**“千言：问题匹配鲁棒性评测”**

**#富婆来相亲#队伍技术报告[[1]](#footnote-1)**

陈怡雯1, † 王利蕾2

重庆邮电大学，重庆 南岸 400065

† 陈怡雯，s210201010@stu.cqupt.edu.cn

|  |
| --- |
| 摘要 千言-问题匹配鲁棒性评测是一个问题匹配的文本分类任务，本文分为问题分析、模型设计、鲁棒性优化、实验过程四个部分。从数据处理、数据增强、模型选择、参数调整、网络微调等各个方面完整阐述了本团队的比赛过程、思路和尝试。 |
| 关键词 自然语言处理，问题匹配，鲁棒性优化 |
|  |

# 问题分析

千言-问题匹配鲁棒性评测是一个文本分类任务，赛题需要判断两个自然问句之间的语义是否等价。本赛题给出三个用于训练的数据集，分别是LCQMC（哈工大文本匹配数据集）、BQ（银行疑问句）和OPPO（小布对话短文本数据集），测试集中包含百度DuQM鲁棒性数据集和OPPO小布对话短文本数据集。数据分布如表1所示。

根据数据分析可知本赛题数据集文本属于短文本，训练数据覆盖领域丰富包含金融、日常对话等领域，但测试集相对更加日常。其次，数据集分布较为均匀。通过观察数据样本的长度发现，测试集与训练集相比长度更小，粒度更细。

表1 数据集分析

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **数据集** | | **数据量** | **缺省值** | **正负样本比** | **query1文本长度** | | | | | | **query2文本长度** | | | | | |
| mean | std | min | 50% | 99% | max | mean | std | min | 50% | 99% | max |
| **训练集** | LCQMC | 238766 | 无 | 1.3830 |  |  |  |  |  |  |  |  |  |  |  |  |
| BQ | 100000 | 无 | 1 |  |  |  |  |  |  |  |  |  |  |  |  |
| OPPO | 167173 | 无 | 0.4486 |  |  |  |  |  |  |  |  |  |  |  |  |
| 总计： | **505939** | 无 | **0.9049** | 9.83 | 4.84 | 1 | 9 | 29 | **123** | 10.2 | 5.4 | 1 | 9 | 33 | **153** |
| **验证集** | LCQMC | 8802 | 无 | 1.0005 |  |  |  |  |  |  |  |  |  |  |  |  |
| BQ | 10000 | 无 | 1 |  |  |  |  |  |  |  |  |  |  |  |  |
| OPPO | 10000 | 无 | 0.4362 |  |  |  |  |  |  |  |  |  |  |  |  |
| 总计： | 28802 | 无 | **0.7602** | 10.5 | 5.45 | 1 | 10 | 30 | **130** | 10.7 | 5.5 | 1 | 10 | 30 | **112** |
| **测试集** | test\_A | **50000** | 无 | / | 8.62 | 2.88 | 3 | 8 | 18 | **47** | 8.51 | 3.1 | 3 | 8 | 18 | **48** |

赛题评分以宏平均准确率进行评测，其中评测细粒度包含5种，分别有Lexical Semantics、Syntactic Structure、Misspelling、Speech Filler、Conversational Semantics，评测公式为：

本赛题要求增强模型的鲁棒性，即当测试集和训练集的领域相差较大或者分布差异较大时也能得到很好的预测效果，所以在构建模型和生成特征时应该分析句子深层关系，提取出句子的根本含义。

# 模型设计

在数据处理方面，根据query1与query2等价和query1与query3等价可以推断出query1与query3等价，所以通过传递性可对数据进行数据增强。根据query1与query2等价可推断出query2与query1可对数据进行增强。

通过文本分析可以发现数据中存在大量文本大致相同但是有同音同调、同音异调的情况出现，所以通过随机替换文本中的字为同音同调、同音异调的字，增加模型的泛化性。

通过训练文本和测试文本的数据观察与分析，发现存在替换近义词反义词的情况，所以也对文本进行了近义词替换操作，替换后文本的label不变。

在预训练模型选择方面，团队通过对ERINE-Gram、NEZHA、RoBERTA等预训练模型进行对比后，选择ERINE-Gram为方案预训练模型。

在网络微调方面，为了提取到更多语义信息增强模型对句子理解，尝试了BN网络、多任务学习、PET等不同解决方案，对文本提取词性特征句子显式/隐式依存句法分析等特征。最优方案是以ERINE-Gram作为预训练模型，提取sequence\_output和pooled\_output向量，并结合GRU网络得到分类结果，具体模型设计如图1所示。

