搜狐 SOHU.com | 比武招新

# 2021 搜狐校园 文本匹配 算法大赛 总决赛

“分比我们低的都是帅哥！”


搜狐 SOHU.com | 比武招新

1 团队介绍

2 赛题解读

3 数据探索

4 方案设计

5 总结思考


搜狐 SOHU.com | 比武招新

# Chapter 1

## 团队介绍


# 团队介绍：分比我们低的都是帅哥

“友谊第一，比赛第二。”

**王晨跃**

北京大学

信息科学与技术学院

研究生一年级

**王明杰**

北京师范大学

计算机科学与技术学院

本科四年级

**张毅**

东莞理工学院

计算机科学与技术学院

本科二年级

搜狐 SOHU.com | 比武招新

# Chapter 2

## 赛题解读

# 赛题解读

- 本次比赛的赛题可以概括为 **“长短不一的文本对在不同粒度下的匹配问题”**。具体而言，任务数据中包含A类（**话题**）与B类（**事件**）两种评判标准，同时有三项细分任务，分别是：短文本-短文本匹配，短文本-长文本匹配，长文本-长文本匹配。
- 判断句对是否匹配是**二分类任务**，以macro F1方法作为评判标准，即对A、B两类任务分别计算F1值然后求平均。
- 比赛同时对资源进行了限制，要求模型大小小于2G且单个句对推理时间不超过500ms。考虑到文本匹配在许多在线服务中应用广泛，我们进一步希望**探索出符合工业界线上需求的解决方案**，增加其实际落地的可能性。
- 综上，我们认为比赛的难点包括：
  1. 不同评判颗粒度和多个子任务
  2. 需处理大量超长文本
  3. 正负样例存在类别不平衡问题
  4. 探索符合线上场景的解决方案

