多语言阿里小蜜
七步构建跨越语言鸿沟的对话机器人

张佶

阿里巴巴 高级算法专家

Geekbang 极客邦科技 | TGO 鲲鹏會

TGO鲲鹏会

汇聚全球科技领导者的高端社群

全球12大城市

850+高端科技领导者

使命 Mission

为社会输送更多优秀的科技领导者

愿景 Vision

构建全球领先的有技术背景优秀人才的学习成长平台

扫描二维码，了解更多内容

Geekbang. 极客邦科技 InfoQ

极客邦科技
会议推荐2019

5月

QCon 北京
全球软件开发大会
大会：5月6–8日
培训：5月9–10日

QCon 广州
全球软件开发大会
培训：5月25–26日
大会：5月27–28日

6月

GTLC GLOBAL TECH LEADERSHIP CONFERENCE 上海
技术领导力峰会
时间：6月14–15日

GMTC 北京
全球大前端技术大会
大会：6月20–21日
培训：6月22–23日

7月

ArchSummit 深圳
全球架构师峰会
大会：7月12–13日
培训：7月14–15日

10月

QCon 上海
全球软件开发大会
大会：10月17–19日
培训：10月20–21日

11月

GMTC 深圳
全球大前端技术大会
大会：11月8–9日
培训：11月10–11日

AiCon 北京
全球人工智能与机器学习大会
大会：11月21–22日
培训：11月23–24日

12月

ArchSummit 北京
全球架构师峰会
大会：12月6–7日
培训：12月8–9日

阿里小蜜的问答处理流程

知识问答 QA Bot

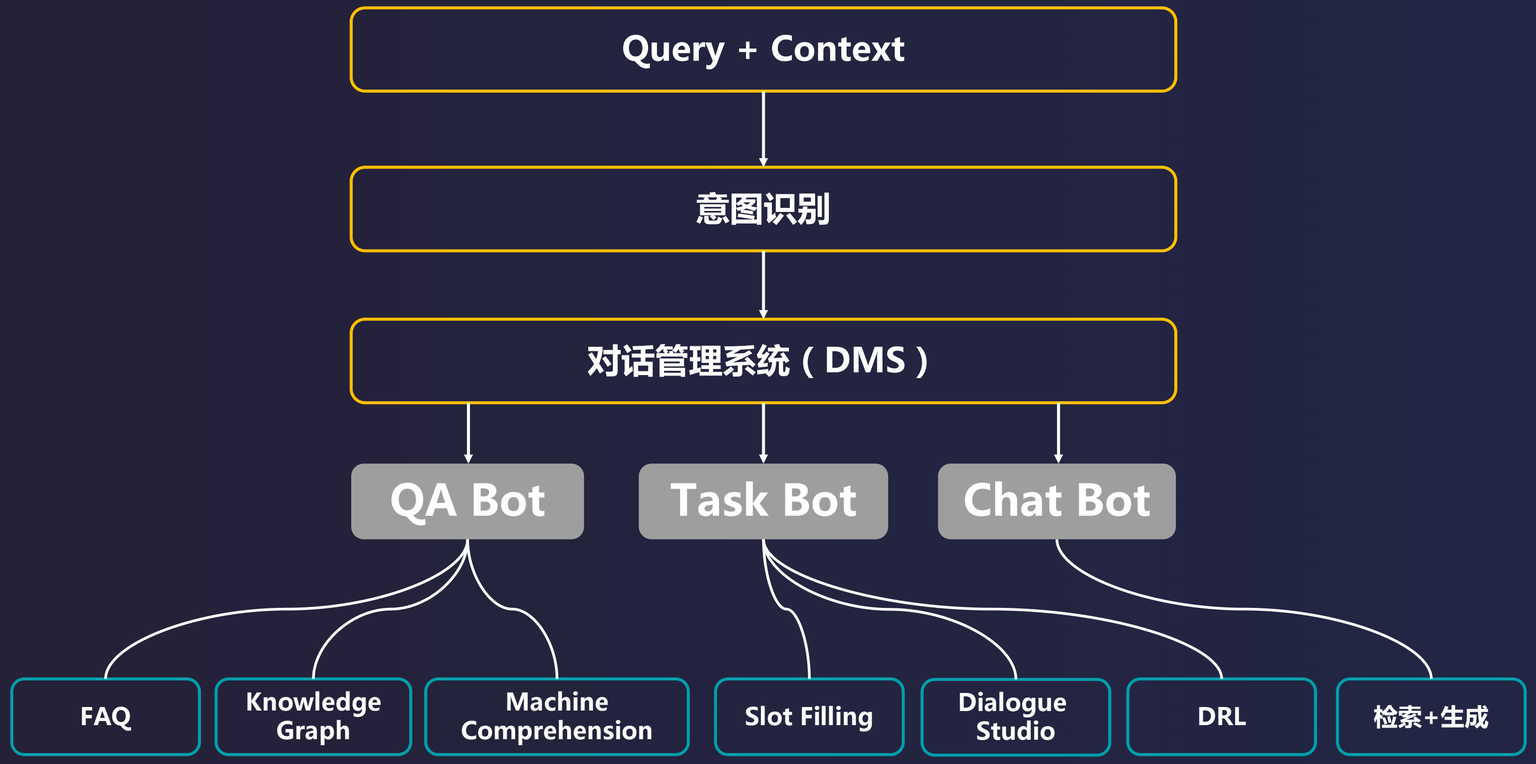
- 我双十一买的东西什么时候能收到？
- 淘宝几天无理由退款？

任务解决 Task Bot

- 我要买一张明天从北京到杭州的机票
- 帮我推荐一款笔记本电脑

闲聊 Chat Bot

- 来讲个笑话
- 你觉得我美吗
- 帮我写一首藏头诗



智能服务走出海外 — Lazada 多语言机器人

LAZADA
Effortless Shopping

6 个国家：越南，泰国，新加坡，马来西亚，印度尼西亚，菲律宾

5种语言： 英语，越南语，泰语，马来西亚语，印度尼西亚语

覆盖5.6亿消费者

一个月上线一个新语言

2017

