爱奇艺
悦 享 品 质

爱奇艺多语言台词机器翻译技术实践

--张轩玮

大纲

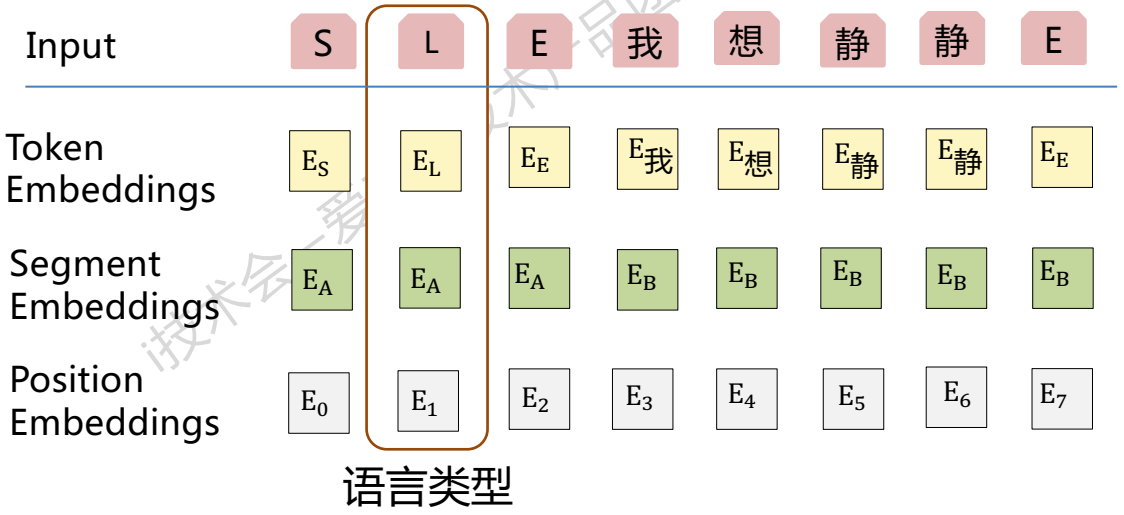
- **背景**
- **多语言台词机器翻译模型**
  - One-to-Many模型
    - 融合台词上下文信息
    - 增强编码能力
    - 增强解码能力
    - 欠翻译/过翻译
    - 增加容错能力
    - 代词翻译
    - 成语翻译
    - 角色翻译
- **在爱奇艺的落地应用**

爱奇艺多语言台词机器翻译

- **背景**
  - 爱奇艺布局海外市场，长视频出海，重要的一环就是台词翻译
  - 多语言台词翻译
    - 中文台词翻译为：*泰语、越南语、印尼、马来、西班牙、阿拉伯、……*
  - 台词翻译特点：
    - 句子短，上下文信息不足，歧义性大
    - 很多台词来源于ocr或asr识别的结果，会有错误
    - 角色名、代词的正确翻译很重要
    - 视频场景信息对语义消歧有帮助

多语言台词机器翻译模型

- **One-to-Many模型**
  - 目标语言很多，使得训练和部署模型的代价很大
  - One-to-Many模型
    - 一个模型，翻译多种目标语言，不同目标语言之间参数共享
    - 优点：
      - 极大的减轻了模型的训练，部署和维护的成本
      - 充分利用不同语言之前迁移学习的特点，起到相互促进的作用，提高效果



