千言数据集：问题匹配鲁棒性竞赛技术报告

田昱锟[1)]

[1)](计算机科学与工程学院, 东南大学, 南京 210000)

摘　要　本文探讨了一种基于机器学习的文本匹配方法，旨在解决自然语言处理中的问答匹配问题。该问题在问答系统、对话系统及信息检索等多个领域具有广泛的应用价值，本文所研究的任务背景为百度千言问题匹配鲁棒性竞赛。技术方案的核心在于结合数据增强、预训练模型、GRU 网络和后处理修正方法，以提升系统的整体性能。文章详细分析了多种预训练模型及技术方案在该任务中的表现，比较了不同模型在语法结构、词汇变化及语义变化方面的鲁棒性。通过在复杂测试数据集上的实验，研究展现了模型在语义理解、泛化能力及处理速度方面的优势，并探讨了进一步优化模型性能的可能路径。最后，本文展望了未来的改进方向，提出了提升问答匹配系统准确性和效率的潜在方法。截至 2024 年 10 月 17 日，本方案在竞赛总榜上排名第 6（东南大学参赛选手中排名第 1），总分 89.901。

关键词　自然语言处理；机器学习；文本匹配

Qianyan Dataset: Technical Report on the Question Matching Robustness Competition

Yukun Tian[1)]

[1)](School of Computer Science and Engineering, Southeast University, Nanjing 210000)

Abstract　This paper explores a machine learning-based text matching approach aimed at addressing the question matching problem in natural language processing (NLP). This task holds significant value in various fields, including question-answering systems, dialogue systems, and information retrieval, with the research background rooted in the Baidu Qianyan Question Matching Robustness Competition. The core of the proposed technical solution combines data augmentation, pre-trained models, GRU networks, and post-processing correction methods to enhance overall system performance. The paper provides a detailed analysis of multiple pre-trained models and technical solutions in this task, comparing their robustness to syntactic structures, lexical variations, and semantic changes. Experiments on complex test datasets demonstrate the models' performance in terms of semantic understanding, generalization ability, and processing speed, while also discussing potential avenues for further optimization. Finally, the paper outlines future work and proposes potential methods to improve the accuracy and efficiency of question matching systems. As of October 17, 2024, this solution ranks 6th overall in the competition (1st among Southeast University participants), with a total score of 89.901.

Key words　Natural Language Processing; Machine Learning; Question Matching

1　问题分析

1.1　问题背景

问题匹配（Question Matching）任务旨在判断两个自然问句之间的语义是否等价，是自然语言处理领域一个重要研究方向。问题匹配同时也具有很高的商业价值，在信息检索、智能客服等领域发挥重要作用。近年来，神经网络模型虽然在一些标准的问题匹配评测集合上已经取得与人类相仿甚至超越人类的准确性，但是在处理真实应用场景问题时，这些模型鲁棒性较差，在非常简单（人类很容易判断）的问题上无法做出正确判断，造成了极差的产品体验和经济损失。

当前大多数问题匹配任务在与训练集同分布的测试集上进行测试，夸大了模型能力，缺乏对模型细粒度优势和劣势的评测。因此，本次评测关注问题匹配模型在真实应用场景中的鲁棒性，从词汇、句法、语用等多个维度检测模型的能力，发现模型的不足之处，推动语义匹配技术的发展。本次评测集中的样本均来自于搜索问答和对话型问答两个场景，难度大，考察点丰富，覆盖了真实应用中诸多难以解决的问题。

收稿日期：2024-10-17；修改日期：2024-10-17　田昱锟（通信作者），男，2004 年生，本科在读，计算机学会 (CCF) 学生会员（会员号：U6088G），主要研究领域为计算机视觉、多模态学习等. E-mail: 21322187@seu.edu.cn.

