# From 0 to 1: The Evolution of Our Retrieval-Augmented Generation Paradigm

Presenter：王琰
NLP Center, Tencent AI Lab

# Effect of Retrieval-Augment Methods

技术创造未来 科技TEG

**万字长文！DeepMind科学家总结2021年的15个高能研究**

云脑智库 · 2022-02-16 00:00 · 145浏览 · 0评论 · 0点赞

Universal Models 通用模型

Massive Multi-task Learning 大规模多任务学习

Beyond the Transformer 超越Transformer的方法

Prompting 提示

Efficient Methods 高效方法

Benchmarking 基准测试

Conditional Image Generation 条件性图像生成

ML for Science 用于科学的机器学习

Program Synthesis 程序合成

Bias 偏见

Retrieval Augmentation 检索增强

Token-free Models 无Token模型

Our Research:

- 从2018年开始持续推进Retrieval-Augment Generation研究
- NAACL 2019：Skeleton-to-Response: Dialogue Generation Guided by Retrieval Memory
- EMNLP 2019: Retrieval-guided Dialogue Response Generation via a Matching-to-Generation Framework
- TASLP: Prototype-to-Style: Dialogue Generation with Style-Aware Editing on Retrieval Memory
- ACL 2020: Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation
- ACL 2021: Neural machine translation with monolingual translation memory (Outstanding Paper Award)
- Arxiv: Exploring Dense Retrieval for Dialogue Response Selection
- IJCAI 2022 & SIGIR 2022 & CCL 2022: Organizing a tutorial on Retrieval-Augmented Text Generation

# Motivation(In 2018)

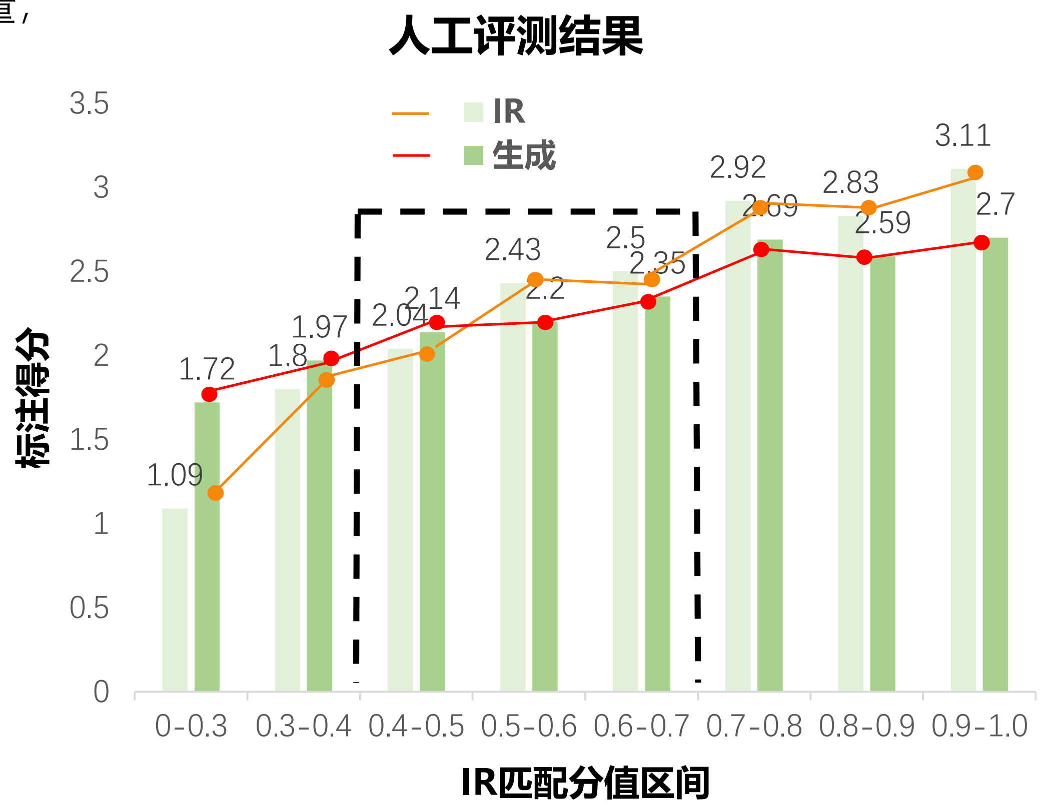


检索vs生成:
- **检索**: 信息量丰富，但失之毫厘容易谬以千里，没有扩展性
- **生成**: 神经网络储存知识有限，生成的回复比较单调，缺乏信息量;

**工程基础**: 线上检索系统性能良好，用生成模型替代代价较大;

| Query $x$ | Retrieved query ($x'$) and response($y'$) | Generation result |
| --- | --- | --- |
| 再也不吃肯德基了 | $x'$ : 再也不吃麦当劳了<br>$y'$ : 那就去吃肯德基，最近的十三鲜小龙虾汉堡不错 | 那吃什么 |
| 深圳今天天气怎么样 | $x'$ : 北京最近天气怎么样<br>$y'$ : 久违的阅兵蓝，也不知道能持续几天 | 今天天气不错 |

- 目标: **Is Deep Combination (1+1>1) Possible?**



# Retrieval-Augmented：问题，及1.0版本



**美好的愿望：强强联合，1+1>1**

**残酷的现实：1-to-N映射问题**

Context：我周末跟我儿子一起去迪士尼，他要玩那个钢铁侠飞行之旅。

Retrieve

R1: 最近北京还开了个环球影城啊，啥时候去玩玩

R2: 钢铁侠那个应该是5D电影吧，有点东西

R3: 那里的饭又贵又不好吃，千万自备干粮偷偷带进去

Refine?

Ground-truth

那干脆一起溜娃呗？我女儿也想去那里

**训练的结果：**退化为二者之一，1+1=1

**检索引导生成1.0**

Context：我周末跟我儿子一起去迪士尼，他要玩那个钢铁侠飞行之旅。

Retrieve

Retrieved Context: 迪士尼真是太好玩了, 我特别喜欢米奇

Retrieved Response: 我女儿也超级喜欢米奇，她总吵着要去米奇的奇妙魔法屋

Remove

Skeleton: 我女儿也超级喜欢___，她总吵着要去___

Rewrite

我女儿也超级喜欢迪士尼，她总吵着要去迪士尼看米奇，周末一起吗？

去芜存菁：只保留检索结果中有用的部分，形成回复骨架

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. Skeleton-to-Response: Dialogue Generation Guided by Retrieval Memory. *NAACL 2019*.

# 去芜存菁



- **解决方案**：根据检索结果生成骨架→ 根据骨架生成回复
- **骨架抽取**：目标是从检索结果中抽取最合适的回复骨架
- **回复生成**：目标是根据任意一个骨架，能生成最合适的回复
- **联合训练**：利用强化学习联合微调骨架抽取和回复生成模型
- **实验结果**：

| model | human score | dist-1 | dist-2 |
|---|---|---|---|
| IR | 2.093 | **0.238** | **0.723** |
| IR+rerank | 2.520 | 0.208 | 0.586 |
| Seq2Seq | 2.433 | 0.156 | 0.336 |
| MMI | 2.554 | 0.170 | 0.464 |
| EditVec | 2.588† | 0.154 | 0.394 |
| SKP | 2.581 | 0.152 | 0.406 |
| JNT | 2.612† | 0.147 | 0.377 |
| CAS | **2.747** | 0.156 | 0.411 |

Table 1: Response performance of different models. Sign tests on human score show that the CAS is significantly better than all other methods with p-value < 0.05, and the p-value < 0.01 except for those marked by †.

- **Case Study**

| | | |
|---|---|---|
| Query | 善良的妹子会幸福一生的<br>May you kind girls have a happy life | 好的到时候有时间骑车出来玩嘛，<br>我觉得在大理最适合骑车玩了<br>OK, Let's ride when free. It is Dali. |
| Retrieved query | 谢谢善良的妹子<br>Thanks to the kind girl | 改天一起出来骑车吧,我们都喜欢玩单车<br>Let's ride when free, as all of us like riding |
| Retrieved response | 我是汉子哈哈<br>I am a boy, LOL | ok 加微信abcdefg<br>ok add my wechat abcdefg |
| Skeleton | _汉子_<br>_boy_ | ok加微信_<br>OK, add my wechat _ |
| Generated response (CAS) | 哈哈哈，我是善良的**汉子**<br>LOL, I am a kind **boy** | 你**加我微信**吧，我也在大理<br>**Add** my **wechat**, I am in Dali, too |
| IR+rerank | 谢谢嘛<br>Thanks | ok 加微信: abcdefg<br>OK, add my Wechat: abcdefg |
| MMI | 我不是善良的妹纸<br>I am not a kind girl | 我也想骑在大理<br>I want to ride Dali, too |
| EditVec | 谢谢，善良的妹子也会幸福的<br>Thanks, a kind girl will be happy too | 有时间可以去玩<br>Let's go when free |
| JNT | 我不是善良的妹子<br>I am not a kind girl | 我也在大理<br>I am in Dali, too |