8月 马来西亚、菲律宾上线

2018

2-3月 新加坡、印尼上线

4月 泰国、越南上线

5月 6个国家全量放量

11月 支持6个国家的第一次双十一

2019

5月 6个国家模型自动训练上线

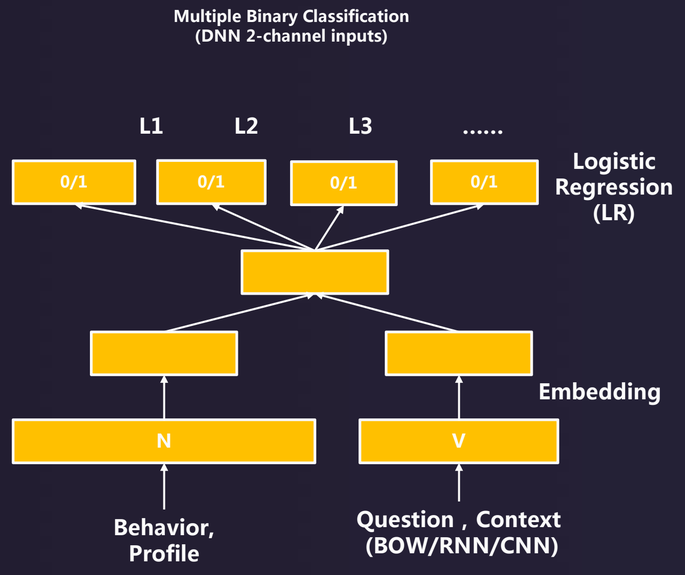
智能服务走出海外 — AliExpress 多语言机器人

AliExpress 全球速卖通

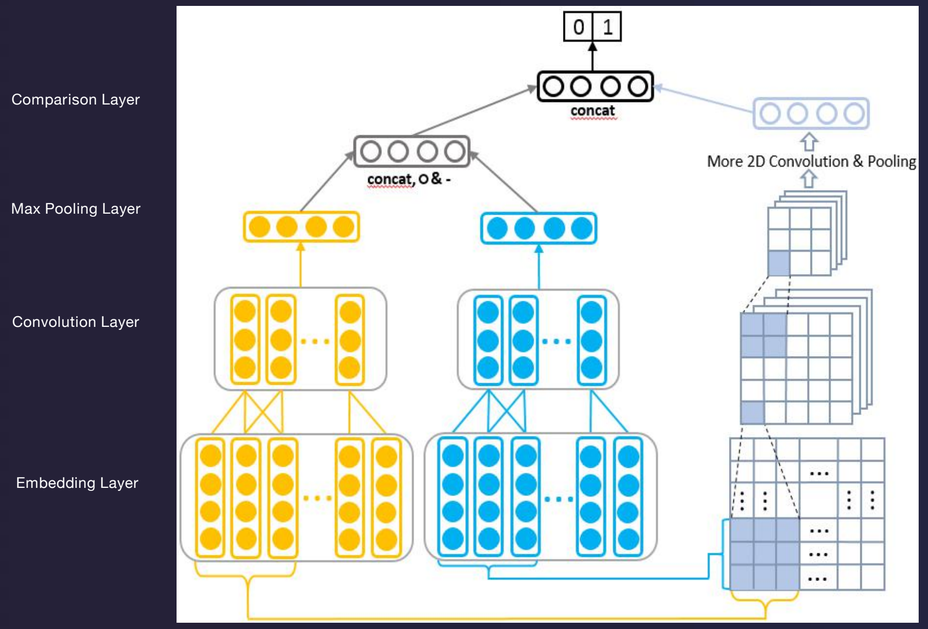
5种语言支持: 英语, 西班牙语, 俄语, 阿拉伯语, 法语

200个国家和地区

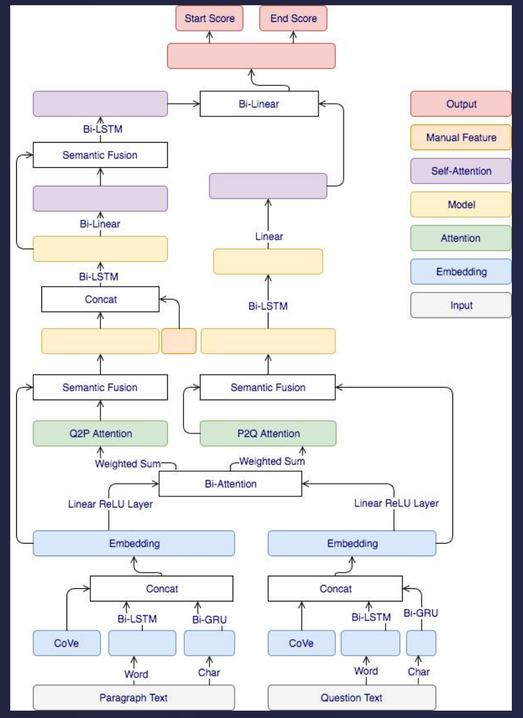
需要将中文对话机器人已有的NLP能力拓展到更多语言



意图识别



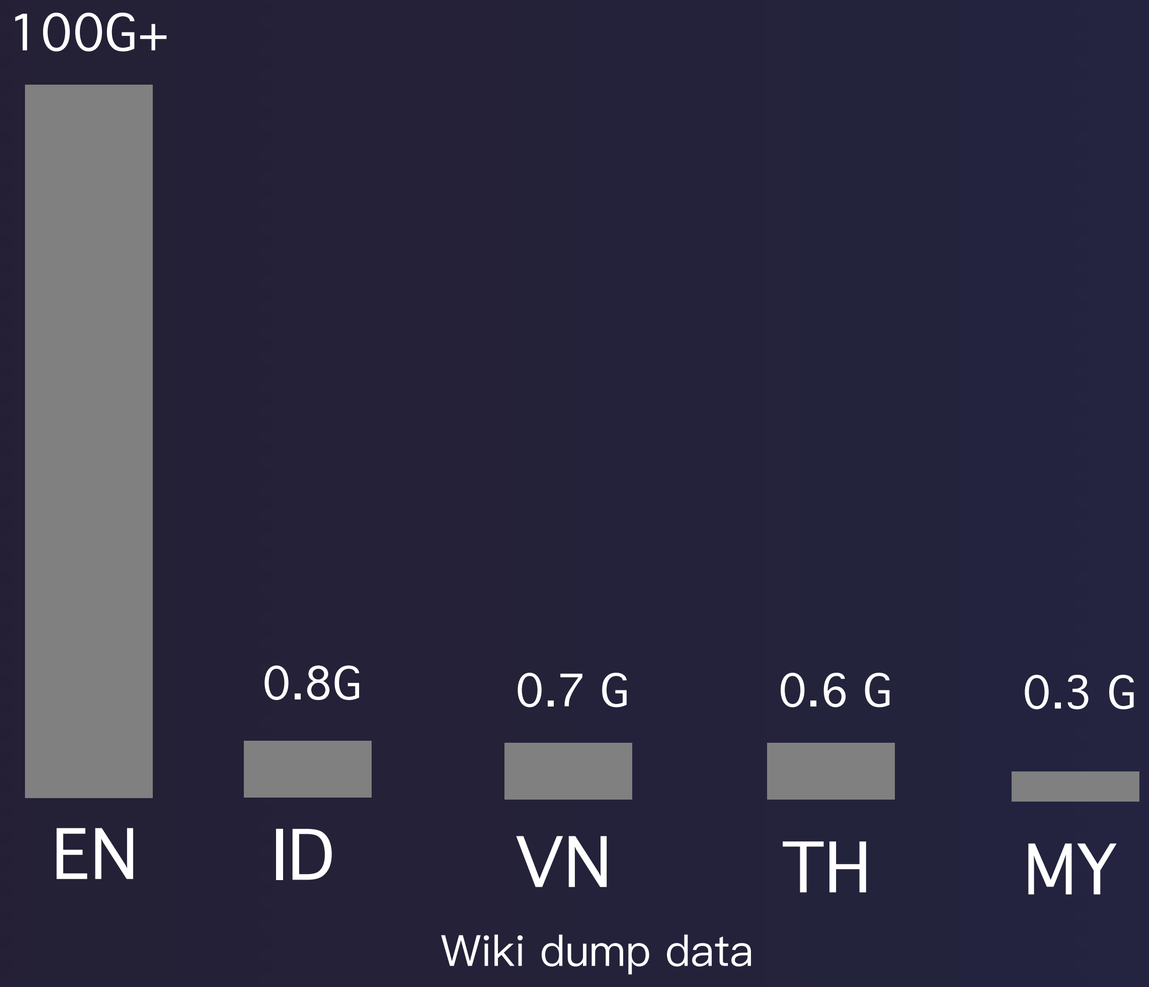
文本语义匹配



机器阅读理解

多语言机器人的挑战

1. 新语种、低资源语言导致数据不足（Low-Resource）



多语言机器人的挑战

2. 不熟悉东南亚语言

CN: **我要取消订单**
EN: I want to cancel order
ID: Saya ingin membatalkan pesanan
VN: Tôi muốn hủy đơn đặt hang
TH: ฉันต้องการยกเลิกคำสั่งซื้อ

3. 混合语言现象

อยากได้ tracking ของ order 2063279 ครับ　　（泰语+英语 混合）
en boleh check sebab apa order sy cancel（马来语+英语 混合）

多语言机器人的挑战

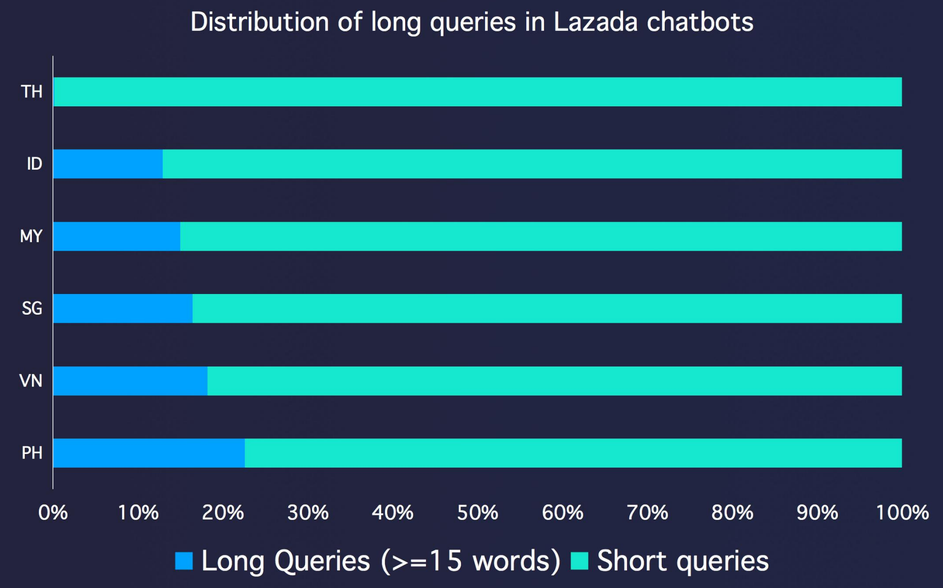
4. 不同国家、不同地区、不同语言文化

不同的打招呼方式:

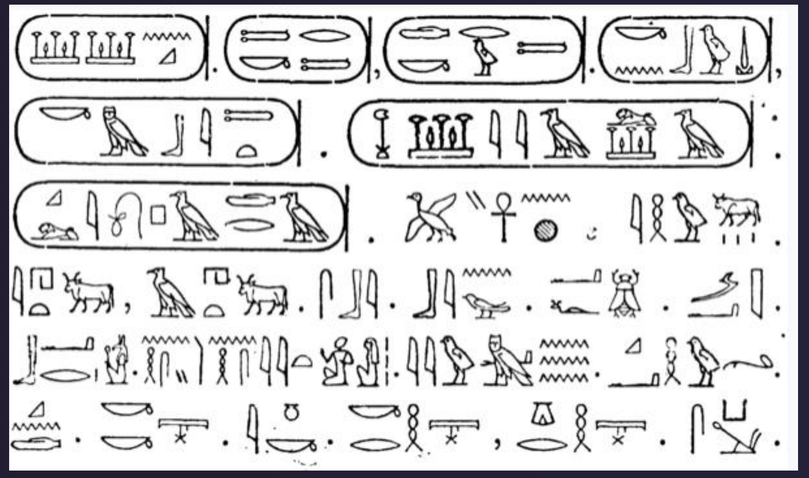
CN: 你好，在吗?
EN: Hi, Hello, How are you
ID: kk, kakak, mbk, mbak (older sister, older brother)
VN: chào anh, chào chị (older brother, older sister), 222 (hi hi hi)