搜狐 SOHU.com | 比武招新

# Chapter 3

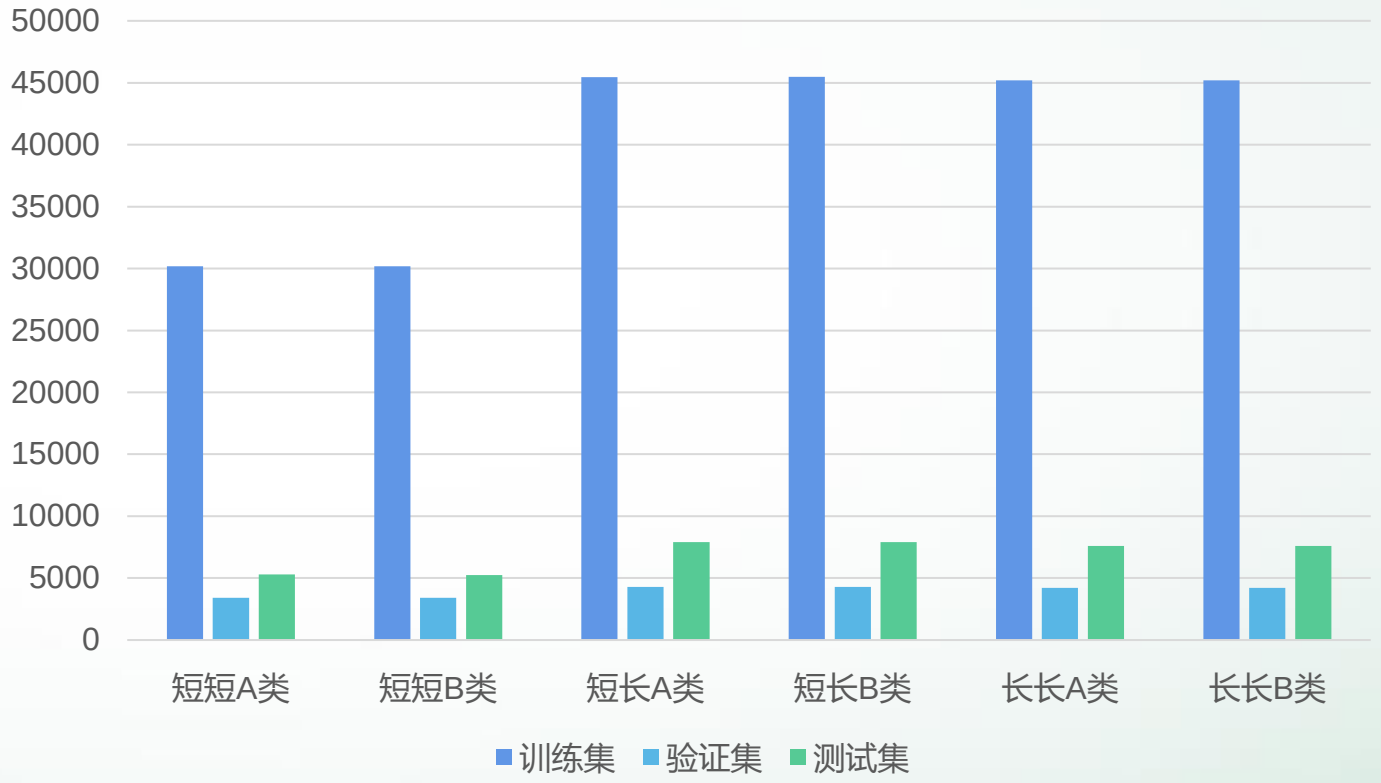
## 数据探索

# 数据探索：样本数目

在复赛/决赛环节中，我们采用了初赛与复赛的训练集（共计241726条）和复赛的验证集（共计13825条），并在复赛测试集（共计41480条）上推理结果。

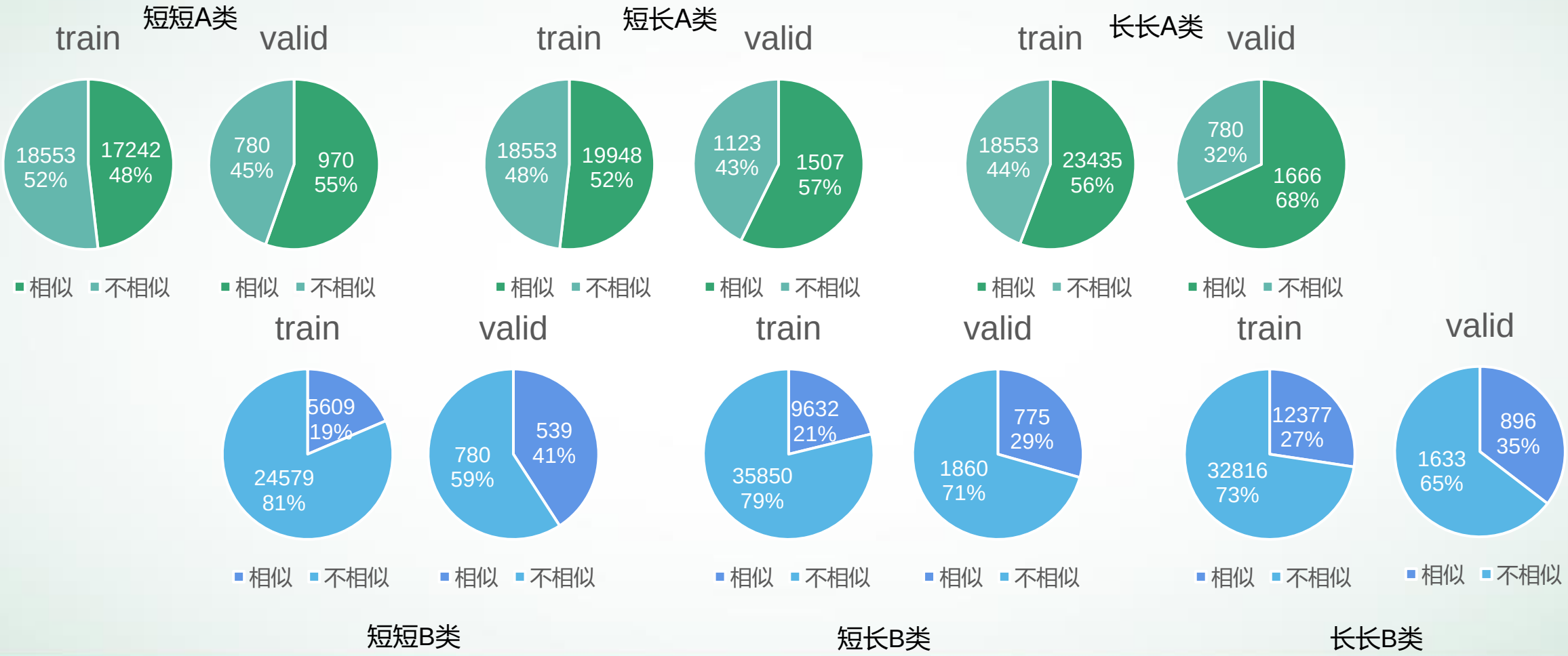
**1. 各类任务的样本数目：**

- 短短匹配任务样本数量略少
- 训练集、验证集与测试集的样本数目比例基本一致



# 数据探索：正负样本比例

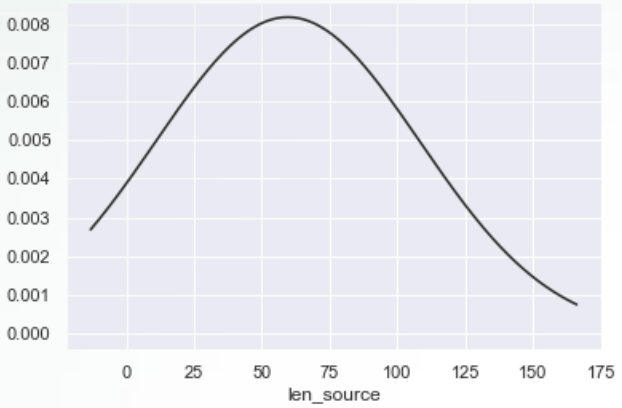
**2. 各类任务的正负样本比例：** B类类别不平衡现象明显，训练集与评估集分布各不一致