1.2 数据分析

1.2.1 数据源

本次竞赛主要基于千言数据集，采用的数据集包括了哈尔滨工业大学（深圳）的 LCQMC 和 BQ 数据集、OPPO 的小布对话短文本数据集、谷歌 PAWS 数据集，以及百度的 DuQM 数据集。本次评测，训练集由 LCQMC、BQ、小布对话短文本、PAWS 数据集组成，测试集由 DuQM、小布对话短文本数据集组成，从词汇、句法、语用 3 大维度评估模型，期望从多个维度、多个领域的角度评价模型的鲁棒性，进一步提升问题匹配技术的研究水平。

训练集：包含四个文本相似度数据集，分别为哈尔滨工业大学（深圳）的 LCQMC、BQ Corpus、谷歌 PAWS 数据集以及 OPPO 小布对话短文本数据集。4 个数据集的任务一致，都是判断两段文本在语义上是否匹配的二分类任务。

测试集：百度 DuQM 测试集：通过对搜索问答场景中的原始问题进行替换、插入等操作，并过滤掉真实场景中未出现过的问题，保证扰动后问题的自然性和流畅性，然后进行人工筛选和语义匹配标注，得到最终的评测集。OPPO 小布对话短文本测试集：采样自 OPPO 语音助手小布的真实对话场景数据，进行人工筛选和语义匹配标注，得到最终的评测集。

1.2.2 数据分析结论

题目特点:

1）题目测试集和训练集分布不同（大部分来自于不同数据集），任务难度很大，需要模型非常强的通用语义处理能力和泛化能力；

2）本竞赛采用比较少见的测试集公开方式测评，通过对测试集的大概观察，不难发现，测试集的语义粒度很细，很多问题匹配的难度较大，而且由于采用了不同的数据集，OPPO 数据集上的匹配判定标准和 baidu 数据集也有些微的区别。

3) 根据[1] 介绍的几种观测方法观察测试集可知：如图 1(a)、图 1(b) 所示，训练集和测试集中句对间细粒度差异主要有替换、插入、交换三种。图 1(c) 图 1(d) 所示，训练集和测试集中大部分文本句对的编辑距离得分较高，文本内容比较类似。其中，训练集测试集分布有所差异，测试集更注重细粒度变化的考察。

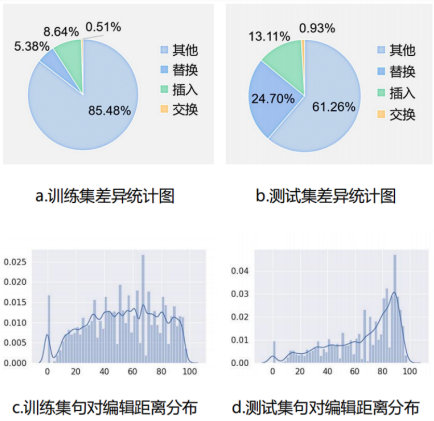

图 1 训练集和测试集数据概况[1]

2 方案设计

2.1 数据预处理方案

主要是针对自然语言的文本特性，针对训练集进行数据增强：在自然语言处理任务中，数据增强是提高模型泛化能力和鲁棒性的重要手段。以下针对等价性推断、文本同音替换、近义词替换等数据增强策略进行详实的描述。通过传递性推断、同音替换以及近义词替换等多种数据增强方法，可以有效提高模型的训练数据丰富性和模型的泛化性能。这些策略从不同角度出发，模拟了文本中语义、语音或表达方式的多样性，进而使模型在面对复杂的真实场景时具备更强的鲁棒性与适应性。

1. 传递性推断增强

通过等价性传递性对数据进行增强。若已知 query_1 与 query_2 等价且 query_2 与 query_3 等价，则根据传递性可推断出 query_1 与 query_3 等价。此推断过程能够显著增加训练数据的多样性，进而提升模型的泛化能力。与此同时，还可以通过对等价性进行逆推，若 query_1 与 query_2 等价，则自然可以推断出 query_2 与 query_1 等价，这进一步为数据增强提供了新的方向。

$$\text{If } q_1 \equiv q_2 \text{ and } q_2 \equiv q_3, \text{ then } q_1 \equiv q_3$$

2. 同音同调、同音异调替换

在文本分析过程中，发现大量文本中存在同音同调或同音异调的字词，表明音素相同但字符有细微变化。这类现象在某些语言（如汉语）中尤为常见。为提高模型对这些变体文本的泛化能力，可以通过随机替换文本中的字为相应的同音同调或同音异调的字。例如，将“调”替换为“条”（同音同调）或“掉”（同音异调），并保持原文本的标签不变。

