这3个尺寸，其中 405B 模型在与 GPT-4o[22]等模型的对比中，展现出了强大的竞争力。ChatGLM 系列模型中，GLM-4 系列是能力最强的模型，最新的 GLM-4-Plus 在多个关键指标上实现了与 GPT-4o 等顶尖模型相媲美的性能。TrueTeacher 通过利用 FLAN-PaLM 540B[31]注释不同模型生成的摘要来生成合成数据，并在这些数据上训练了最先进的一致性评估模型。

4.3.2 实验结果与分析

实验结果如表3所列，其中，“1 st”表示模型经过第一阶段训练，“2 st”表示模型经过两阶段训练，下同。分析可得出以下结论：

1）在 SFETS 数据集上，FaithEval（2 st）取得了明显的性能提升，BACC 值提高了1.6个百分点，AUC 值提高了5.2个百分点。实验结果验证了模型在提升摘要忠实度评估性能

方面的作用。在 Rank19 数据集上，FaithEval(2 st)的表现良好，仅略低于 BERTScore，推测原因在于，CNN/DM 数据集的平均原文长度超出了 BERT 模型的最大输入长度，在拼接不忠实摘要后，可供模型训练的原文长度进一步减少，这可能限制了模型对原文信息的充分学习。

2)与 BERTScore 相比，FaithEval(2 st)在 SFETS 数据集上 AUC 值提高了 8.6 个百分点，表明了利用 BERTScore 检索构建不忠实摘要，并以此训练忠实度评估模型的有效性。

3)与 QuestEval 相比，FaithEval(1 st)在 SFETS 数据集上 AUC 值提高了 20.6 个百分点，在 Rank19 数据集上 ACC 值提高了 1.3 个百分点。推测性能提升的原因在于 QuestEval 主要从实体角度评估摘要的忠实度，而 FaithEval 不仅涵盖实体，还支持评估其他类别的不忠实错误，展现出更强的可扩展性。

4)FECASM 通过替换参考摘要中的关键词构建不忠实摘要，而 MFMA 则通过遮蔽原文和摘要的部分内容并进行填充构建不忠实摘要。SFETS 数据集上的实验结果表明，MFMA 的 BACC 值相比 FECASM 高出 20.9 个百分点，这表明高质量的不忠实摘要对于提升模型性能至关重要。

5)FactCC 通过替换原文单句中的关键词构建不忠实摘要，与之相比，FaithEval(1 st)即便在数据量不到 FactCC 一半的情况下，在 SFETS 数据集上 BACC 值仍高出 9.9 个百分点，AUC 值高出 14 个百分点。在 Rank19 数据集上，FaithEval(1 st)的 ACC 值也提高了 0.2 个百分点。上述实验结果验证了，所提出的数据增强方法在构建高质量不忠实摘要方面的有效性。

表 3 中英文对比的实验结果

Table 3 Results of comparative experiments in Chinese and English

类别	方法	中文		英文
		BACC	*AUC*	*ACC*
ROUGE	ROUGE-1	—	0.821	0.568
	ROUGE-2	—	0.915	0.630
	ROUGE-L	—	0.854	0.587
语义相似度	BERTScore	—	0.881	**0.713**
问答	QuestEval	—	0.602	0.689
文本蕴含	FactCC	0.625	0.668	0.700
	FECASM	0.545	—	—
	MFMA	0.754	—	—
LLM	TrueTeacher	0.882	—	—
	Llama-3.1-405B	0.860	0.881	0.416
	GLM-4-Plus	0.803	0.892	0.574
	FaithEval(1 st)	0.724	0.808	0.702
	FaithEval(2 st)	**0.898**	**0.967**	0.702

6)在 SFETS 数据集上，FaithEval(2 st)的 BACC 值相比 FactCC，FECASM 和 MEMA 分别提高了 27.3 个百分点、35.3 个百分点和 14.4 个百分点，表明了结合原文和参考摘要进行协同训练对提升模型性能的重要性。

7)在 SFETS 数据集上，LLM 系列方法整体表现良好，表明了 LLM 在评估摘要事实一致性方面的有效性。TrueTeacher 的 BACC 值高于 Llama-3.1-405B 和 GLM-4-Plus，这表明利用 LLM 生成的合成数据训练一致性评估模型可能是一个有潜力的方向。在 Rank19 数据集上，Llama-3.1-405B 和 GLM-4-Plus 的 ACC 值略低，这表明两者在精确区分忠实摘要与不忠实摘要方面可能存在一定局限。