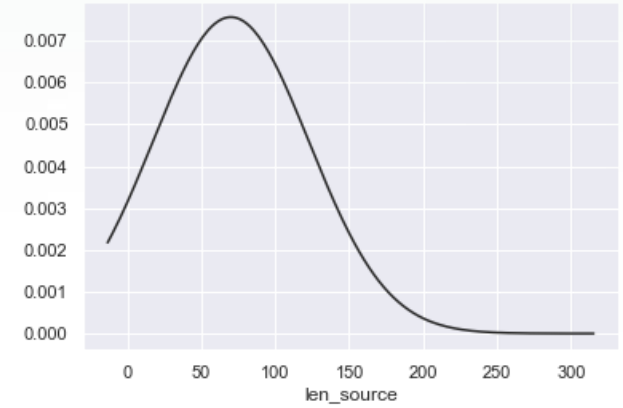


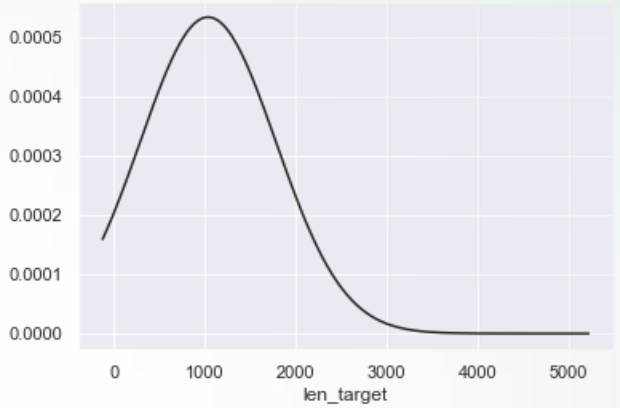
# 数据探索：文本长度

**3. 在各类任务的文本长度上：**短文本$\mu \approx 60$，长文本$\mu \approx 1000$；部分文本超长，达数千字

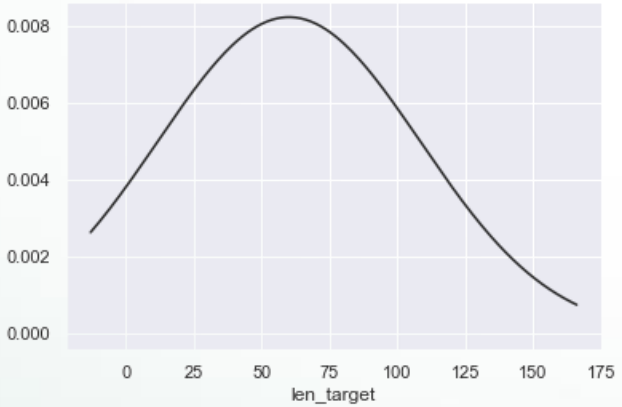
**source文本**

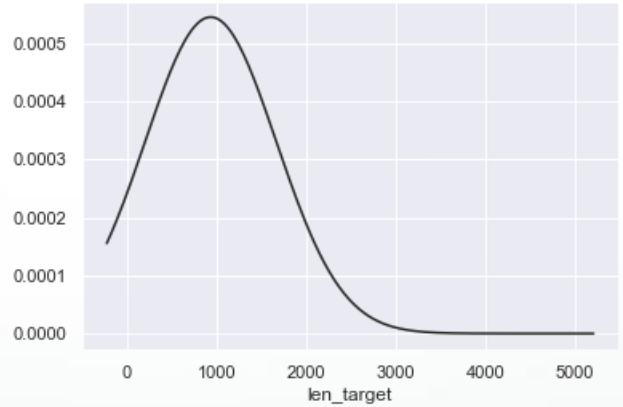


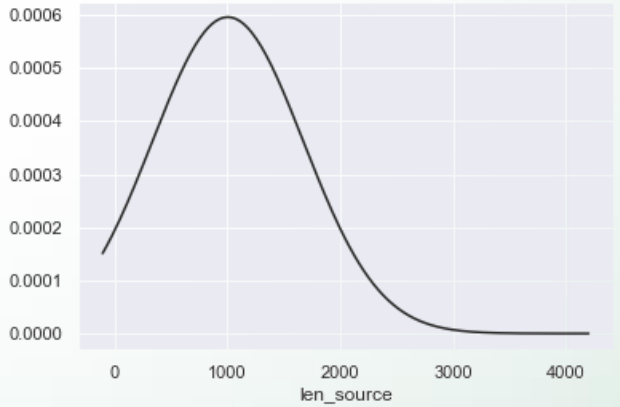




**target文本**




**短短匹配** **短长匹配** **长长匹配**

搜狐 SOHU.com | 比武招新

# Chapter 4

## 方案设计

# 方案设计：整体思路

**我们的方案主要包含以下几个流程**

**数据预处理**
- 句对拼接，对长文本截断，简单的数据增强策略

**模型定义与训练**
- 包含cross-encoder与bi-encoder两类模型，尝试多组预训练模型，结合更优的训练方式与提分技巧

**模型推理**
- 在验证集上对AB两类任务进行阈值优化后推理

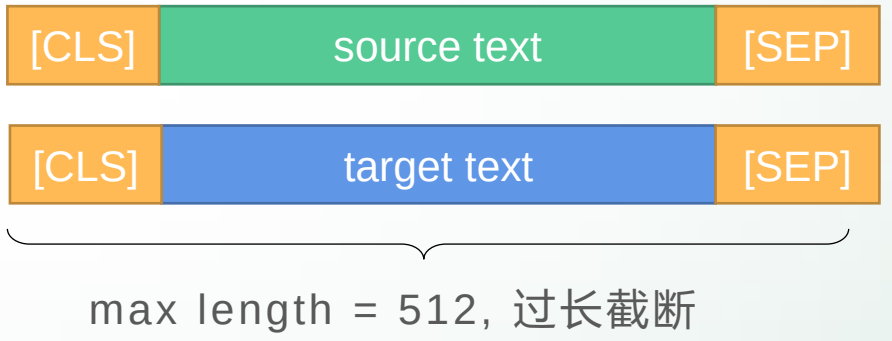
**模型融合**
- 基于输出结果投票，在验证集上寻找最佳组合

# 方案设计：数据预处理

## 针对匹配任务的数据处理

- cross-encoder类：拼接source与target,返回token_ids, token_types, label/id与任务编号（AB大类0-1，或子任务0-5）


