Cross-Lingual Overview

Linzhi Wu

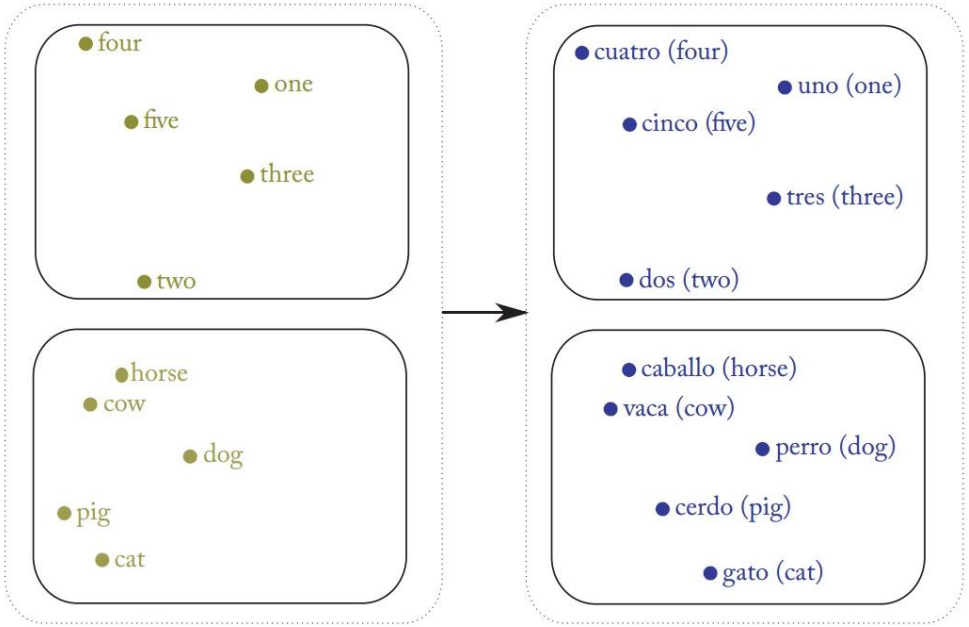
2020 / 9 / 21

跨语言词向量

- 词映射(对齐): 通过简单地线性变换来完成源语言向量空间到目标语言向量空间的转换; 目标在于学习一个从源语言到目标语言的线性变换矩阵$W^{s\to t}$

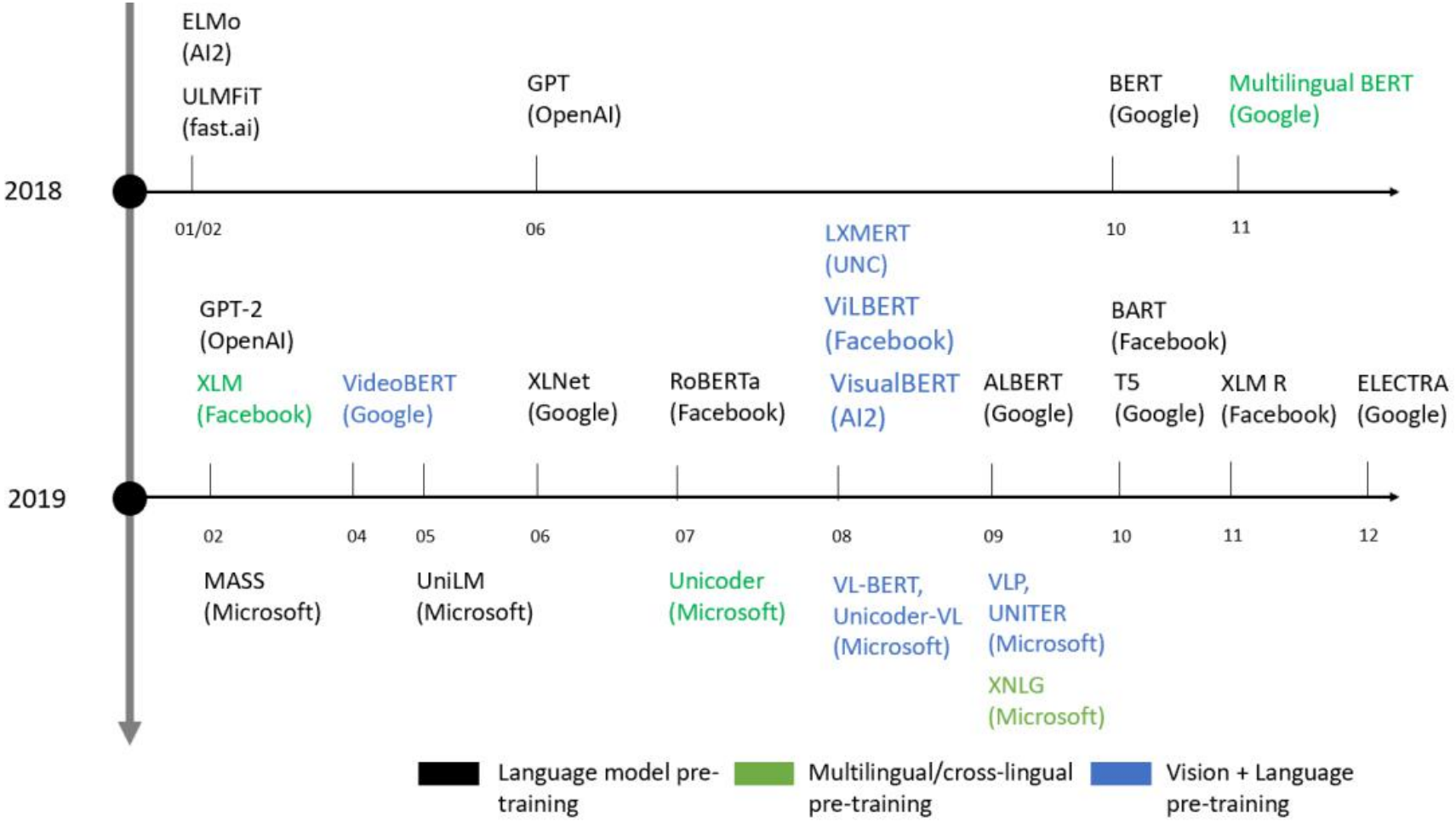
$$W^* = \arg\min_{W \in O_d(\mathbf{R})} \left\| WX^s - X^t \right\|_{\mathrm{F}} = VU^{\mathrm{T}},\quad U\Sigma V^{\mathrm{T}} = \mathrm{SVD}(X^{t\mathrm{T}} X^s)$$

X^s为源语言向量，X^t为目标语言向量，$O_d(\mathbf{R})$为d维正交矩阵集



互为翻译的词的向量表示(相同含义)在向量空间中有着相似的几何排列!

预训练模型发展趋势


2018
2019
ELMo (AI2)
ULMFiT (fast.ai)
GPT (OpenAI)
BERT (Google)
Multilingual BERT (Google)
01/02
06
10
11
LXMERT (UNC)
ViLBERT (Facebook)
BART (Facebook)
GPT-2 (OpenAI)
XLM (Facebook)
VideoBERT (Google)
XLNet (Google)
RoBERTa (Facebook)
VisualBERT (AI2)
ALBERT (Google)
T5 (Google)
XLM R (Facebook)
ELECTRA (Google)
02
04
05
06
07
08
09
10
11
12
MASS (Microsoft)
UniLM (Microsoft)
Unicoder (Microsoft)
VL-BERT, Unicoder-VL (Microsoft)
VLP, UNITER (Microsoft)
XNLG (Microsoft)
Language model pre-training
Multilingual/cross-lingual pre-training
Vision + Language pre-training

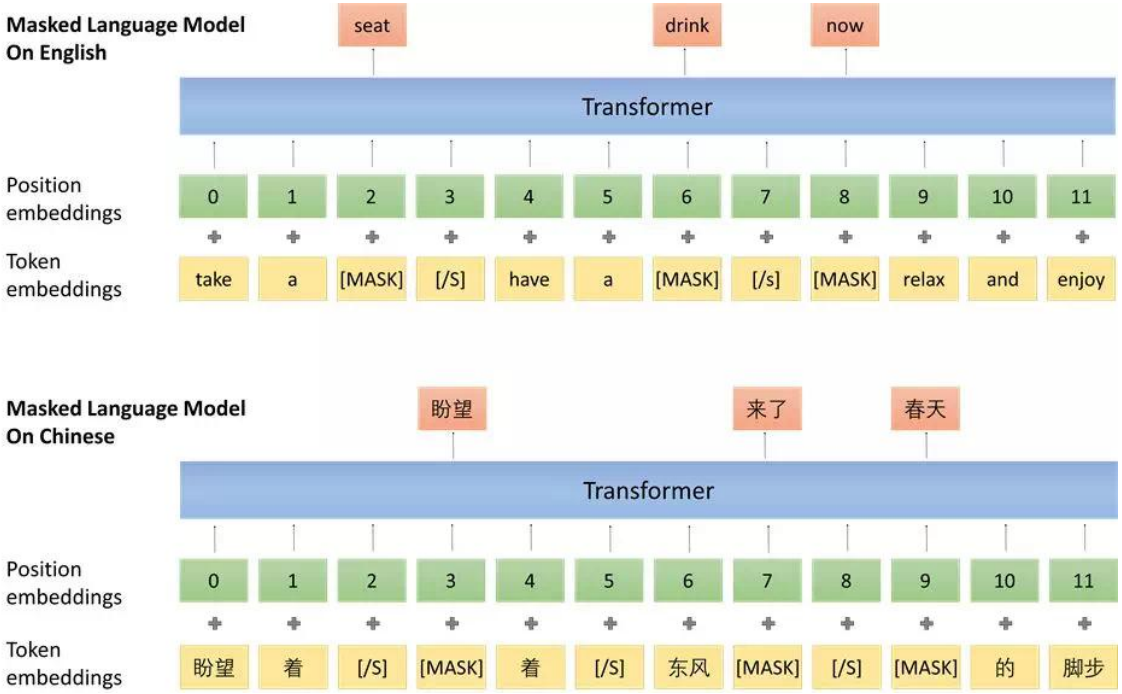
跨语言预训练

- 预训练模型能够缓解跨任务或跨语言中出现的低资源问题
- 很多NLP任务往往只在少数语言(如英文)上存在足够的标注数据，而在其他语言上并没有或仅有少量的标注数据
- 目标是利用特定任务在某种语言的标注数据上训练模型，并将学到的知识迁移到其他语言上去

- 给定多种语言的单语语料和不同语言对之间的双语语料，跨语言预训练模型能学习不同语言之间的对应关系，并保证不同语言的向量表示存在于同一个语义空间
- 该类模型使用某种语言上充足的标注数据进行下游任务微调，由此产生的任务模型能够直接作用于其他语言的输入
- 如果该任务在其他语言上同样存在少量的标注数据，则可以通过继续微调获得更好的效果