# Retrieval-Augmented 2.0：解耦

技术创造未来 科技TEG

## 1.0版本的瑕疵：缺少骨架数据

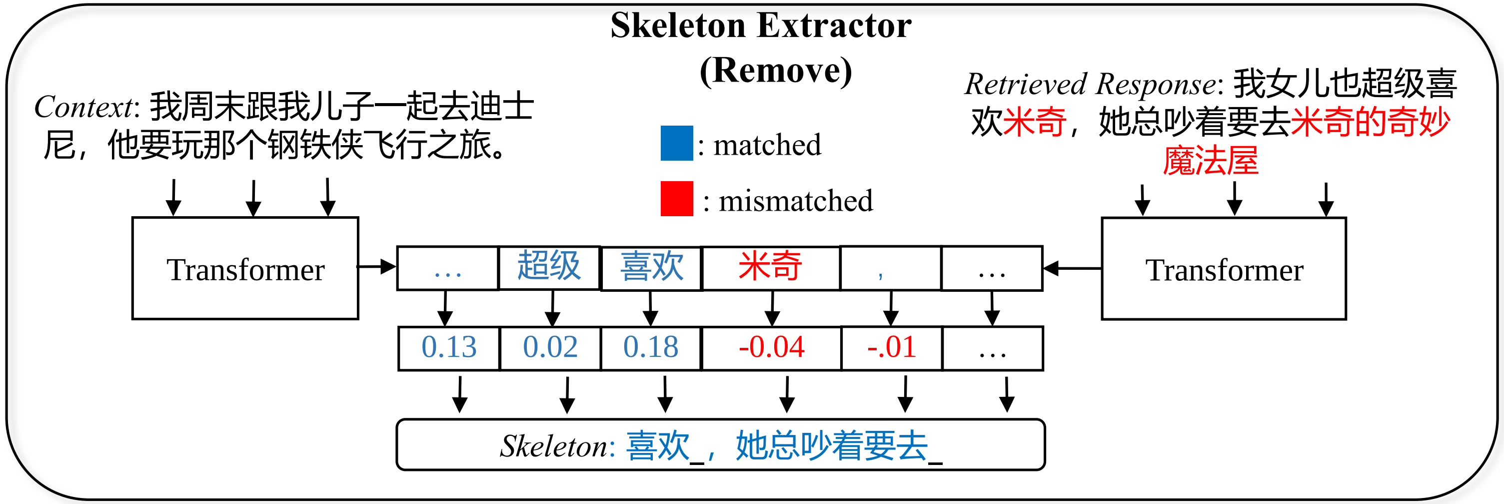
## 2.0版本：把回复生成解耦两个子任务：骨架提取 + 回复生成

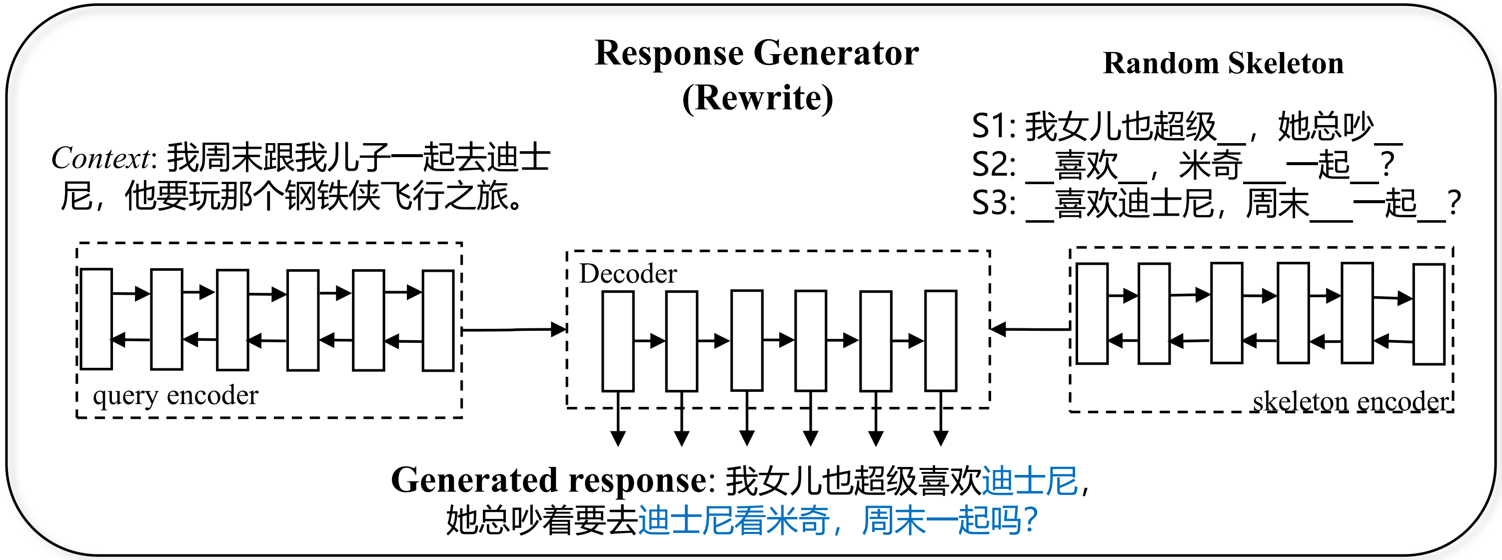
- 骨架提取：对于任意检索结果，提取尽量好的骨架
- 回复生成：对于任意骨架，生成尽量好的回复 (Conditional denoising autoencoder)

被普遍认为是对话领域SOTA的非预训练模型[1]

- **LSTM-Tokens** (Cai et al., 2019b) The state-of-the-art exemplar-conditioned open-domain response generation model. It uses the dialogue context along with tokens extracted from an exemplar response (using a transformer-based matching framework) to inform generation. LSTM with attention is used as the decoder.

[1] Prakhar Gupta, Jeffrey Bigham, Yulia Tsvetkov, Amy Pavel. Controlling Dialogue Generation with Semantic Exemplars. NAACL 2021.




Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. Retrieval-guided dialogue response generation via a matching-to-generation framework. *EMNLP 2019*.

# Retrieval-Augmented 2.0：解耦

技术创造未来 科技TEG

- **人工评测结果：**

| Models | Informativeness | Relevance | Fluency | Dist-1(%) | Dist-2(%) |
|---|---|---|---|---|---|
| *Retrieval* | 2.65 (0.90)† | 2.58 (0.86) | 2.96 (0.72) | **49.10** | **84.19** |
| *Seq2Seq* | 2.01 (0.65) | 2.58 (0.53) | 2.71 (0.43) | 30.38 | 54.52 |
| *Seq2Seq-MMI* | 2.47 (0.70) | 2.79 (0.67) | 2.99 (0.61) | 30.98 | 62.85 |
| *RetrieveNRefine*++ | 2.30 (0.79) | 2.62 (0.63) | 2.82 (0.51) | 29.83 | 61.07 |
| *EditVec* | 2.29 (0.61) | 2.62 (0.60) | 2.83 (0.47) | 35.30 | 67.57 |
| *Skeleton-Lex* | 2.45 (0.61) | 2.80 (0.56) | 2.99 (0.46) | 25.70 | 56.61 |
| Ours | **2.69** (0.87) | **3.11** (0.55) | **3.20** (0.55) | **49.01** | **80.36** |

- **Case study：**

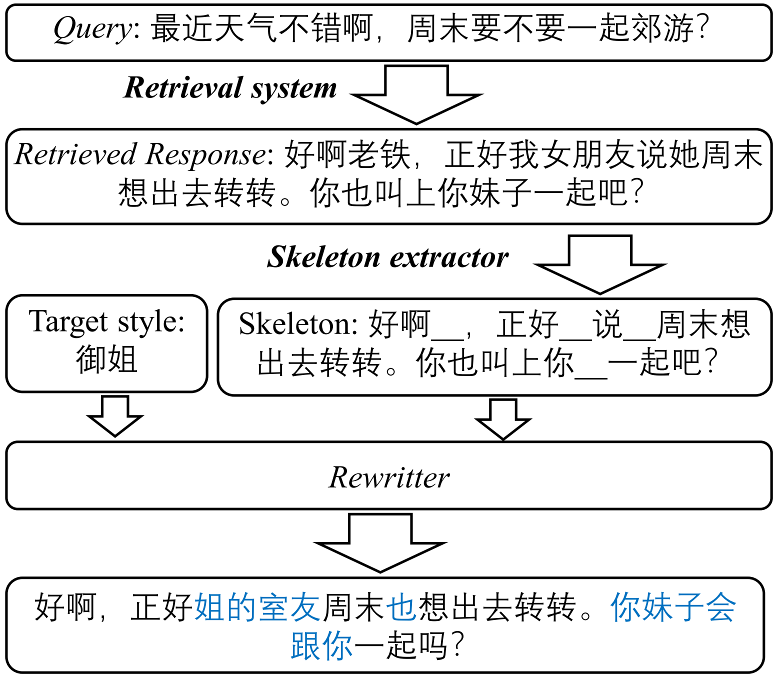
| | |
|---|---|
| Query | 喜欢长头发的女生 I like girls with long hair |
| Retrieved response | 可以 把 头发 后面 汤 大 卷 ， 前面 是 直 刘海 ， 这样 看 的 很 可爱<br>You can try perming the hair, hair with curly ends and straight bangs is very cute. |
| Ours | 我也喜欢长**头发**的女生,**卷**了**直刘海**, 头发就不好看了 I also like girl with long **hair**,<br>I think once the **straight bangs** are **curled**, it doesn’t look good any more. |
| *Seq2Seq-MMI* | 我也喜欢长头发的女生 I also like girls with long hair. |
| *RetrieveNRefine*++ | 我喜欢长头发的女生 I like girls with long hair. |
| *EditVec* | 我也喜欢长头发的女生 I also like girls with long hair. |
| *Skeleton-Lex* | 我喜欢长头发的女生 I like girls with long hair. |
| Query | 我提现怎么还没到 Why hasn’t my withdrawal arrived yet? |
| Retrieved response | 就是 提现 到 你 绑定 的 银行卡 上 了 Withdrawal is in the bound bank card. |
| Ours | 你**提现**的时候**绑定**了**银行卡**吗 Have you **bounded** the **bank card** when **withdrawing** ? |
| *Seq2Seq-MMI* | 我提现也到了 My withdrawal has arrived too. |
| *RetrieveNRefine*++ | 你要支付宝干嘛 Why do you need Alipay? |
| *EditVec* | 你是提现的吗 Do you want to withdraw? |
| *Skeleton-Lex* | 你不是已经到了吗 Haven’t you arrived ? |

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. Retrieval-guided dialogue response generation via a matching-to-generation framework. *EMNLP 2019*.