这种方法可以模拟实际应用中的语音输入或口语表达的多样性，增强模型对变体表达的容错性，进而提高模型的鲁棒性。

For example: 调 → 条 (同音同调), 调 → 掉 (同音异调)

3. 近义词替换

通过对训练文本和测试文本的观察与分析，发现存在大量词汇替换的现象，特别是近义词和反义词的替换。在数据增强过程中，可利用近义词库对文本进行近义词替换操作，且替换后的文本标签保持不变。这种策略增加了训练样本的多样性，模拟了用户可能用不同表达方式提出相同问题的场景，从而提升模型对语言多样性的适应能力。

例如，将“迅速”替换为“快捷”或“快速”，这些近义词替换不会改变原文本的语义或标签。

For example: 迅速 → 快捷, 快速

2.2 模型 & 算法设计

2.2.1 预训练模型

此处由于使用百度的 AI Studio 作为算力平台，故主要选择 paddlepaddle 系列的 ERNIE 模型，备选的方案包括 ERNIE 3.0[2], ERNIE-Gram[3] 等。同时还尝试了在自己的算力设备上训练部分知名文本处理模型，包括 SBERT[4], RoBERTa[5] 等。最终经过测试和 Baseline 的结果（ERNIE-Gram）来进行对比，最终在本任务场景（中文语义匹配）中 ERNIE-Gram 模型的性能表现最佳，故使用 ERNIE-Gram 作为预训练模型 Backbone 的选择。

2.2.2 模型设计方案

1. **模型选择**：使用了 **ERNIE-Gram** 作为预训练模型，该模型能够通过对文本进行预训练来获取丰富的词法和句法特征，进而增强句子的理解能力。

2. **特征提取与融合**：从 **ERNIE-Gram** 模型中提取 *sequence_output* 和 *pooled_output* 两类特征。

- *sequence_output* 表示每个序列位置的隐藏状态向量。
- *pooled_output* 则捕捉整个序列的全局特征。

进一步使用 **GRU（门控循环单元）网络**来处理 *sequence_output*，捕获文本序列中的长短期依赖信息。

3. **池化操作**：GRU 网络输出之后，进行 **最大池化（Max Pooling）**和 **平均池化（Avg Pooling）**操作，分别提取特征图中的最大值和平均值。这两种池化操作确保模型能从不同角度获取文本特征，提升模型的泛化能力。

4. **分类层**：池化后的特征通过一个 **线性层（Linear Layer）**进行进一步的特征转换与降维。最后，通过 **Sigmoid** 激活函数输出分类结果。Sigmoid 函数常用于二分类任务，其输出范围为 $[0, 1]$，可解释为属于某类别的概率。

综上所述，该模型设计通过结合预训练模型特征、序列建模的长短期依赖处理、池化操作与多任务学习等方法，有效提升了文本分类任务的准确性与鲁棒性。

2.3 训练 & 调参策略

1. 概览

本方案详细介绍了使用 PaddlePaddle 框架进行深度学习模型训练与调参的过程，主要涵盖了学习率调度机制、优化器配置、损失函数定义、模型训练与验证策略，以及模型的保存逻辑。通过合理的超参数设置和模型调优手段，提升模型的收敛速度与泛化性能。

2. 学习率调度与优化器配置

模型采用线性衰减和热身策略对学习率进行动态调整，旨在通过初期的较小学习率提高模型的稳定性，并在后期进行逐步衰减。总训练步数 num_training_steps 由以下公式确定：

$$\text{num\_training\_steps} = \text{len(train\_data\_loader)} \times \text{epochs}$$