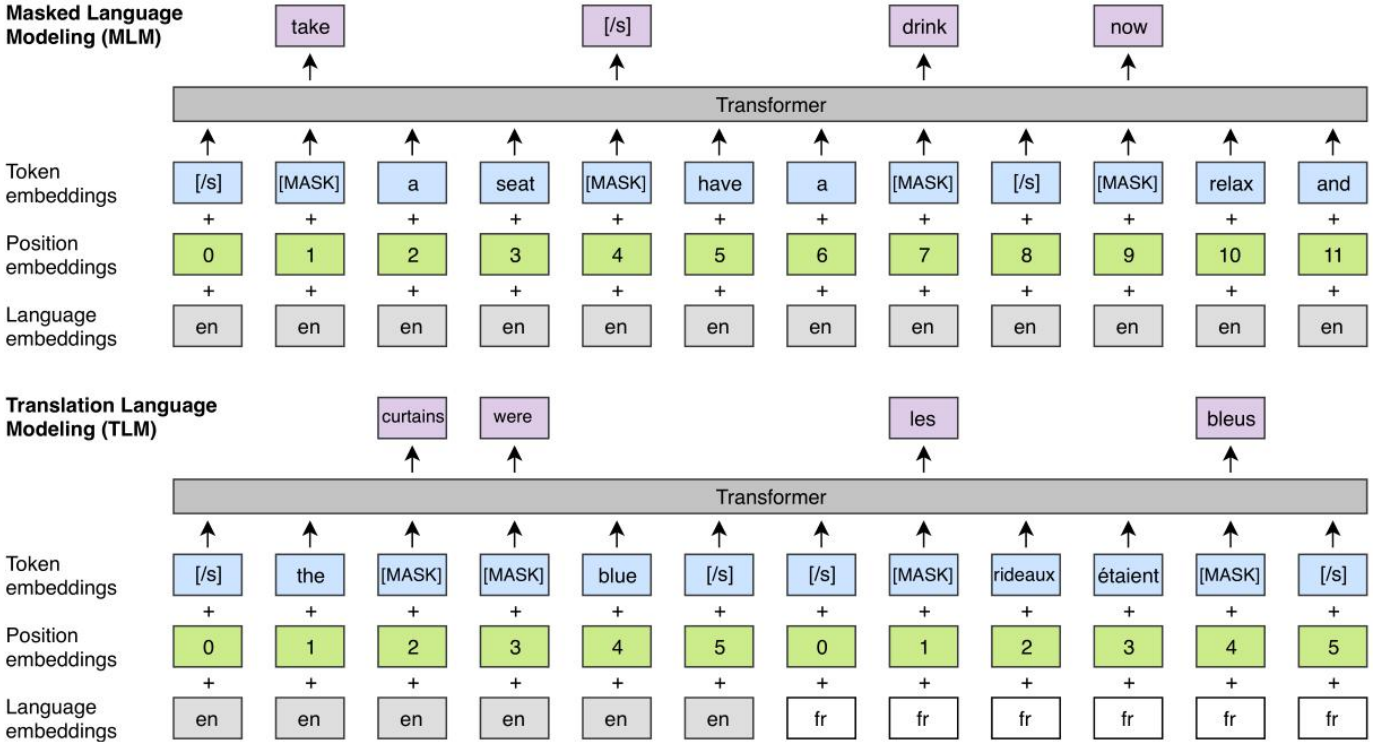
M-BERT

- 核心思想：用相同的模型和权重来处理所有的目标语言！
- 直接使用104种语言的维基单语数据以多任务方式交替训练
  - 语言之间共享相同的WordPiece (共同的词汇表)
  - 很多语言混杂在一起 (Code-switching)
- 问题
  - 对于语言顺序(主谓宾或形容词名词)不同的语言，效果不是很好
  - 准确率不如单语BERT
  - 不适用远距离的语言对



XLM

- 将互为翻译的句子对作为BERT的输入
- 随机Mask掉句子对中的双语词 (TLM)
- 可以较好地学到两种语言之间的对应关系 (天然的对齐)
- 问题
  - 1. 依赖大规模双语语料库　2. 需要大规模计算资源



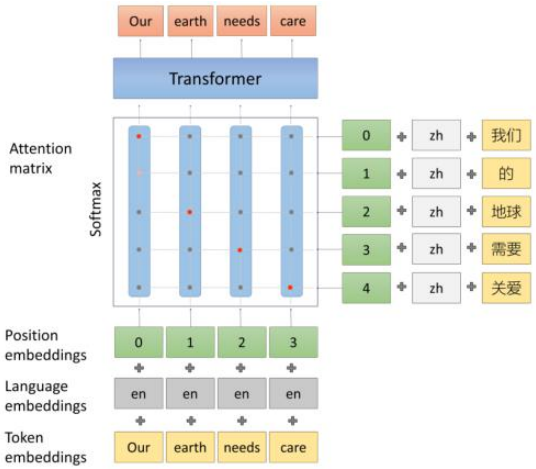
Unicoder

- 更多跨语言预训练任务

(a) 跨语言的词语恢复

(b) 跨语言的同义句子分类

(c) 跨语言的MLM



(a) Cross-lingual word recovery

(b) Cross-lingual paraphrase classification

(c) Cross-lingual masked language model

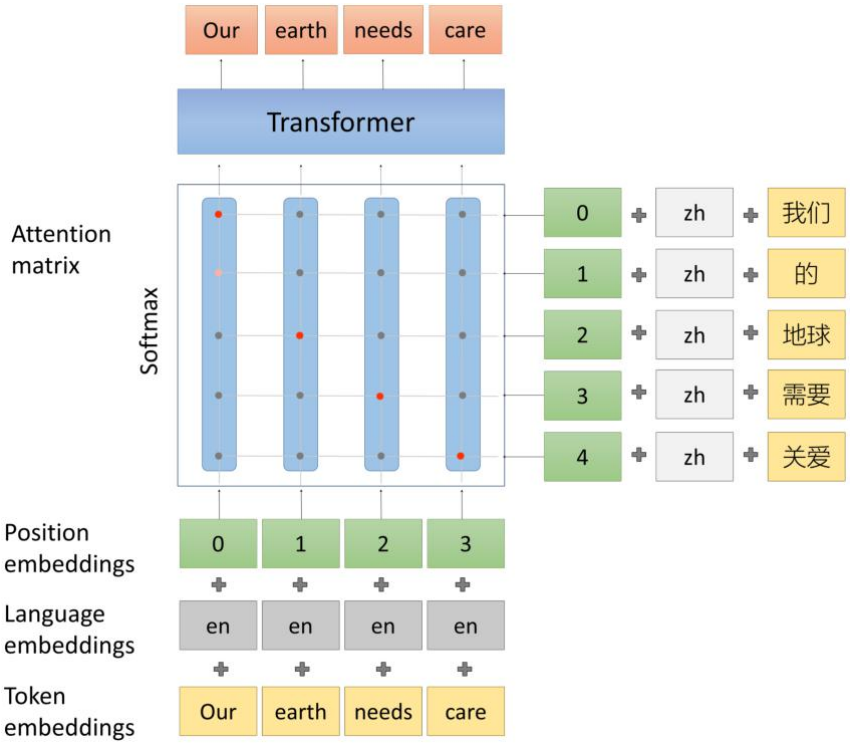
Unicoder

- 跨语言的词语恢复
  - 输入是一个同义的双语句对；模型首先用英文词的表示对中文所有词做一次attention，得到一个隐变量空间的表示，该表示是中文词表示的加权和, 然后将得到的表示输入Transformer, 尝试预测出原始的英文词序列。通过该结构，模型在不引入词对齐工具的情况下，学习中-英词之间的对应关系
  - 类似于翻译语言模型，该任务旨在让预训练模型学习两种语言之间潜在的词对齐

$$en: X = (x_1, x_2, \ldots, x_m), zh: Y = (y_1, y_2, \ldots, y_n)$$

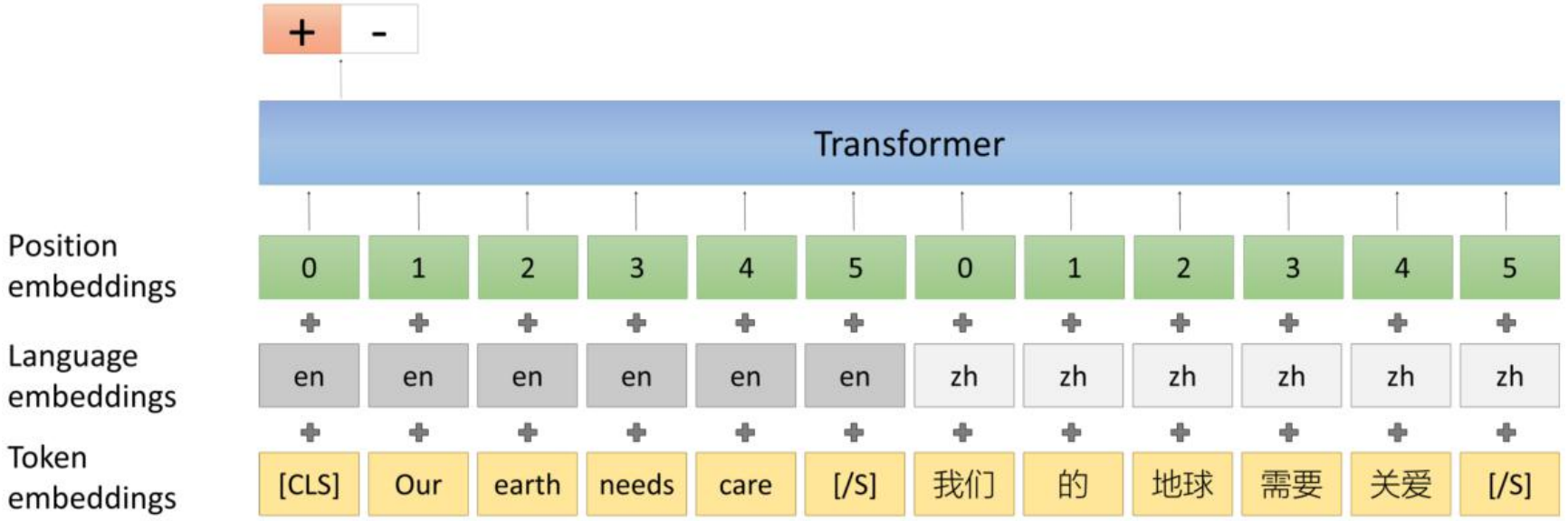
$$x_i^t = \sum_{j=1}^{n} \text{softmax}(A_{ij}) y_j^t, A_{ij} = W[x_i^s, y_j^t, x_i^s \otimes y_j^t]$$

$$X^t = (x_1^t, x_2^t, \ldots, x_m^t) \xrightarrow[\text{(Recovery)}]{Transforme\ r} X$$



Unicoder

- 跨语言的同义句子分类
  - 输入是两个不同语言的句子；目标是判定这两个句子意思是否相同。模型将两个句子拼接作为输入, 用一个词对应的表示训练一个二分类器。通过这个基础的句子级别任务，模型可以学习两个语言在句子层面的对应关系