4.4 消融实验

4.4.1 两阶段训练分析

为探究两阶段训练策略的有效性，以 BERT-base-Chinese 模型作为起始检查点，设计以下模型变体进行比较：FaithEval-T 表示仅使用原文单句进行第一阶段训练；FaithEval-TT 表示使用原文单句进行第一阶段和第二阶段训练；FaithEval-TS 表示第一阶段使用原文单句，第二阶段使用参考摘要单句进行训练；FaithEval-MX 表示混合原文单句和摘要单句进行训练。在 SFETS 数据集上的实验结果如表 4 所列，分析后可得出以下结论：

表 4 两阶段训练的实验结果

Table 4 Experimental results of two-stage training

模型	*BACC*
BERT-base	0.522
FaithEval-T	0.721
FaithEval-TT	0.721
FaithEval-TS	**0.893**
FaithEval-MX	0.882

1)与起始检查点相比，FaithEval-T 的性能得到明显提升，这表明基于原文单句构建的训练集使模型得到了有效的学习。

2)与 FaithEval-T 相比，FaithEval-TT 的性能并未进一步提升，这表明仅利用原文单句进行训练对模型性能的提升作用有限。进一步地，与 FaithEval-T 相比，FaithEval-TS 的 BACC 值提高了 17.2 个百分点，这验证了在原文单句训练的基础上引入摘要单句进行第二阶段训练的正确性和有效性。

3)FaithEval-MX 相比于分阶段训练的 FaithEval-TS，性能下降了 1.1 个百分点。这一结果强调了将原文数据和摘要数据分开进行两阶段递进训练的必要性。

4.4.2 数据增强方法分析

探究 3 种数据增强方法对模型性能的影响。实验结果如表 5 所列，可得出以下结论：

1)ES，S3T 和 IFEM 这 3 种方法在单独使用时均提升了模型性能，验证了每种方法的有效性。

2)IFEM 方法表现最佳，原因在于其能够生成与原文内容高度相关的不忠实摘要，这对于训练忠实度分类模型极为有利。

3)S3T 方法表现欠佳，推测原因在于其对单句的转换幅度较大，且检索质量依赖数据集中文章间内容的相关性，导致整体效果缺乏稳定性。

4)ES 方法表现介于 IFEM 和 S3T 方法之间。该方法通过交换实体位置引入不忠实错误，与 IFEM 方法相比，这种错误较为明显和固定，更易于模型识别。而与 S3T 方法相比，基于原文的实体替换使得生成的不忠实摘要在质量上更加稳定和可控。

5)3 种方法混合使用时，总体效果略低于单独使用 IFEM 和 ES 方法的效果。原因在于，为保持不忠实样本总数不变，混合使用时各类样本的数量仅为单独使用时的 1/3，所以效果有所下降，符合预期。

表 5 数据增强方法的实验结果

Table 5 Experimental results of data augmentation methods

模型	*BACC*
BERT-base	0.522
+ES	0.655
+S3T	0.599
+IFEM	**0.831**
+ES+S3T+IFEM	0.651

进一步，探究以不同比例混合 S3T，ES，IFEM 这 3 种数据增强方法构建的训练数据时，模型性能的变化情况。鉴于 3 种方法的性能表现为 S3T＜ES＜IFEM，故实验中保持 ES 方法的比例不变，逐步改变数据集中 S3T 方法和 IFEM 方法的比例。实验结果如图 4 所示，分析可得出以下结论：

1）当数据增强方法的混合比例为 $S3T:ES:IFEM=5:5:5$ 时，模型的 BACC 值最高，这表明 3 种方法的均等使用可以最大程度地提升模型性能。推测原因，可能是该比例能够充分利用各种数据增强的优势，提高模型的泛化能力。

2）当 $S3T:ES:IFEM=2:5:8$ 时，模型的 BACC 值位居第二，这表明在该比例下，ES 和 IFEM 的组合对模型性能有较大的贡献，同时也进一步证实了表 5 中的实验结果。