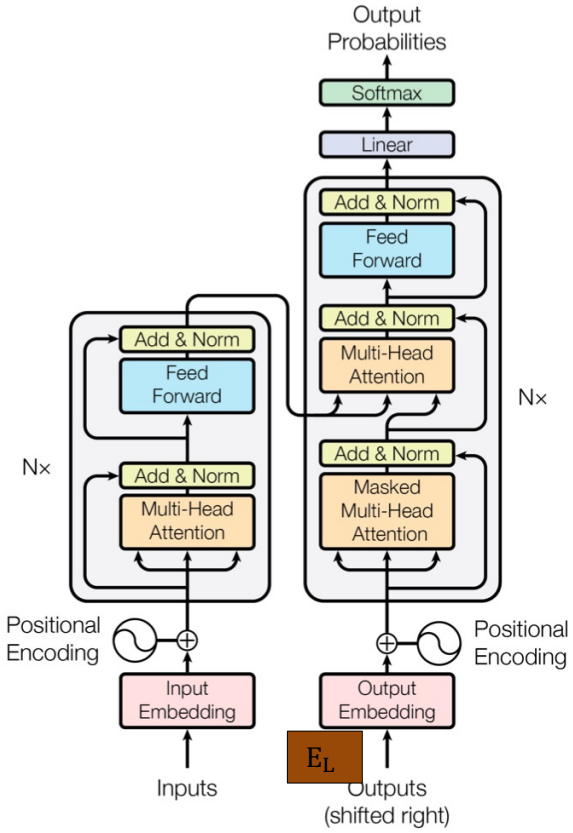


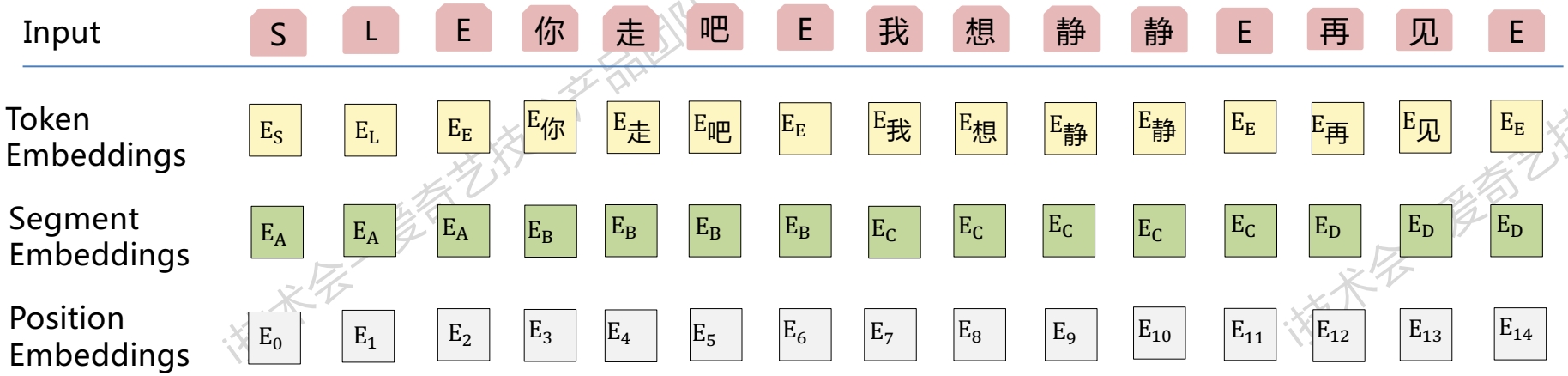
Figure 1: The Transformer - model architecture.

多语言台词机器翻译模型

- **融合台词上下文信息**
  - 台词一般较短，上下文信息不足，会有歧义
  - Context-Fusion by BERT style（Encoder）
    - 输入：上文+分隔符 + 中心句 + 分隔符 + 下文
    - 输出：对上文和下文进行MASK