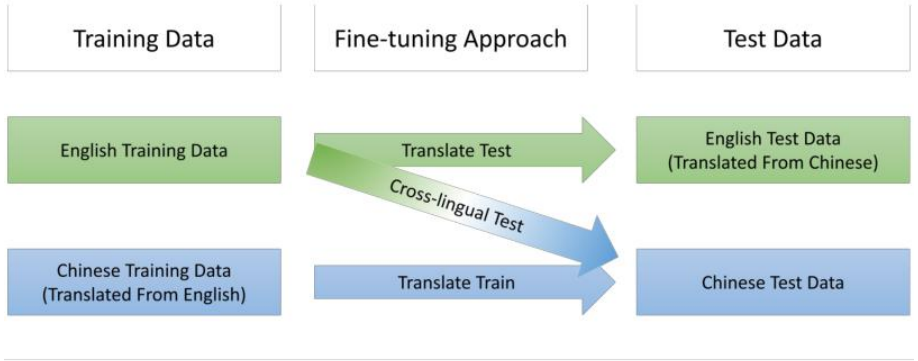
Unicoder

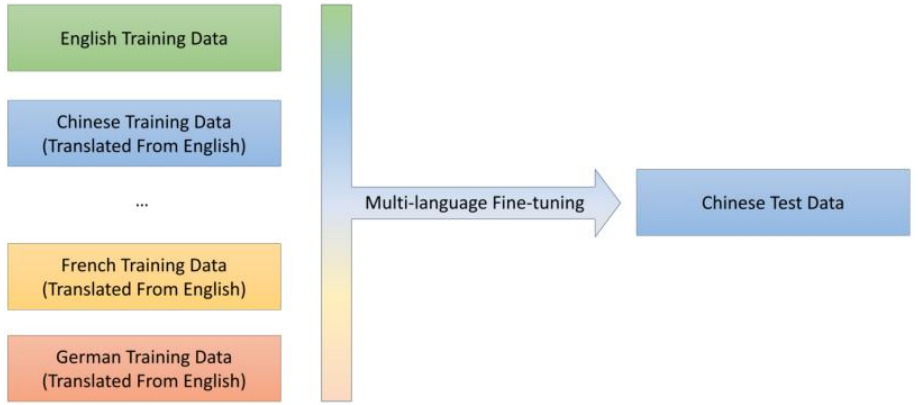
- 跨语言MLM
  - 输入是一篇用多种语言写成的文章；文章中相邻句子的语言不同，但是仍然保持通顺的承接关系。该任务在多语言的文章上进行Masked Language Model，可以模糊语言之间的边界，更好的将多种语言混合成一种语言

Unicoder

- 预训练好的模型需在跨语言的任务上微调, 目前有3种方法:
  - Translate Test: 将中文测试集翻译成英文测试集, 转换成英文训练、英文测试的问题
  - Translate Train: 将英文训练集翻译成中文训练集, 转换成中文训练、中文测试的问题
  - Cross-lingual Test: 直接将在英文训练集上训练得到的模型在中文上进行测试 (要求模型能同时编码多种语言且有较强的跨语言迁移能力)

- 多语言Fine-tuning
1. 训练集被翻译成多种语言
2. 以多任务的方式共同训练(微调)
3. 直接在中文测试集上进行测试

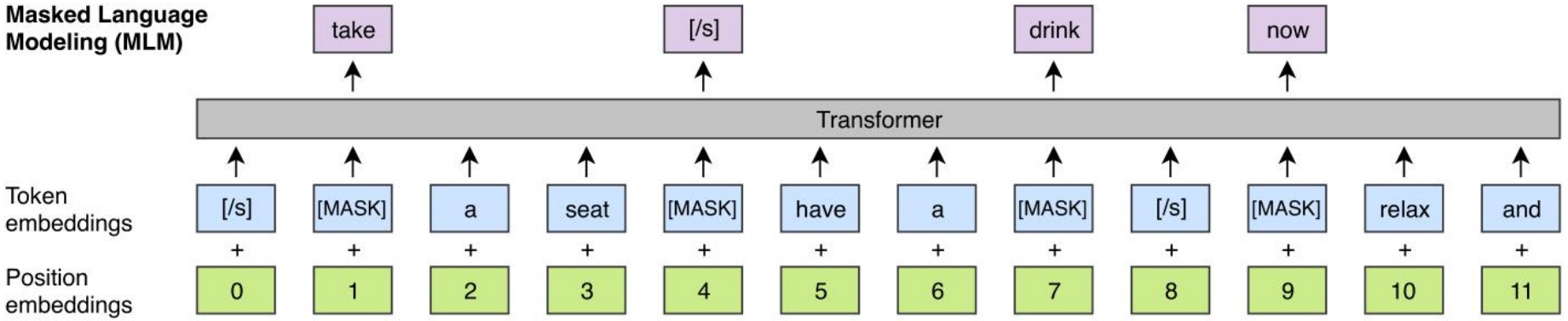




多语言微调能稳定地提升模型在各个语言上的测试效果; 对于一个确定的预训练模型, 微调的语言越多, 效果越好!

XLM-R

- 跨语言方法类似于XLM+RoBERTa
  - 从每个语言采样文本, 基于MLM训练目标, 不同于XLM的是去掉语言嵌入层
  - 直接采用SentencePiece工具对原始文本数据进行切分
- 在100种语言上使用2.5TB清洗过的CommonCrawl语料训练
- 多语言模型通过微调时利用多语言的训练集, 可以超越单语BERT模型
- 效果
  - ✓ 在四项跨语言理解(XLU)基准测试中取得迄今为止最好的结果
  - ✓ 首次实现了在不牺牲每种语言性能的情况下进行多语言建模



XLM-R

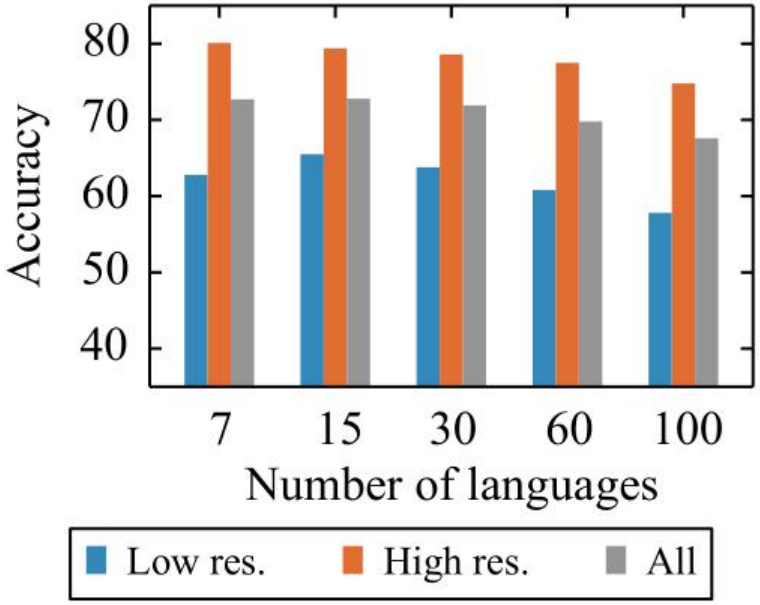
- “能力稀释”(Capacity Dilution)

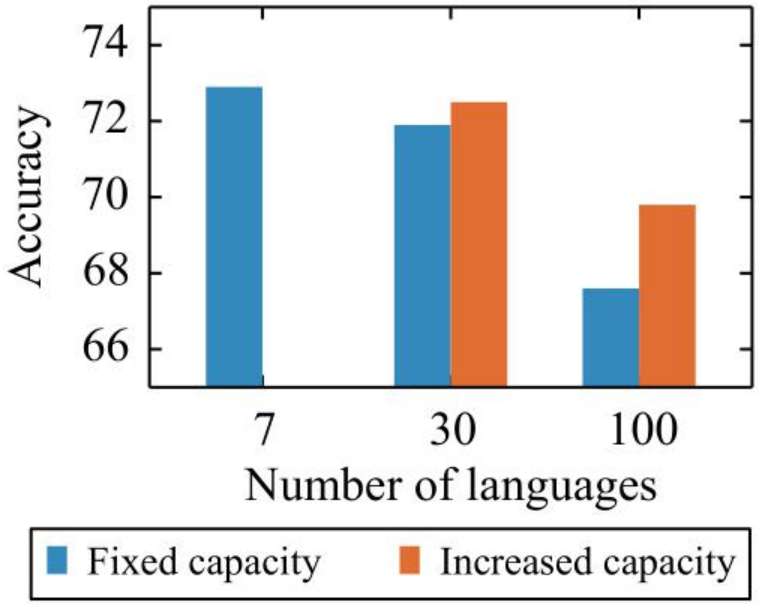
模型参数量固定, 随着参与训练的语种数量的增加, 模型在每个语种上的性能表现会下降！(能力稀释可能源于词典的稀释)

- “多语言诅咒”(Curse of Multilinguality)

在一定程度上, 更多的语言可以提高低资源语言的跨语言性能, 超过这个临界点后(稀释作用开始), 在单语和跨语言基准测试上的整体性能将下降！

- 通过简单地增大模型可以有效缓解多语言诅咒





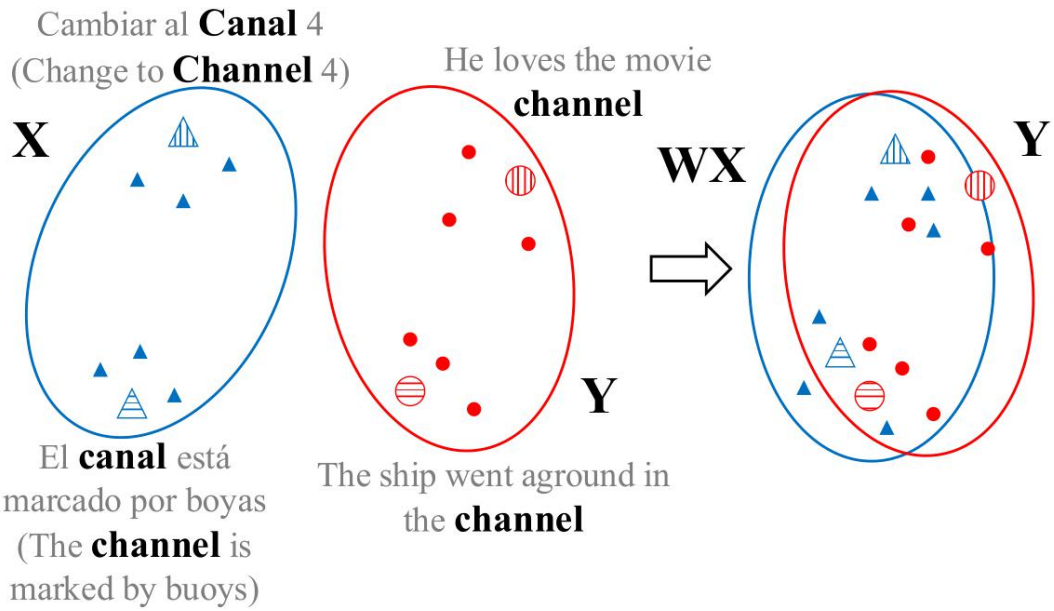
跨语言预训练任务-小结

- 在共享模型参数和多语言词汇表的基础上, 在不同语言的输入序列上进行MLM任务。该任务能够保证将不同语言的向量表示映射到同一个语义空间 ***M-BERT***
- 将双语句对拼接成一个新的输入序列, 并在该序列上进行MLM任务。通过显式引入双语对齐信息作为监督信号, 保证能够更好地学习不同语言之间的对应关系, 从而获得更好的跨语言理解能力 ***XLM***
- 输入同样是一个双语句对, 首先对该句中每个源语言-目标语言单词对计算一个attention score; 然后, 将每个源语言单词表示为全部目标语言单词向量表示的加权求和； 最后, 基于新生成的源语言表示序列, 恢复原始的源语言序列 ***Unicoder***
- 输入是两个不同语言的句子, 训练目标是判定这两个句子是否互译, 可以通过该任务学习得到不同语言在句子层面的对应关系 ***Unicoder***
- 输入是一篇由多种语言句子构成的段落, 并在此基础上进行MLM任 ***Unicoder***