3）从整体分析来看，数据增强方法及其比例对模型性能有显著影响。不同比例的组合会导致模型性能出现明显差异，因此，选择适当的混合比例对于提高模型性能至关重要。

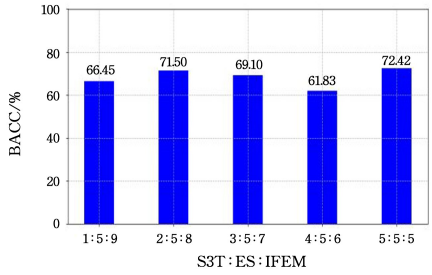

图 4 数据增强方法的混合比例对模型性能的影响

Fig. 4 Effect of mixed ratio of data augmentation methods on model performance

4.4.3 多任务学习分析

本小节探究多任务学习中损失权重 α 对模型性能的影响。实验结果如图 5 所示，分析可得出以下结论。

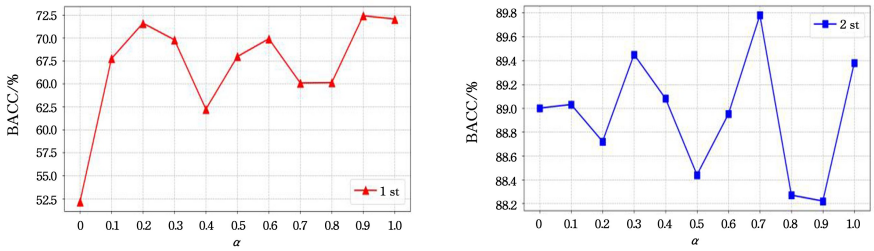

图 5 α 对模型性能的影响

Fig. 5 Effect of α on model performance

1）在模型训练的两个阶段中，与单一的二分类任务（$\alpha=1.0$）相比，引入多分类任务后模型性能得到显著提升，证明了多任务学习在提升模型训练效果方面的有效性。

2）第一阶段训练中，当 $\alpha=0.9$ 时，模型表现出最佳性能。推测原因为训练初期，模型更依赖于二分类任务的指导，有助于为模型的基础分类能力奠定基础。

3）第二阶段训练中，当 $\alpha=0.7$ 时，模型表现出最佳性能。该阶段的训练数据相比第一阶段更为复杂，模型需要更细致地识别样本间的潜在差异，增加了对多分类任务的依赖程度。

4.5 数据量影响分析实验

本节探究训练数据量对模型性能的影响，验证模型的鲁棒性。以训练步数（Step）作为训练数据量的表征，图 6 中展示了模型在不同训练步数下的性能变化，分析可得出以下结论。

1）在 0－8 000 步期间，模型性能提升较为迅速，从 72.42%快速提升到 85.27%，这表明在训练初期，模型从较少的数据中学习到了较多的有效信息，性能显著提高。

2）在 8000－28 000 步期间，模型性能提升相对平缓，从 85.27%逐步提升到 87.68%，说明随着训练的推进，模型逐渐收敛，新数据带来的性能提升幅度变小。

3）28 000 步之后，模型性能有一个较为明显的提升，在达到 89.78%后基本保持稳定，这表明模型在达到一定训练量后，具备了较好的稳定性和鲁棒性，对后续数据量的增加不再敏感，能够保持良好的性能表现。

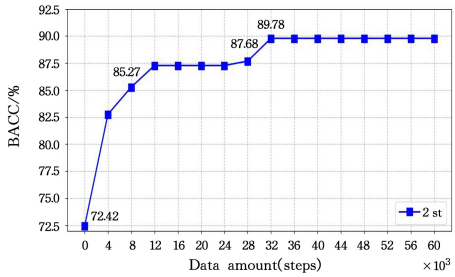

图 6 训练数据量对模型性能的影响

Fig. 6 Effect of training data amount on model performance

结束语 本文提出了一种基于数据增强和两阶段训练的摘要忠实度评估模型 FaithEval。首先，定义了同主题相似检索、外插掩码填充两种数据增强方法，并运用这些方法从文本摘要数据集中提取训练数据；然后，利用基于原文和参考摘要构建的训练数据，对模型进行两阶段训练，逐步增强其对摘要忠实度的理解能力；最后，人工构建摘要忠实度评估测试集 SFETS，为模型性能评估提供基准。实验结果表明，FaithEval 在摘要忠实度评估任务上性能卓越，在 SFETS 数据集上超越了现有基线方法，达到了当前最优水平。

