1. Introduction

1.1 tower模型策略：三个步骤

- 扩展LLaMA-2模型的多语言能力，具体策略是通过一个包含有20B个token的数据集对其进行持续的预训练，这个步骤最终的产物是TOWERBASE模型。（另外这个预训练和之前不同的是，这个数据集包含有平行语料）为什么语料训练使用了平行语料库？
- 制作了一个数据集TOWERBLOCKS，用来针对性地提升LLM的翻译相关任务的能力
- 通过监督性的微调训练，得到指令型的在翻译领域的模型TOWERINSTRUCT

1.2 论文成果

- 发布了TOWER家族的模型
- 发布了专用于翻译相关任务的数据集TOWERBLOCKS
- 发布了针对评估翻译相关任务表现的评估框架TOWEREVAL
- 发布了基准模型（预训练模型），鼓励未来探索

2. TOWER

2.1 数据清洗

2.1.1 单语数据

数据集：mC4，对10个语言均匀采样；同时去重，语言识别，困惑都度筛选提高数据质量

2.1.2 平行数据

对（xx->en）和（en->xx）的数据均匀采样，数据均来自公共数据源；通过质量阈值去除不合格的平行句子对

2.1.3 模型训练

略

2.2 TOWERBLOCKS

2.2.1 多样性（diversity）

- 领域多样性（domain）
- 模板多样性（template）
- 任务多样性（task）

2.2.2 质量（quality）

2.2.3 TOWERINSTRUCT

通过TOWERBLOCKS数据集对TOWERBASE模型进行微调，得到TOWERINSTRUCT模型

- 对话模板

User	Model
<|im start|>user Translate the following text from Portuguese into English. Portuguese: Ontem a minha amiga foi ao supermercado mas estava fechado. Queria comprar legumes e fruta. English: <|im end|>	<|im start|>assistant Yesteday, my friend went to the supermarket but it was closed. She wanted to buy vegetables and fruit. <|im end|>

细节补充

- 受到另一篇论文的启发，之前的工作都是单语语料，虽然都充分利用了单语语料的优势，但也有很多不足。最后论文是是有了1/3的平行语料的句子，2/3的单语的句子，这个方法在后面的结果将会得到验证，大大提高了翻译质量 _^