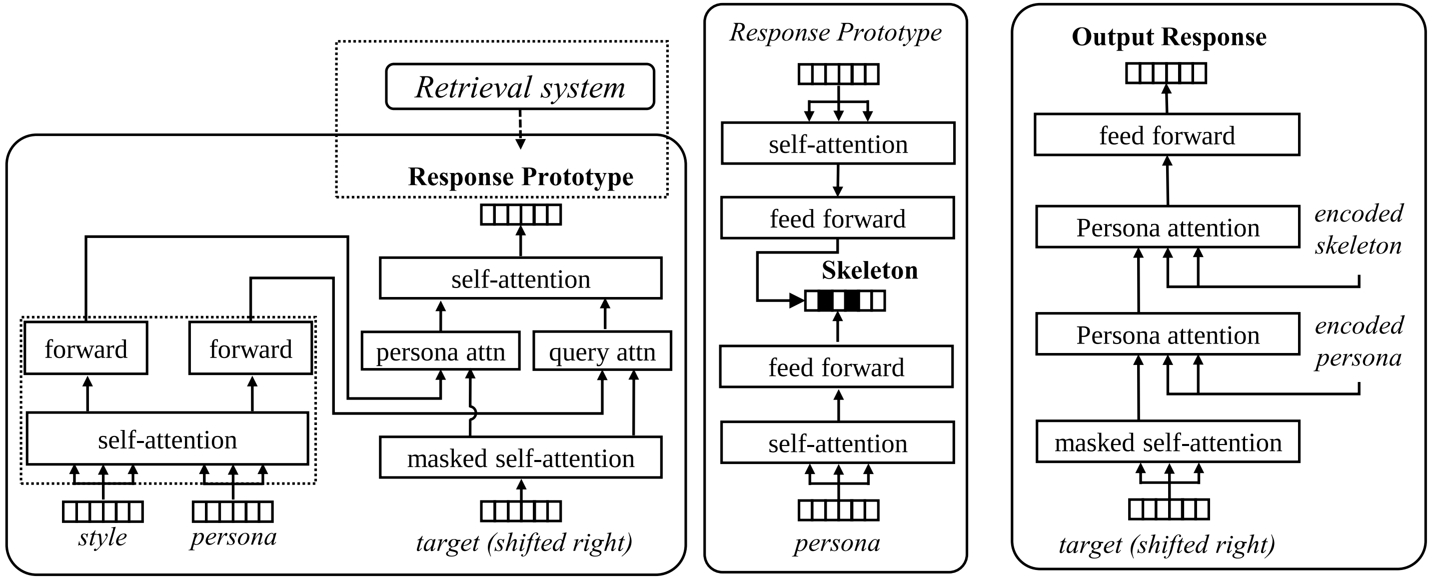
# Retrieval-Augmented 2.0：可控对话生成



- **目标**：在检索系统回复的基础上，生成符合人设并带特定风格的回复（暖男、萌妹子、御姐等）
- **意义**：
  - 支持业务部门对回复风格的需求（例如游戏npc以及IP化对话机器人）
  - 保证质量的前提下实现风格可控对话生成
- **效果**：完美平衡了回复质量与回复风格，能生成高质量并且风格化的回复




**模型**：Generate, Delete, Rewrite三级生成结构。实际系统使用时，Delete & Rewrite模块可以跟任意检索系统搭配使用




(1) Generate | Retrieval (2) Delete (3) Rewrite

Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation. *ACL 2020.*

# 可控对话生成实验结果



**实验结果：**

- 同时在回复质量和回复风格上超过了SOTA
- **回复质量 (Quality)**：提出的模型 (PS) 回复质量超过了SOTA模型GPT2 (3.45 vs 3.32);
- **语言风格** (Style Expression)： PS模型风格表达的准确度超过了SOTA模型ECM (3.69 vs 3.35)

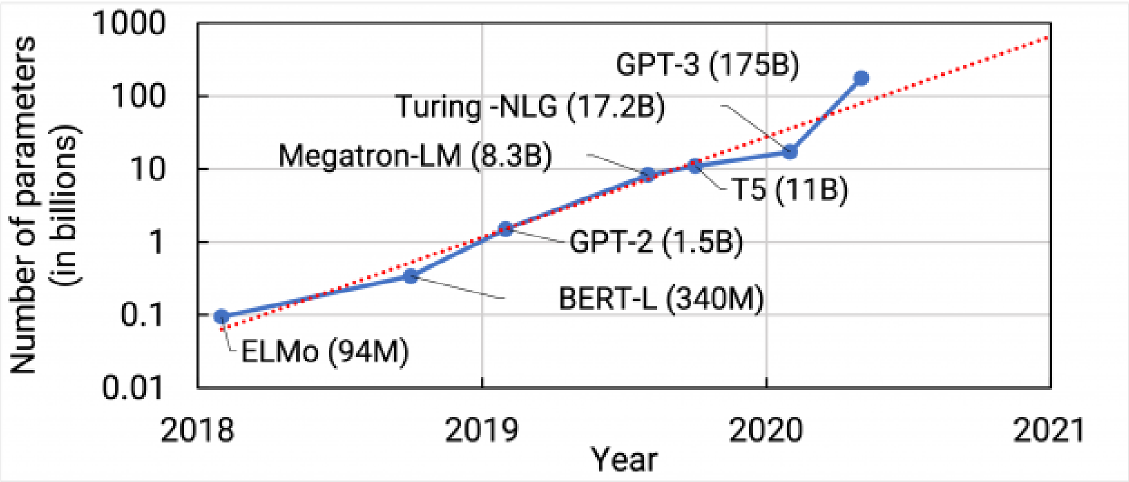
| Style | Metrics | Generative | | | | Retrieval-Based | | | Ours | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Seq2seq | GPT2-FT | Speaker | ECM | SR | RST | RRe | PS w/o R | PS |
| Male | Quality↑ | 2.97 | 3.33 | 2.49 | 2.56 | 2.58 | 2.15 | 2.78 | 2.94 | **3.48** |
| | Style Expression↑ | 2.93 | 2.99 | 3.51 | 3.60 | 2.98 | 3.21 | 3.01 | 3.36 | **3.75** |
| | Ranking↓ | 3.04 | 2.71 | 3.42 | 3.15 | 3.89 | 4.01 | 3.43 | 2.34 | **1.56** |
| Female | Quality↑ | 2.97 | 3.31 | 2.86 | 2.81 | 2.60 | 2.16 | 3.11 | 3.01 | **3.42** |
| | Style Expression↑ | 3.07 | 3.02 | 3.01 | 3.09 | 3.02 | 3.14 | 3.09 | 3.49 | **3.64** |
| | Ranking↓ | 2.94 | 2.62 | 3.18 | 3.20 | 3.66 | 3.86 | 2.89 | 2.28 | **1.52** |
| Overall | Quality↑ | 2.98 | 3.32 | 2.68 | 2.67 | 2.59 | 2.14 | 2.94 | 2.98 | **3.45** |
| | Style Expression↑ | 3.00 | 3.05 | 3.26 | 3.35 | 3.03 | 3.17 | 3.01 | 3.43 | **3.69** |
| | Ranking↓ | 2.99 | 2.66 | 3.30 | 3.17 | 3.78 | 3.94 | 3.16 | 2.31 | **1.54** |
| | Distinct-1(%)↑ | 27.64 | 36.42 | 26.15 | 12.45 | 37.62 | 33.12 | **48.52**† | 29.98 | **40.88** |
| | Distinct-2(%)↑ | 72.33 | 74.30 | 50.40 | 31.64 | 84.33 | 85.63 | **94.11**† | 78.54 | **90.82** |

| **Query** | **Retrieved Response** | **Style** | **Generated Response** |
|---|---|---|---|
| 没有做完的梦最痛。<br>Unfinished dreams hurt the most. | 这几天有时候做噩梦。<br>I sometimes have nightmares at these days. | Male | 哥这几天一直都在做噩梦 。 |
| | | Female | 大姨妈来前几天我老做噩梦。 |
| | | Like | 这几天我很享受做噩梦。 |
| | | Disgust | 最近我就烦的很，天天做噩梦。 |
| | | Happy | 哈哈。这几天我经常梦到你。 |
| | | Anger | 靠！我要去死！这几天我老是做噩梦。 |
| | | Sad | 唉， 日子没法过了，老是做噩梦。 |

# Motivation (In 2020)



## 背景：参数大爆炸的时代




Trend of SOTA Text Generation Model Sizes with Time

为何如此之大?