其中，优化器采用了 AdamW[6] 算法，能够有效地控制权重衰减，防止过拟合。优化器对除去“bias”和“norm”相关参数的权重应用衰减，进一步优化模型性能：

$$\text{apply\_decay\_param\_fun} = \lambda x : x \in \text{decay\_params}$$

3. 损失函数与评价指标

为优化分类模型，损失函数选用了交叉熵损失（CrossEntropyLoss），常用于多类分类问题。评价指标使用分类准确率（Accuracy），其计算与更新流程贯穿于训练与验证过程：

$$\text{criterion} = \text{paddle.nn.loss.CrossEntropyLoss()}$$

4. 训练与验证流程

每个训练轮次包含多个训练步，在每一步中，模型通过前向传播计算输出，并使用交叉熵损失

函数计算损失值 *ce_loss*。如果存在 KL 散度损失 *kl_loss*，则总损失公式如下：

$$\text{loss} = ce\_loss + kl\_loss \times \text{args.rdrop\_coef}$$

通过反向传播机制 backward()，损失函数的梯度被传播至模型的各个参数，从而通过优化器进行参数更新：

optimizer.step() and lr_scheduler.step()

5. 模型验证与保存

每隔固定步数，模型在开发集（dev set）上进行验证，输出验证损失和准确率。当验证集的准确率优于历史最佳值时，模型的参数状态会被保存至预定的路径，以保证模型最优状态的存储：

6. 结果监控

为了实时监控模型的训练过程，系统定期打印全局步数、训练轮次、当前批次、损失、交叉熵损失、KL 散度损失、准确率以及训练速度等关键信息，有助于用户随时掌握模型的训练进展。

2.4 后处理（结果修正）算法设计

由于模型为了保证整体的语义提取能力，很多时候实际上并不能针对特定类型的文本给出比较好的匹配判断，例如通过上述方法设计和训练的模型，在当前测试集上的表现如下表中第一行所示，模型在有明显弱项的情况下整体的表现性能相当好，说明模型已经具备较强的语义提取能力，在大部分的任务上表现很好。所以，后续工作主要集中于针对模型弱项进行补强，通过单独校正的方式修正模型判断。此处主要基于[7] 的修正方法进行了改良，主要从拼写、语法结构和语义内容等角度进行校正，通过针对几种典型的高难样本设计专门的后处理，来提高最终在特定类型文本对上的匹配正确率。

表 1 修正前后模型表现对比

模型	score	asymmetry	neg_asymmetry	misspelling
无修正	82.126	70.624	36.735	82.906
修正	89.899	86.72	65.306	99.359

2.4.1 拼写错误识别与校正

该模块的主要任务是识别和校正句子中的拼写错误。为此，导入了一些必要的工具库，并依次进行以下操作：

1. **找出完全相同的句子对**：找到 `text_1 == text_2` 的句子对，直接排除这些句子对，因其不涉及拼写错误。
2. **检测仅有同音字不同的句子对**：使用 `lazypinyin` 识别仅存在同音字不同的句子对，这些句子可能因拼音相似而存在拼写错误。
3. **检测包含拼写错误的句子对**：利用 `pycorrector` 库对句子进行拼写错误检测，并标记出存在错误的句子对。该步骤预计耗时 1.5 小时。
4. **去除部首后的拼写错误检测**：针对汉字的特殊性，去除部首后再对句子进行比较，以进一步排查拼写错误。该步骤预计耗时 0.7 小时。
5. **同音名与同音地名不作为拼写错误处理**：对同音的名字和地名进行特殊处理，这类同音异义的词汇不作为拼写错误来处理。

2.4.2 词汇语义识别与校正

该模块主要负责对句子的词汇语义进行识别和校正，处理涉及名词、动词、形容词、副词等不同词类的操作。具体步骤如下：

1. **插入名词/动词/形容词/副词等词汇**：在适当的句法位置插入新的词汇，评估对句子语义的影响并确保句法的合理性。
2. **替换名词/动词/形容词/副词等词汇**：对句子中的词汇进行替换（例如近义词替换），并评估替换对句子整体语义的影响。此步骤预计运行时长为 0.1 小时。
3. **校正词汇语义的预测值**：根据词汇语义模型的预测结果，对其进行调整，以保证词汇的语义在上下文中的一致性与准确性。

2.4.3 句法结构识别与校正

该模块的主要任务是识别并校正不同句法结构的句子。具体操作如下：

1. **校正对称性与不对称性**：对句子中对称性和不对称性的结构进行校正，保证语法结构的平衡。
2. **校正负不对称性**：处理句子中否定表达的不对称性，确保句意一致性。