举例：我想静静 { Let me alone / I miss Jingjing }

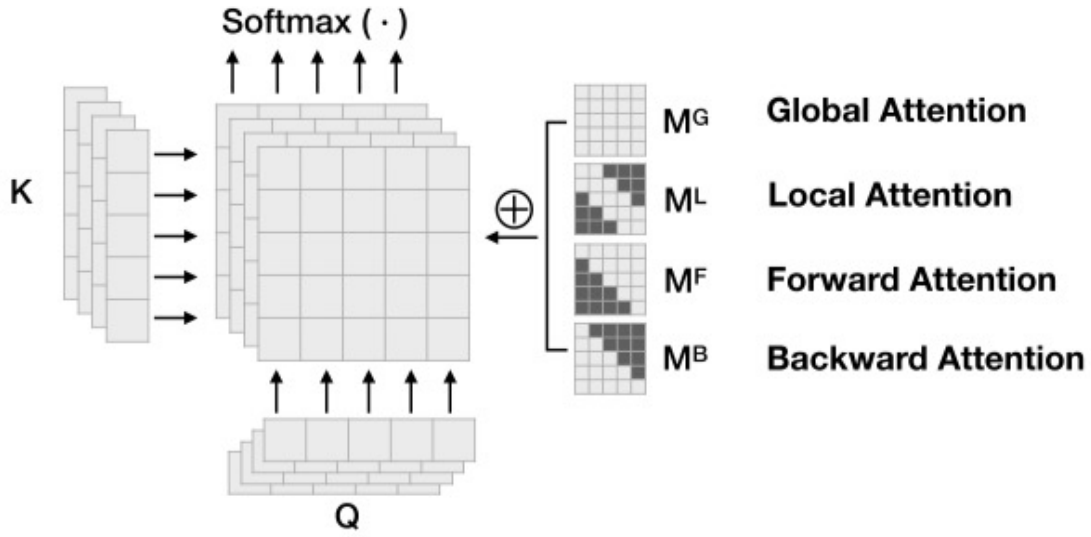
(你走吧) 我想静静 (再见) → Let me alone



多语言台词机器翻译模型

- **增强编码能力**

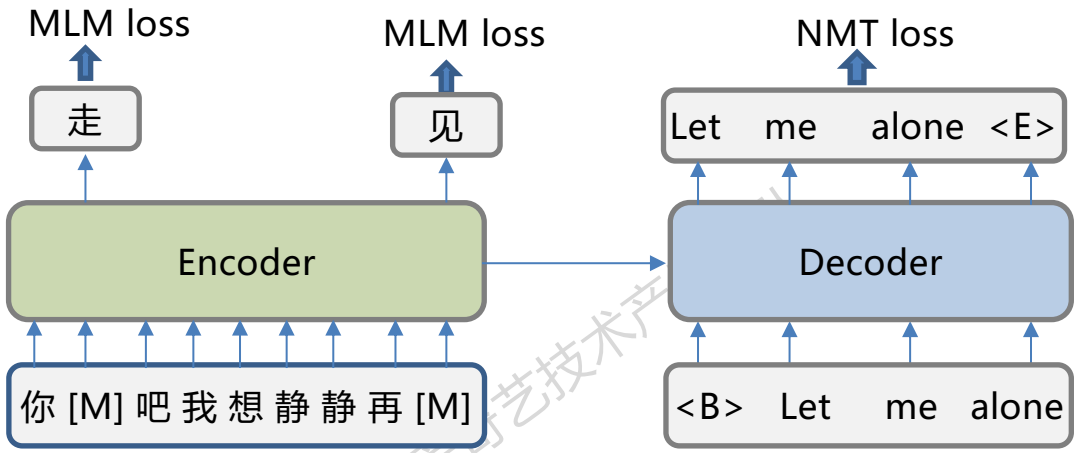
强化Attention：鼓励不同的head学习不同的特征，从而丰富模型的表征能力



a) 强化Attention

- Global Attention：建模任意词之间的依赖关系
- Local Attention ： 强制模型发掘局部的信息特征
- Forward and Backward Attention： 建模模型序列顺序信息