基于上述跨语言预训练任务, 模型能够学习到同一语义在不同语言中的对应关系, 模糊不同语言之间的差异和边界, 并由此获得进行跨语言下游任务模型训练的能力

CLBT

- 直接使用单语言预训练的BERT, 而非从头开始利用跨语言数据训练语言模型 (离线学习)
- 假设双语句对中互为翻译的词具有相同的词向量
- 通过线性变换，将目标语言的上下文词向量映射到源语言
- 优势：仅需少量双语语料库和计算资源

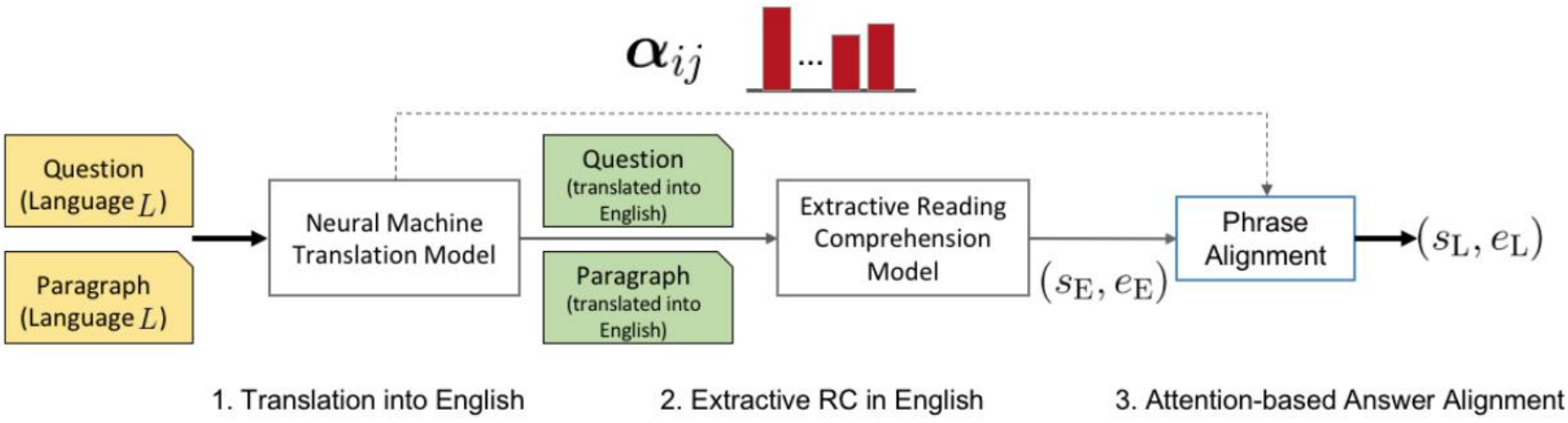


西班牙语中的canal一词的上下文表示映射到英语的语义空间，两种语言中具有相近语义的向量会更接近

跨语言阅读理解 (CLMRC)

- 着重解决基于篇章片段抽取的MRC (Span-Extraction MRC)
- 基于翻译系统实现跨语言机器阅读理解(MRC)，可以简要归纳为:

1）将目标语言输入<篇章，问题>翻译成源语言
2）通过源语言的阅读理解系统得到一个源语言的答案
3）将源语言答案回译成目标语言
4）使用NMT的注意力机制将源语言答案映射到目标语言



跨语言阅读理解-改进

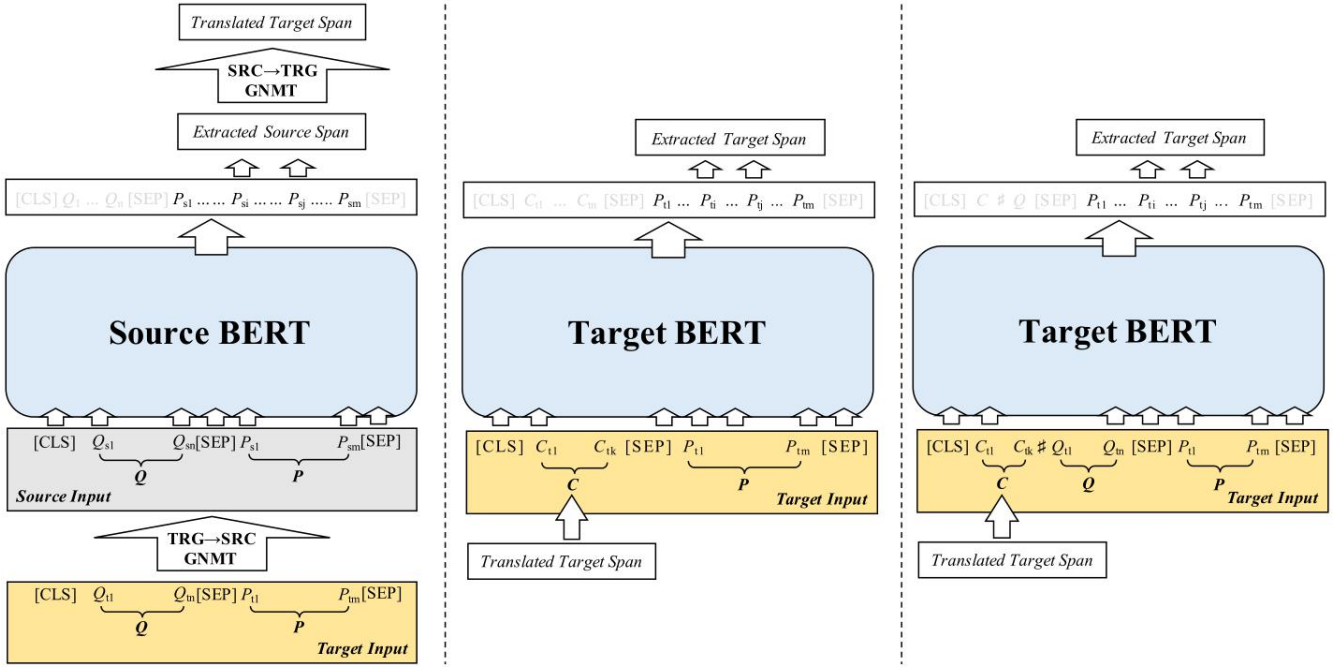
- 除英文外其他语言缺乏大规模阅读理解数据
- 将英语阅读理解模型应用于其他语言
- 方法
- **改进回翻技术**

✓ 解决目标语言没有训练数据的情况

➢ GNMT (左)

➢ 答案对齐器 (中)

➢ 答案验证器 (右)



跨语言阅读理解

- **Dual BERT**
  - ✓ 适用于目标语言有一定的训练数据的情况
  - ➢ 双通道编码器 (Dual Encoder): 源语言和目标语言均由BERT编码
  - ➢ 双语解码器 (Bilingual Decoder)

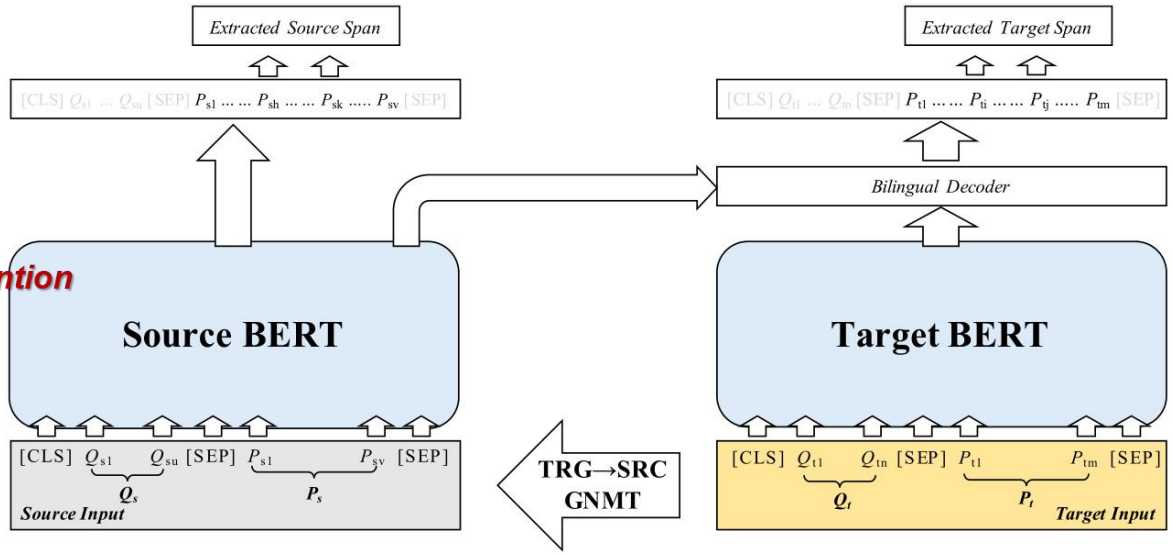
设目标语言表示B_T和源语言表示B_S

$$A_T = \text{softmax}(B_T \cdot B_T^{\mathrm{T}})$$
$$A_S = \text{softmax}(B_S \cdot B_S^{\mathrm{T}})$$
$$A_{TS} = B_T \cdot B_S^{\mathrm{T}}, A_{TS} \in \mathbf{R}^{L_T * L_S}$$
$$\widetilde{A}_{TS} = A_T \cdot A_{TS} \cdot A_S^{\mathrm{T}}, \widetilde{A}_{TS} \in \mathbf{R}^{L_T * L_S}$$