3. **寻找包含不同词语或反义词的句子对**（阈值为0.5）：通过计算语义相似度或差异，识别包含不同词或反义词的句子对，并分析其影响。
4. **处理包含添加或删除减乘除等副词的句子对**：处理句子中由于副词（如增加或删除修饰词）导致的句法结构变化。
5. **处理主动语态/被动语态的句子对**：针对主动和被动结构的句子对，分析和校正句子结构。

3 实验与结果验证

上述各项技术插入后最终表现如下，组织了消融实验：

表 2 模型性能总览

模型	score
ERNIE-Gram-baseline	78.584
Baseline + Corrector	86.569
Augmentation + Mymodel	86.342
Augmentation + Mymodel + Corrector	**89.901**

4 结果分析与讨论

目前本技术方案最终结果在**“千言数据集：问题匹配鲁棒性”竞赛中排名总榜第 6 名，所有东南大学参赛者中排名第 1 名**，总分达 89.901 分。

图 2 竞赛排行榜（截止 10.17 12:00）

实际上，还有一种能够显著提高最终表现的策略[1] 但是由于算力限制并未采用：

模型集成——通过集成多个不同模型的表现，达到最终取得较高表现的结果。此处需要多个基础模型和一个权重判断模型，对多个基础模型进行训练和调参对于整体算力上的要求太高了，故并未采用。我个人猜测前 5 名选手大部分应该都采用了这个方案，此处提出供读者参考。

致 谢 非常感谢魏通老师和各位助教学长的帮助和指导。

参考文献

[1] DataFountain Discussion Article: 千言数据集：问题匹配（Question Matching）赛道介绍[EB/OL]. https://discussion.datafountain.cn/articles/detail/3813.

[2] SUN Y, WANG S, FENG S, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation[EB/OL]. 2021. https://arxiv.org/abs/2107.02137.

[3] XIAO D, LI Y K, ZHANG H, et al. ERNIE-gram: Pre-training with explicitly n-gram masked language modeling for natural language understanding[C/OL]//TOUTANOVA K, RUMSHISKY A, ZETTLEMOYER L, et al. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 1702-1715. https://aclanthology.org/2021.naacl-main.136. DOI: 10.18653/v1/2021.naacl-main.136.

[4] REIMERS N, GUREVYCH I. Sentence-bert: Sentence embeddings using siamese bert-networks[EB/OL]. 2019. https://arxiv.org/abs/1908.10084.

[5] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[EB/OL]. 2019. https://arxiv.org/abs/1907.11692.

[6] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization [EB/OL]. 2019. https://arxiv.org/abs/1711.05101.

[7] AI Studio Project on 千言问题匹配[EB/OL]. https://aistudio.baidu.com/projectdetail/2384565?searchKeyword=%E5%8D%83%E8%A8%80%E9%97%AE%E9%A2%98%E5%8C%B9%E9%85%8D&searchTab=ALL.

附录 A

源代码和数据集见附件。

Yukun Tian, **Male**, born in 2004, currently an undergraduate student, student member of the China Computer Federation (CCF) (ID: U6088G). His main research areas include computer vision and multimodal learning.

Background

Question Matching (QM) is a task aimed at determining whether two natural language questions are semantically equivalent. It is an important research direction in the field of natural language processing (NLP). Question matching also holds significant commercial value, playing a key role in areas such as information retrieval and intelligent customer service.

In recent years, while neural network models have achieved human-like or even superhuman accuracy on some standard question matching benchmark datasets, they exhibit poor robustness in real-world applications. These models often fail to make correct judgments on very simple questions (which are easily distinguishable by humans), leading to poor product experience and economic losses.

Currently, most question matching tasks are tested on test sets that share the same distribution as the training data, which overestimates the model's capabilities and lacks fine-grained evaluation of the model's strengths and weaknesses. Therefore, this evaluation focuses on the robustness of question matching models in real-world application scenarios. The evaluation assesses the model's performance across multiple dimensions, including lexical, syntactic, and pragmatic variations, aiming to identify shortcomings and promote the development of semantic matching technology. The samples in this evaluation set are derived from search-based QA and conversational QA scenarios, featuring high difficulty, diverse evaluation points, and covering many challenges encountered in real applications.