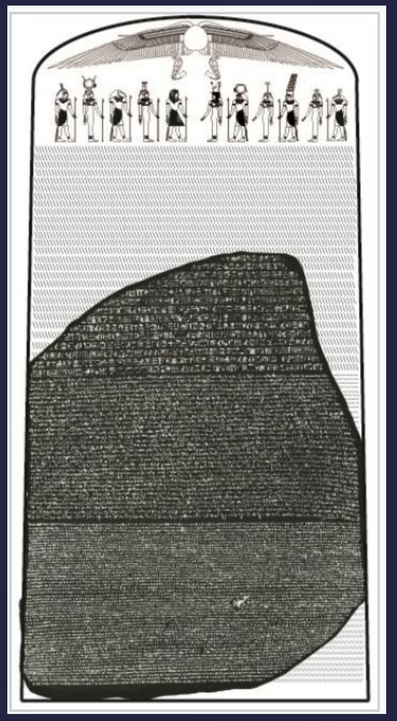
不同长度的问题描述:

中文：取消订单
英文（菲律宾）：I want to cancel my order, i got a confirmation that its been cancelled due to out of stock, then i purchase a new one, then now they will told me that previous order has been shipped. how can that be happen?



从一个历史故事看多语言场景 — 破解罗塞塔石碑





公元5世纪后古埃及文字已无人能解读，直到罗塞塔石碑在1799年被发现

上面用三种语言记述了同一件事：古埃及圣书体、世俗体和古希腊文

从一个历史故事看多语言场景 — 破解罗塞塔石碑

资源稀缺的语言

↓

↑

资源丰富的语言

罗塞塔石碑的解读

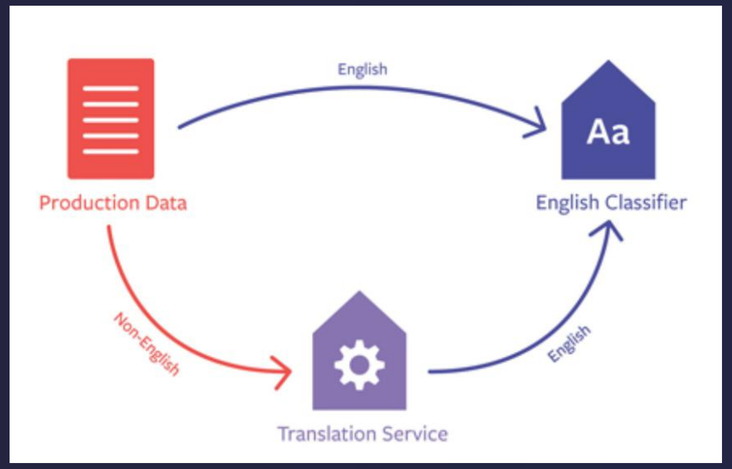
- 1830年代石碑的解读才有大的飞跃
- 法老的名字在古埃及文字中是被框起来的，容易识别
- 人们意识到古埃及文字可能是部分表音的
- 由法老名字开始得到了读音对照表，解读了更多法老名字，并推广到其他文字上

从罗塞塔石碑破解引发的思考：

- 利用资源丰富的语言帮助算法模型理解资源稀缺的语言
- 在多语言场景中，我们的“罗塞塔石碑”在哪里?
- 如何有效地利用平行文本?

机器翻译是否可行?

- 将所有的训练语料翻译成新语种语言
- 将新语种的线上Query翻译成中文或英文，用大语言的模型来预测
- 但效果**往往很差**，原因是小语种、口语化的对话翻译误差不断累积，导致最终模型训练和预测偏差较大。



举例：

越南语: dạ em không biết là mình còn theo dõi cuộc nói chuyện này không ạ
机器翻译: I don't know if I'm watching this conversation
人工翻译: I am wondering whether you are still there ?

七步构建多语言机器人

- 第一步：了解一个语言的特性并进行预处理
- 第二步：对齐不同语言的词汇
- 第三步：从词汇到句子的表示
- 第四步：理解混合语言
- 第五步：充分利用多语数据
- 第六步：本地化
- 第七步：多语言模型自动迭代

第一步：了解新语言特性并做预处理

是否连写语言

大部分语言是空格分词，少数连写语言需要特有分词器
中文，泰语都是连写语言，越南语空格区分的是单字，需要连接起来成单词

TH : ฉันต้องการยกเลิกคำสั่งซื้อ VN: Tôi muốn hủy đơn đặt hang

表音还是表意

了解语言是表音还是表意，设计不同的纠错算法
表意语言: 我的帐号 (zh)
表音语言: Tôi muốn trả lại đơn đặt hàng (vn)
表音语言更容易出现拼写错误，拼写纠错至关重要

词形是否丰富

词形变换丰富的语言会导致大词表，需要进行词型归一
整个维基统计下来，法语在40多万词，阿拉伯语是100多万词，屈折变化多。
派生: happy – happiness（后缀，派生出新的词性）
屈折: produk – produk-produk（单复数等不同的语法关系）

第一步：了解新语言特性并做预处理

输入	Bagaimana saya membelinya produk-produck
1. 分词	bagaimana | saya | membelinya | produk | produck
2. 拼写纠错	bagaimana | saya | membelinya | produk | produk
3. 词形归一	bagaimana | saya | membeli | produk
4. 去停词	bagaimana | ~~saya~~ | membeli | produk
输出	bagaimana | membeli | produk

第一步：了解新语言特性并做预处理

语言越来越多，词形归一有没有通用的方案？—— 先分词，后接BPE

BPE是什么，解决了什么：

- **BPE是通过统计方法将单词进一步地分解成subword的方法，可以拆解常见的前缀后缀等，使得能对屈折变化多的语言用较小的词表去表示**
- **BPE处理下不会有未登入词（OOV）的产生**

示例：

- 原句：comment désactiver ma carte de crédit du site aliexpress?
- 分词：comment désactiver ma carte de crédit du site aliexpress ?
- BPE：comment dés@@ activer ma carte de crédit du site ali@@ express ?（派生和组合词）