- bi-encoder类：为source和target分别处理，返回source_token_ids, target_token_ids，label/id与任务编号



# 方案设计：模型定义

## 针对匹配任务的两类模型思路

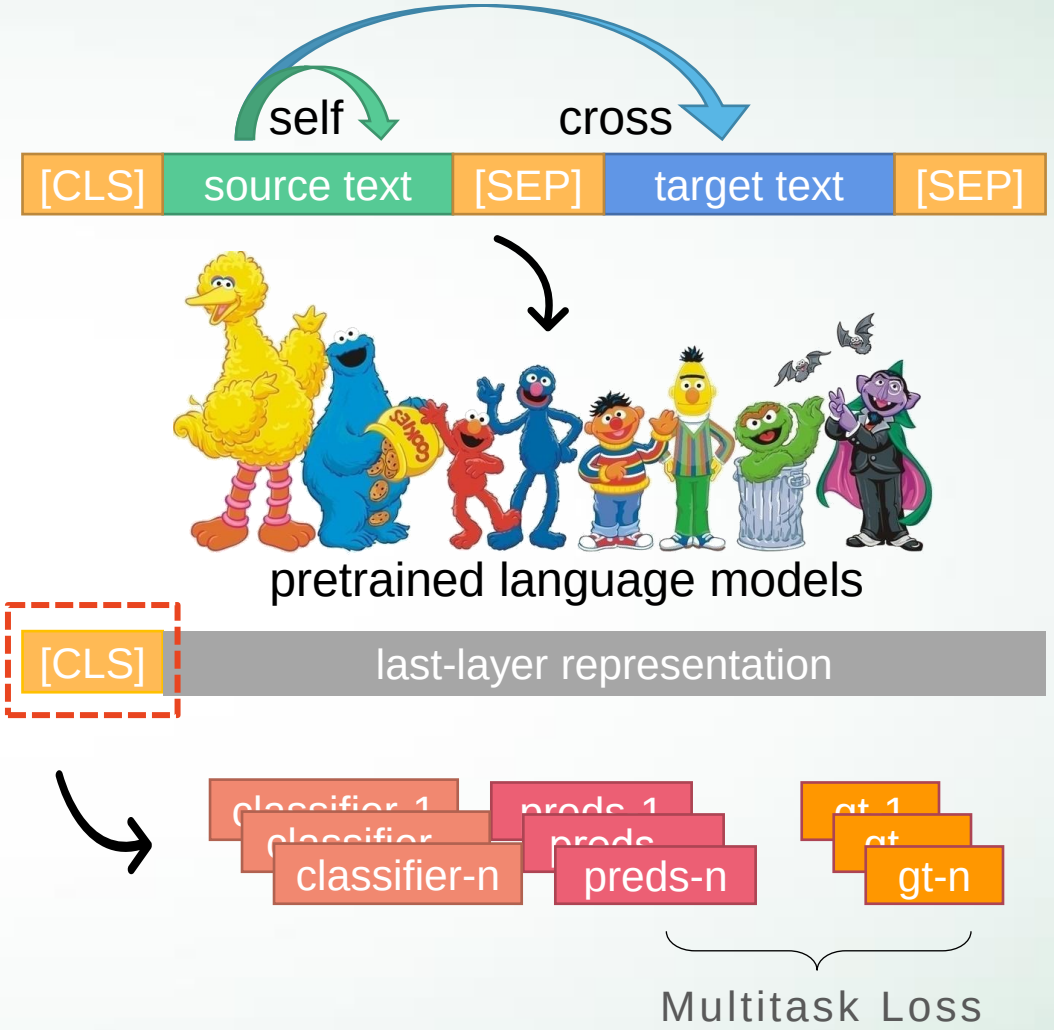
- cross-encoder类[1]：对一组句对进行编码，编码过程中可以进行句内及句间的信息交互（**interaction-based**）