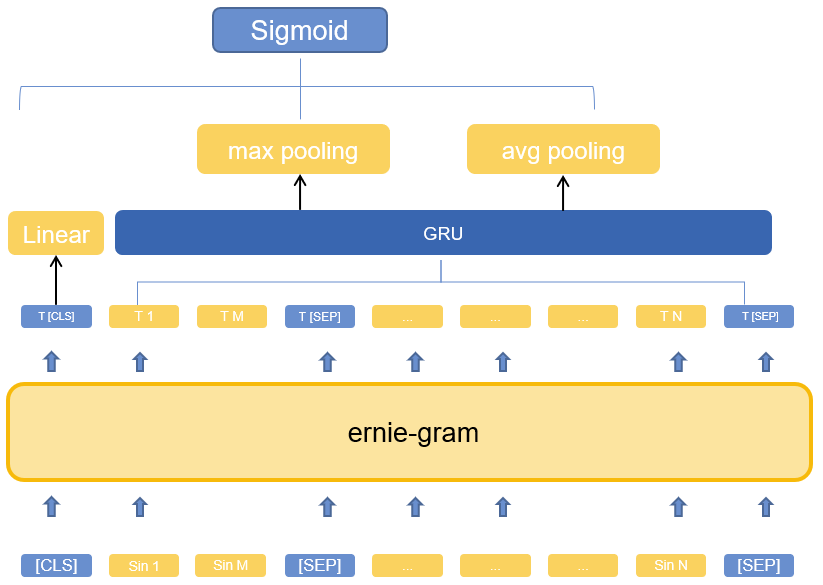


图1 模型设计

# 问题匹配鲁棒性评测

在增强模型鲁棒性要求上，数据处理方面使用上文提到的数据增强方法，即同音异调、同音同调近义词同义词等替换方法，如图2、图3所示。另外，获取文本词性信息、文本依存句法信息、拼音信息等特征，提高了文本的差异性。