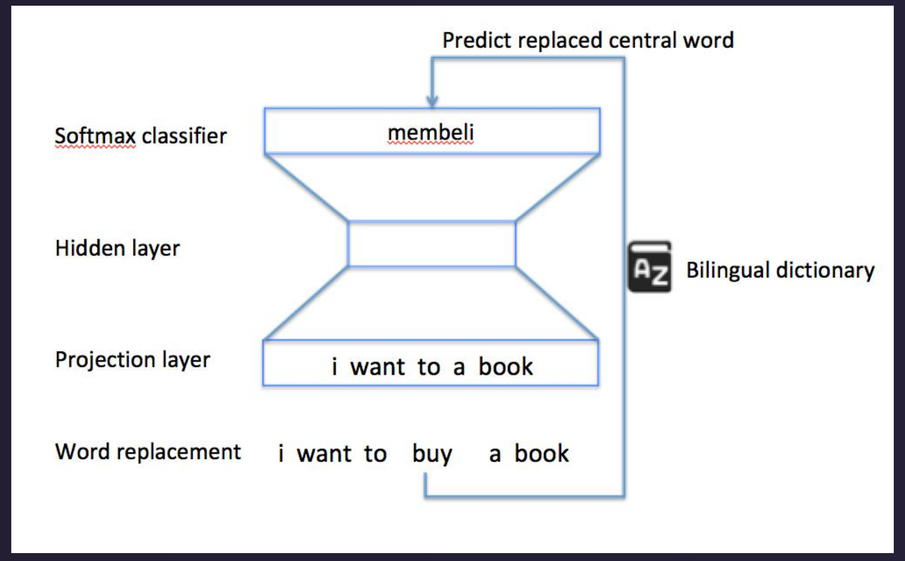
七步构建多语言机器人

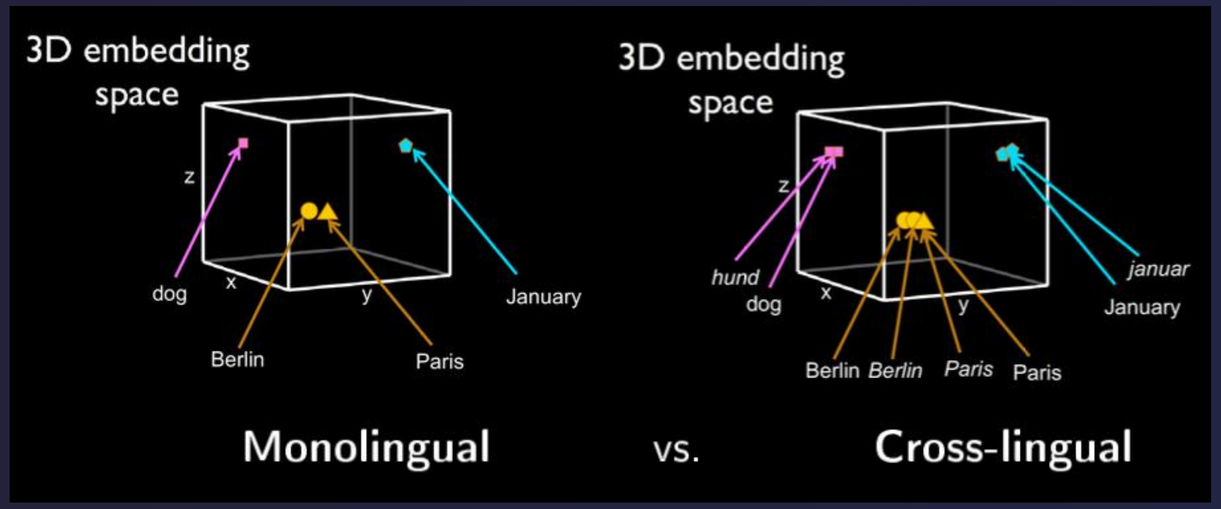
- 第一步：了解一个语言的特性并进行预处理
- 第二步：对齐不同语言的词汇
- 第三步：从词汇到句子的表示
- 第四步：理解混合语言
- 第五步：充分利用多语数据
- 第六步：本地化
- 第七步：多语言模型自动迭代

第三步：对齐不同语言的词汇

- 词向量是很多NLP应用中基本的特征
- 新语种语料少难以训练高质量词向量
- 多语言词向量使得各个语言的语料可以共享，用资源丰富语言的语料提升资源缺乏语言的词向量质量
- 可以应对混合语言的场景

词汇	同义词
id_membeli	en_purchase, en_buy, id_menjual
id_produk	en_product, id_barang-barang
id_kemana	Id_kesana, id_pulang, id_kerumah





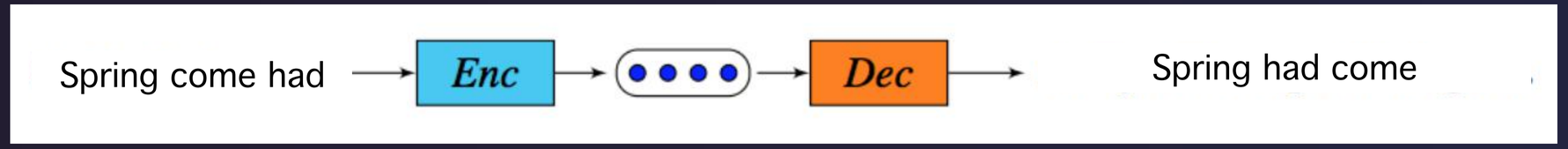
七步构建多语言机器人

- 第一步：了解一个语言的特性并进行预处理
- 第二步：对齐不同语言的词汇
- 第三步：从词汇到句子的表示
- 第四步：理解混合语言
- 第五步：充分利用多语数据
- 第六步：本地化
- 第七步：多语言模型自动迭代

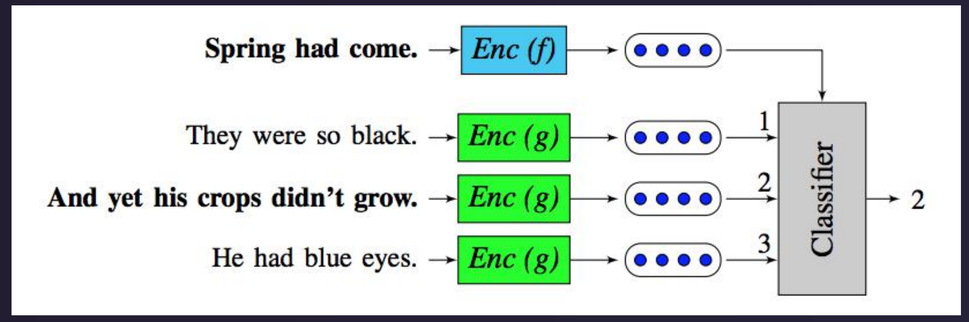
第三步：从词汇到句子的表示

新语种标注数据较少情况下的句子表示

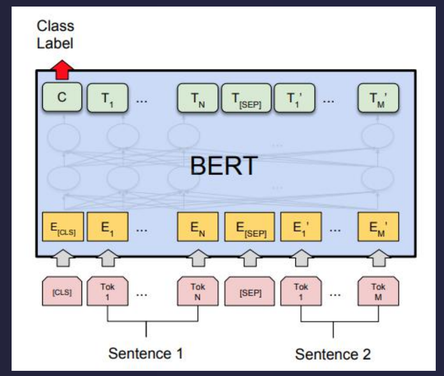
- Word Average: 词向量平均得到句子向量 （AUC 70%）
- Denoise-auto encoding: 恢复一个乱序的句子 （AUC 75%，充分利用单句、非连贯的语料）
- Quick-thought: 上一个句子预测下一个句子（三选一） （AUC 80%）
- Language Model: （AUC 90%）



Denoise-auto encoding http://www.jmlr.org/papers/volume11/vincent10a/vincent10a.pdf



Quick-thought https://arxiv.org/pdf/1803.02893.pdf



Language Model https://arxiv.org/pdf/1810.04805.pdf

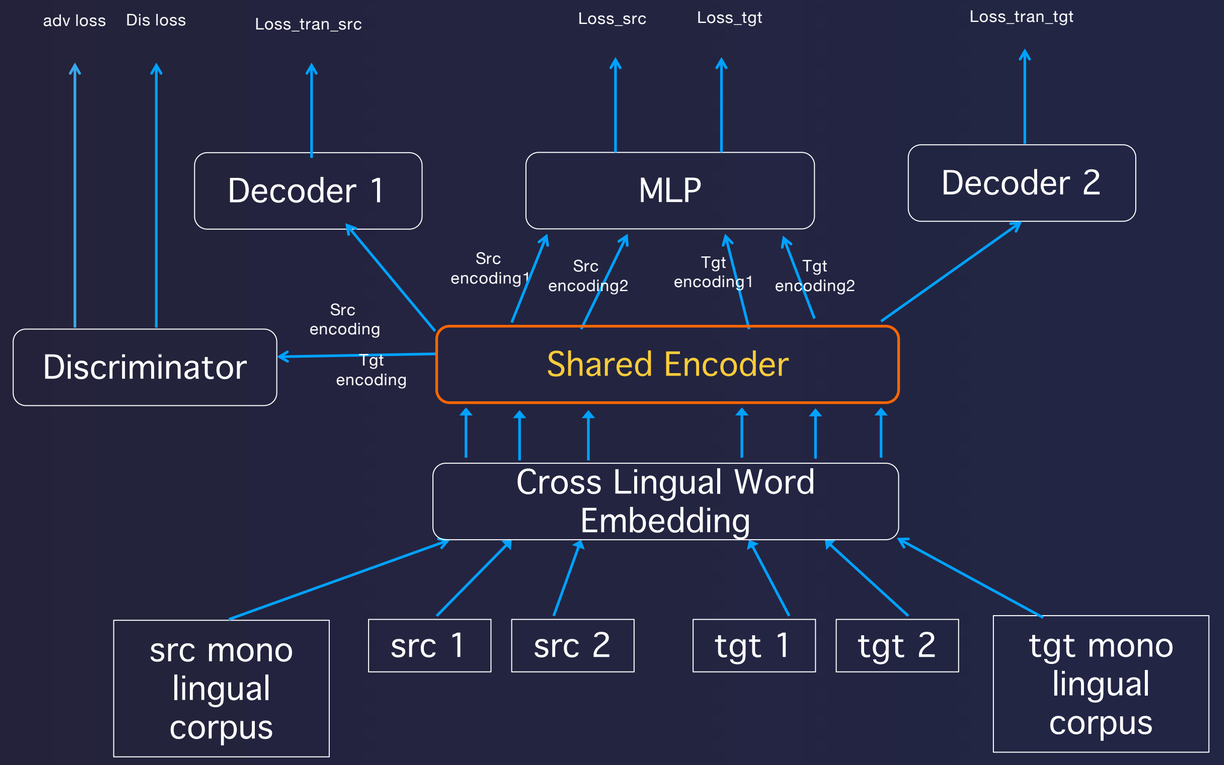
七步构建多语言机器人

- 第一步：了解一个语言的特性并进行预处理
- 第二步：对齐不同语言的词汇
- 第三步：从词汇到句子的表示
- 第四步：理解混合语言
- 第五步：充分利用多语数据
- 第六步：本地化
- 第七步：多语言模型自动迭代