MLM：借鉴BERT，使用Masked LM任务增强模型对文本的理解能力



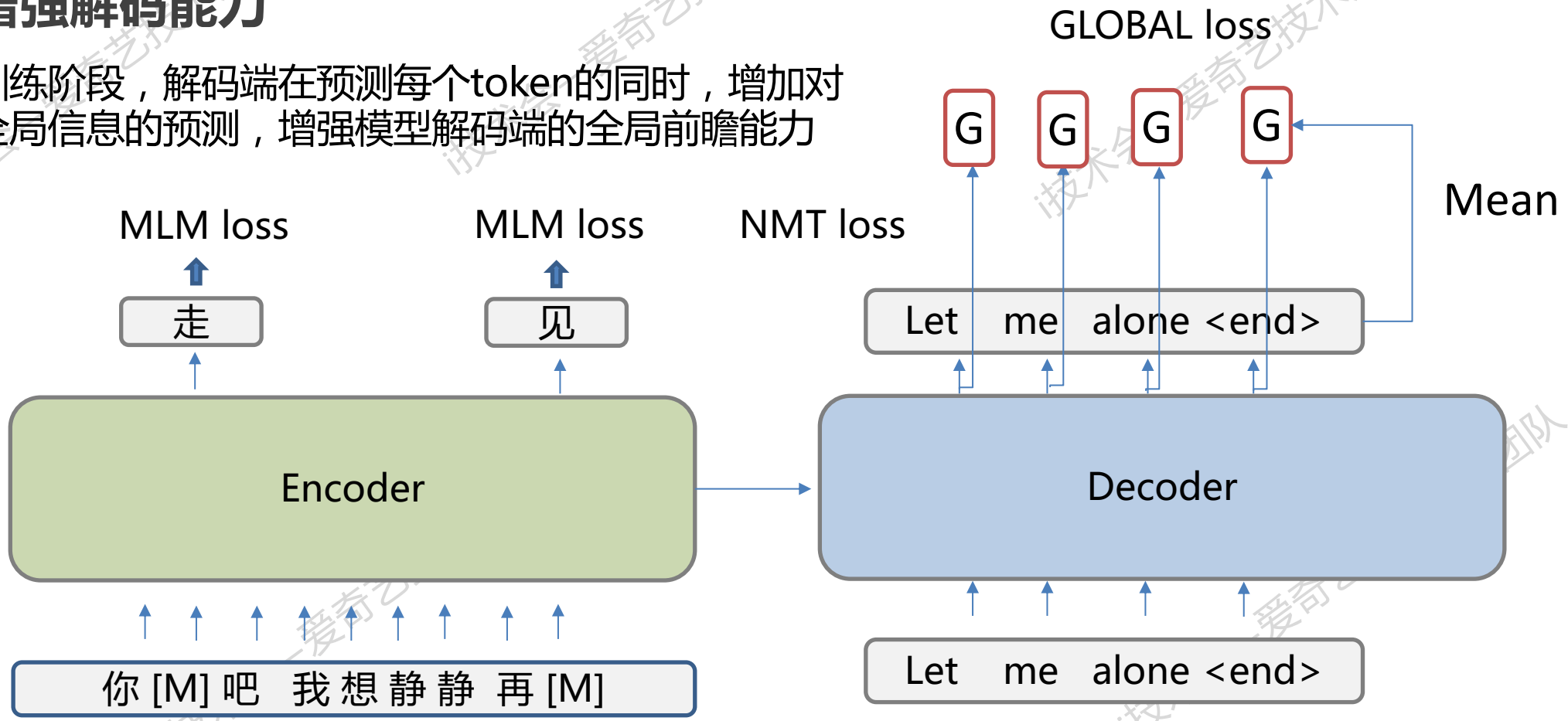
b) Masked LM

$$loss = NMT\ loss + \alpha * MLM\ loss \quad (\alpha < 1)$$

多语言台词机器翻译模型

- **增强解码能力**

  训练阶段，解码端在预测每个token的同时，增加对全局信息的预测，增强模型解码端的全局前瞻能力

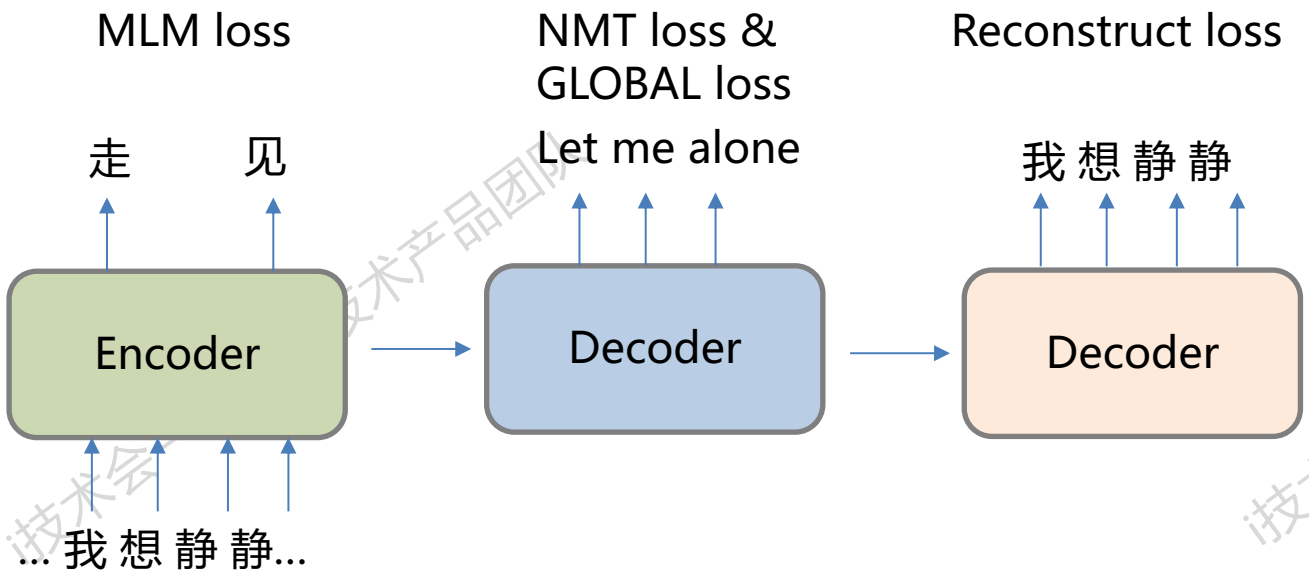


$$\text{loss} = \text{NMT loss} + \alpha * \text{MLM loss} + \beta * GLOBAL\ loss \quad (\alpha, \beta < 1)$$

多语言台词机器翻译模型

- **欠翻译/过翻译**
  - 欠翻译：目标语言词语缺失
  - 过翻译：目标语言词语冗余
  - 增加Reconstruct loss

（你走吧）我想静静（再见）
- Let alone
- Let me me alone

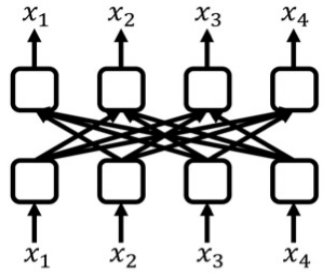


$$\text{loss} = \text{NMT loss} + \alpha * \text{MLM loss} + \beta * GLOBAL\ loss + \gamma * \text{Reconstruct loss} \quad (\alpha, \beta, \gamma < 1)$$

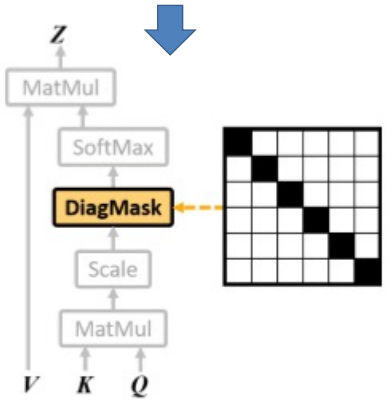
多语言台词机器翻译模型

- **增加容错能力**
  - 台词字幕有很大部分来源于OCR或ASR，会出现一些错误
  - T-TA（Transformer-based Text Autoencoder）：