一个黑盒模型，两种不同任务:

- 储存知识（结构化/非结构化）
- 根据知识进行推理

## 疑惑：人类可以把所有知识存在大脑里吗?


- 显然不行，大家都喜欢开卷考试，憎恨闭卷考试
- 模型的开卷考试：parametric neural networks + non-parametric index
  - 迁移性：通过改变知识库快速切换领域，更新知识
  - 可解释性：检索结果即推理的依据

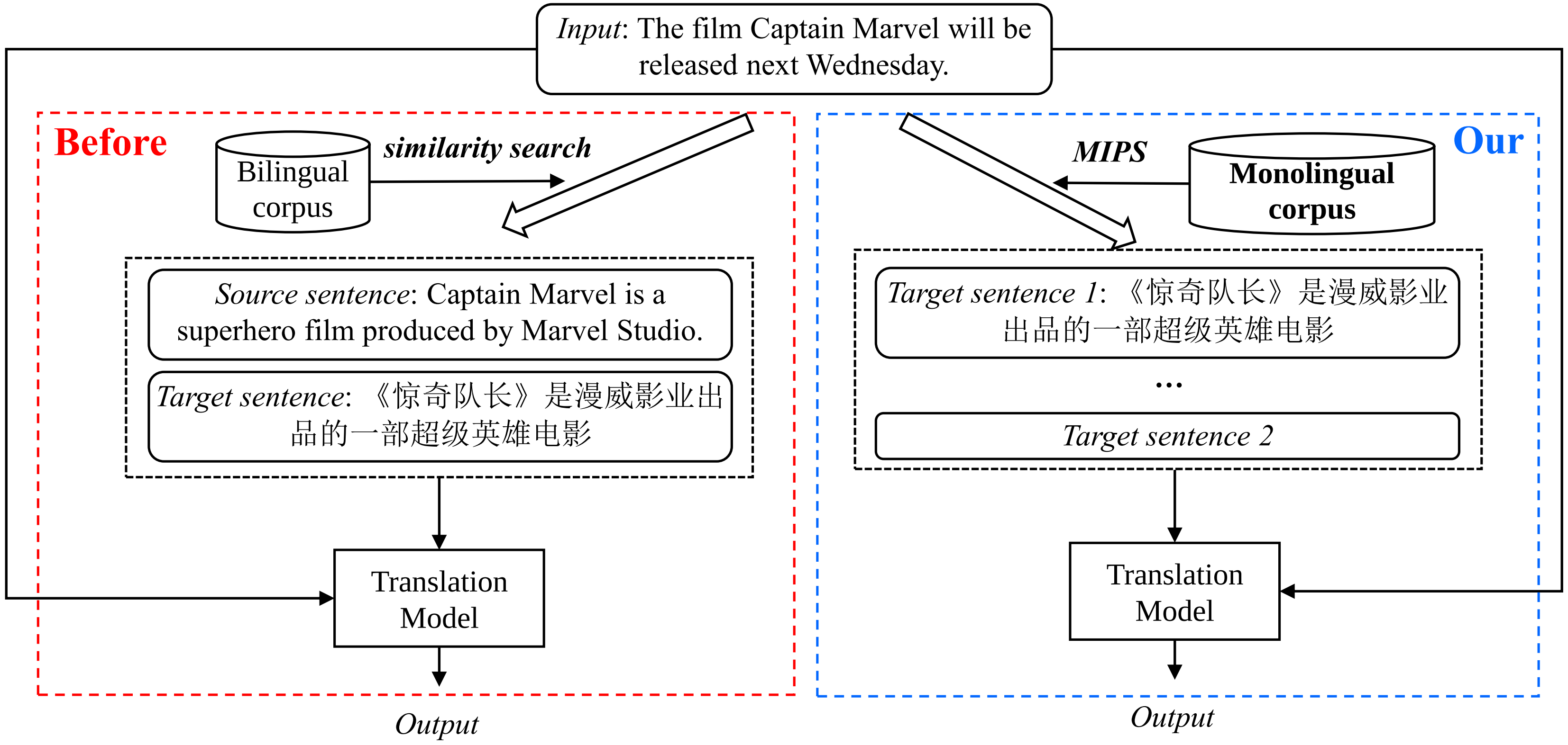
# Retrieval-Augmented 3.0：联合训练+无监督数据

技术创造未来 科技TEG

**2.0版本的问题：**训练推理不统一，检索依赖off-the-shelf系统，不能利用无监督数据

**3.0版本：**

- 单语翻译记忆：从单语语料而不是双语中获取知识 （摆脱数据依赖，更加符合人类习惯）
- 联合训练: 检索模型和生成模型统一到同一个可学习框架中




Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural machine translation with monolingual translation memory. *ACL 2021*. (Outstanding Paper Award)

# Model Overview

技术创造未来 科技TEG

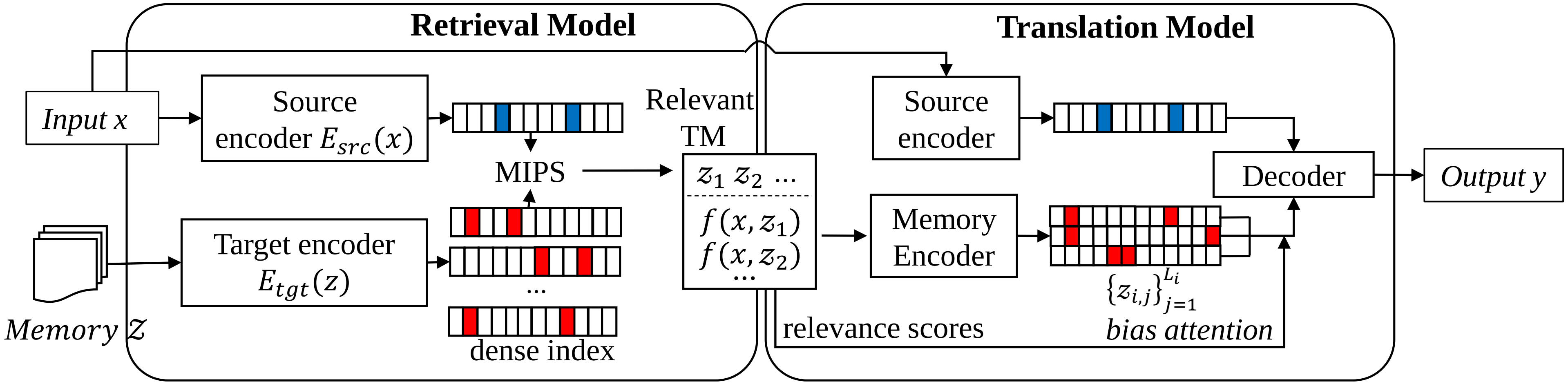
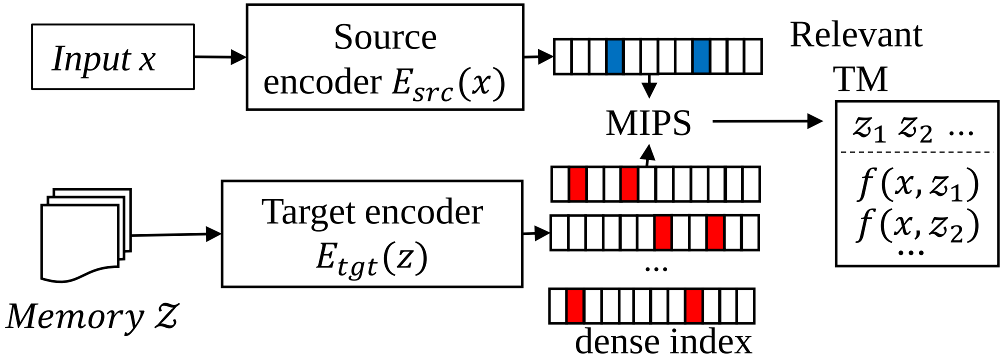



Figure 1: Overall framework. For an input sentence $x$ in the source language, the retrieval model uses Maximum Inner Product Search (MIPS) to find the top-$M$ TM sentences $\{z_i\}_{i=1}^{M}$ in the target language. The translation model takes $\{z_i\}_{i=1}^{M}$ and corresponding relevance scores $\{f(x, z_i)\}_{i=1}^{M}$ as input and generate the translation $y$.