第四步：理解混合语言

Shared Encoder 混合句子表示

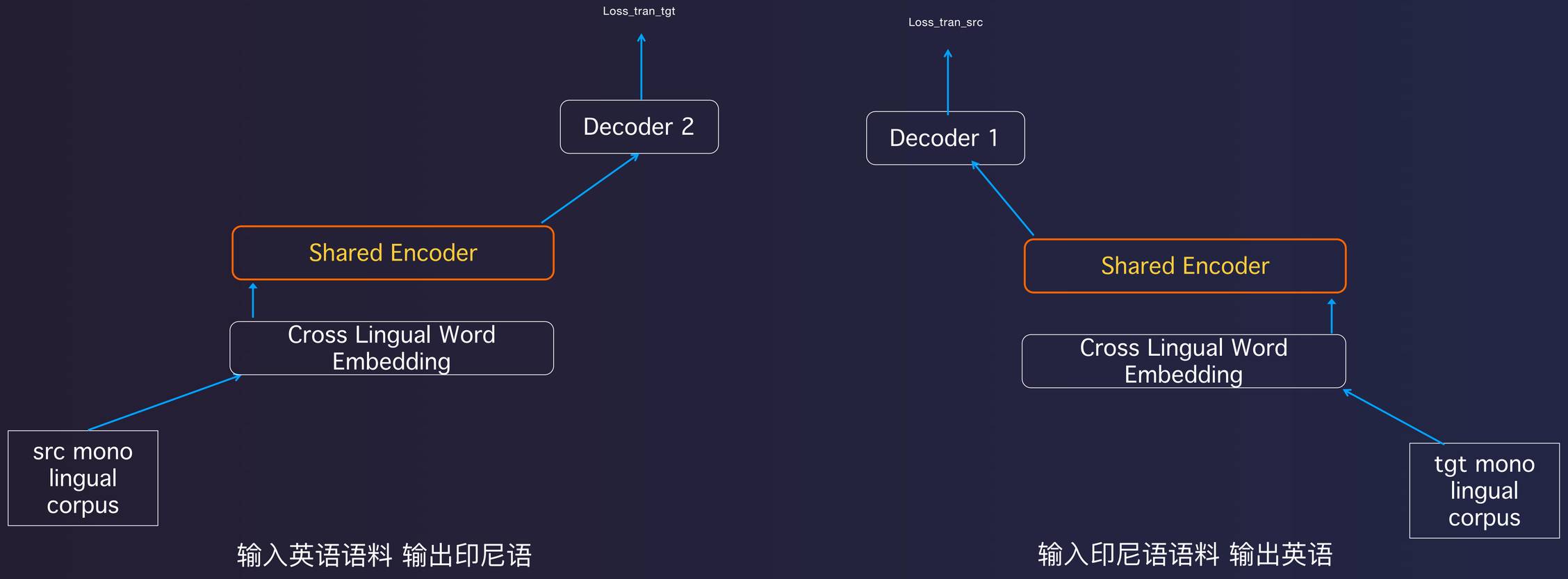
- 建立Shared Encoder，对齐跨语言句子
- 多任务联合训练翻译模型，充分利用无监督语料来使得Shared Encoder拉近不同的语言，包括单语翻译、单语Back-Translation、双语翻译等任务。
- 引入判别器，通过对抗学习，提升Shared Encoder的语言混合能力，让判别器无法区分来源语言。
- 训练英语等同于训练新语种。
- 支持混合语言理解。