然而，受模型最大输入长度的限制，当原文内容超出该长度时，超出部分将被截断，部分上下文信息丢失，这可能影响 FaithEval 在长文本摘要评估中的表现。因此，未来研究将着

重探索有效的长文本处理策略，如关键信息提取或文本内容的层次化表示，以增强模型对长文本的理解与评估能力。此外，未来研究还将进一步探索 FaithEval 在多语言环境中的应用潜力，通过跨语言迁移学习或者构建多语言数据集的方式，使模型满足不同语言场景下的摘要忠实度评估需求。

参 考 文 献

[1] KRYSCINSKI W，KESKAR N S，MCCANN B，et al. Neural Text Summarization：A Critical Evaluation[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019：540-551.

[2] WU S X，HUANG D G，LI J Y. Abstractive Text Summarization Based on Semantic Alignment Network[J]. Acta Scientiarum Naturalium Universitatis Pekinensis，2021，57(1)：1-6.

[3] CHEANG C，CHAN H，WONG D，et al. TempoSum：Evaluating the Temporal Generalization of Abstractive Summarization[J]. arXiv：2305. 01951v1，2023.

[4] SUN K L，LUO X D，LUO Y R. Survey of Applications of Pretrained Language Models[J]. Computer Science，2023，50(1)：176-184.

[5] CAO Z，WEI F，LI W，et al. Faithful to the Original：Fact Aware Neural Abstractive Summarization[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. AAAI，2018：4784-4791.

[6] PAGNONI A，BALACHANDRAN V，TSVETKOV Y. Understanding Factuality in Abstractive Summarization with FRANK：A Benchmark for Factuality Metrics[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. ACL，2021：4812-4829.

[7] KRYSCINSKI W，MCCANN B，XIONG C，et al. Evaluating the Factual Consistency of Abstractive Text Summarization[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. ACL，2020：9332-9346.

[8] CAO M，DONG Y，WU J，et al. Factual Error Correction for Abstractive Summarization Models[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. ACL，2020：6251-6258.

[9] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. 2019：4171-4186.

[10] LEE H，YOO K M，PARK J，et al. Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking[C]// Proceedings of the Findings of the Association for Computational Linguistics. ACL，2022：1019-1030.

[11] FALKE T，RIBEIRO L F R，UTAMA P A，et al. Ranking Generated Summaries by Correctness：An Interesting but Challenging Application for Natural Language Inference[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL，2020：2214-2220.

[12] HUANG Y，FENG X，FENG X，et al. The Factual Inconsistency Problem in Abstractive Text Summarization：A Survey[J]. arXiv：2104. 14839，2021.

[13] LUO Z，XIE Q，ANANIADOU S. ChatGPT as a Factual Inconsistency Evaluator for Abstractive Text Summarization[J]. arXiv：2303. 15621v1，2023.

[14] GOODRICH B，RAO V，LIU P J，et al. Assessing The Factual Accuracy of Generated Text[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019：166-175.

[15] SCIALOM T，DRAY P A，GALLINARI P，et al. QuestEval：Summarization Asks for Fact-based Evaluation[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics，2021：6594-6604.

[16] DURMUS E，HE H，DIAB M. FEQA：A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization[C]// Proceedings of the Annual Meeting of the Association for Computational Linguistics. ACL，2020：5055-5070.

[17] LIN C. ROUGE：A Package for Automatic Evaluation of Summaries[C]// Proceedings of the Meeting of the Association for Computational Linguistics. 2004：74-81.

[18] ZHANG T，KISHORE V，WU F，et al. BERTScore：Evaluating Text Generation with BERT[C]// Proceedings of the 8th International Conference on Learning Representations. 2020.

[19] KOCMI T，FEDERMANN C. Large Language Models Are State-of-the-Art Evaluators of Translation Quality[C]// Proceedings of the 24th Annual Conference of the European Association for Machine Translation. Tampere，Finland：European Association for Machine Translation，2023：193-203.