✓ 更加充分的信息交互

✓ 更好的效果（~2-3%）

× 加剧了文本长度限制

× 较难满足工业场景在线需求




[1] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

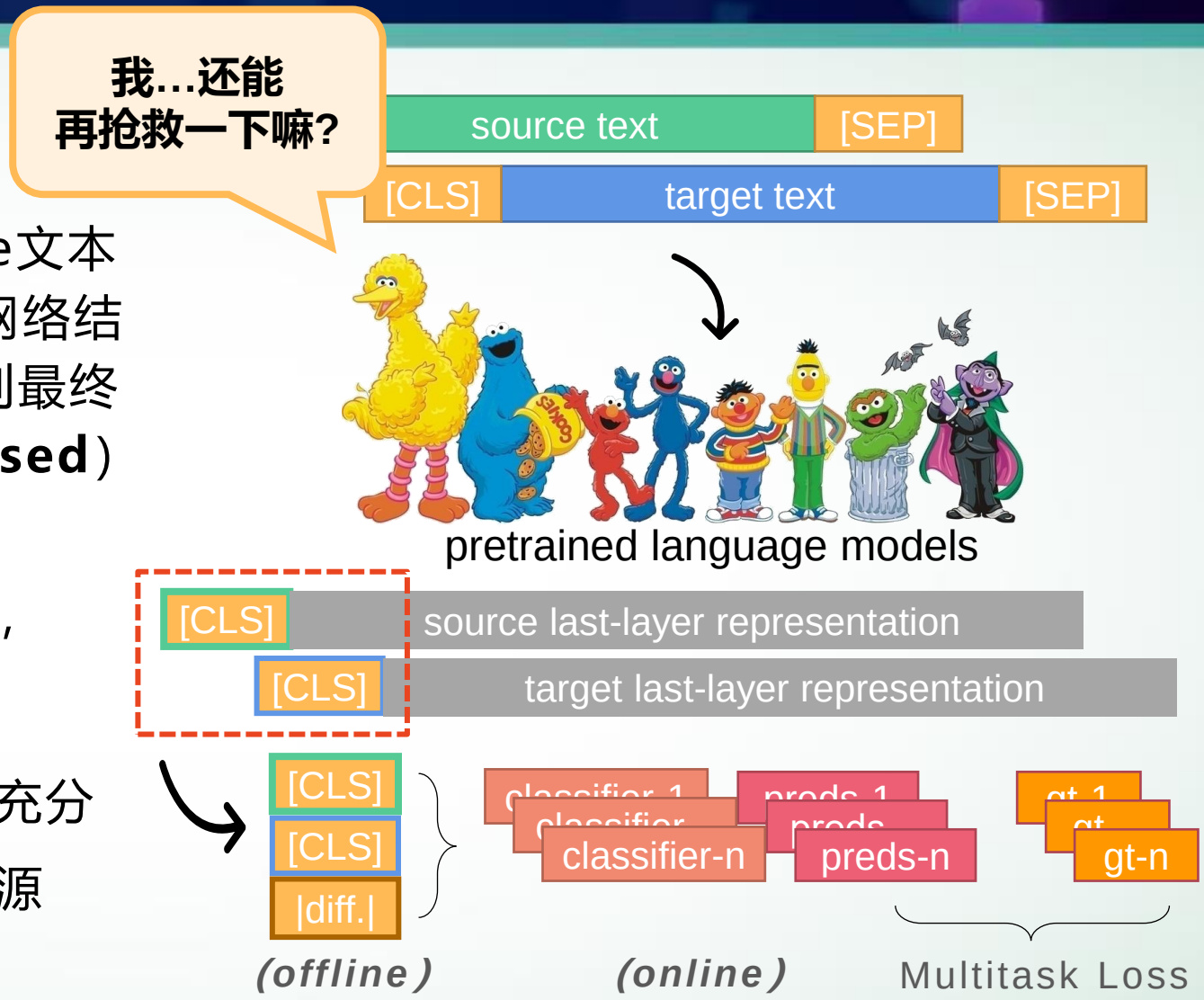
# 方案设计：模型定义

## 针对匹配任务的两类模型思路

- bi-encoder类[2]：分别对source文本和target文本进行编码，再通过网络结构进行表示间的交互和计算，得到最终分类结果（**representation-based**）