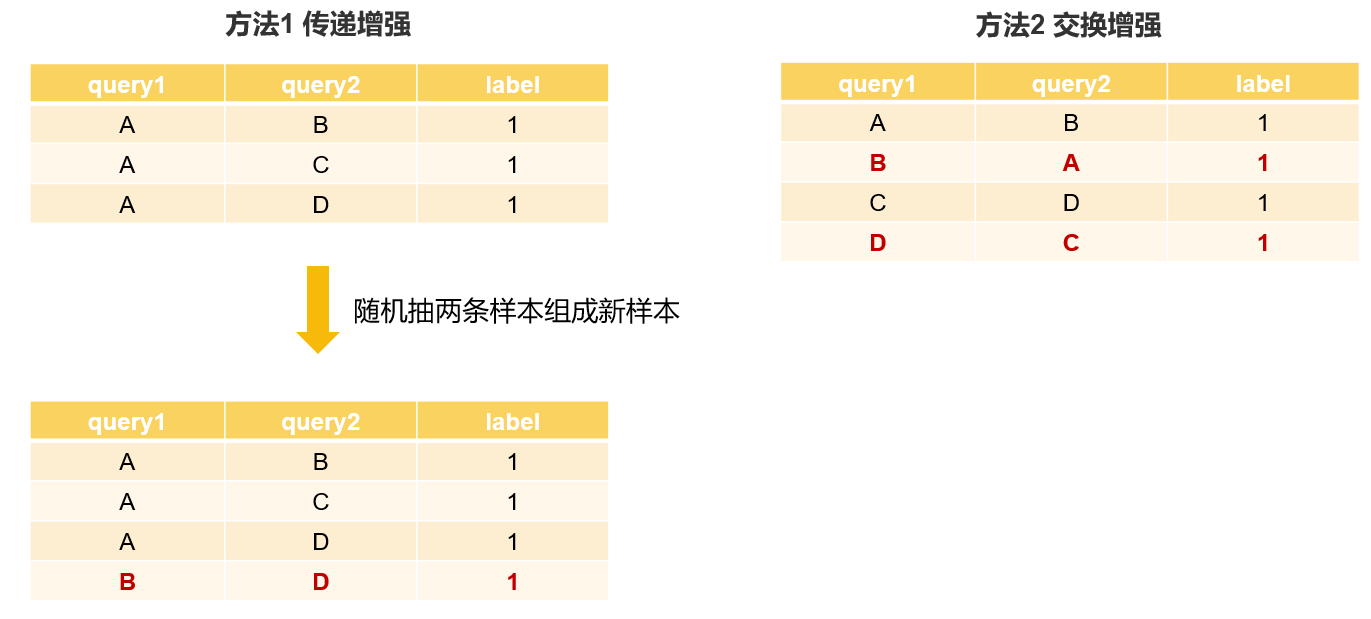


图2数据传递/交换增强



图3数据传递/交换增强

在网络结构方面通过调整参数等方法，在ERINE-Gram后提高drop\_out比例至0.2，使用最大池化和平均池化，在对减少参数量的同时，对特征进行提纯，在分类输出前继续使用drop\_out防止模型过拟合。具体提升如表2所示。

表2 问题匹配鲁棒性评测

|  |  |
| --- | --- |
| **方案对比** | **A榜线上成绩** |
| ernie-gram-baseline | 80.1 |
| ernie-gram-base-trans | 80.9 |
| ernie-gram-GRU-base | 86.653 |
| ernie-gram-GRU-dropout | 86.849 |
| ernie-gram-GRU-pool-dropout | 87.416 |
| ernie-gram-GRU-pool-dropout-eda | 88.151 |
| ernie-gram-GRU-pool-dropout-eda-lac-ddparser | 90.863 |

# 实验和结果分析

4.1预训练模型

在比赛初期对baseline理解后，对预训练模型进行了更换和对比，分别比较了baseline的ERINE-Gram预训练模型、NEZHA预训练模型[1]和RoBERTA[2]预训练模型，在同样的参数下各个预训练模型的效果如表3所示。由结果可知，在本赛题中，使用ERINE-Gram预训练更有优越性。

表3 预训练模型对比

|  |  |
| --- | --- |
| **model** | **score** |
| ernie-gram | 87.038 |
| nezha | 83.752 |
| roberta | 84.713 |

继2018年BERT预训练模型提出后，陆续有针对BERT模型的不足提出改进的变体模型出现。ERINE-Gram[3]通过一种显式gram掩蔽方法，其中n-gram用单个[MASK]符号掩码,并且直接使用显式的gram标识来屏蔽和预测的，以此增强粗粒度信息在预训练中的集成。此外，ERNIE-Gram使用从生成器模型中采样的似然n-gram标识掩盖n-gram，然后用似然和原n-gram之间的成对关系将它们恢复到原来的n-gram，以实现综合的n-gram预测和关系建模。另外，ERINE-Gram的训练预料包含贴吧、搜索等日常生活，语料领域丰富，所以ERINE-Gram相比于其他两种预训练模型获得了更好的效果，因为时间有限，没有与其他的预训练模型进行比较。

4.2双塔结构

在语义匹配领域中，有两种结构应用较广泛，单塔模型和双塔模型。单塔指的是将两个query先拼接再进入网络，双塔[4]指的是两个query分别进入网络结构再进行拼接，如图4所示，但因为双塔结构没有两个句子之间的交互，语义信息不够准确导致学习的效果不佳，所以最后我们没有采取这种结构进行预测。