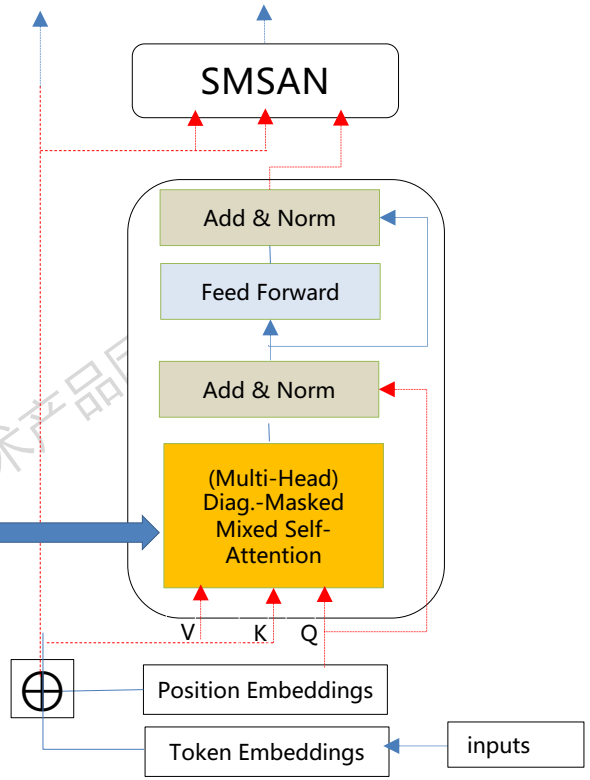
$$\text{loss} = \text{NMT loss} + \alpha * \text{MLM loss} + \beta * GLOBAL\ loss + \gamma * \text{Reconstruct loss} + \delta * \text{TTA loss} \quad (\alpha, \beta, \gamma, \delta < 1)$$