✓ 更大的文本长度上限
✓ 可先离线储存文本表示，在线分类或计算相似度

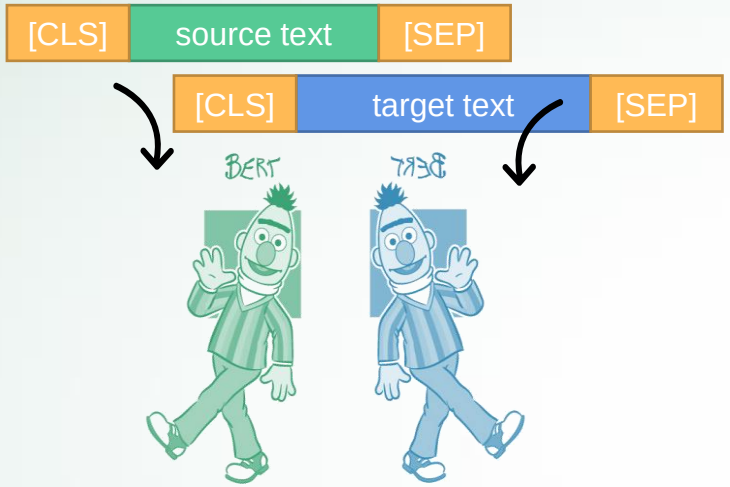
× 文本之间的信息交互不充分
× 训练时需要更多计算资源



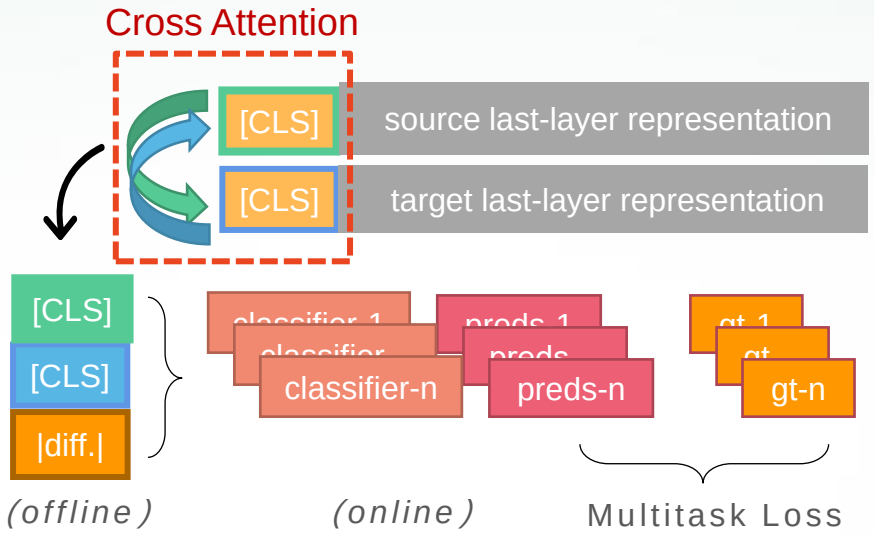

[2] Reimers, N. and Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

# 方案设计：模型定义

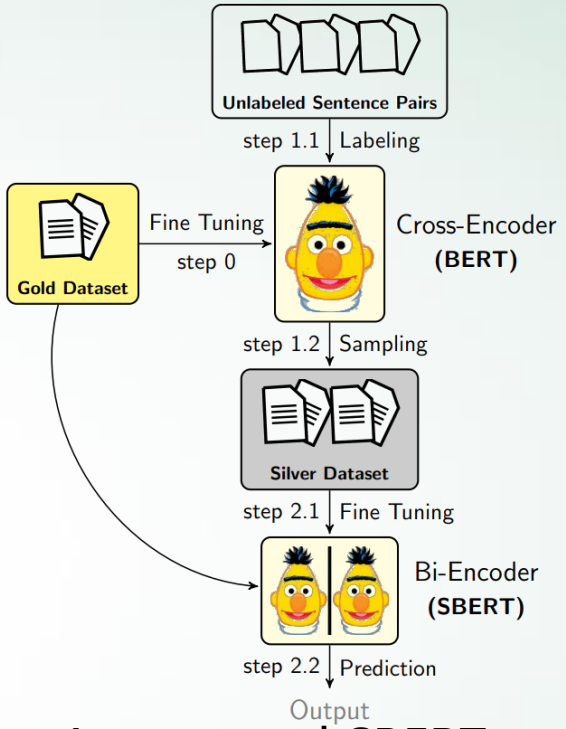
## 如何抢救bi-encoder类模型？




- “双兄弟”：设置两个单独的encoder，分别对source文本和target文本进行编码
- 对于匹配文本**领域差距明显**的匹配任务，或许会有更好的效果




- Cross Attention：加入跨句子表示的注意力模块，显式促进文本之间的信息交互
- 与原方案相比得分**略微增长**，但整体提升不明显



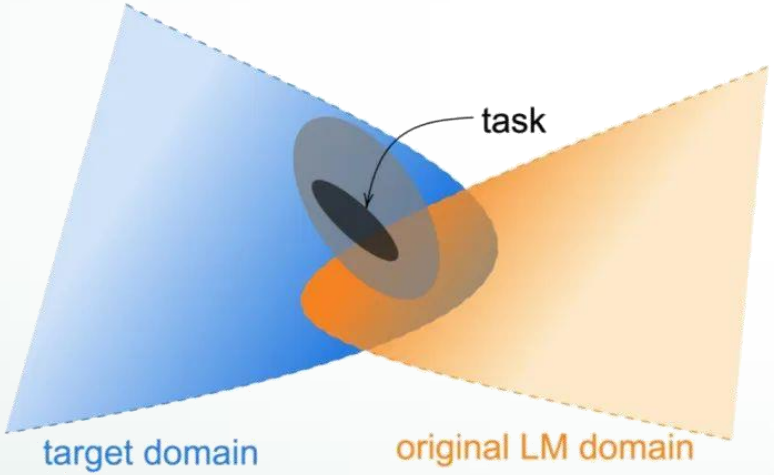

- Augmented SBERT[3]：利用cross-encoder为构造的句对打分作为**伪标签**训练bi-encoder模型
- 噪声过大，未进行实现

[3] Thakur, N., Reimers, N., Daxenberger, J. and Gurevych, I., 2020. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. arXiv preprint arXiv:2010.08240.

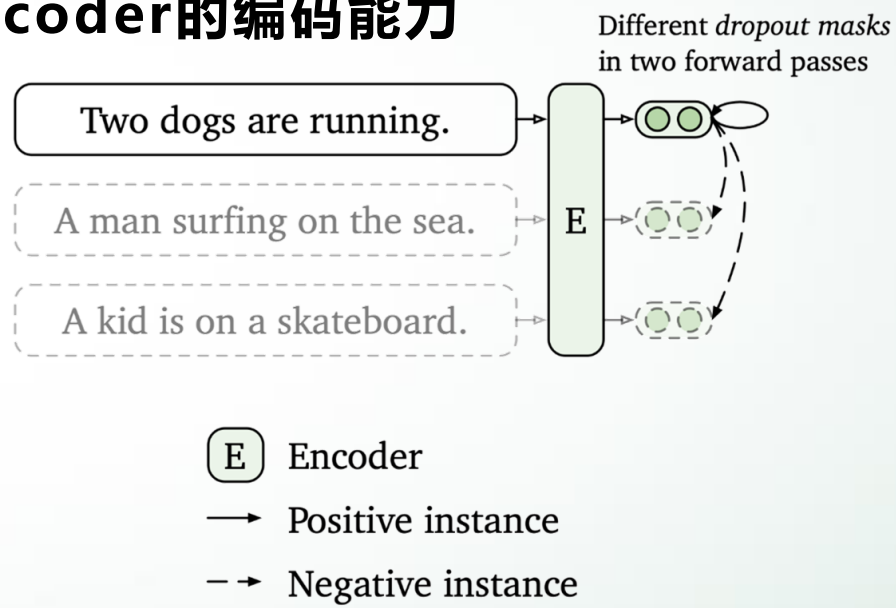
# 方案设计：模型训练

## 在预训练和下游任务微调之间……

- Task-adaptive Pretraining[4]：在比赛提供的文本语料上继续MLM任务，**使语言模型更贴近任务领域的语料分布**
- 效果：新闻领域与预训练文本领域较为接近，提升有限




- SimCSE对比学习[5]（待尝试）：通过dropout构建正例，其余句对为负例，**拉开正负样例表示的距离，从而提升encoder的编码能力**




[4] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D. and Smith, N.A., 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. arXiv preprint arXiv:2004.10964.