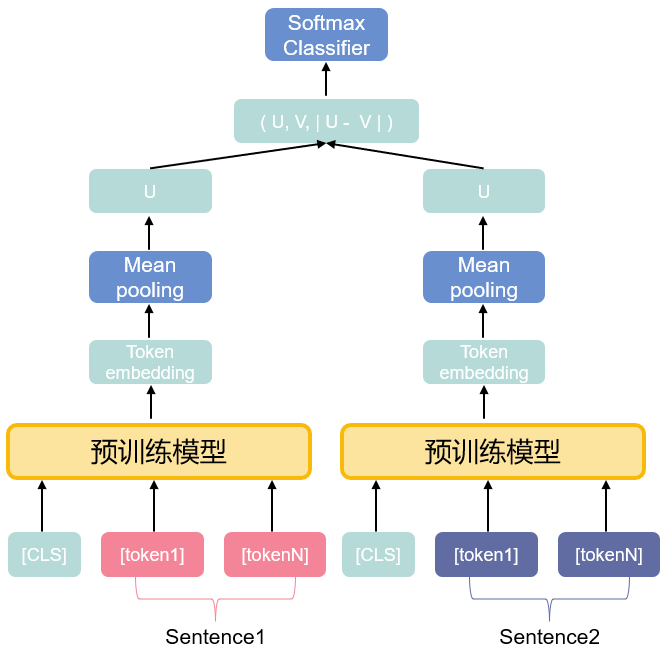


图4双塔结构

4.3 ERINE-Gram+Bilstm/mean+pool

在baseline中，提取的pooled\_output向量进行分类，这样可能会造成一些语义特征的丢失，所以我们提取了sequence\_output向量，并分别进入双向的LSTM网络，得到最后一个时间步的结果进行拼接再分类。双向LSTM的好处在于可以联系上下文，对上文的词汇向量有一定的记忆作用。另外还尝试将sequence\_output向量mean之后直接和pooled\_output进行拼接再分类。具体模型如图5所示。

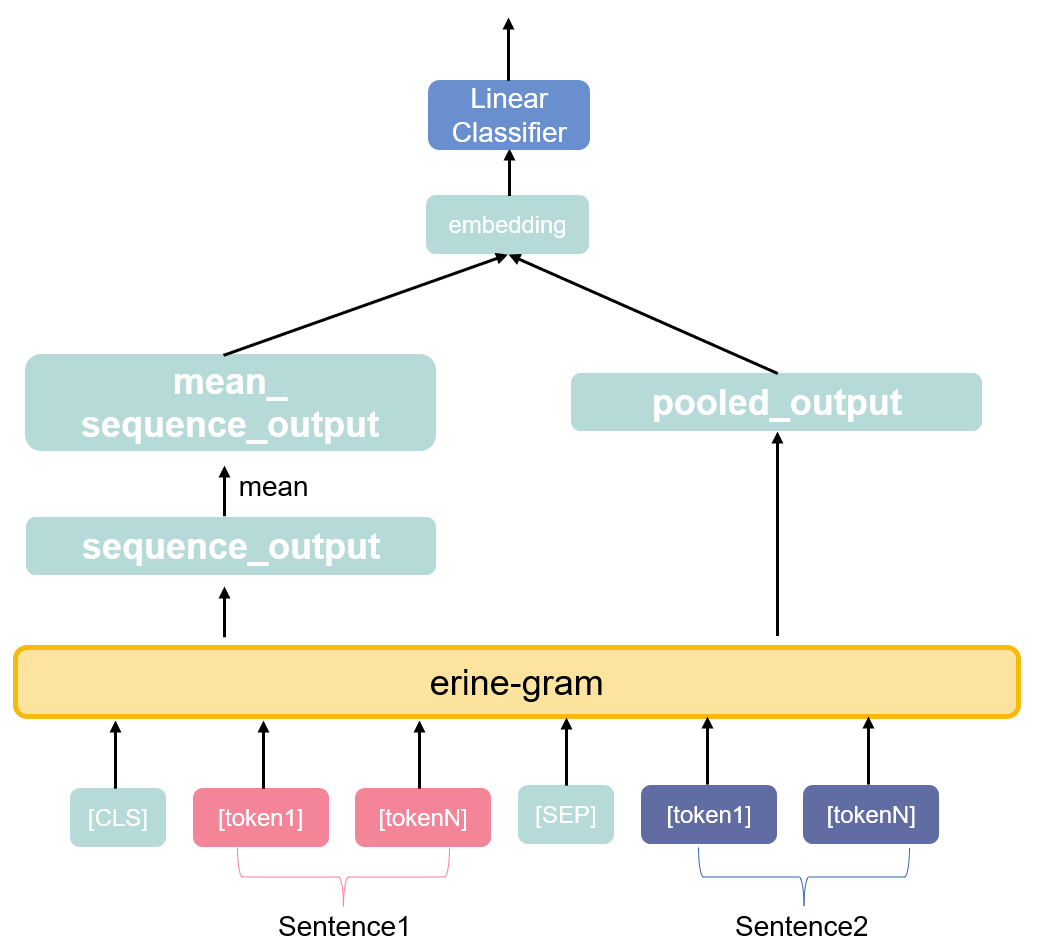
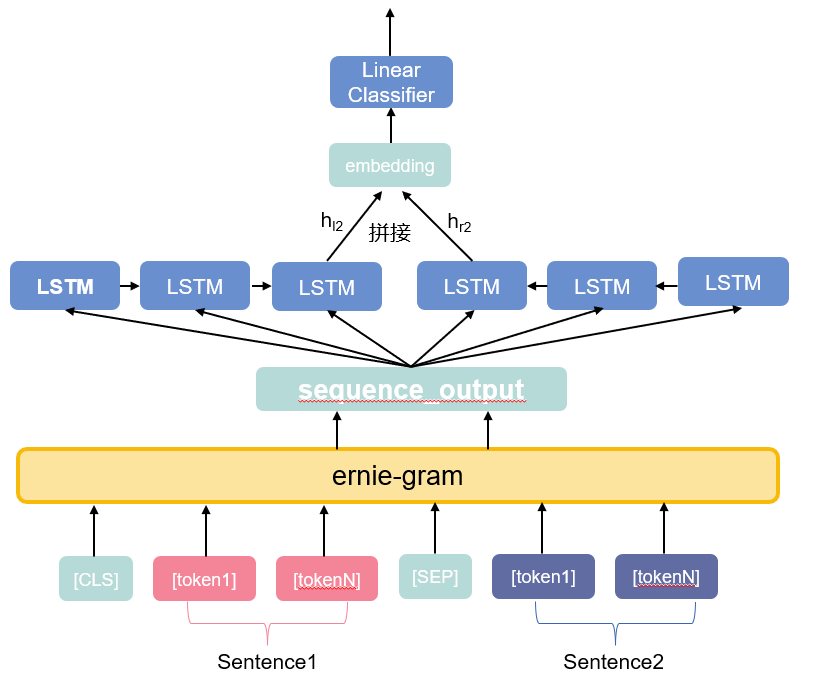