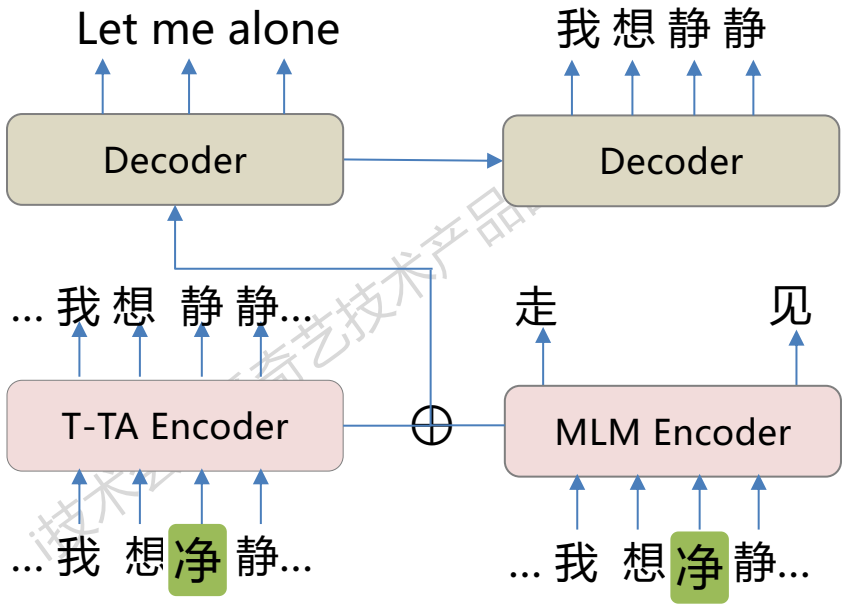
a) Language Autoencoding



b) Diagonal Masking



c) T-TA



d) 融合T-TA Encoder

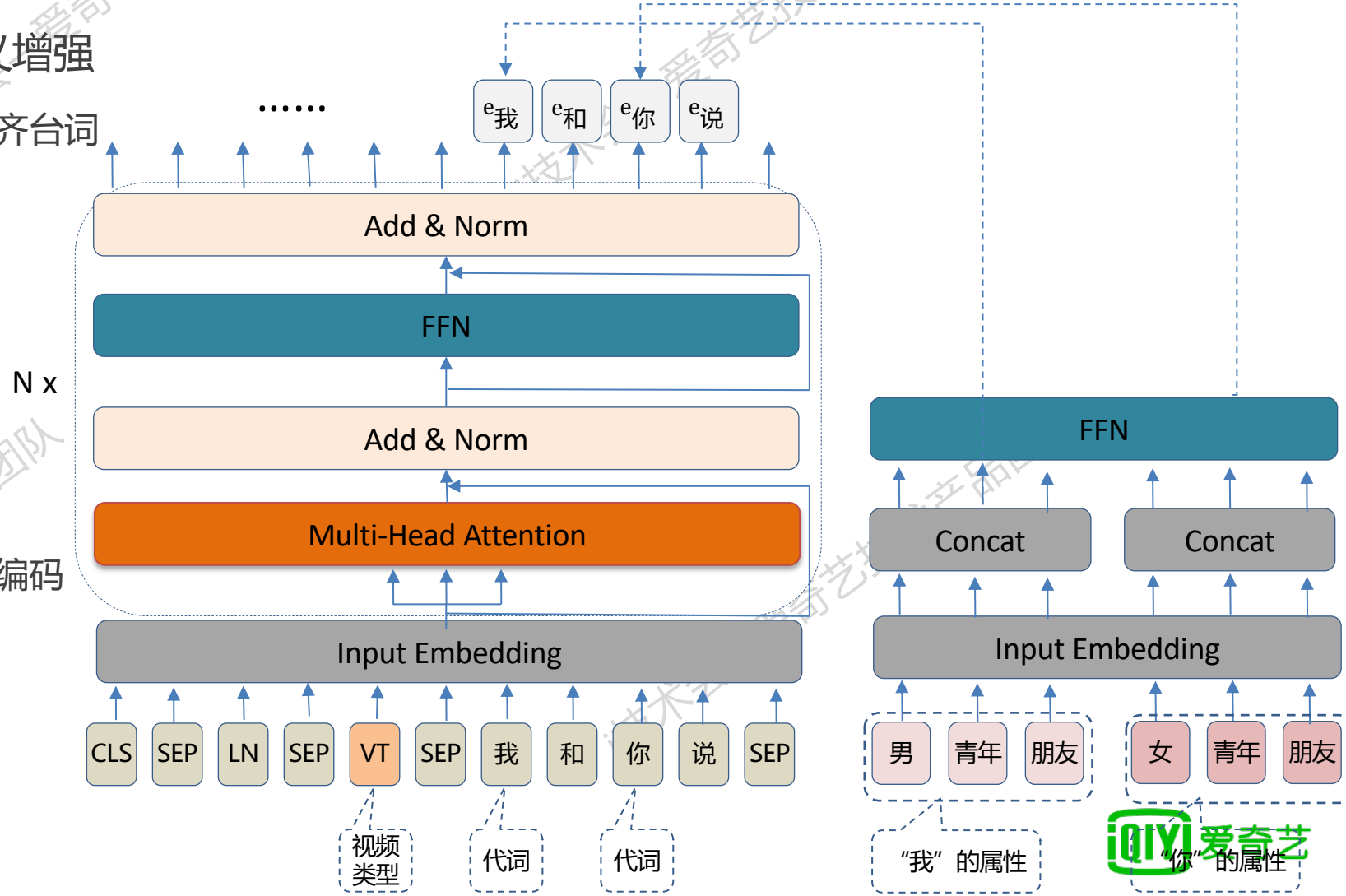
多语言台词机器翻译模型

- **代词翻译**
  - 对于某些目标语言，一个中文代词对应多种翻译，以泰语为例：

汉语人称代词	对应泰语人称代词数量	泰语人称代词
第一人称代词(我)	12	ผม: 男性自称语，多用于正式以及非正式的场合 กระผม: 男性自称语，多用于古代时期多用于正式的场合，对长辈和地位高的人使用，多为自谦 ดิฉัน: 女性自称语，多用于正式的场合，对长辈和地位高的人使用，多为自谦建议在非常正式的场合使用，如下级对上级 ฉัน: 一般为女性自称，但沿用至今很多口语中已经不分男女，适用于多种场合 หนู: 多用于女性在长辈面前的自称 ข้าพเจ้า: 正式场合和祭拜时的自称 เรา: 男女通用的自谦语，相当于女性自谦语的หนู，但หนู更礼貌正式，其中有“我们”的意思 อู: 男女通用自称语，用于在好朋友面前的自称（多译为：老子），其他场合用会显得很不礼貌 ข้า: 古语，通常是自降身份的自称，用于古装剧 พี่/น้อง/ลูก: 以身份的不同来称呼，年纪大的人可以在年幼的人面前自称พี่，年幼的人可以在年长的人面前自称น้อง，在父母面前自称ลูก或以小名自称
第二人称代词(你,您)	15	พระคุณเจ้า: 对高僧的称呼 พระองค์ท่าน: 对君主的称呼 ท่าน: 对长辈的礼貌称呼 คุณ: 适用于多场合，也可以是长辈，平辈，晚辈中的称呼 เธอ: 对女性朋友的称呼 เจ้า: 长辈对晚辈的亲切称呼，或者用于亲密朋友之间多用于古装剧，常用的第二人称代词 หล่อน: 女性对女性的称呼，偏贬义词 นาย: 男性对亲密朋友的称呼，但口语中不常用 มึง: 用于对亲密朋友的称呼，其他场合会显得不礼貌，多为藐视之语 แก: 多用于上司对下属，长辈对晚辈的熟悉称呼，非正式场合使用 หนู: 长辈对晚辈的亲切称呼 ลื้อ: 泰籍华裔男性对好朋友的称呼 พี่/น้อง/ลูก: 说话对象的不同采用不同的称呼，年幼的人称呼年纪略大的人为พี่，年长的人称呼年幼的人为น้อง，长辈称呼晚辈为ลูก
第三人称代词(他,她)	5	เขา: 男性他，适用于多场合 เธอ: 女性她，使用于多场合 มัน: 动物它，也有称呼人，但带有强烈的贬义 พระองค์: 指代地位高的官员或上司 ท่าน: 指代长辈或者尊敬的人