以印尼语文本匹配任务为例

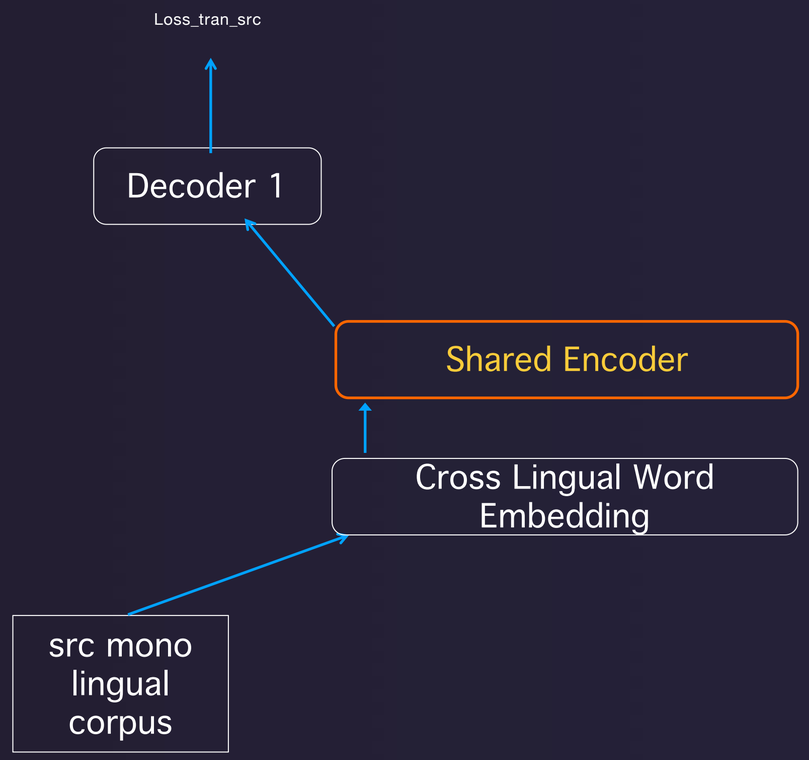
第四步：理解混合语言

任务1: 双语翻译--利用双语平行语料数据进行双向翻译训练

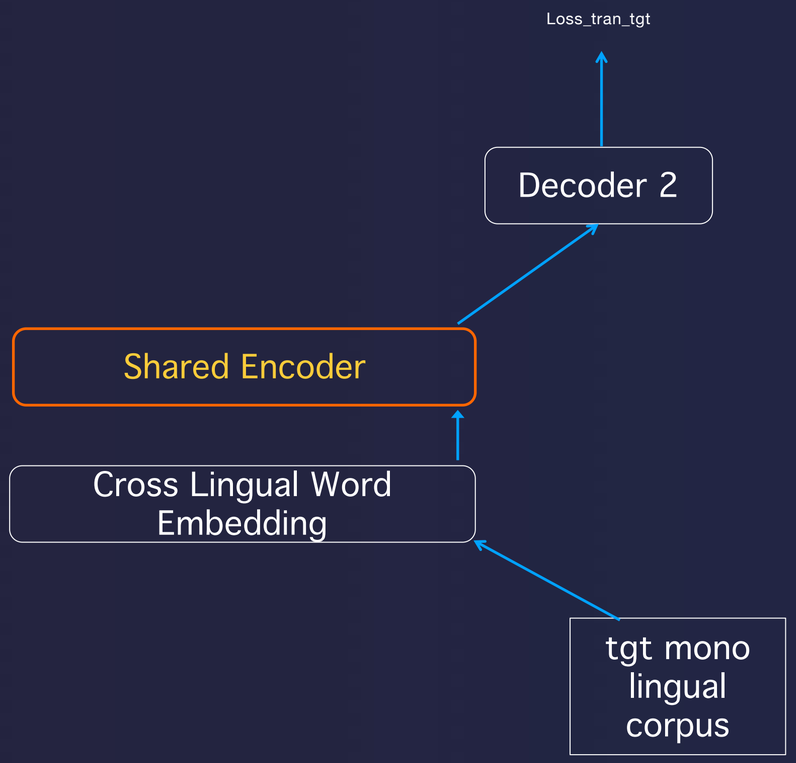


第四步：理解混合语言

任务2: 单语De-noise翻译



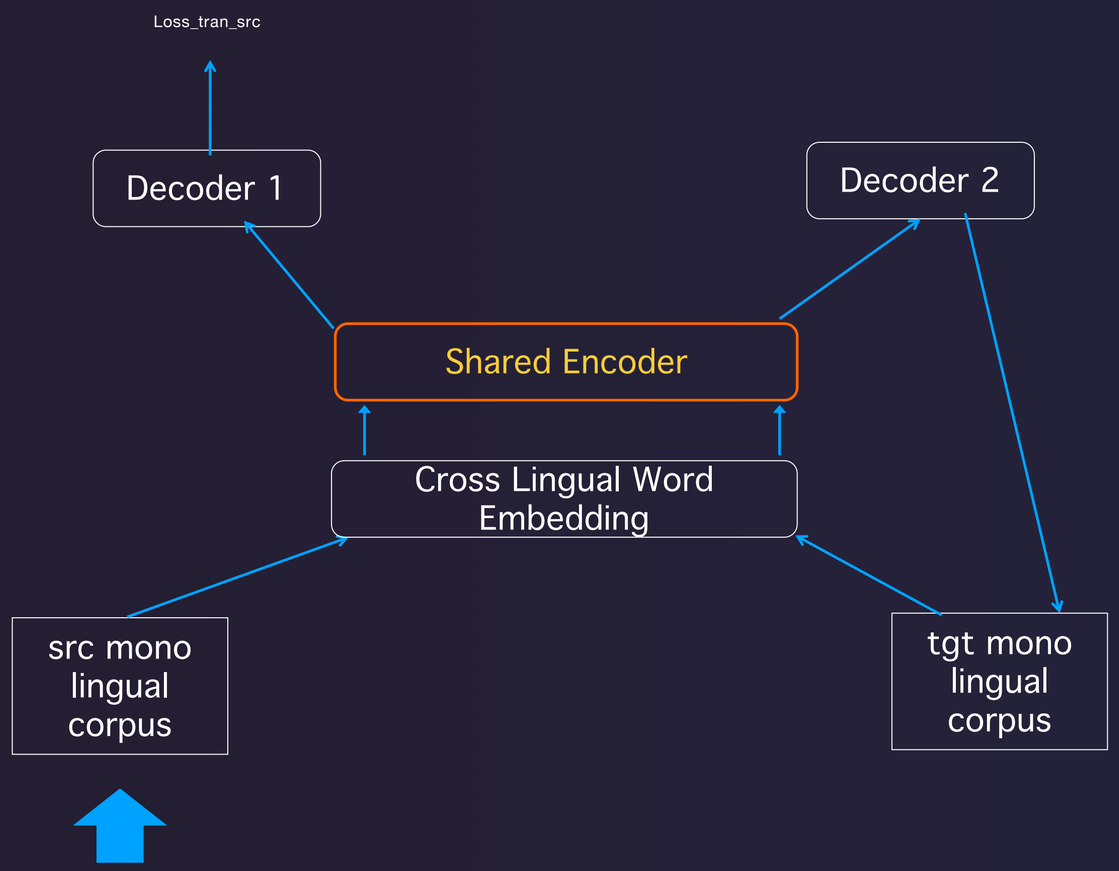
输入打乱的英文语料，输出正常的英文



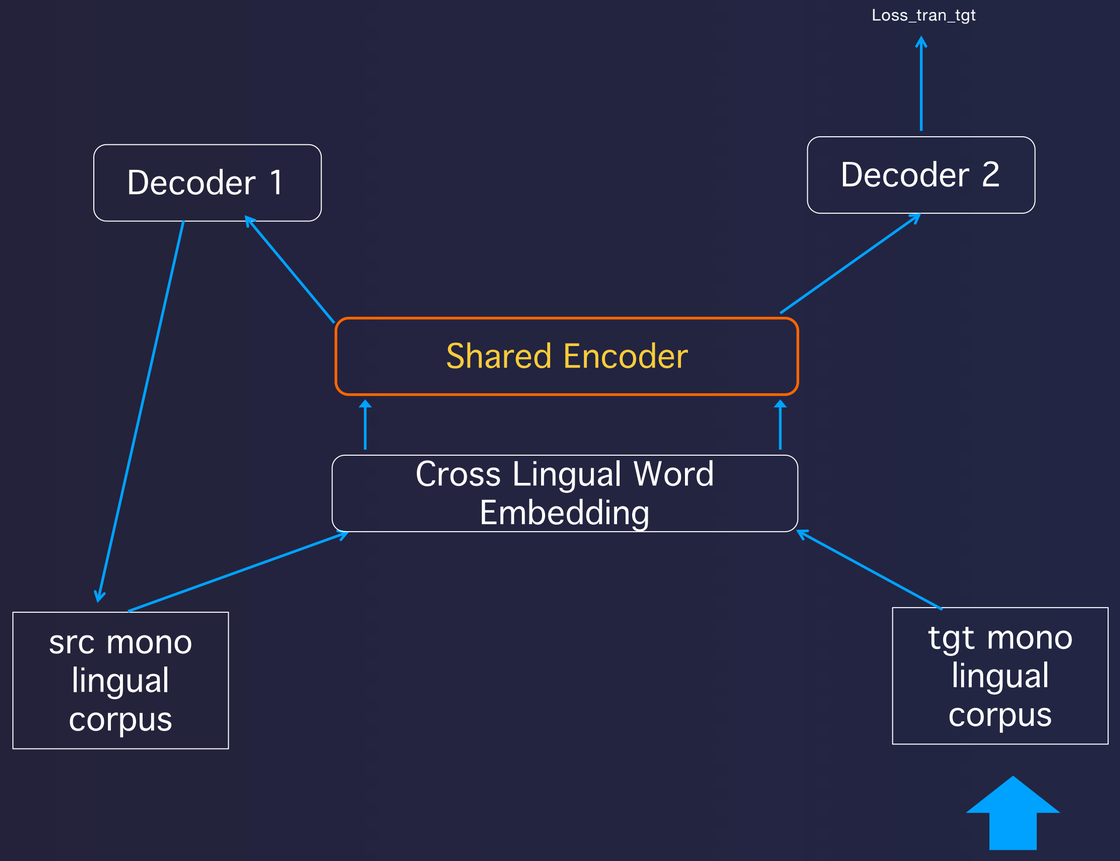
输入打乱的印尼语语料，输出正常的印尼语

第四步：理解混合语言

任务3: 单语Back Translation



输入英语语料 –> 输出印尼语–>输入印尼语–>输出英语

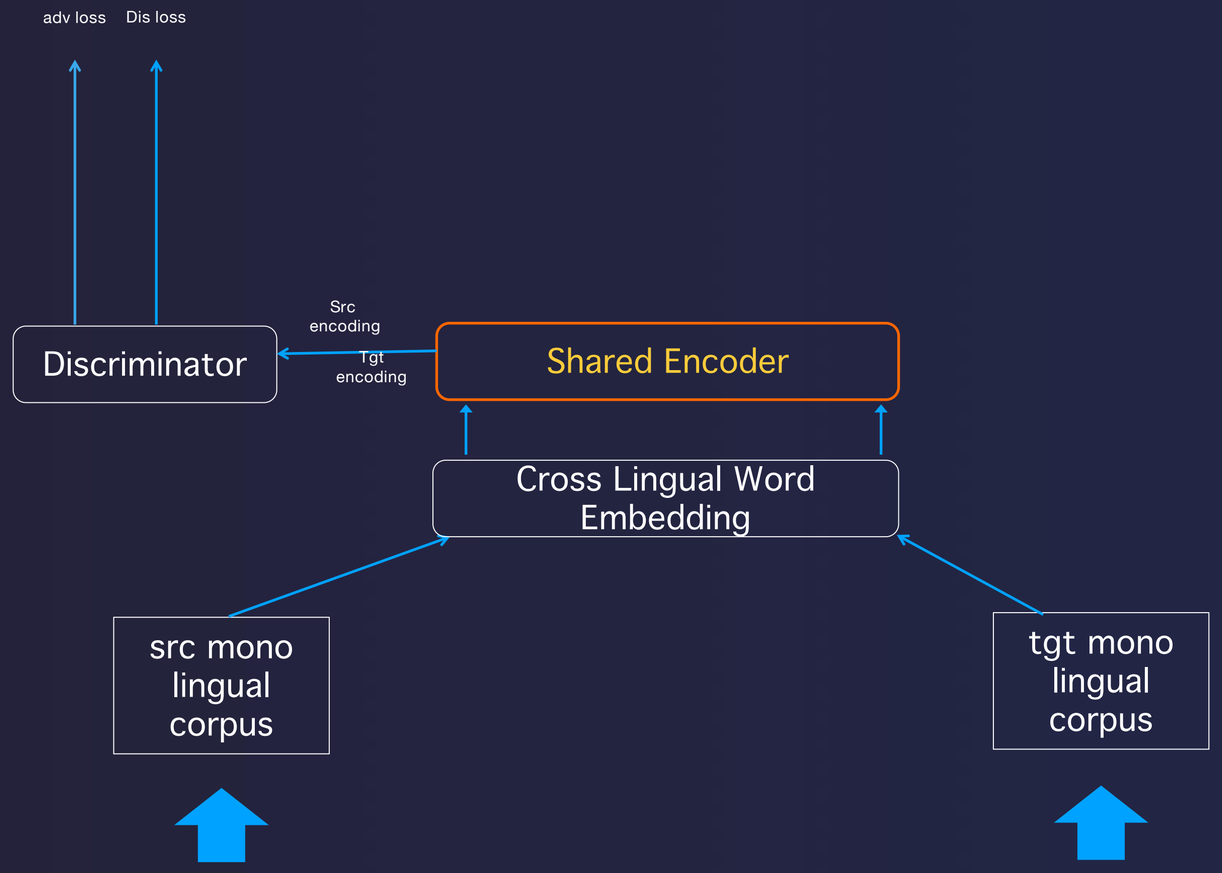


输入印尼语语料 –> 输出英语–>输入英语–>输出印尼语

第四步：理解混合语言

任务4: 对抗训练

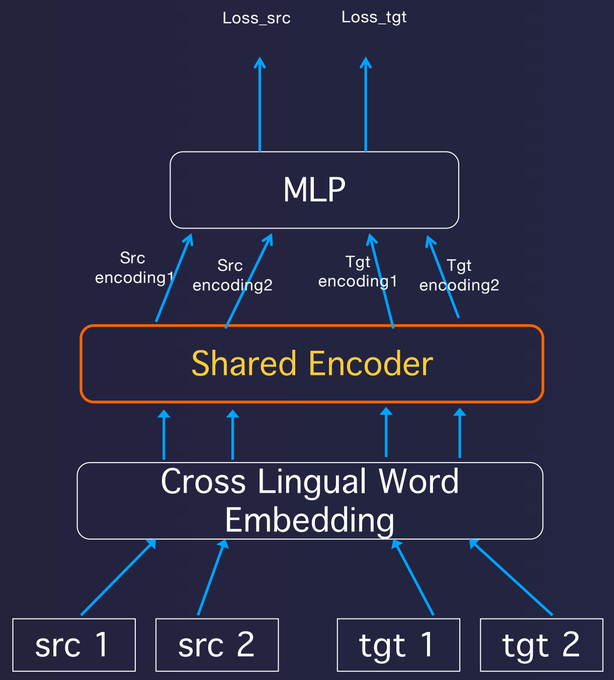
- Discriminator是一个区分来源语言的判别器，期望能区分英语为0，印尼语为1。
- 首先Fix住Discriminator，训练Share Encoder，期望输入英语和印尼语时，让判别去无法区分，输出都是0.5，实现语言无关。
- 然后Fix住Share Encoder，训练判别器，英语是0，印尼语是0.5.
- 交替训练Shared Encoder和Discriminator实现对抗。