# Retrieval Model






$$E_{src}(x) = normalize(W_{src} Trans_{src}(x))$$

$$E_{tgt}(z) = normalize(W_{tgt} Trans_{src}(z))$$

$$f(x, z) = E_{src}(x)^T E_{tgt}(z)$$

★ **Monolingual Memory**：

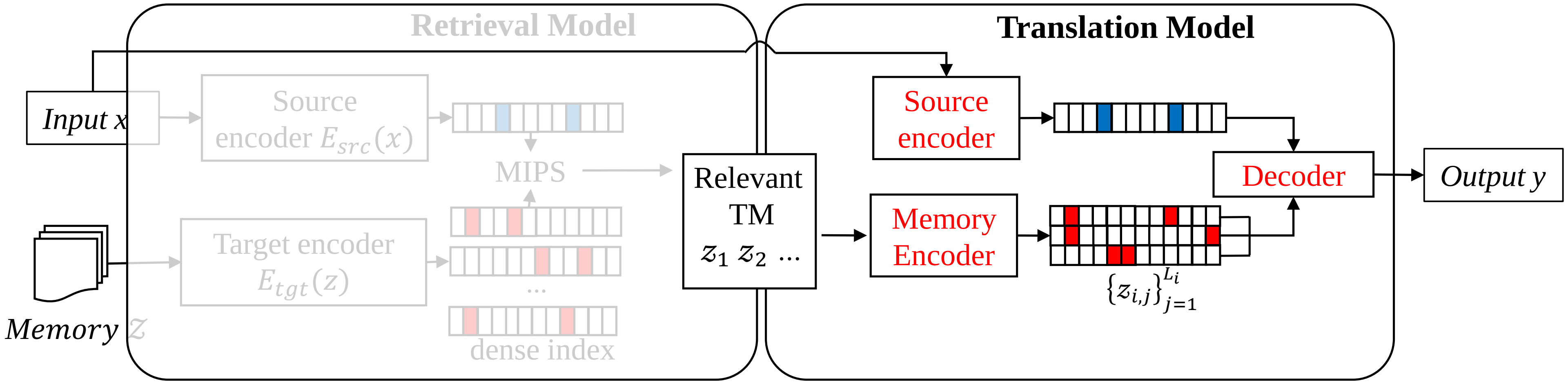
- Connects source-side and target-side
- Abundant data in target language can be used as TM

★ **Fast Retrieval**:

- The selection can be reduced to Maximum Inner Product Search (MIPS)
- Efficient search with off-the-shelf vector search toolkit (FAISS)

# Generation/Translation

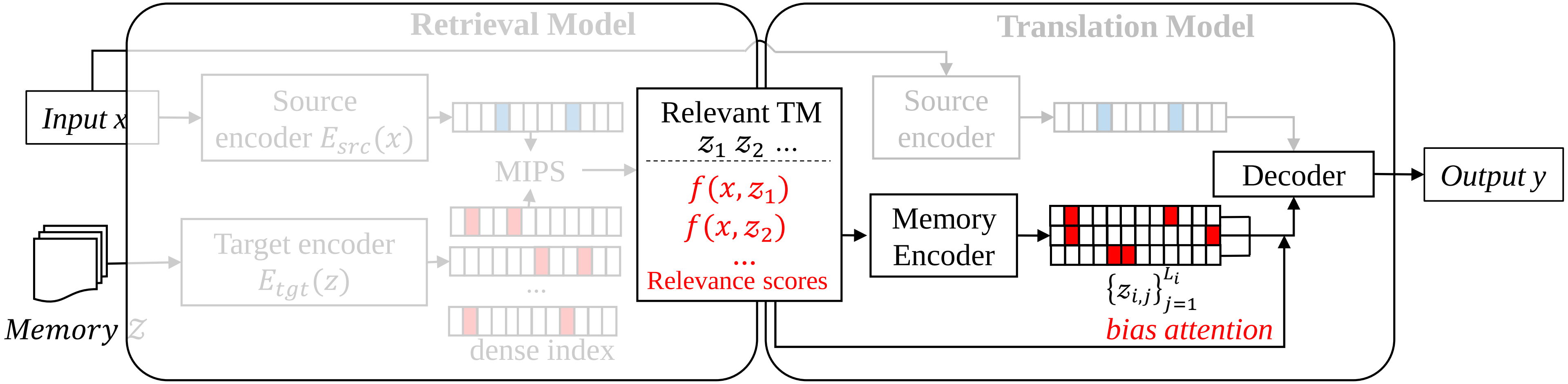





- **Changes to standard Transformer:**
  - A separate memory encoder for TM
  - The decoder attends over the output of both source encoder and memory encoder

# Training: gradient

技术创造未来 科技TEG




$$\alpha_{ij} = \frac{exp(h_t^T W_m z_{i,j} + \boxed{\beta f(x, z_i)})}{\sum_{i=1}^{M} \sum_{k=1}^{L_i} exp(h_t^T W_m z_{i,k} + \boxed{\beta f(x, z_i)})}$$

★ **Task-Specific Retrieval**:

- Unifies the memory retriever and the downstream NMT model into a learnable whole
- Memory retrieval can be end-to-end optimized for the translation objective.

# Experiment 1: Conventional Setting



**Experiment 1: Use bilingual corpus only (to verify the effectiveness of joint training)**

| # | System | Retriever | Es⇒En | | En⇒Es | | De⇒En | | En⇒De | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Dev | Test | Dev | Test | Dev | Test | Dev | Test |
| | *Existing NMT systems** | | | | | | | | | |
| | Gu et al. (2018) | source similarity | 63.16 | 62.94 | - | - | - | - | - | - |
| | Zhang et al. (2018) | source similarity | 63.97 | 64.30 | 61.50 | 61.56 | 60.10 | 60.26 | 55.54 | 55.14 |
| | Xia et al. (2019) | source similarity | 66.37 | 66.21 | 62.50 | 62.76 | 61.85 | 61.72 | 57.43 | 56.88 |
| | *Our NMT systems* | | | | | | | | | |
| 1 | *this work* | None | 64.25 | 64.07 | 62.27 | 61.54 | 59.82 | 60.76 | 55.01 | 54.90 |
| 2 | | source similarity | 66.98 | 66.48 | 63.04 | 62.76 | 63.62 | 63.85 | 57.88 | 57.53 |
| 3 | | cross-lingual (fixed) | 66.68 | 66.24 | 63.06 | 62.73 | 63.25 | 63.06 | 57.61 | 56.97 |
| 4 | | cross-lingual (fixed $E_{\text{tgt}}$)† | 67.66 | 67.16 | 63.73 | 63.22 | 64.39 | 64.01 | 58.12 | 57.92 |
| 5 | | cross-lingual† | **67.73** | **67.42** | **64.18** | **63.86** | **64.48** | **64.62** | **58.77** | **58.42** |

Table 2: Experimental results (BLEU scores) on four translation tasks. *Results are from Xia et al. (2019). †The two variants of our method (model #4 and model #5) are significantly better than other baselines with $p$-value $<$ 0.01, tested by bootstrap re-sampling (Koehn, 2004).

- Our best model (model #5) surpasses the best reported model (Xia et al., 2019) by 1.69 BLEU points in average and up to 2.9 BLEU points (De⇒En)

# Experiment 2: Low-Resource Setting

技术创造未来 科技TEG

## Experiment 2: Plug-and-Play domain adaption with monolingual data

| | Medical | Law | IT | Koran | Subtitle | Avg. | Avg. Δ |
|---|---|---|---|---|---|---|---|
| #Bilingual Pairs | 61,388 | 114,930 | 55,060 | 4,458 | 124,992 | - | - |
| #Monolingual Sents | 184,165 | 344,791 | 165,181 | 13,375 | 374,977 | - | - |
| Using Bilingual Pairs Only | | | | | | | |
| Transformer Base | 47.81 | 51.40 | 33.90 | 14.64 | 21.64 | 33.88 | - |
| Ours | 47.52 | 51.17 | 34.64 | 15.49 | 22.66 | 34.30 | +0.42 |
| + Monolingual Memory | | | | | | | |
| Ours + domain-specific | **50.32** | 53.97 | **35.33** | **16.26** | **22.78** | **35.73** | **+1.85** |
| Ours + all-domains | 50.23 | **54.12** | 35.24 | 16.24 | **22.78** | 35.72 | +1.84 |

Table 4: Test results on domain adaptation.

- 1/4 bilingual data + 3/4 monolingual data
- Monolingual data improves 1.85 BLEU in average
- Strong Cross-domain transferability by hot-swapping domain-specific monolingual TM

# Summary



★ **Monolingual Memory**： Abundant data in target language can be used as TM

★ **Task-Specific Retrieval**： Memory retrieval can be end-to-end optimized for the translation objective.

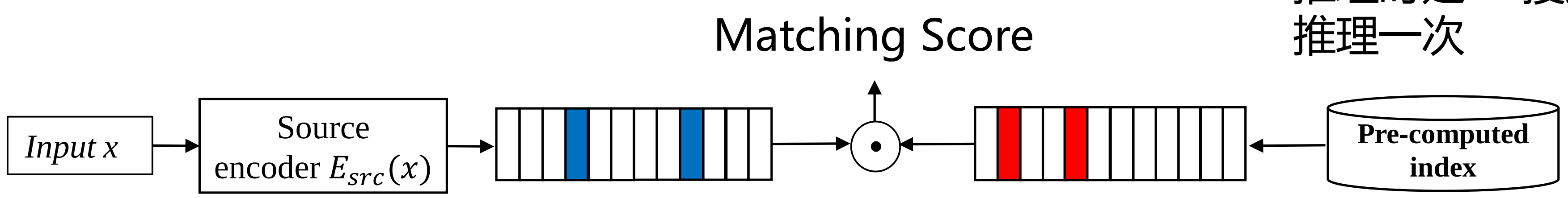
★ **Fast Retrieval**： Efficient search with FAISS-based MIPS