[20] WANG J，LIANG Y，MENG F，et al. Is ChatGPT a Good NLG Evaluator? A Preliminary Study[C]// Proceedings of the 4th New Frontiers in Summarization Workshop. ACL，2023：1-11.

[21] LIU Y，ITER D，XU Y，et al. G-EVAL：NLG Evaluation using GPT-4 with Better Human Alignment[C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. ACL，2023：2511-2522.

[22] OPENAI. GPT-4 Technical Report [J]. arXiv：2303. 08774，2023.

[23] WANG P，LI L，CHEN L，et al. Large Language Models are not Fair Evaluators[J]. arXiv：2305. 17926v2，2023.

[24] GEKHMAN Z，HERZIG J，AHARONI R，et al. TrueTeacher：Learning Factual Consistency Evaluation with Large Language Models[C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. ACL，2023：2053-2070.

[25] BLEI D M，NG A Y，JORDAN M T. Latent Dirichlet Allocation[C]// Proceedings of the 15th Annual Neural Information Processing Systems Conference. Vancouver，BC，Neural Information

Processing Systems Foundation, 2002: 601-608.
[26] LEWIS M, LIU Y, GOYAL N, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[C]// Proceedings of the Annual Meeting of the Association for Computational Linguistics. ACL, 2020: 7871-7880.
[27] HU B, CHEN Q, ZHU F. LCSTS: A Large Scale Chinese Short Text Summarization Dataset[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. ACL, 2015: 1967-1972.
[28] HERMANN K M, KOISKY T, GREFENSTETTE E, et al. Teaching Machines to Read and Comprehend[C]// Proceedings of the 29th Annual Conference on Neural Information Processing Systems. Montreal, QC: Neural Information Processing Systems Foundation, 2015: 1693-1701.
[29] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and Efficient Foundation Language Models[J]. arXiv: 2302.13971, 2023.
[30] GLM T, ZENG A, XU B, et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools[J]. arXiv: 2406.12793, 2024.
[31] CHUNG H W, HOU L, LONGPRE S, et al. Scaling Instruction-Finetuned Language Models[J]. arXiv: 2210.11416, 2022.

ZHAO Jinshuang, born in 2000, postgraduate, is a member of CCF (No. Z2722G). Her main research interests include natural language processing and text summarization.

HUANG Degen, born in 1965, Ph.D, professor, is a member of CCF (No. 17961S). His main research interests include natural language processing, machine translation and text summarization.

(责任编辑:何杨)

2025 年 CCF 会士提名将于 11 月 1 日截止

根据《中国计算机学会会士条例》的规定,2025 年度 CCF 会士候选人提名工作将于 11 月 1 日截止。

CCF 会士候选人的资格

1)候选人在提名截止日前(2025 年 11 月 1 日)在计算机或相关领域从业 15 年以上(受高等教育期间的从业时间按如下方式计算:学士 2 年、硕士 4 年、博士 6 年,按最高学历计算,不累计)、本学会会龄 5 年以上,并在计算机及相关领域有重大发明创造及有重要贡献、或对本学会发展有重要贡献的人士。

2)候选人须得到一名主提名人和两名附议人的提名。

提名人的资格

CCF 会士和杰出会员为有效提名人(如会员资格失效,则提名无效)。

提名要求

1)每位提名人作为主提名人提名的会士候选人不得超过 2 人,作为附议提名人不得超过 3 人(即作为主提名+附议提名共不超过 5 人)。

2)作为主提名人提名时,应保证所提名的会士候选人获得另外两名附议提名人的提名。

提名方式(二选一):

线下提名:主提名人将提名表以 word 文档格式(切勿用 PDF 格式)通过邮件附件发送到会员部指定邮箱(xliu@ccf.org.cn),主题为:会士提名+被提名人姓名,邮件需同时抄送两名附议提名人,并请主提名人在邮件正文里署名,收到邮件回复后视为提交成功。

线上提名:由主提名人填写提名表后登录 CCF OA(http://oa.ccf.org.cn)系统提交,附议提名人收到提醒邮件通过链接登录 CCF OA 系统提交意见后完成附议提名。请点击文末"阅读原文"下载提名表及线上具体操作流程。

联系人:

CCF 会员部 刘霞 xliu@ccf.org.cn/010-6264 8654

据 CCF 微信公众号