图5 ERINE-Gram+Bilstm/mean+pool模型结构

4.4 ERNIE-Gram+LAC

为了更好的表达文本，将每条文本的词性加入训练，文本向量进入预训练模型，词性向量进入GRU网络，GRU网络是BiLSTM的进阶版，拥有更少的参数和近似的训练效果，选择GRU模型可以更好的分析两个文本词性向量之间的关系，最后进入分类器输出结果，具体模型如图6所示，

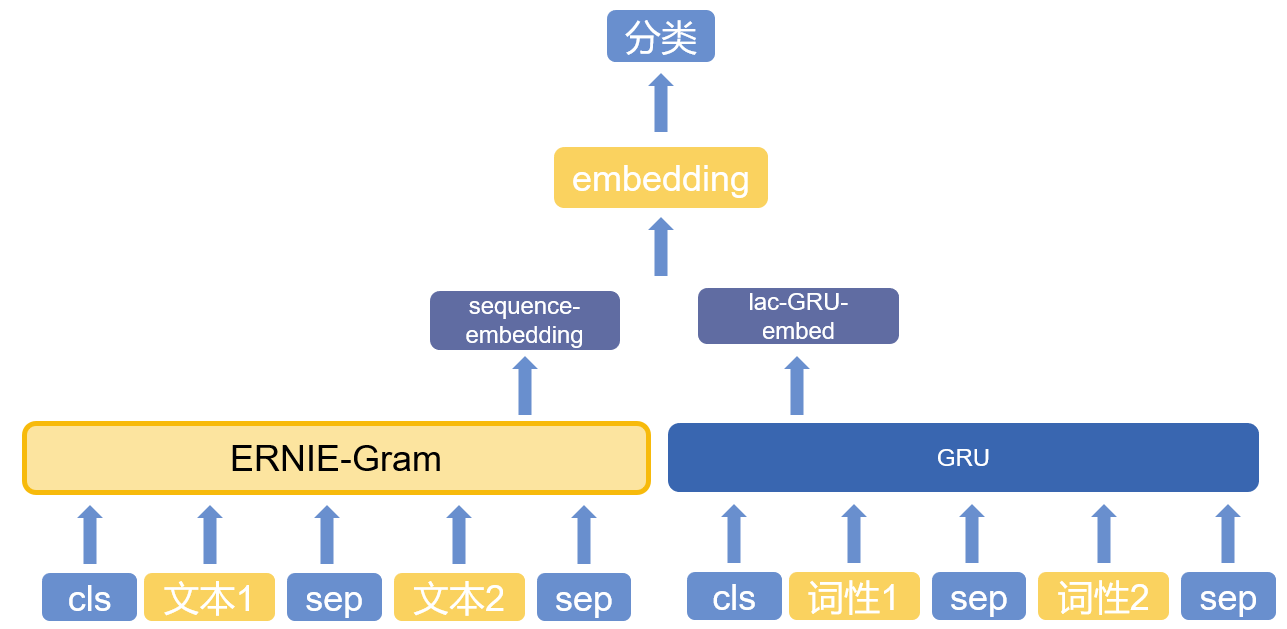


图6 ERNIE-Gram+LAC模型结构

4.5 Pattern-Exploiting Training

Pattern-Exploiting Training [5]是由prompt范式启发而来的，其主要思想是通过一个固定的模板范式，将文本的标签转为指令式文本，作为原来文本的前缀或后缀，指导模型进行相应任务。简而言之，就是预训练模型的NSP任务转为MLM任务。如本次比赛可以添加一个前缀pattern：“\_相似句子对：”，其中“很相似句子对/不相似句子对”分别对应label 0 / 1，pattern的得分只需看第一个位置中“不”/“很”两个token的概率谁高即可。

