文本相关性实验报告

实验背景（问题描述）

在自然语言处理（NLP）中，文本相关性任务旨在评估两段文本之间的相关程度，这对于搜索引擎、推荐系统、问答系统等应用至关重要。BERT（Bidirectional Encoder Representations from Transformers）作为一种强大的语言表示模型，已被广泛应用于多种NLP任务中，包括文本相关性评估。本实验旨在探索BERT模型在文本相关性任务上的表现。

实验方法

数据集

中文自然语言推理数据集（A large-scale Chinese Nature language inference and Semantic similarity calculation Dataset） 本数据及通过翻译加部分人工修正的方法，从英文原数据集生成，可以一定程度缓解中文自然语言推理和语义相似度计算数据集不够的问题。

本实验采用其中用于文本相似度的数据集Chinese-STS-B，如下：

https://github.com/pluto-junzeng/CNSD?tab=readme-ov-file

	Train	Dev	Test	Sum
Chinese-STS-B	5.7k	1.5k	1.3k	8.5k

模型介绍

BERT

BERT是一种预训练语言表示模型，使用Transformer作为基础架构，通过在大规模文本语料库上进行预训练，学习深层次的语言特征。BERT的双向训练机制使其能够有效捕获文本中的上下文信息，这对于理解两段文本之间的相关性非常重要。具体结构在这里不细展开，论文链接如下：

https://arxiv.org/abs/1810.04805

预训练模型选择

使用开源`bert-base-chinese`以下为模型介绍：

- **模型描述**

本模型为专门针对中文预训练的BERT模型。在预训练过程中，我们采用了原始BERT论文中描述的方法，对词片（word pieces）独立应用了训练和随机遮蔽。这种预训练策略旨在使模型能够更好地理解和生成中文语言。

- **训练过程**
  - `type_vocab_size`: 2
  - `vocab_size`: 21128
  - `num_hidden_layers(encoder)`: 12
- **模型类型**
  - **类型**：填充-掩码（Fill-Mask）
  - **适用语言**：中文
  - **基础模型：**该中文BERT模型建立在BERT基础（uncased）模型之上。

实验设置

- **模型训练参数**：预训练模型在数据集的训练集上进行微调，学习率为2e-5，批大小为64，训练10个epoch，取验证集上准确率表现最好的一个epoch进行测试操作。
- **评价指标**：使用模型在测试集上的准确率、召回率和F1分数来评估模型性能。
- **硬件配置**：训练过程在NVIDIA RTX3070 GPU的本地电脑上进行。
- **数据处理：**
  - 对数据进行裁切，最长长度为97（100-3），在这个条件下数据集中有119条过长数据被裁切，使用最长优先裁剪的方法来处理过长数据。
  - 原数据格式text1 text2 label，训练前将text1和text2拼接成一个句子并且加入标记词、分割词、padding等：`[CLS] text1 [SEP] text2 [SEP] [PAD]` padding的填充为最大长度，也就是到100长度。
  - 完成上述操作以后用Transformer库中自带的tokenizer进行分词
- 优化器&损失函数：优化器使用AdamW，损失函数使用交叉熵函数，用结果序列第一个token也就是[CLS]位置的token过一层softmax层得到每一类的概率，再用最大概率的类别序号和label作为损失函数的输入进行训练。