Self-Adaptive Attention

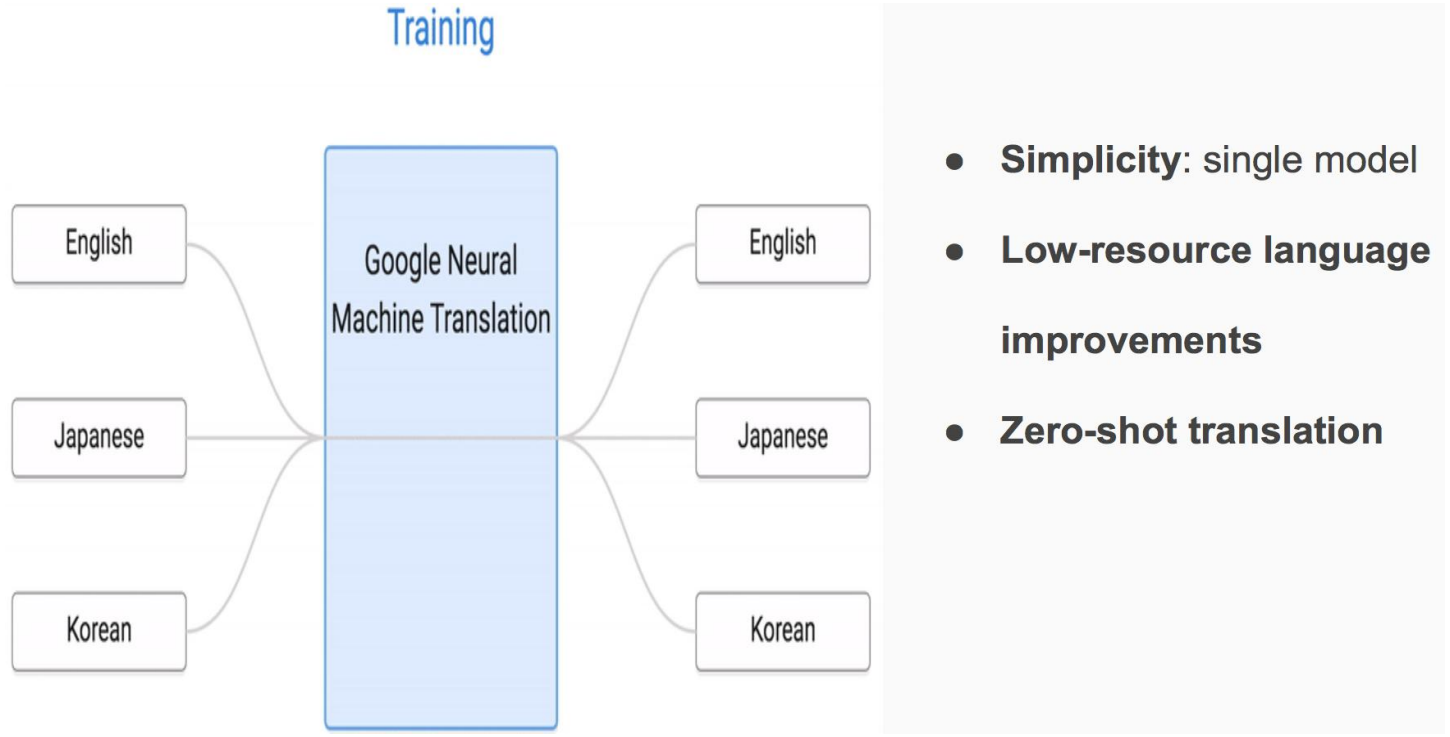
$$R' = \text{softmax}(\widetilde{A}_{TS}) \cdot B_S$$
$$R = W_r R' + b_r, W_r \in \mathbf{R}^{h*h}$$
$$H_T = \text{concat}[B_T, \text{LayerNorm}(B_T + R)]$$
$$P_T^s = \text{softmax}(W_s^{\mathrm{T}} H_T + b_s), W_s \in \mathbf{R}^{2h}$$
$$P_T^e = \text{softmax}(W_e^{\mathrm{T}} H_T + b_e), W_e \in \mathbf{R}^{2h}$$

(预测最终span的起止位置)



多语言机器翻译(M-NMT)

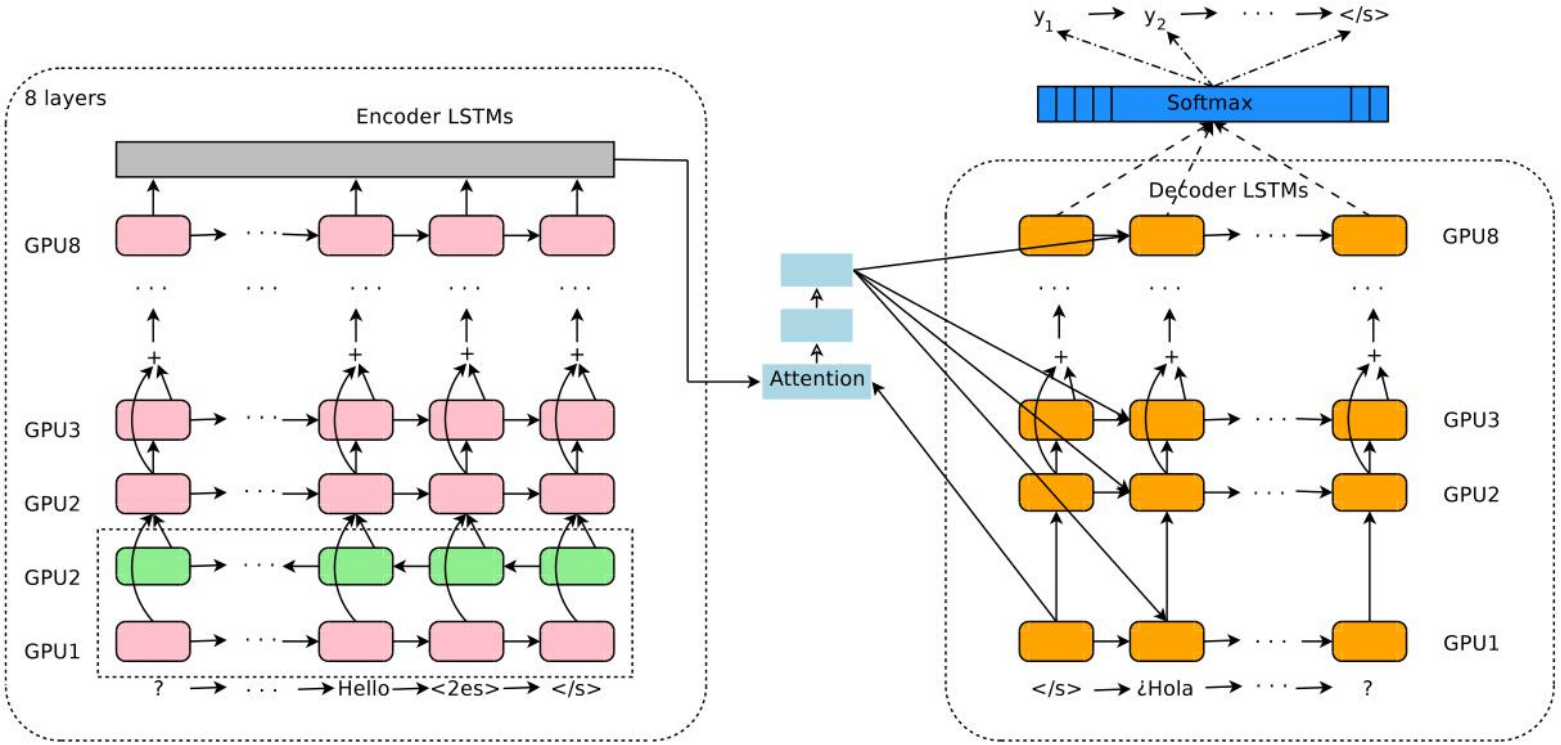
- 使用单个统一的模型来完成多种语言的翻译任务
- 通过对其他语种的学习可以大幅提升低资源语种的翻译效果
- 可以直接翻译训练语料中不存在(zero-shot)的语种组合(通过fr→de, de→en学习fr→en的翻译)



Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

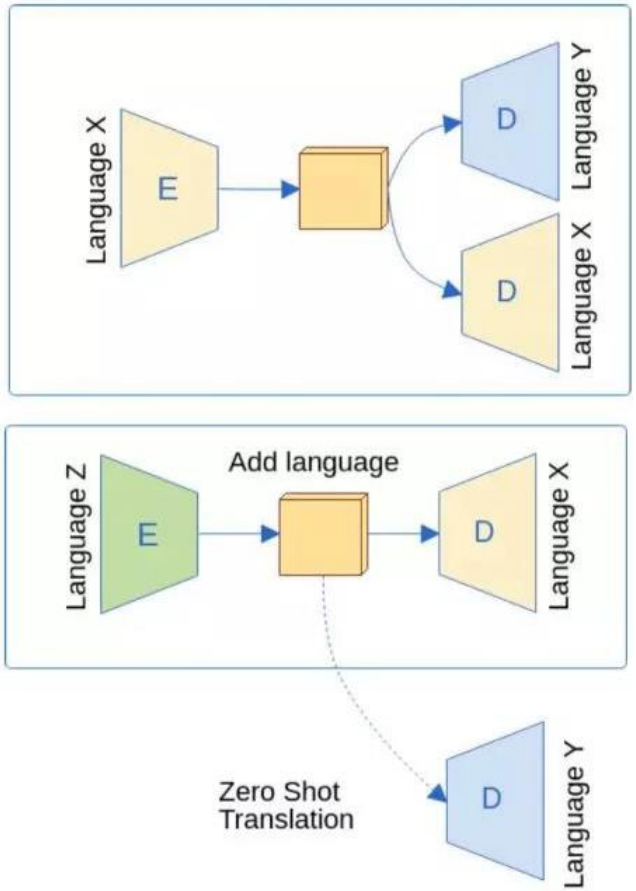
多语言机器翻译(M-NMT)

- 核心思想
  - 只修改输入数据: 在输入数据的开始人工加入目标语言的标识符
    如: **<2es>** Hello, how are you? → Hola, ¿cómo estás? (表明目标语言是Spanish)
  - 使用共享的BPE WordPiece
  - 训练时, 每个mini-batch中混合多个语言的平行语料



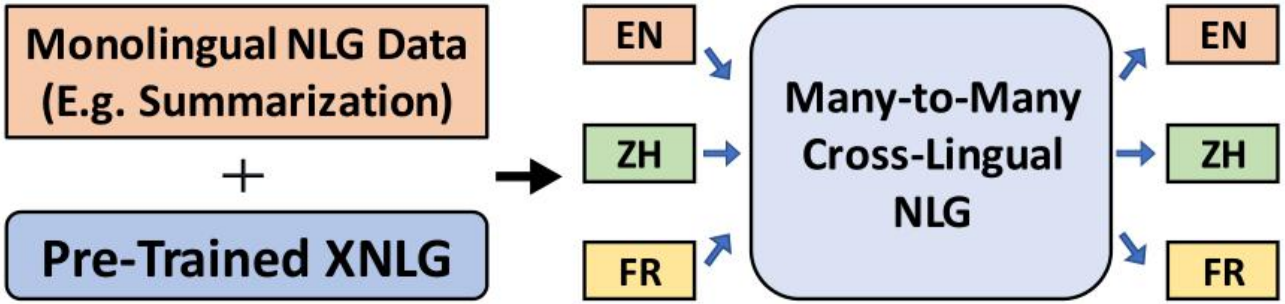
多语言机器翻译-挑战

- 不同语言的词表显著不同, 增加的新语言的词汇不在现有词表中(Vocabulary Mismatch)
- 方案1: Encoder和Decoder不共享参数
- 给定翻译对X-Y, 需要完成自编码任务(X-X, Y-Y)和翻译任务(X-Y, Y-X), 同时要求Encoder得到的两种表示相近
- 对于新语种Z, 假设有Z-X平行语料, 只需添加Z的Encoder, 然后固定X的Decoder参数进行训练, 此过程只需更新Z的Encoder参数
- 方案2: 在向量空间完成隐式翻译
- 先训练新的语言的单语词向量, 然后将已经训练好的翻译模型的词向量参数矩阵取出, 在二者之间学习一个线性映射矩阵, 用于将新语言转换到模型的语义空间



XNLG

- 在单语NLG训练数据上微调预训练生成模型, 然后在多种其他语言上进行评估，即将文本生成任务的监督信号迁移到其他预训练语言，实现NLG模型的跨语言的零样本或少样本学习