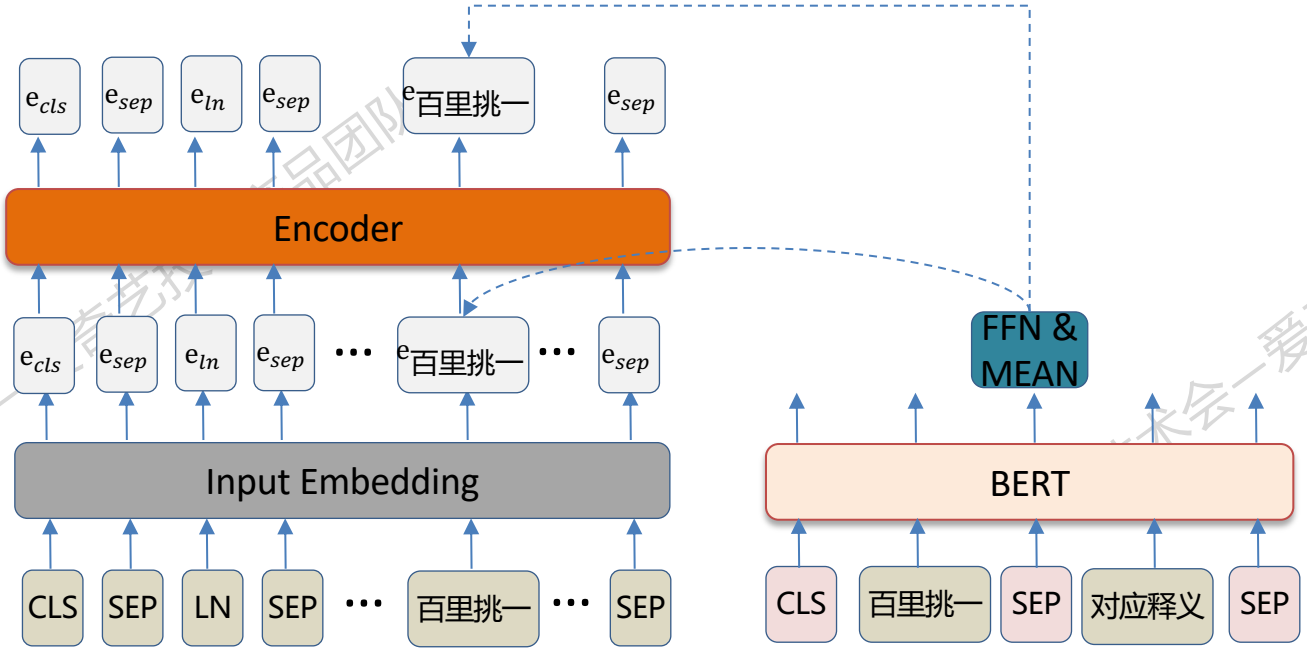
多语言台词机器翻译模型

- **代词翻译**
  - 融合视频场景信息的代词语义增强
    - 通过人脸识别和声纹识别对齐台词和角色
    - 标注角色的人物属性
      - 性别
      - 年龄
      - 人物关系
      - 身份
    - 人物属性编码作为对应代词编码



多语言台词机器翻译模型

- **成语翻译**
  - 成语翻译特点：
    - 语义独立：跟上下文无关
    - 无法按字面意思翻译，需借助其它辅助信息（释义，人工词表等）
  - 使用BERT编码中文释义，直接替换Encoder的成语输入和添加到encoder输出