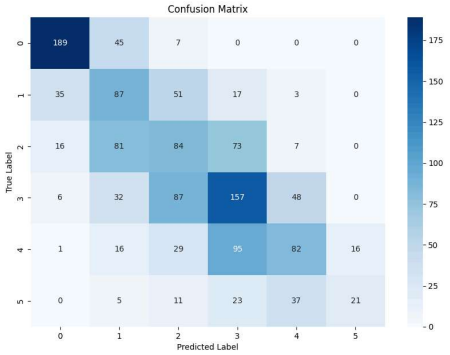
实验结果

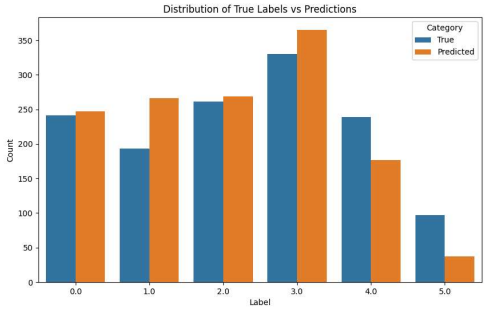
性能比较

注：此处2类和6类用的是同一个模型，因为考虑到接受度，把相关度为0的作为高相关内容推送比，相关度为1的错误推送要更加不可接受，在此希望保留对正负样本的内部比较。

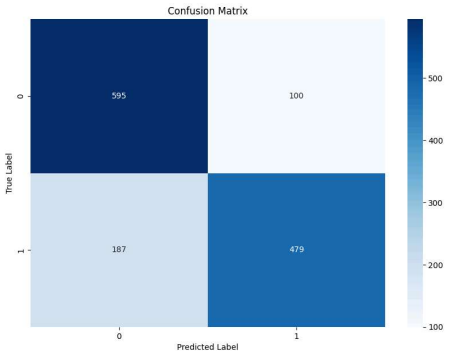
	Accuracy	F1 Score	Recall	Bad Case
bert(6类)	0.4555	0.4528	0.4555	0.1271（相差绝对值大于1）
bert(2类)	0.7891	0.7695	0.7192	0.2109

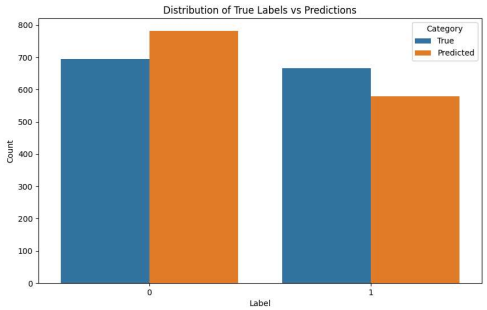
可视化分析





混淆矩阵和数据分布的可视化分析表明，BERT模型在绝大多数情况下能够准确预测文本之间的相关性等级。尤其是在高度相关（4-5分）的句子对中，BERT展现出了较高的敏感度。另外，原数据大体符合正态分布，可以考虑成另类的回归问题。





2分类用来评判展示数据的可接受读，所以错误的负例非常关键，模型中标签为0但预测为1的值所占比例不多，这是比较好的，商业场景不容易降低用户体验。

Badcase分析（详见bert_output_badcase.txt)

常见错误类型（靠前的是标签，靠后的是预测值，这里定义分数上下浮动超过1分算badcase)

语义问题：

- 但该中心说，奥黛特是第一个在12月在加勒比海上空形成的。　　这是第一次命名风暴在加勒比发展在12月。　4　2
- 商业报纸新品种上市　　好莱坞贸易刊物“综艺”挂牌销售　4　0
- 我想你在找米奇(1992)。 我想你在找那部电影　　3　1

分析：

语义问题通常涉及对句子含义的深层理解，以及如何准确地捕捉到两个句子之间的相似度或差异度。这些例子显示了在处理具体信息时，模型可能未能准确把握细节之间的联系，导致评分不准确。如：

- 无法得知奥黛特是一个风暴的名字
- 无法得知该刊物是一种商业报纸。
- 无法得知米奇（1992）是一部电影。

改进点：

1. **使用更丰富的训练数据**：引入具有丰富语义关系的句子对，特别是包含细节对比的数据，有助于模型学习更细微的语义差异。
2. **利用外部知识库**：对于涉及具体知识（如人物、事件）的语义理解，可以结合外部知识库进行语义增强，以改善模型的理解和推理能力。
3. **增加模型规模：**增加模型的规模可以让模型更加充分的学习隐含的语义信息。

翻译问题:

- 一个人正在进行劳动。　今天有个人要表演。　　2　0
- 我要冒个险.　我在这里会很直接的。　2　0
- 猫自己打扫卫生。　猫在舔自己。　3　1

分析：

翻译问题主要源于语言间表达方式的差异，数据中提及了有一部分数据是通过翻译英文数据得到的，但是直接翻译可能无法准确反映原意，特别是在文化或习语表达上的差异。如：

- 劳动在英文中是working，其中包含所有工作内容，但是在中文中更偏向体力劳动。
- 很直接straightly有直接做果断做的意思，但是中文没有做事的含义。
- clean在英文中有自我清洁的隐含，尤其是对动物，但是翻译成打扫卫生和猫在舔自己就完全不一样了。