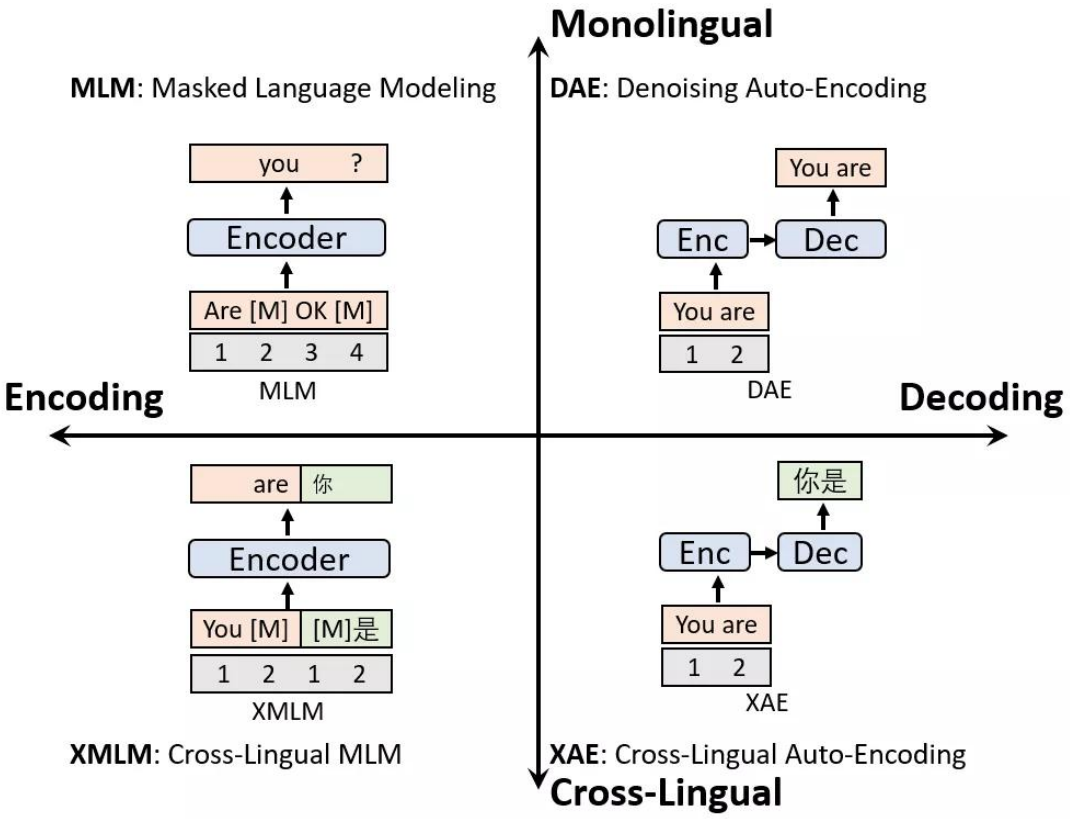


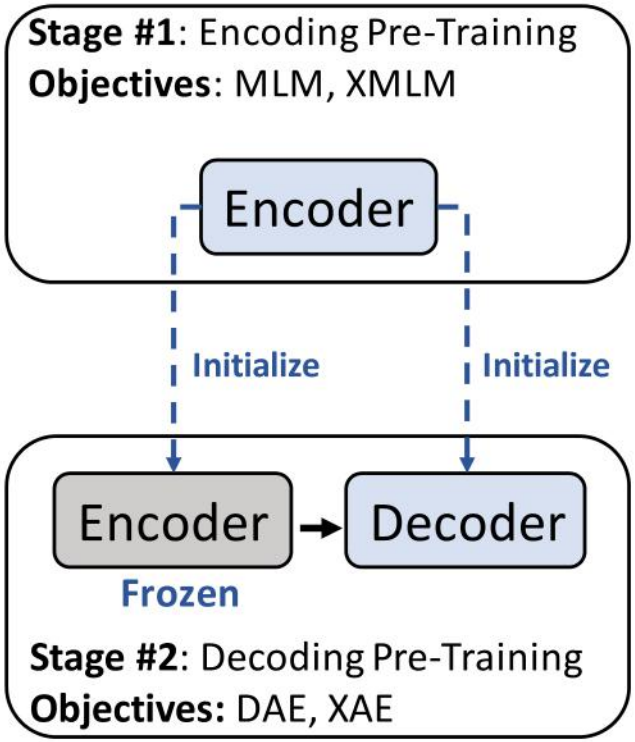
训练数据	测试数据
输入：The former queen Homaira of Afghanistan, wife of Mohammed Zahir Shah, died in Rome on Wednesday at the age of 86, two days after being hospitalised for heart troubles, her family said. 输出：Wife of Afghanistan 's former king dies in Rome	输入：纽约股市显著下跌美国纽约股票市场 2 日急剧下跌，道·琼斯 30 种工业股票平均指数下泻 了 68.63 点，降至 3370.81 点. 输出：纽约股市显著下跌

(文本摘要示例)

XNLG

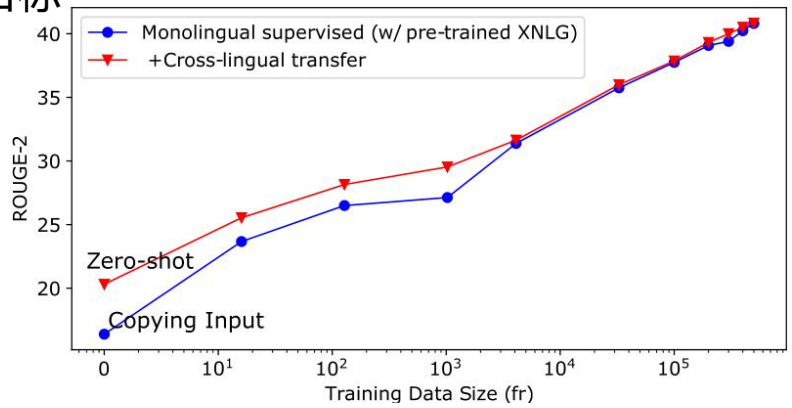
- XNLG是一个Seq2Seq的Transformer模型
- 预训练任务
  - 两个阶段 (编码预训练+解码预训练)
  - 两个维度 (单语预训练+跨语言预训练)

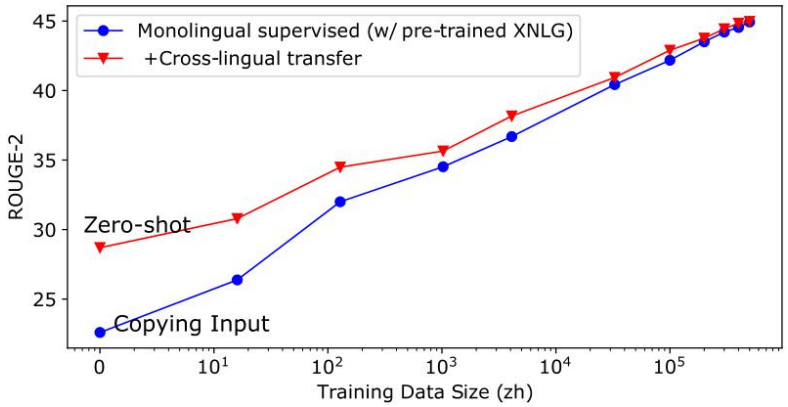




XNLG

- 实验: 跨语言零样本问题生成/文本摘要任务(英文训练, 其他语言测试)
- 效果
  - XNLG可以超越基于机器翻译的流水线(pipeline)模型
  - 在各种训练数据量上, XNLG都能将源语言的知识迁移到目标语言上并提升目标语言上的性能, 尤其在目标语言训练数据较少时

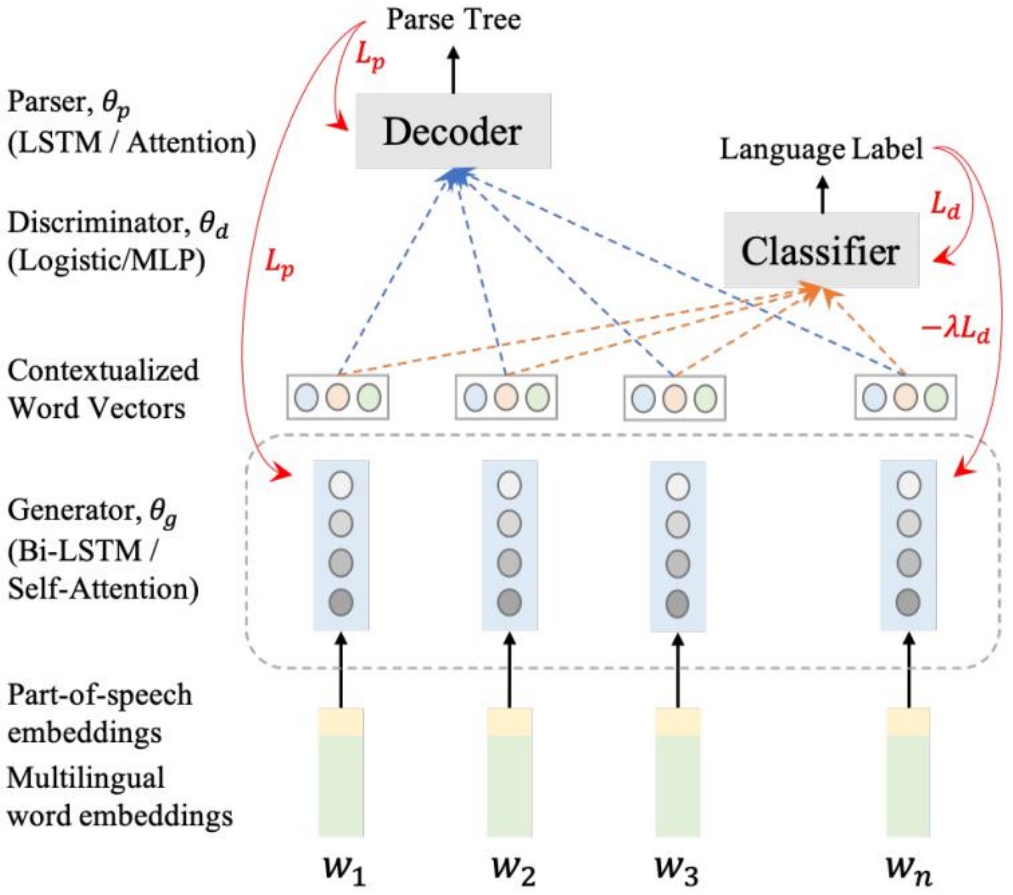




基于对抗训练的跨语言句法分析

- 对抗策略
  - Gradient Reveral (梯度反转)
  - GAN
  - Wasserstein GAN (WGAN)
- 方法
  - ✓ 基于上下文的编码器编码不同语言的特征 (生成器)
  - ✓ 解码器解码依存树 (训练目标)
  - ✓ 预测无标注的辅助语言标签来对抗训练 (判别器)
- 核心思想

以辅助语言识别任务作为对抗，引导编码器学习捕捉与语言无关的特征！



基于对抗训练的跨语言句法分析

- 训练过程
- 依存解析

Graph: $\mathcal{L}_p = -\sum_m \log p(h(m)|x, m),$

Transition: $\mathcal{L}_p = -\sum_i \log p(t_i|x, t_{<i}).$

- 语言识别
  - GAN $\mathcal{L}_d = \mathbb{E}_{x\sim X^a}[\log\ D(G(x)] + \mathbb{E}_{x\sim X^b}[\log\ (1 - D(G(x))]$
  - WGAN $\mathcal{L}_d = \mathbb{E}_{x\sim X^a}[D(G(x)] - \mathbb{E}_{x\sim X^b}[D(G(x)]$
  - GR multi-class cross-entropy loss
- 目标 (最小化)

$$\mathcal{L} = \mathcal{L}_p - \lambda\mathcal{L}_d$$

*对于句法解析任务, 采用正常的训练方式更新**encoder**和**decoder***
*对于语言识别任务, 采用对抗训练的方式更新**encoder**和**classifier***

Algorithm 1 Training procedure.