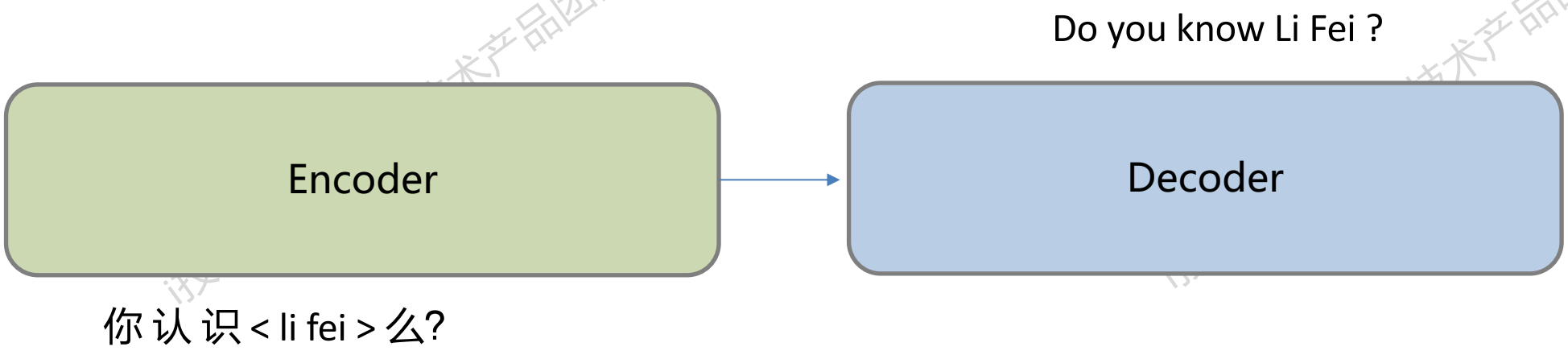


多语言台词机器翻译模型

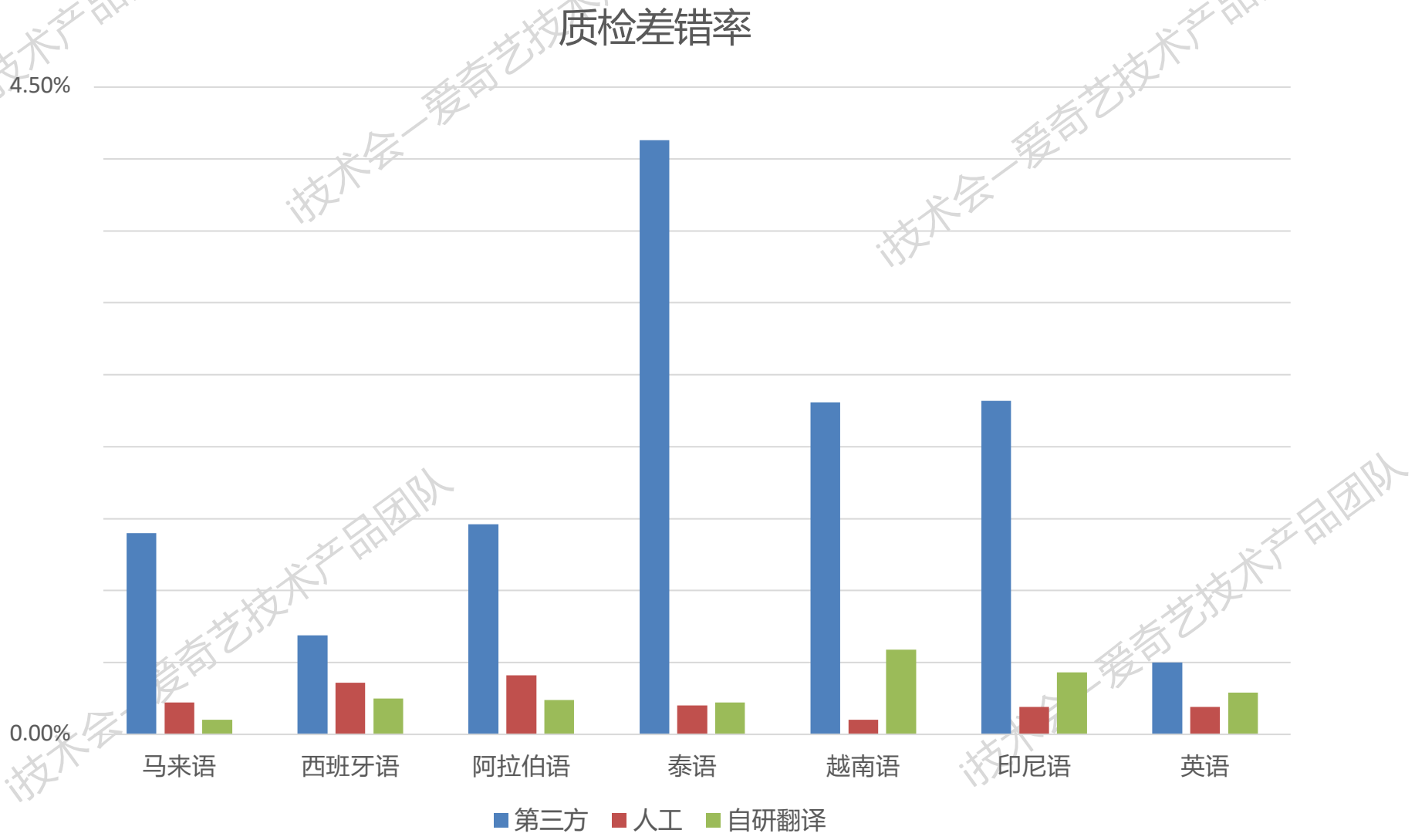
- **角色名的翻译**

通过增加特殊标识的方式以及数据增强，使得模型学习到特定copy能力，大多数语言以中文拼音作为角色的对应翻译，这里以此为例

- ✓ 将人名替换成拼音以及特殊标识<>(如李飞->< li fei >)
- ✓ 通过训练集挖掘人名以及姓氏模版和伪名字合并成伪数据
- ✓ 将增强的数据加到原来数据中进行训练



在爱奇艺的落地应用



在爱奇艺的落地应用

国际站

Simplified Chinese

esian

ysia

Tranditional Chinese

Arabic

Vietnamese

Thai

References

1. Fast and Accurate Deep Bidirectional Language Representations for Unsupervised Learning. 2020
2. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. 2020
3. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019
4. Mixed Multi-Head Self-Attention for Neural Machine Translation 2019
5. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. 2017
6. Attention Is All You Need. 2017
7. Neural Machine Translation with Reconstruction. 2016

悦　　享　　品　　质

Q & A

提问送爱奇艺会员卡！

爱奇艺