第四步：理解混合语言

任务5: 文本相似度任务调优

- 使用两种语言的相似度标注数据，混合在一起
- Fix Shared Encoder, 仅训练上层MLP



七步构建多语言机器人

- 第一步：了解一个语言的特性并进行预处理
- 第二步：对齐不同语言的词汇
- 第三步：从词汇到句子的表示
- 第四步：理解混合语言
- 第五步：充分利用多语数据
- 第六步：本地化
- 第七步：多语言模型自动迭代

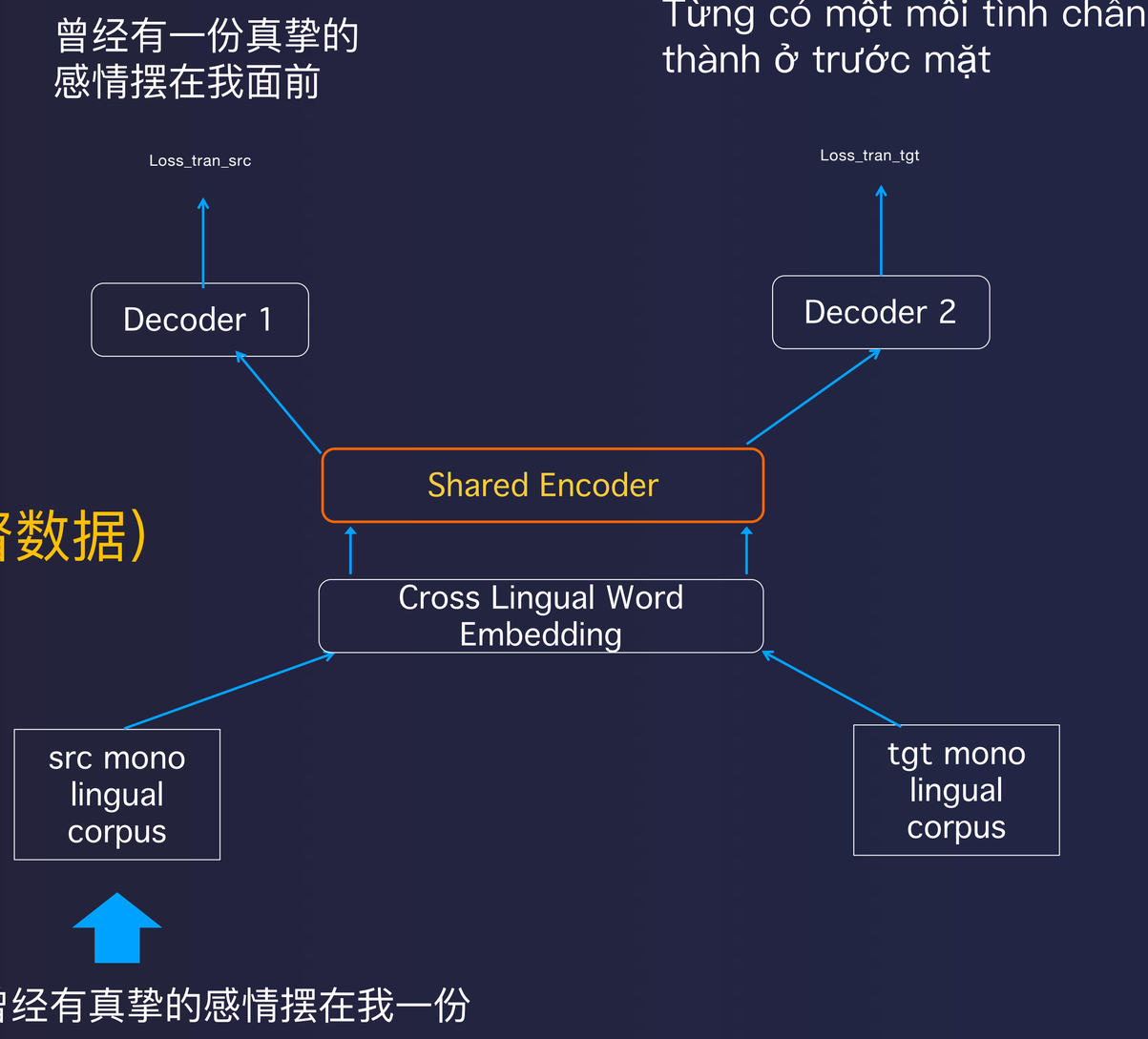
第五步：充分利用多语数据

电影字幕 – 高质量的平行语料数据

zh: 曾经有一份真挚的感情摆在我面前
id: Sudah temukan Cinta namun Kucampakan
vn: Từng có một mối tình chân thành ở trước mặt

Wikipedia –表述正式的数据来源（单语无监督数据）

English:15GB
Indonesian: 800MB
Vietnamese: 600MB
Thai: 700MB

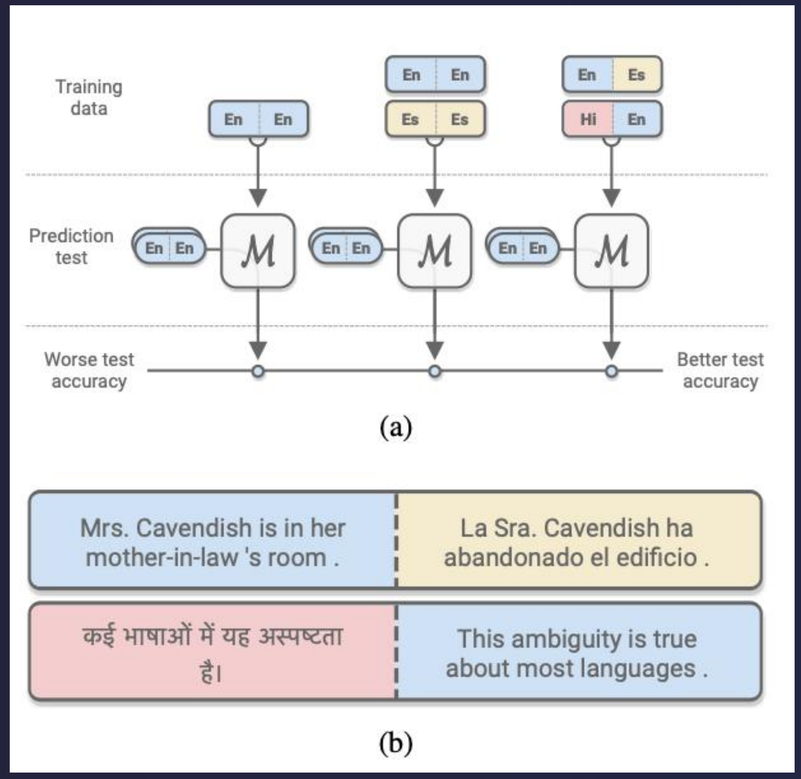


第五步：充分利用多语数据

通过 Back Translation 实现数据增广

原句: 曾经有一份真挚的感情摆在我面前
zh to en to zh:在我面前有一种真诚的感觉
zh to th to zh:在我面前有一种真正的爱
zh to fr to zh:在我面前表现出真诚的爱