改进点：

1. **优化翻译流程**：在模型训练之前，对翻译文本进行人工审核或使用更高质量的翻译工具，以确保翻译的准确性和自然性。
2. **跨语言预训练**：使用跨语言预训练模型（如mBERT、XLM-R）作为基础，这些模型在多语言文本上进行预训练，对处理翻译文本有天然的优势。
3. **不对数据进行翻译：**工业上的文本相关性任务往往需要同时能够接受中文和英文，不进行翻译，而训练一个多语料的模型更加使用可行。

标准制定问题：

- 你也可以用它。　　是的，你能做到的。　　3　1
- 你必须决定你想从中得到什么。　　你得找出适合你的方法。　　4　1
- 我也有同样的想法。　　我也有同样的问题。　　4　2

分析：

标准制定问题反映了在给定标签或评分时存在的主观性和不一致性，特别是在判断两个句子之间的相关性或相似度时。虽然我不能得知标注的标准，但是我确信有一些存在标注错误。如：

- 用它和能做到不是一致的，居然给到3分。
- 目的和方法一致性也不够强，居然给到了4分。
- 想法和问题不是很相似，居然给到了4分。

改进点：

1. **明确评分标准**：制定更详尽、明确的评分指南，对不同级别的语义相关性给出具体的例子和解释，以减少标注时的主观性。

2. **进行标注人员培训**：对参与数据标注的人员进行培训，确保他们理解评分标准，并能够一致地应用这些标准。
3. **引入多人标注和审核**：对每个句子对进行多人标注，然后通过讨论或使用专家审核来解决标注间的不一致问题，以提高标注数据的质量和一致性。

通过这些分析和改进措施，我想可以有效地解决语义理解、翻译准确性以及评分标准一致性方面的问题，从而提高模型在文本相关性任务上的表现。

结论

对BERT的结果分析，可以发现BERT在处理复杂的语义关系和上下文信息时具有显著优势，但是在处理包含隐藏的知识信息的语句时展现出语义深度和广度的缺失，令人有些遗憾。强大的LLM可以很好的解决这个问题，并在这个任务中达到几乎满分的答案，但是不管是堆砌计算资源的成本还是计算所需的时间成本都远远大于BERT。

未来的研究可以探索结合外部知识库或进一步优化模型结构来解决出现的问题，或者在不削弱LLM知识储备的前提下，减少运算所需的计算资源。

实验不足

虽然本实验取得了一定的成果，但在多个方面存在局限性和不足，在该系统投入商业使用前必须进一步改进和深入研究。

计算资源限制

首先，受限于可用的计算资源，BERT模型并没有调整到最佳状态。理想情况下，在有商业要求时，最好进行grid-search以保证模型已经达到了数据集上的最佳状态，但是本实验只进行了简单的最优验证。

数据选择

在数据的选择上，无论是数量还是质量方面，都算不上多或高。一个健壮的实验结果通常需要基于大规模、高质量的数据集来验证。因此，数据规模和质量的不足可能会影响实验结果的稳定性和泛化能力。而且在商业领域还需要单独根据需求划定自己的数据集，以自己的标准标注。

代码整理和实验脚本

从技术实施角度来看，为了方便快速实验，代码没有经过严谨的模块化整理，也没有编写方便实验的脚本。这导致后续进一步实验的复杂度提高，增加了实验复现和扩展的难度。

Badcase分析不全面

最后，对于badcase的分析并不全面，只手动讨论了一些比较典型的问题。一个细致和全面的badcase分析有助于深入理解模型的不足之处，为模型的改进提供具体指导。未能覆盖更多情况的badcase分析限制了我们对模型性能瓶颈的理解。

参考文献

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
2. https://huggingface.co/google-bert/bert-base-chinese