# One drawback



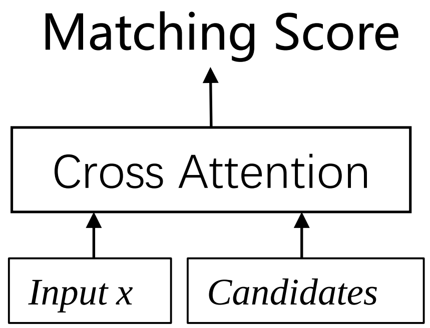
- **Input-Output interaction in dual-encoder architecture relies on inner product only**

推理时延 = 搜索一次+
推理一次



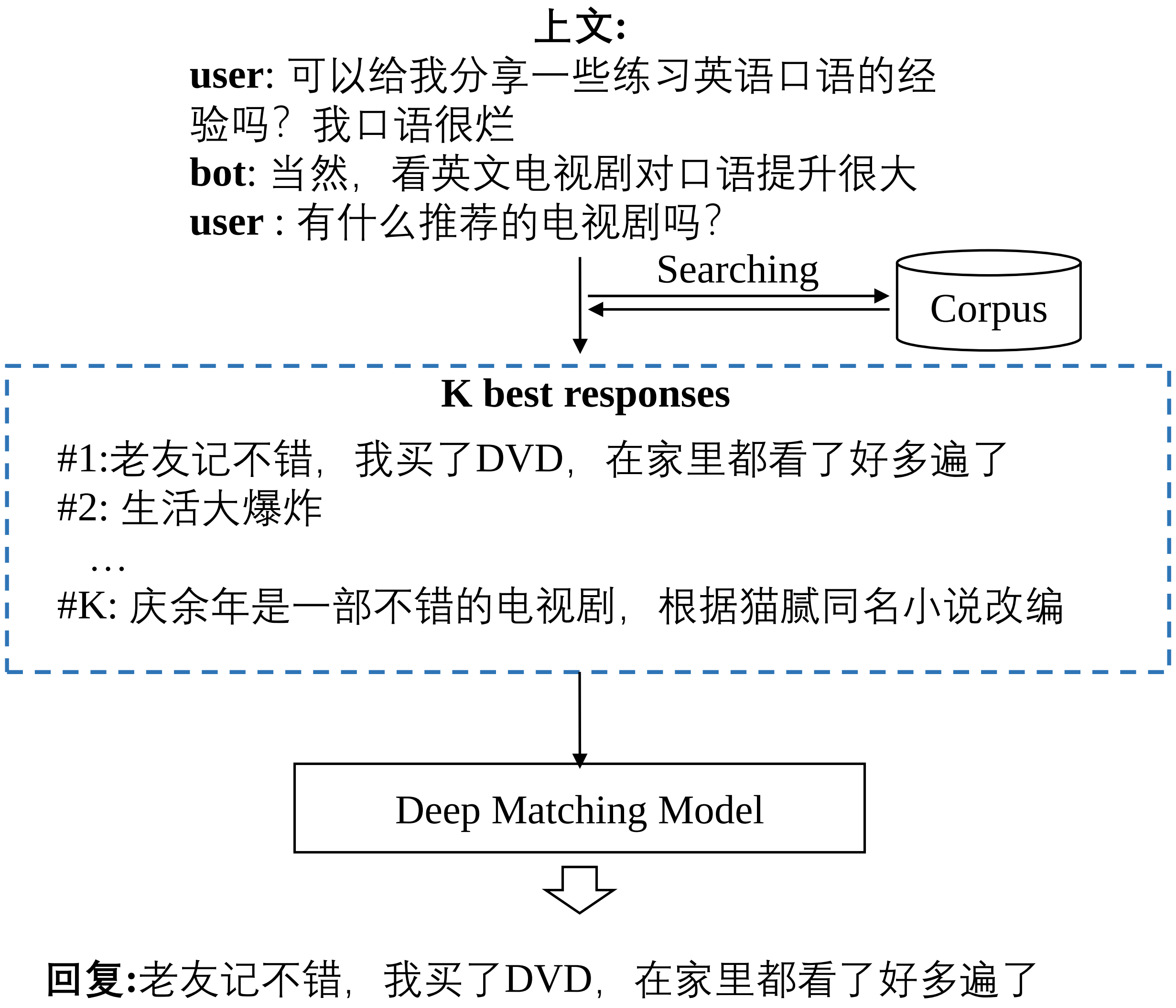

- **More accurate cross-encoder architecture is too slow**

推理时延 = 搜索一次+
推理k次 (k=候选集数目)



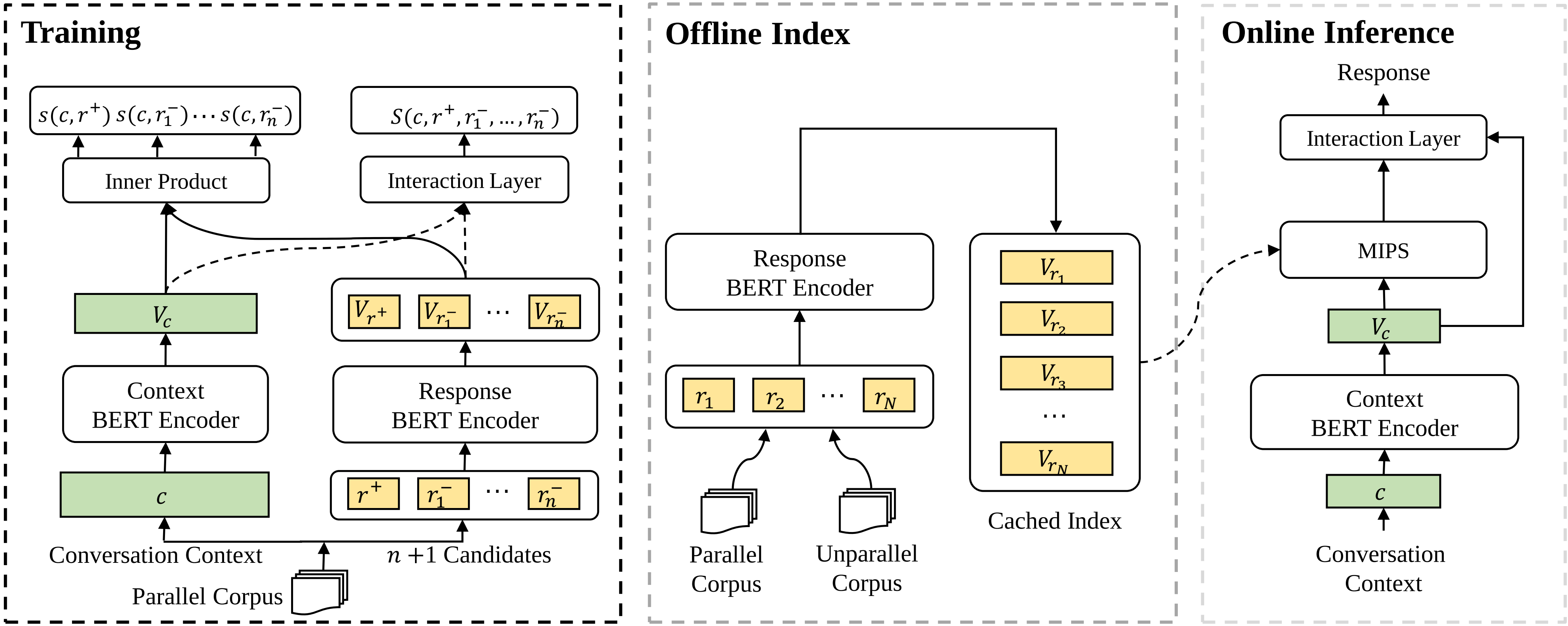
# 近期工作：新的向量检索框架

- 检索经典流程：
  - Recall (Off-the-Shelf Search) -then- Rerank (Learnable Deep Matching Model)
- 结构：
  - Recall：基于query相似度的搜索（TF-IDF，BM25，Dense Vector）
  - Rerank: Cross-encoder Architecture
  - 训练：当作分类任务进行训练（1 positive vs 1 random negative）
- 缺点：
  - 时间复杂度：Recall + k * deep model inference
  - 效果：基于query相似度的recall算法成为了瓶颈