[5] Gao, T., Yao, X. and Chen, D., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. arXiv preprint arXiv:2104.08821.

# 方案设计：模型训练

## 尝试一些更好的中文预训练模型

**RoBERTa-wwm-ext**
- 采用WWM策略，取消NSP任务
- 更大规模的中文训练数据

**ERNIE**
- 融入实体概念等先验语义知识
- 基于贴吧提问-回帖的DLM任务

**NEZHA**
- 完全函数式的相对位置编码
- 加入Span预测任务

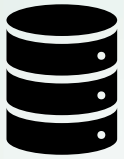
**MacBERT**
- 替换[MASK]为近义词，由模型“纠正”
- 使用SOP任务，预测上下句顺序

**WoBERT**
- 以词为基本单位
- 可处理字数更长的文本

**RoBERTa-base-word**
- 同WoBERT
- 训练语料为CLUECorpusSmall

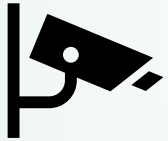
# 方案设计：模型训练

## 结合一些常见的比赛训练技巧



### 数据增强

- 互换source文本与target文本的位置
- A类负例是B类负例，B类正例是A类正例



### Focal Loss

- $FL(p_t) = -(1-p_t)^{\gamma}\log(p_t)$
- 针对类别不平衡问题，降低负样本的权重

### Multi-Sample Dropout

- 对encoder的输出进行多次dropout，得到多组样本，以更好地训练classifier

### 优化器与学习率

- 尝试Lookahead + Radam
- 加入权重衰减，线性规划学习率

### 对抗训练

- 通过FGM对Embedding矩阵进行扰动
- 模型需要抵抗扰动，提高鲁棒性

### 伪标签

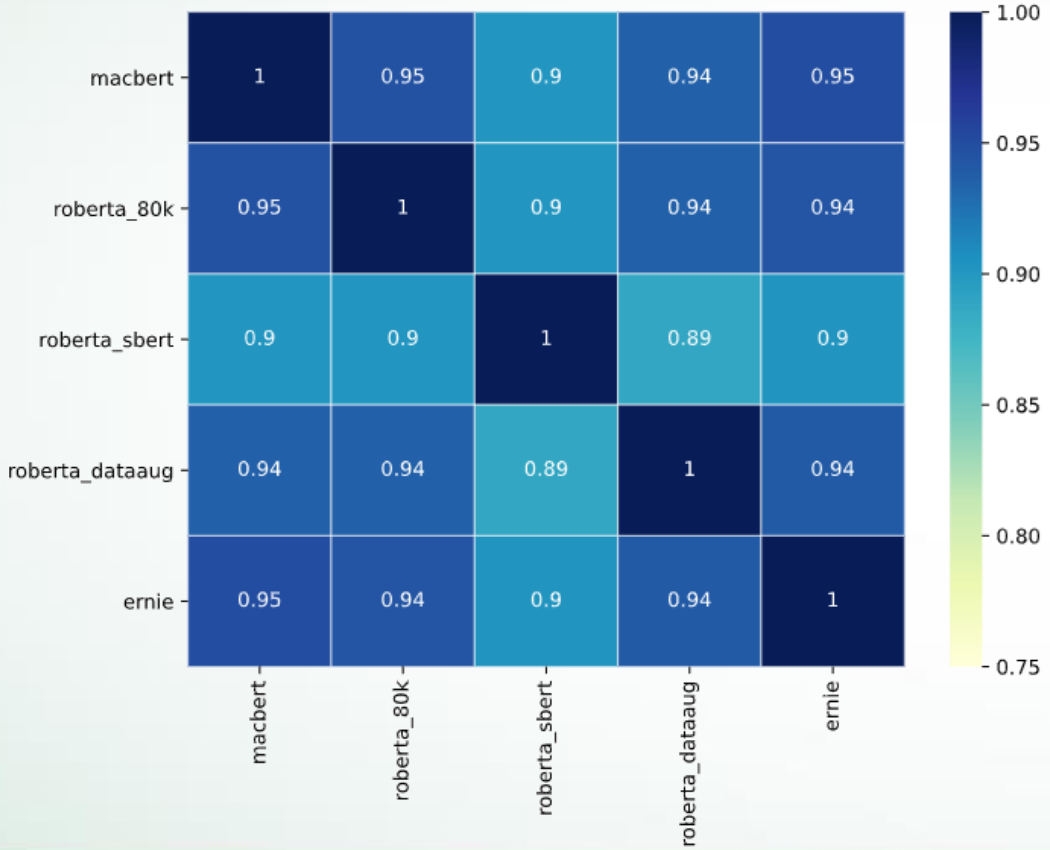
- 模型在测试集上标注得到伪标签
- 将伪标签与真实标签一起训练模型

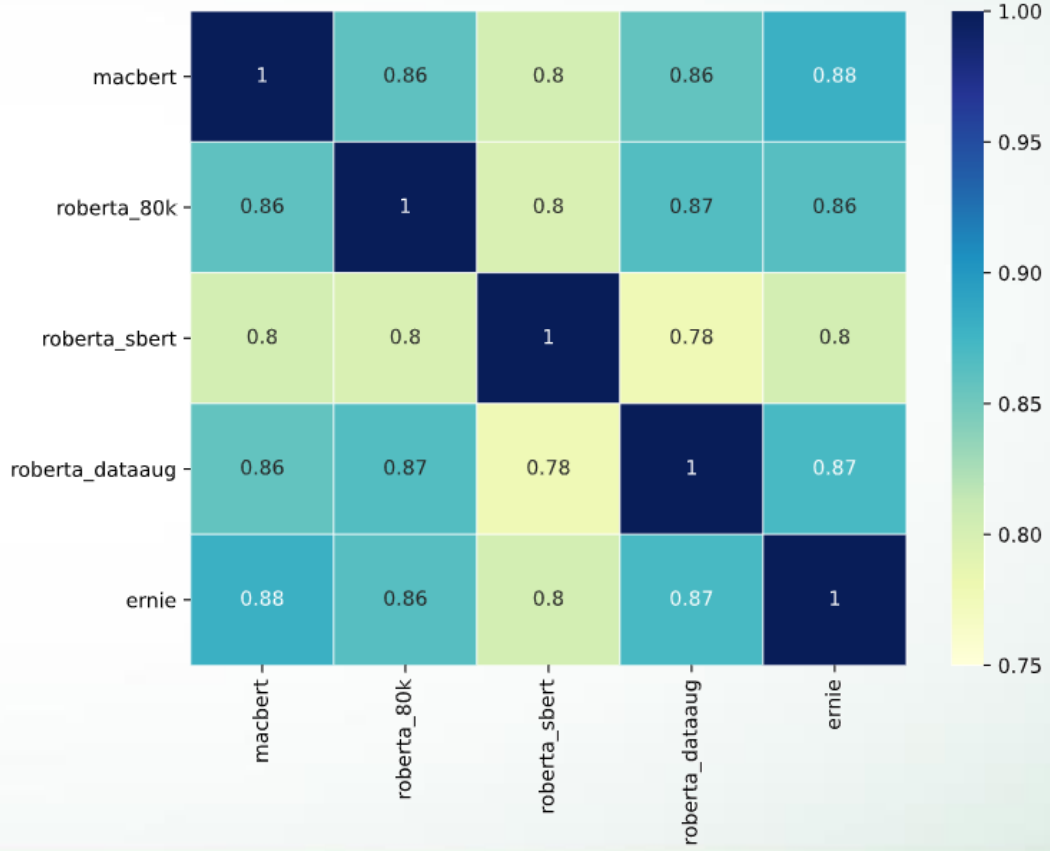
# 方案设计：模型推理与模型融合

- 测试集推理前，在验证集上**搜索AB两类任务最优的正例阈值**，以最大化F1指标