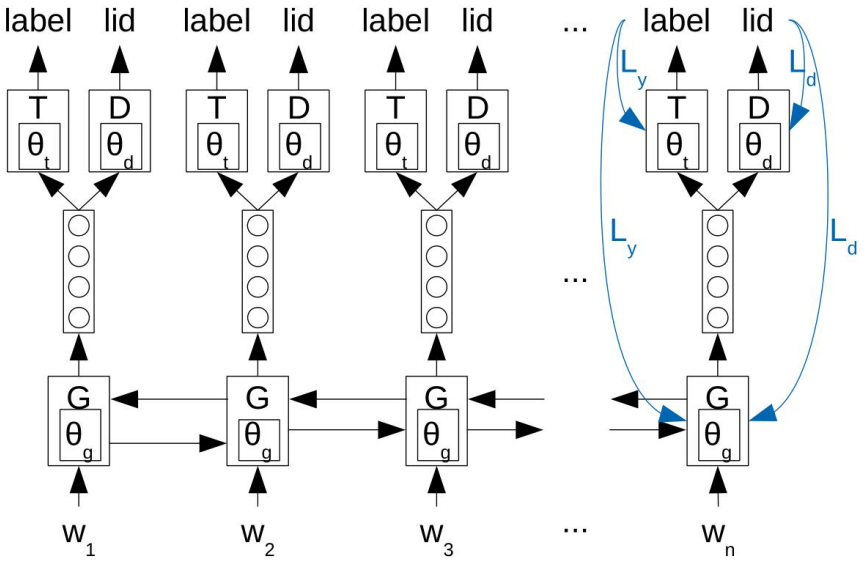
Parameters to be trained: Encoder (θ_g), Decoder (θ_p), and Classifier (θ_d)
X^a = Annotated source language data
X^b = Unlabeled auxiliary language data
I = Number of warm-up iterations
k = Number of learning steps for the discriminator (D) at each iteration
λ = Coefficient of $\mathcal{L}_d$
α_1, α_1 = learning rate; B = Batch size

Require:

for $j = 0, \cdots, I$ **do**
  Update $\theta_g := \theta_g - \alpha_1 \nabla_{\theta_g} \mathcal{L}_p$
  Update $\theta_p := \theta_p - \alpha_1 \nabla_{\theta_p} \mathcal{L}_p$
for $j = I, \cdots, num\_iter$ **do**
  for k steps **do**
    $(x_a^i)_{i=1}^{B/2} \leftarrow$ Sample a batch from X^a
    $(x_b^i)_{i=1}^{B/2} \leftarrow$ Sample a batch from X^b
    Update $\theta_d := \theta_d - \alpha_2 \nabla_{\theta_d} \mathcal{L}_d$
  Total loss $\mathcal{L} := \mathcal{L}_p - \lambda\mathcal{L}_d$
  Update $\theta_g := \theta_g - \alpha_1 \nabla_{\theta_g} \mathcal{L}$
  Update $\theta_p := \theta_p - \alpha_1 \nabla_{\theta_p} \mathcal{L}$

基于对抗训练的序列标注

- 核心思想
  - 特征生成器(G): bi-LSTM (单层)
  - 领域判别器(D): FFN
  - 目标标注器(T): FFN
  - D和T分别预测每个时间步的语言ID和target标签 (在token-level预测语言id比sentence-level更有效)



- GD训练

$$O_t = \max_{\theta_t} \sum_{i=1}^{|y|} \log p(y_i|y_{<i}; G(x))$$

$$O_d = \max_{\theta_d} \log p(lid(x)|G(x))$$

$$\theta_g := \theta_g + \nabla_{\theta_g} O_t - \lambda \nabla_{\theta_g} O_d$$

- GAN / WGAN训练

结论: 对抗训练更有助于句法分析或句子压缩等high-level任务, 而对于low-level任务(如词性标注)则没有效果! high-level任务更能从language-agnostic的特征中获益, 而low-level任务更需要知道language-specific词典信息!

$$O_d = \max_{\theta_d} E_{x\sim p_t}[\log D(G(x)] + E_{x\sim p_s}[\log(1 - D(G(x))]$$

GAN判别器目标

$$O_d = \max_{\theta_d \in L} E_{x\sim p_t}[D(G(x)] - E_{x\sim p_s}[D(G(x_s)]$$

WGAN判别器目标

$$O_g = \max_{\theta_g}[O_t - O_d]$$

生成器目标

跨语言评测数据集-XNLI

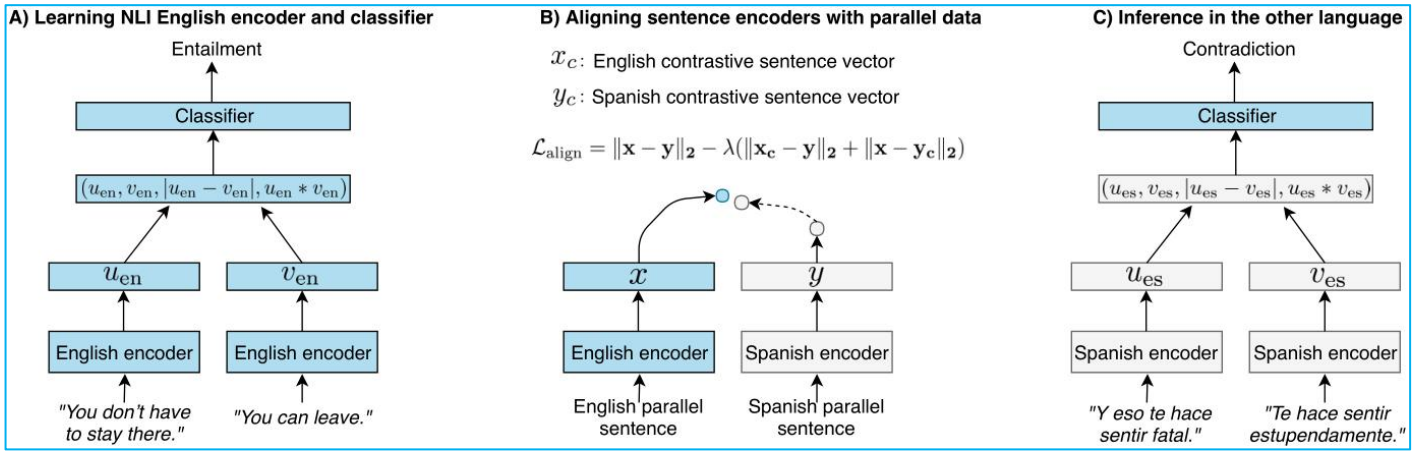
- 一种用于跨语言句子理解(XLU)的基准测试数据集
- MultiNLI语料库的扩展版, 支持15种语言, 包含10个领域
- 用文本蕴含标注的英文句子对, 再(人工)翻译成其他14种语言
- 只有英语有训练集, 其他语言只有开发集和测试集, 考验模型能否将英语训练数据集上学习到的推理知识迁移到其他语言

Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Slate	Contradiction
Arabic	تحتاج الوكالات لأن تكون قادرة على قياس مستويات النجاح. لا يمكنأللوكاالت أ اتعرف ما إذا كانت ناجحة أم لا	Nine-Eleven	Contradiction

XNLI评测任务

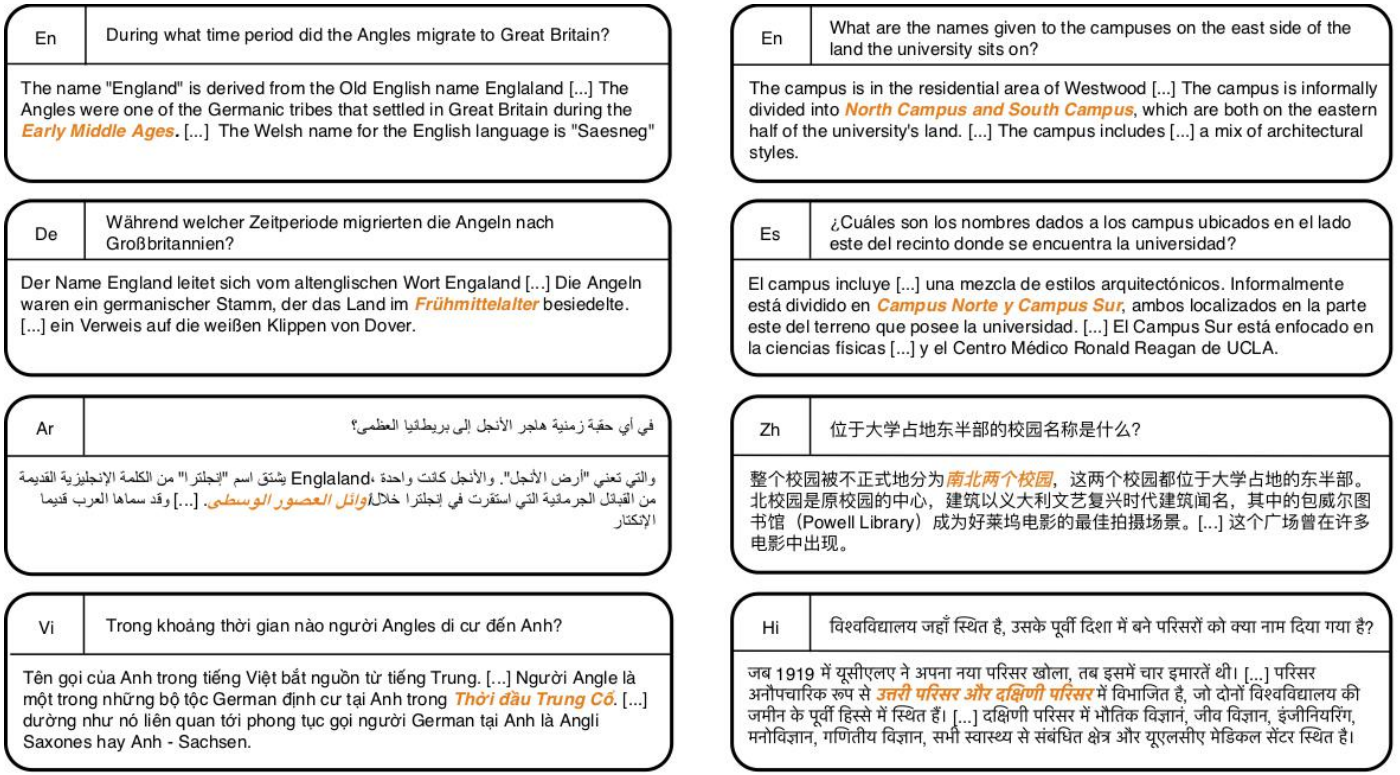
- 基于翻译的方法（依赖翻译系统）
- ✓ Baseline1: 将英文数据集翻译成目标语言, 在翻译后的数据集上训练模型 (Translate Train)
- ✓ Baseline2: 在测试阶段, 将目标语言翻译成训练阶段所用的语言, 并在训练后的模型上进行测试 (Translate Test) ***BETTER***
- 基于跨语言表示的编码器（与语言无关的统一Embedding）
- ✓ Baseline3: X-CBOW, 预训练的统一多语言句子级别的词向量, 基于CBOW方式训练的词向量的平均得到
- ✓ Baseline4: X-BiLSTM, 多语言语料上训练的BiLSTM (使用最终的隐状态或**最大的隐状态**作为特征)