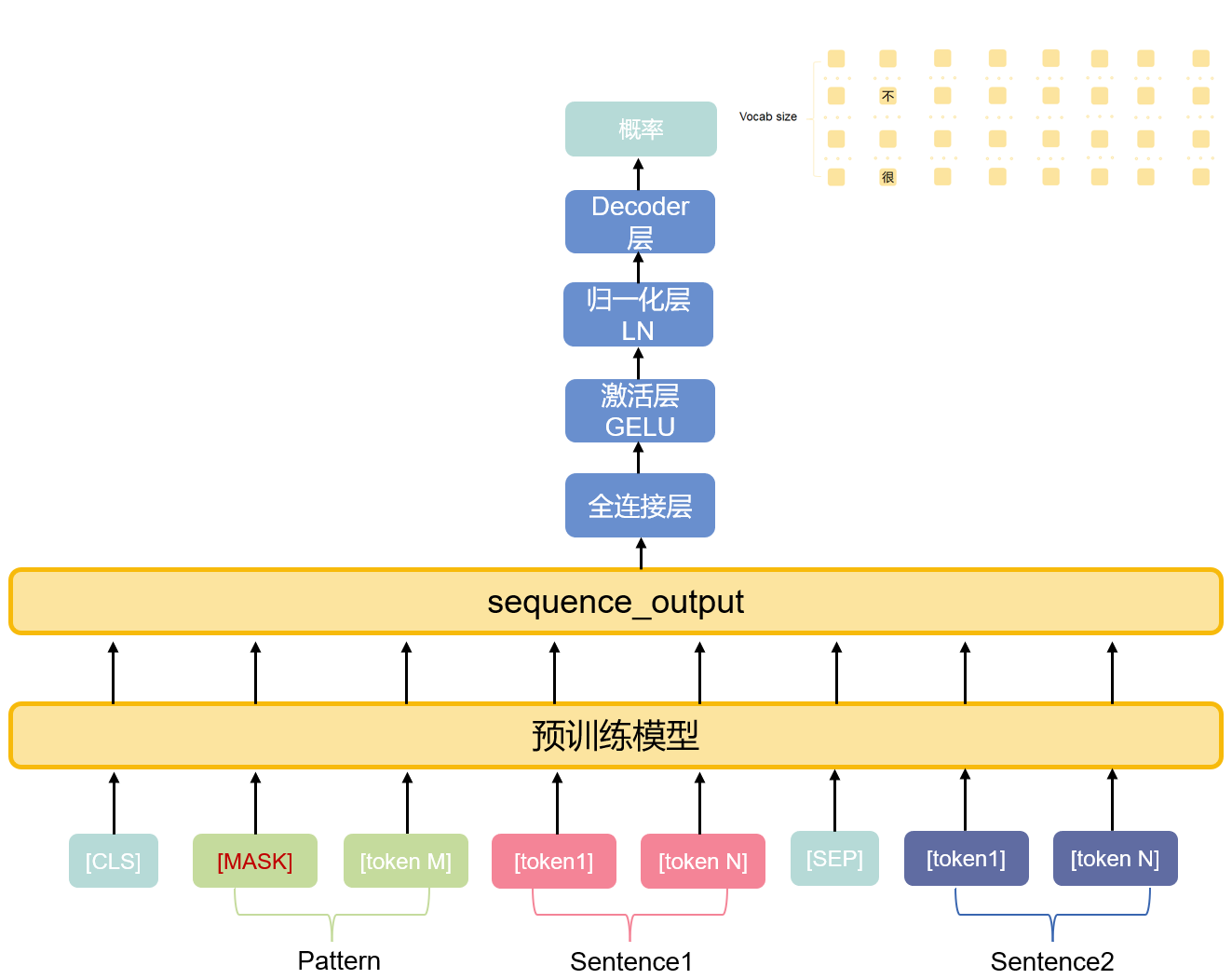


图7 PET结构

4.6 Multi-Task-Learning

为了让模型更好的适应不同的任务，采用多任务学习的方式去微调预训练模型，也就是无论最后有多少个任务，底层参数统一共享，顶层参数各个模型各自独立[6]。由于对于大部分参数进行了共享，模型的过拟合概率会降低，共享的参数越多，过拟合几率越小，共享的参数越少，越趋近于单个任务学习分别学习。再通过微调后的模型放入全量数据进一步训练，从而得到最后的模型。

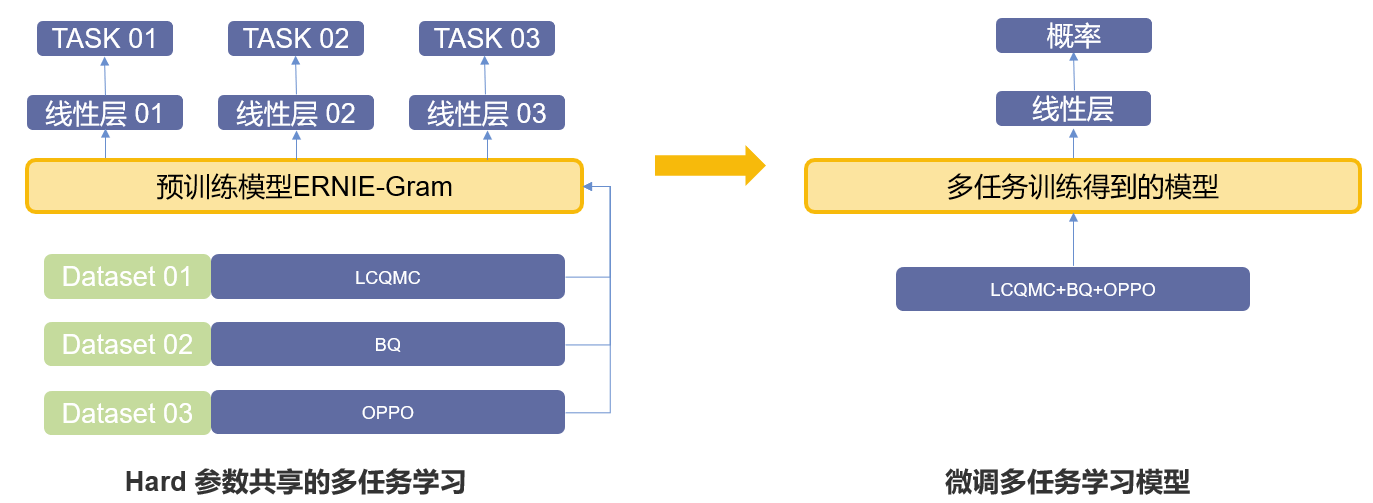


表4 结果对比

|  |  |
| --- | --- |
| **方案** | **A榜线上** |
| 双塔结构 | 82.86 |
| ERINE-Gram+Bilstm | 86.218 |
| ERINE-Gram+mean+pool | 86.509 |
| ERNIE-Gram+LAC | 87.938 |
| Pattern-Exploiting Training | 87.313 |
| Multi-Task-Learning | 87.629 |
| ernie-gram-GRU-pool-dropout-  eda-lac-ddparser | 90.863 |

参考文献

1. Wei, J., “NEZHA: Neural Contextualized Representation for Chinese Language Understanding”[J] .arXiv preprint arXiv:2107.13586, 2019.
2. Liu Y , Ott M , Goyal N , et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. 2019.
3. Xiao, D.:ERNIE-Gram: Pre-Training with Explicitly N-Gram Masked Language Modeling for Natural Language Understanding[J].arXiv preprint arXiv:1901.07291, 2019.
4. Chopra S , Hadsell R , Lecun Y . Learning a similarity metric discriminatively, with application to face verification[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005.
5. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G.:Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing [J] .arXiv preprint arXiv:2107.13586
6. Caruana. R. Multitask Learning: A Knowledge based Source of Inductive Bias. Proceedings of the Tenth International Conference on Machine Learning. 1993.

1. [↑](#footnote-ref-1)