  - 由于正负样例类别不平衡，最后阈值在0.4左右，B类小于A类
  - 如果验证集与测试集分布一致，可考虑按子任务进行阈值优化（未尝试）
- 模型以优化后的阈值进行推理，各模型推理后的**标签输出通过投票进行融合**
  - 在验证集上搜索最佳的模型组合，融合模型输出作为最终提交

# 方案设计：模型推理与模型融合

- **“和而不同”**：不同模型方案、预训练模型、训练技巧的模型融合可提升1%+
  - 相比于取概率分布，取0-1标签放大了模型之间的差异，输出的相关系数明显降低



搜狐 SOHU.com | 比武招新

# Chapter 5

## 总结思考

# 总结思考：技术方向

- ✓ 结合文本匹配任务中cross-encoder（实现句对间充分的信息交互）和bi-encoder（可处理更长文本且更符合工业落地需要）**两种思路**。
- ✓ 针对bi-encoder模型，提出了一些**有益的改进**。
- ✓ 在预训练模型和下游任务微调之间，尝试采取**无监督方法**提升语言模型的语义理解及句子编码能力。
- ✓ 尝试更优的中文语言模型及提分技巧，在F1得分上**取得较好效果**。

- ➢ 虽然对**长文本**直接截断较为有效，但仍需更优雅、有效地处理方式。
- ➢ 主要依赖预训练语言模型的隐式编码能力，可考虑**人工加入其他特征**。
- ➢ 针对各个**子任务的不同表现**，详尽分析bad case并使用不同的策略。
- ➢ 继续尝试针对文本匹配的**无监督方法**，如BERT-whitening等。
- ➢ 由于赛题很吃计算资源，诸多想法有待**多次实验**以进行更准确的论证。

# 总结思考：个人感言

## 王晨跃

比赛过程中，**组队这件事情收益最大**，因为自己参加比赛的经历很有限，组队后在身经百战的队友们的交流和讨论后确实开阔了思路，也有了继续尝试和提升的动力，能够有效避免自己一个人犯懒、不想努力了的问题。

## 王明杰

我觉得**主动去寻找队友收益最大**，也认识了编程能力很强的晨跃和思维能力很强的张毅，通过不断的尝试和努力，才能走到最后的决赛。

## 张毅

我觉得**勇于试错对自己和队友收益最大**，作为一个NLP和比赛的新手，编程能力和比赛经验有限，所以自己会尝试着踩坑，积累经验，也因此极大地提升了自己的能力，也为整个团队后续的进展提供了有力的帮助。

搜狐 SOHU.com | 比武招新

# Q&A

# Thanks