# Model Overview

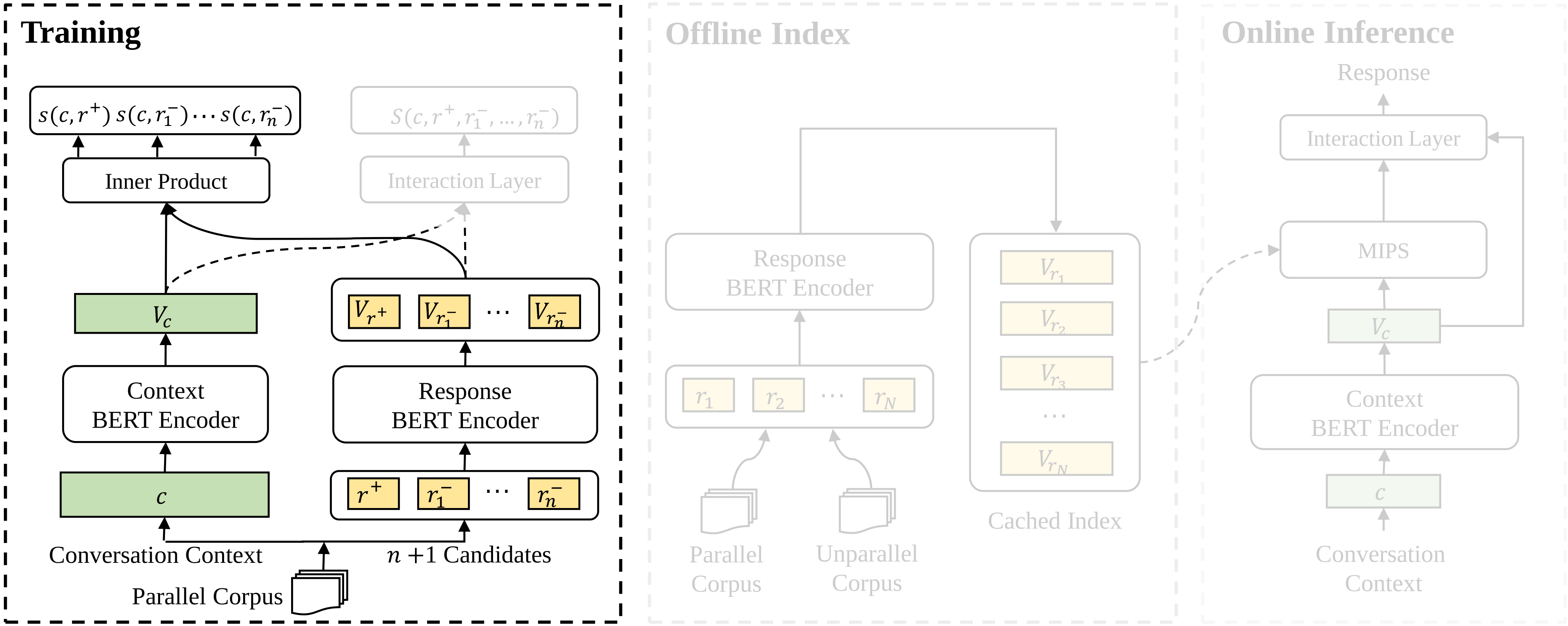
技术创造未来 科技TEG




- Multi-task training achieves SOTA performance on both recall and rerank
- Fast Recall and Rerank via Offline Index
- Select responses from nonparallel corpus

# The Recall Module

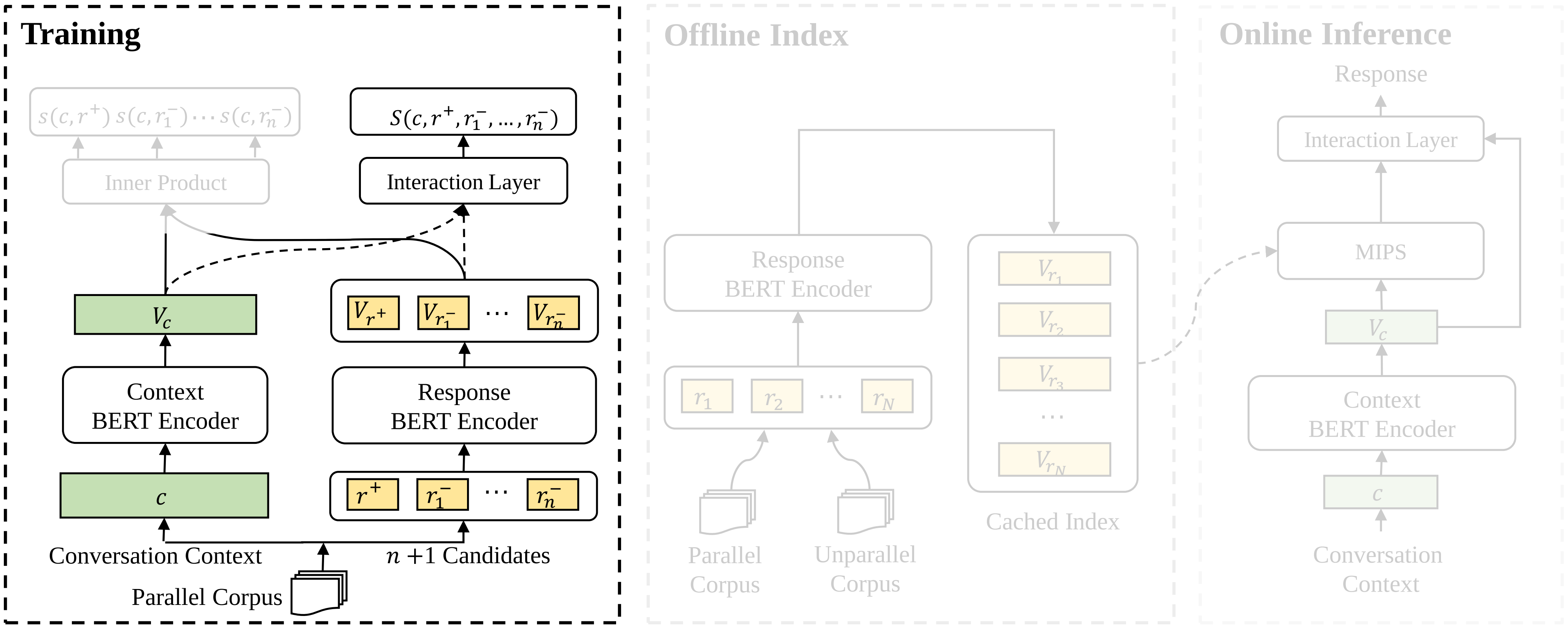
技术创造未来 科技TEG




- Recall module: A typical dual-encoder architecture

# The Rerank Module

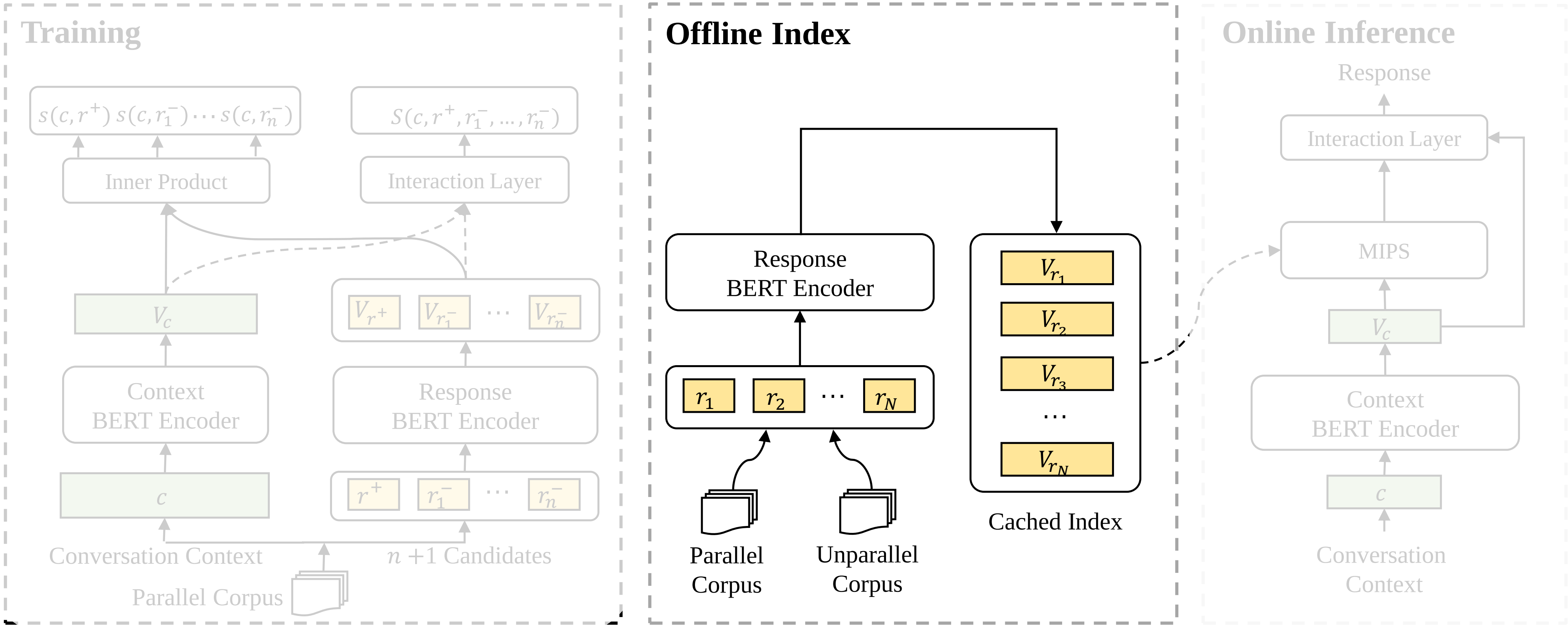
技术创造未来 科技TEG




- Rerank module: an light-weight interaction layer; from point-wise score to list-wise scores
- Training: Multi-task training