多语言数据增广XLDA



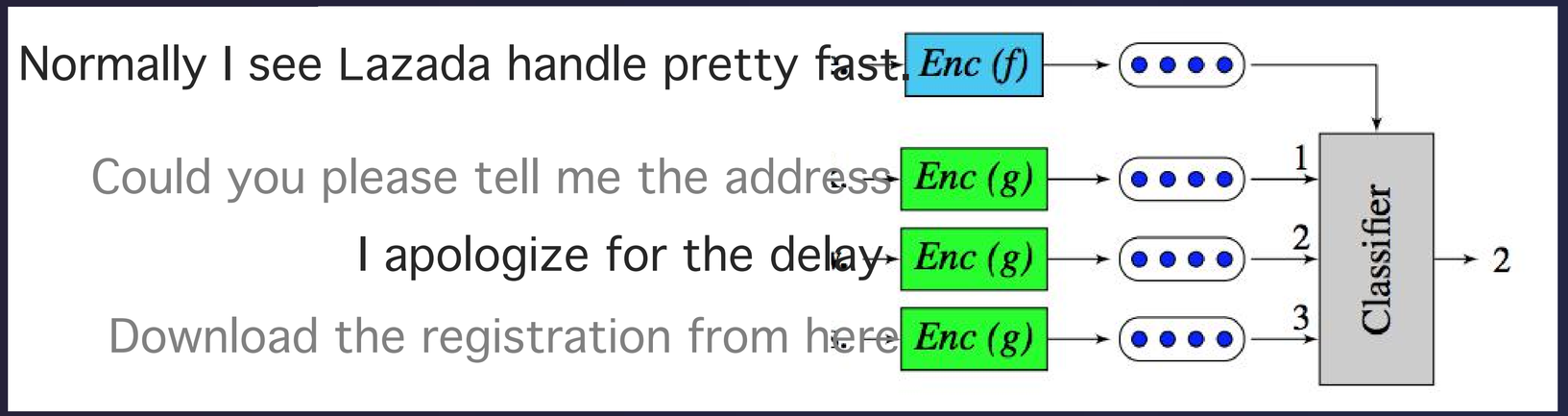
https://arxiv.org/pdf/1905.11471v1.pdf

第五步：充分利用多语数据

历史客服对话日志 — 领域相关的知识

Language	EN	TH	ID	VN
Chatlog	1GB	600MB	900MB	600MB

对话记录示例：
Client: Normally I see Lazada handle pretty fast.
Agent: I apologize for the delay
Agent: Due to the promotion activity, large of orders are waiting to be delivered



七步构建多语言机器人

- 第一步：了解一个语言的特性并进行预处理
- 第二步：对齐不同语言的词汇
- 第三步：从词汇到句子的表示
- 第四步：理解混合语言
- 第五步：充分利用多语数据
- 第六步：本地化
- 第七步：多语言模型自动迭代

第六步：本地化

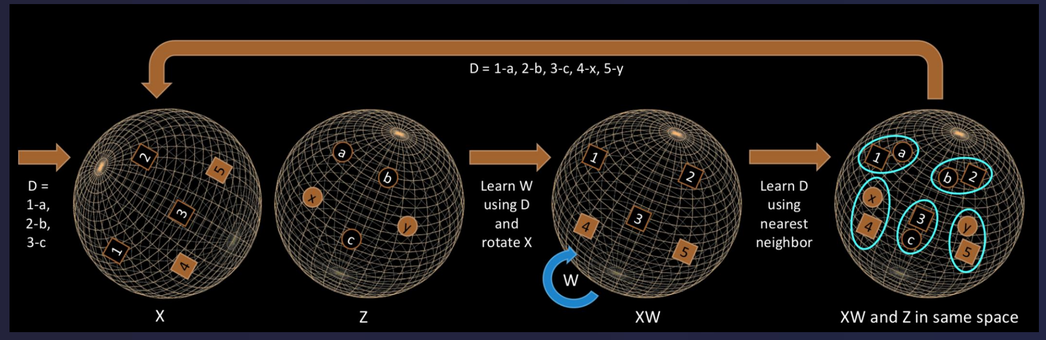
- **部分国家的用户习惯使用长文本Query，可以通过句子改写进行缩写，适配到当地语言习惯**

- **部分语言表达较长，用户容易拆分输入，需要通过上下文进行拼接**

  VN: sao mai minh van chua nhan lai dc ma tien dien tu tu don hang nay
  EN: Why haven’t I received the e-voucher from this order
  ZH: 为什么我还没有收到订单优惠券

- **惯用说法的挖掘（如不同地区的中文、英文等）**

简体	打印机	内存	算法	起床出门找吃的
香港繁体	打印機	内存	算法	起身出去搵野食
台湾繁体	印表機	記憶體	演算法	起床出門找吃的



通过向量空间映射，发现更多对应表达

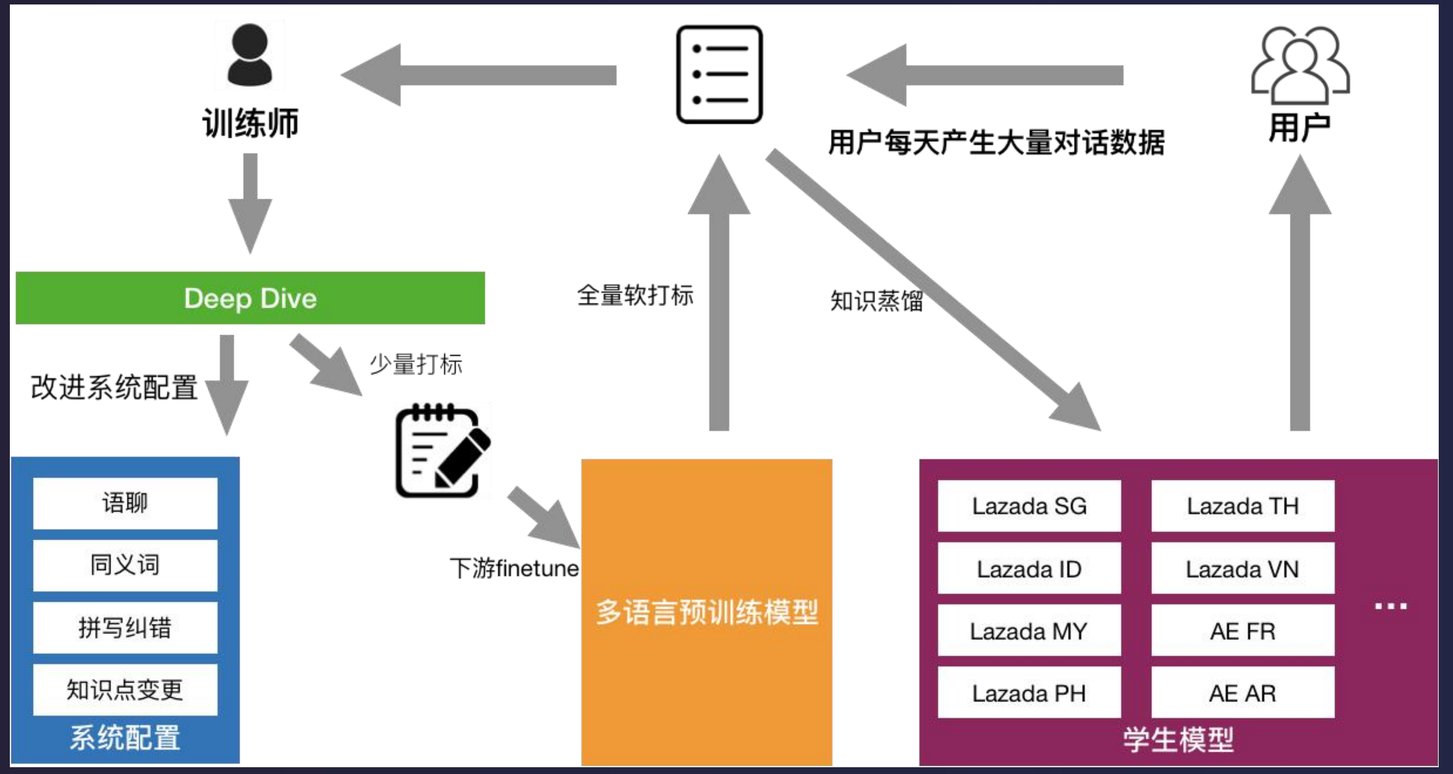
[Artetxe et al., ACL 2017]

七步构建多语言机器人

- 第一步：了解一个语言的特性并进行预处理
- 第二步：对齐不同语言的词汇
- 第三步：从词汇到句子的表示
- 第四步：理解混合语言
- 第五步：充分利用多语数据
- 第六步：本地化
- 第七步：多语言模型自动迭代

第七步：多语言模型自动迭代

- 大量语言上线服务，带来新的挑战：线上模型数量爆炸，难以维护和持续提升
- 对于算法工程师
  - 引入跨语言语言模型+知识蒸馏的Teacher–Student模式，算法工程师仅需要维护Teacher模型。
- 对于标注训练师
  - 建立人机协同的主动学习方式，以最小的人工维护量来保持模型效果的持续增长。
  - 每条标注数据都可被所有语言共享。



总结

- 把多种语言混合在一起，让资源积累充足的语言帮助提升新语种模型效果。
- 多语言场景的复杂性不仅仅在于支持语言的数量，还有各个语言的特性和文化背景
- 利用大数据去驱动学习语言的特征，比如词表、词频、词向量、语言模型，减少对人为特征的依赖；因此对有限的语料进行充分利用，决定了对一个语言的理解能力
- 基于多语言语言模型+知识蒸馏技术，建立人机协同的多语言模型自动迭代机制，通过小成本维护带来持续的效果提升。

极客时间全部课程任学

喊老板来买单！

- 精选 13+ 热门职位的学习路径，包括架构、运维、前端工程师等
- 根据不同技术岗位能力模型匹配合适的课程
- 一键设置购买条件，成员按需选课，自主制定学习计划
- 享充值满赠优惠，帮老板省钱，团队免费学习

立即申请

THANKS

Geekbang. 极客邦科技 InfoQ