Sentence Embedding Alignment



跨语言评测数据集-MLQA

- 评估模型的跨语言(抽取式)问答性能
- 支持7种语言, 每种由超过5K的QA实例(英文有12K)组成, 并采用SQuAD格式, 平均每个实例在四个语言之间是平行的

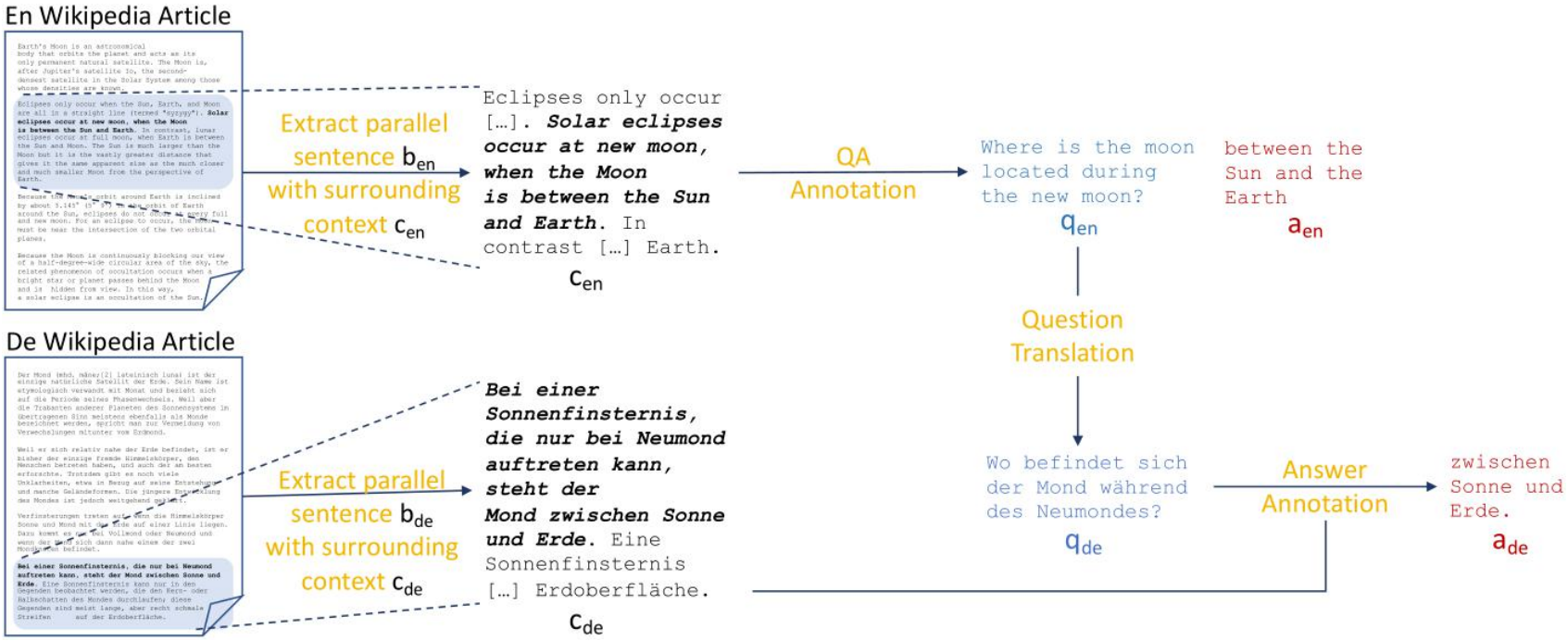


(来自MLQA的两个示例: En-De-Ar-Vi平行和En-Es-Zh-Hi平行)

MLQA标注

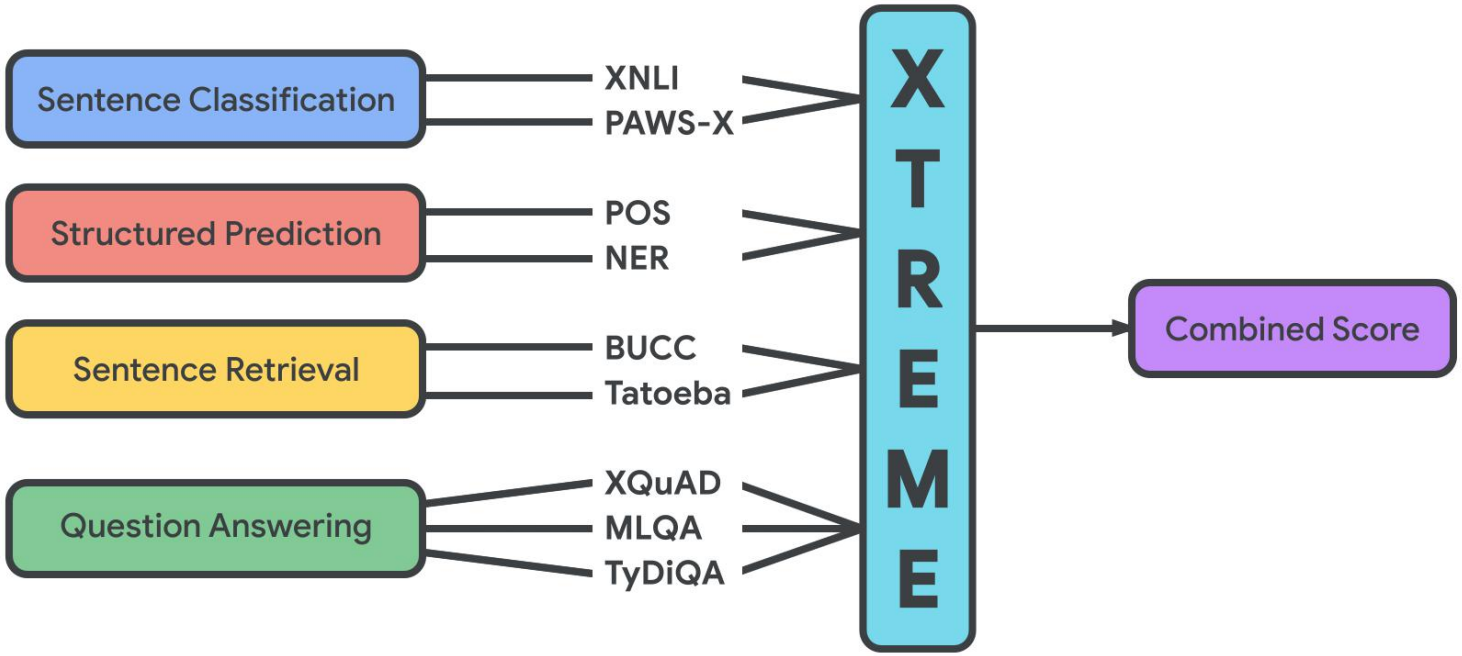
- 平行语料来源: Wikipedia
- 标注流程

1. 从每种语言的同一主题的文章中自动抽取段落 (LASER工具)
2. 众包人员在英文段落上标注问题和答案的区间 (Amazon Mechanical Turk)
3. 专业的翻译人员翻译问题并在目标语言上标注答案的区间 (One Hour Translation平台)



跨语言综合评估-XTREME

- 评估跨语言泛化的大规模、多语言、多任务的基准
- 包括了40种类型多样的语言(涵盖12种语系), 且包含9项跨语言任务
- 任务涵盖了句子分类，结构化预测，句子检索和问答等一系列范式
- 为了使模型在XTREME上获得成功, 必须学习可泛化到多种标准跨语言迁移设置的表征



跨语言综合评估-XTREME

- Zero-shot评估

1. 对多语言文本进行预训练
2. 使用英文对下游任务进行微调
3. 在XTREME非英文任务上进行zero-shot评估

- 好处

- 预训练模型仅需针对每个任务在英文数据上进行微调，便能直接应用于其他语言的评估

Table 1. Characteristics of the datasets in XTREME for the zero-shot transfer setting. For tasks that have training and dev sets in other languages, we only report the English numbers. We report the number of test examples per target language and the nature of the test sets (whether they are translations of English data or independently annotated). The number in brackets is the size of the intersection with our selected languages. For NER and POS, sizes are in sentences. Struct. pred.: structured prediction. Sent. retrieval: sentence retrieval.

Task	Corpus	|Train|	|Dev|	|Test|	Test sets	|Lang.|	Task	Metric	Domain
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI	Acc.	Misc.
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase	Acc.	Wiki / Quora
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS	F1	Misc.
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER	F1	Wikipedia
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction	F1 / EM	Wikipedia
	MLQA			4,517–11,590	translations	7	Span extraction	F1 / EM	Wikipedia
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction	F1 / EM	Wikipedia
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval	F1	Wiki / news
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval	Acc.	misc.

参考文献

- Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. arXiv preprint arXiv:1706.04902, 2017.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformersfor language understanding, NAACL-HLT 2019. ***M-BERT***
- Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. NeurIPS 2019. ***XML***
- Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pretraining with multiple cross-lingual tasks. ACL 2019. ***Unicoder***
- A. Conneau *, K. Khandelwal *, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov Unsupervised Cross-lingual Representation Learning at Scale, 2019 ***XML-R***
- A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, V. Stoyanov XNLI: Evaluating Cross-lingual Sentence Representations, EMNLP 2018 ***XNLI***

参考文献

- Y. Wang, W. Che, J. Guo, Y. Liu, and T. Liu, Cross-lingual bert transformation for zero-shot dependency parsing, EMNLP 2019. ***CLBT***
- P. Lewis, B. O˘guz, R. Rinott, S. Riedel, and H. Schwenk, Mlqa:Evaluating cross-lingual extractive question answering, arXiv 2019. ***MLQA***
- J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization, CoRR 2020. ***XTREME***
- Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. Multilingual extractive reading comprehension by runtime machine translation. arXiv 2018.
- Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, Cross-lingual machine reading comprehension, EMNLP-IJCNLP, 2019. ***CLMRC***
- M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Tho-rat, F. Vi´egas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, arXive 2017

参考文献

- Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT?, ACL 2019.
- Yunsu Kim, Yingbo Gao, and Hermann Ney. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. ACL 2019.
- Z. Chi, L. Dong, F. Wei, W. Wang, X. Mao, and H. Huang, Cross-lingual natural language generation via pre-training, AAAI 2020. ***XNLG***
- Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, Nanyun Peng. Cross-lingual Dependency Parsing with Unlabeled Auxiliary Languages. CoNLL 2019.
- Adel, Heike; Bryl, Anton; Weiss, David; Severyn, Aliaksei. Adversarial Neural Networks for Cross-lingual Sequence Tagging. arXiv 2018