# Offline Index

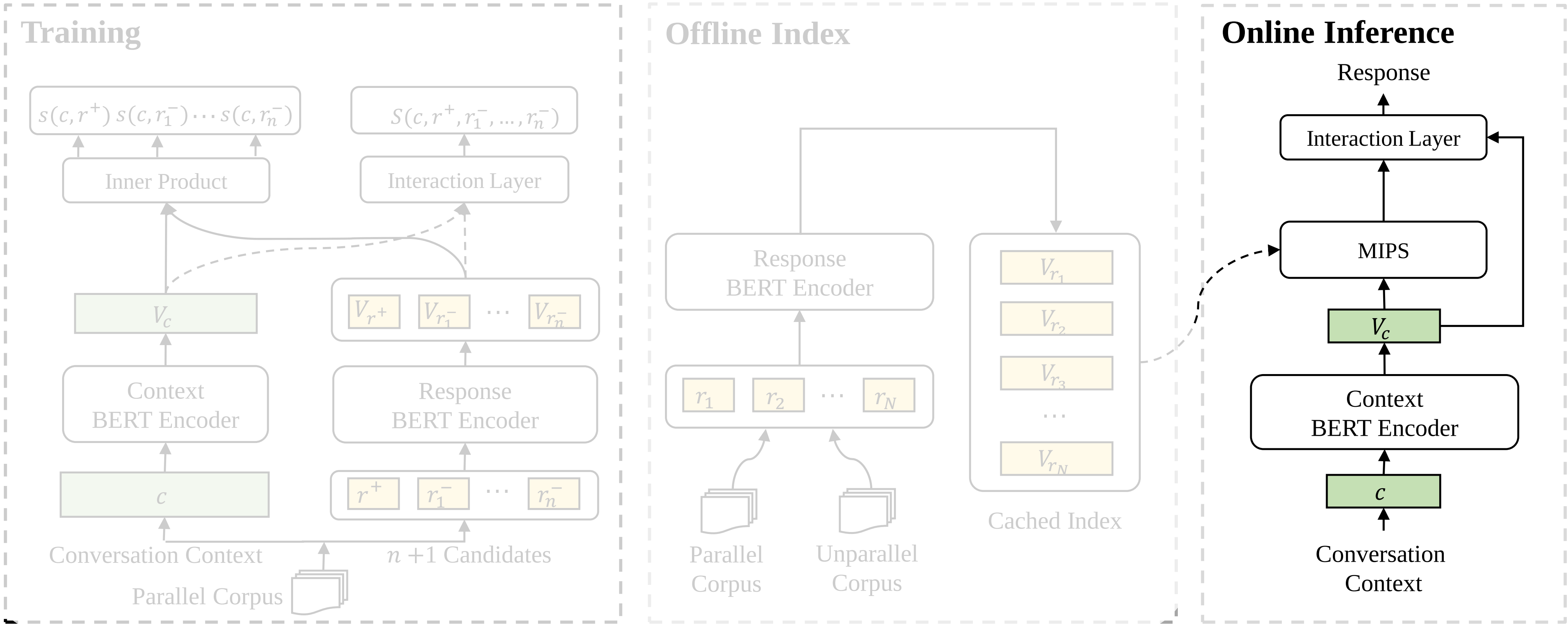





- Pre-compute the representations of all possible responses (both parallel and unparalleled)
- Build index for fast search

# Fast Retrieval






- Recall top k candidates from index
- Fast rerank using the light-weight interaction layer

# 实验一：Rerank

技术创造未来 科技TEG

- Setting: selecting best response from 10 candidates

| Models | Douban | | | | | | Ubuntu | | | RRS | | | | | | E-commerce | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAP | MRR | P@1 | $R_{10}$@1 | $R_{10}$@2 | $R_{10}$@5 | $R_{10}$@1 | $R_{10}$@2 | $R_{10}$@5 | MAP | MRR | P@1 | $R_{10}$@1 | $R_{10}$@2 | $R_{10}$@5 | $R_{10}$@1 | $R_{10}$@2 | $R_{10}$@5 |
| BERT | 0.591 | 0.633 | 0.454 | 0.280 | 0.470 | 0.828 | 0.817 | 0.904 | 0.977 | 0.625 | 0.639 | 0.453 | 0.404 | 0.606 | 0.875 | 0.610 | 0.814 | 0.973 |
| SA-BERT | 0.619 | 0.659 | 0.496 | 0.313 | 0.481 | 0.847 | 0.855 | 0.928 | 0.983 | 0.660 | 0.670 | 0.488 | 0.444 | 0.653 | 0.922 | 0.704 | 0.879 | 0.985 |
| Poly-encoder | 0.608 | 0.650 | 0.475 | 0.299 | 0.494 | 0.822 | 0.882 | 0.949 | 0.990 | 0.715 | 0.729 | 0.578 | 0.518 | 0.708 | 0.925 | 0.924 | 0.963 | 0.992 |
| ColBERT | 0.608 | 0.649 | 0.471 | 0.296 | 0.492 | 0.838 | 0.830 | 0.910 | 0.978 | 0.692 | 0.706 | 0.555 | 0.501 | 0.656 | 0.915 | 0.871 | 0.938 | 0.990 |
| MDFN | 0.624 | 0.663 | 0.498 | 0.325 | 0.511 | 0.855 | 0.866 | 0.932 | 0.984 | - | - | - | - | - | - | 0.639 | 0.829 | 0.971 |
| $UMS_{BERT+}$ | 0.625 | 0.664 | 0.499 | 0.318 | 0.482 | 0.858 | 0.876 | 0.942 | 0.988 | - | - | - | - | - | - | 0.762 | 0.905 | 0.986 |
| BERT-SL | - | - | - | - | - | - | 0.884 | 0.946 | 0.990 | - | - | - | - | - | - | 0.776 | 0.919 | 0.991 |
| SA-BERT+HCL | 0.639 | 0.681 | 0.514 | 0.330 | 0.531 | 0.858 | 0.867 | 0.940 | 0.992 | 0.671 | 0.683 | 0.503 | 0.454 | 0.659 | 0.917 | 0.721 | 0.896 | 0.993 |
| BERT-FP† | 0.644 | 0.680 | 0.512 | 0.324 | 0.542 | 0.870 | **0.911** | **0.962** | **0.994** | 0.709 | 0.724 | 0.565 | 0.505 | 0.705 | 0.932 | 0.870 | 0.956 | 0.993 |
| DR-BERT | **0.659** | **0.695** | **0.520** | **0.338** | **0.572** | **0.880** | 0.910 | **0.962** | **0.993** | **0.758** | **0.771** | **0.648** | **0.584** | **0.744** | 0.928 | **0.971** | **0.987** | **0.997** |
| w/o. IL | 0.648 | 0.685 | 0.516 | 0.331 | 0.550 | 0.868 | **0.913** | **0.961** | **0.993** | 0.733 | 0.746 | 0.606 | 0.542 | 0.727 | **0.933** | 0.960 | 0.984 | 0.996 |
| w/o. NDAP | 0.633 | 0.672 | 0.498 | 0.319 | 0.529 | 0.851 | 0.905 | 0.957 | 0.992 | 0.739 | 0.753 | 0.620 | 0.557 | 0.721 | 0.919 | 0.949 | 0.984 | **0.997** |
| w/o. DA | 0.613 | 0.655 | 0.496 | 0.311 | 0.496 | 0.834 | 0.889 | 0.950 | 0.991 | 0.712 | 0.726 | 0.573 | 0.512 | 0.705 | 0.917 | 0.925 | 0.969 | 0.995 |
| w/o. CL | 0.616 | 0.655 | 0.487 | 0.309 | 0.501 | 0.819 | 0.888 | 0.943 | 0.988 | 0.678 | 0.690 | 0.540 | 0.484 | 0.655 | 0.888 | 0.891 | 0.955 | 0.991 |

- DR-BERT achieves SOTA performance on 4 benchmark datasets
- 联合训练后，只依赖recall model就达到了SOTA（如红框所示）

# 实验二：Full-rank

技术创造未来 科技TEG

- Setting: selecting the best response from whole corpus （human evaluation）

| Baselines | Avg. Human Scores (1-5) |
|---|---|
| docTTTTTquery | 2.12 |
| docTTTTTquery+BERT-FP | 2.80 |
| docTTTTTquery+poly-encoder | 2.92 |
| docTTTTTquery+DR-BERT | 2.96 |
| ColBERT | 2.92 |
| DR-BERT | **3.15** |
| DR-BERT+in-dataset | 3.20 (+1.56%) |
| DR-BERT+out-dataset | **3.24** (+2.78%) |

**Table 4: Full-rank experimental results on our released high-quality RRS test set.**

- 不添加额外数据：DR-BERT效果超过所有baseline
- 添加额外非平行数据：效果进一步提升（如红框所示）

- Inference Speed

| Models | Re-rank Inference Speedup | | | |
|---|---|---|---|---|
| | RRS-10 | RRS-50 | RRS-100 | RRS-1000 |
| **SMN** | 1.0x | 1.0x | 1.0x | 1.0x |
| **MSN** | 2.07x | 2.20x | 2.07x | 1.97x |
| **SA-BERT** | 8.51x | 12.07x | 11.43x | 10.05x |
| **BERT-FP** | 8.81x | 11.91x | 11.27x | 9.99x |
| **ColBERT w/o. cache** | 5.27x | 19.87x | 21.72x | 20.74x |
| **ColBERT** | 11.58x | 52.46x | 76.84x | 217.66x |
| **DR-BERT w/o. cache** | 6.00x | 19.47x | 22.01x | 21.28x |
| **DR-BERT** | 10.48x | 48.26x | 84.79x | 489.68 |
| **DR-BERT w/o cache, IL** | 6.08x | 20.63x | 22.73x | 21.52x |
| **DR-BERT w/o IL** | **13.06x** | **63.53x** | **108.14x** | **502.67x** |

**Table 5: Comparison of re-rank inference speedup with different sizes of candidate set. The batch size for all of the models are the same. IL denotes the interaction layer.**

- Even faster than the very efficient IR model ColBERT

Tencent AI Lab

# Thank You