A Prospective Investigation on the Application of Pre-training Model to Machine Translation

iFlytek htrans

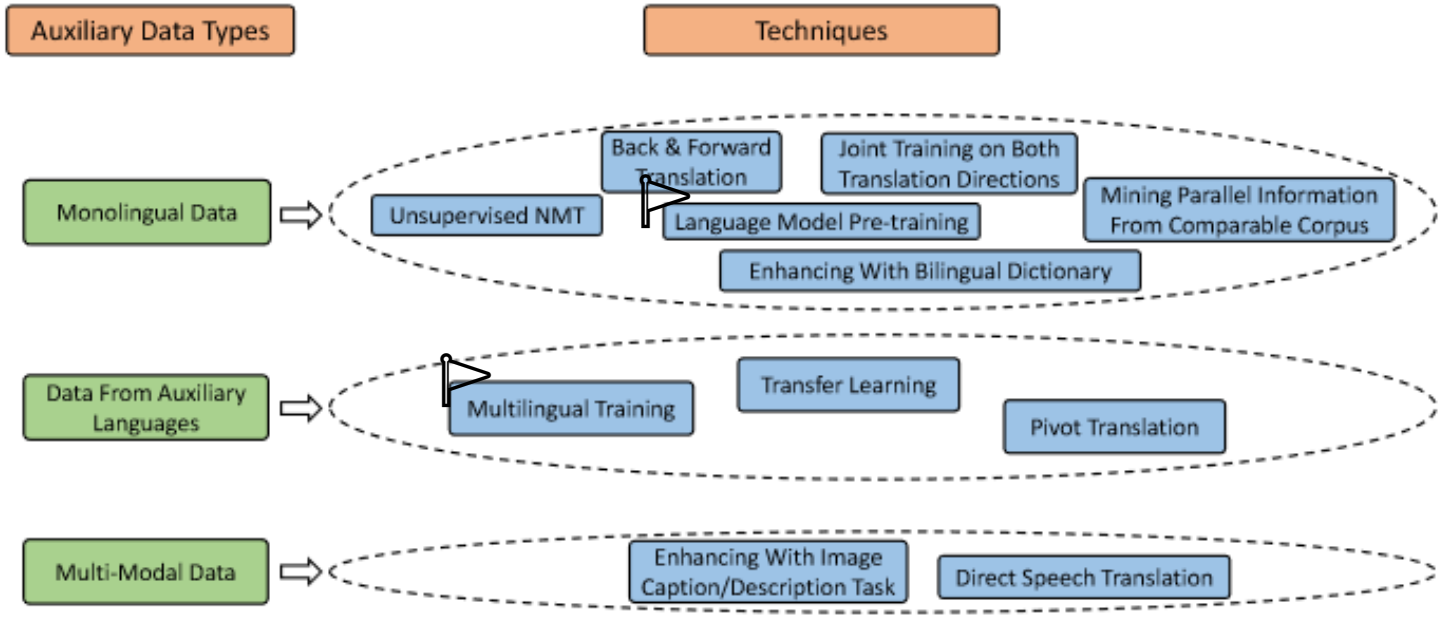
Wei Chen

2022/08

低资源NMT:

- 由于缺乏平行句对，在低资源NMT中利用平行句以外的数据至关重要。
- 因此根据用于帮助低资源语言对的数据，将低资源NMT上的现有算法分为三类 [1]:

  - 单语数据
  - 辅助语言(语法或语义相似)数据
  - 多模态数据



[1] A Survey on Low-Resource Neural Machine Translation. arXiv 21' MSRA Wang et.al

利用单语数据应用在低资源NMT:

- 大量工作已经在NMT系统中利用了单语数据，可将方法分为以下几类 [1]:

  - 正向&反向翻译
  - 两个翻译方向联合训练
  - 无监督NMT
  - 预训练

    利用单语数据和自监督预训练来进行语言理解(encoder)和生成(decoder)，从而提高了NMT模型的质量
  - 可比的单语语料库
  - 双语词典增强

利用单语数据预训练语言模型应用在低资源NMT:

- 关于NMT语言模型预训练的工作，取决于NMT中的编码器和解码器分别或联合预训练可分为两类:
  - 分开预训练:
    - 使用语言模型分别对encoder和decoder预训练并初始化，然后使用监督并行数据微调 [2]
    - XLM, 使用单独的语言模型预训练初始化encoder和decoder，结合MLM遮蔽语言建模，翻译语言模型 [3]
    - [4] 研究了各种模型(BERT/GPT-2/RoBERTa/随机初始化)来初始化encoder和decoder
      (结论: bert初始化encoder&随机初始化decoder 或 bert初始化共享encoder和decoder 在英德上取得最佳)
    - 单独预训练的encoder和decoder初始化后，将弹性权重合并引入微调，以避免忘记语言模型 [5]
    - 通过注意机制将BERT提取的表征融合到encoder和decoder [6]

    存在问题：不能很好地同时训练encoder-decoder的注意力，这在NMT中非常重要，即连接src和tgt表示以进行翻译。
  - 联合预训练:
    - MASS [7]：掩蔽序列到序列学习，随机掩蔽encoder中输入句子中的片段(几个连续token)，在decoder中预测
    - BART [8]：在encoder中添加噪声，随机遮蔽句子中的一些token，在decoder中重建原始序列。
    - T5 [9]：随机遮蔽一些token，并用单个标志token替换连续token

BERT/GPT/XLNET

[2] Unsupervised pretraining for sequence to sequence learning. EMNLP 17' Google Brain Prajit et.al
[3] Cross-lingual language model pretraining. NIPS 19' Facebook AI Research Conneau et.al
[4] Leveraging pre-trained checkpoints for sequence generation tasks. TACL 20' Google Research Sascha et.al
[5] Unsupervised pretraining for neural machine translation using elastic weight consolidation. ACL Workshop 19' Charles University Dusan et.al
[6] Incorporating bert into neural machine translation. ICLR'20 USTC&MSRA Zhu et.al
[7] Masked sequence to sequence pre-training for language generation. ICML'19 NUS&MSRA Song et.al
[8] Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. ACL'20 Fackbook AI Lewis et.al
[9] Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR'20 Google Raffel et.al

利用辅助语言数据应用在低资源NMT:

- 大量工作已经在NMT系统中利用了辅助语言数据，可将方法分为以下几类 [1]:

  - 多语言训练

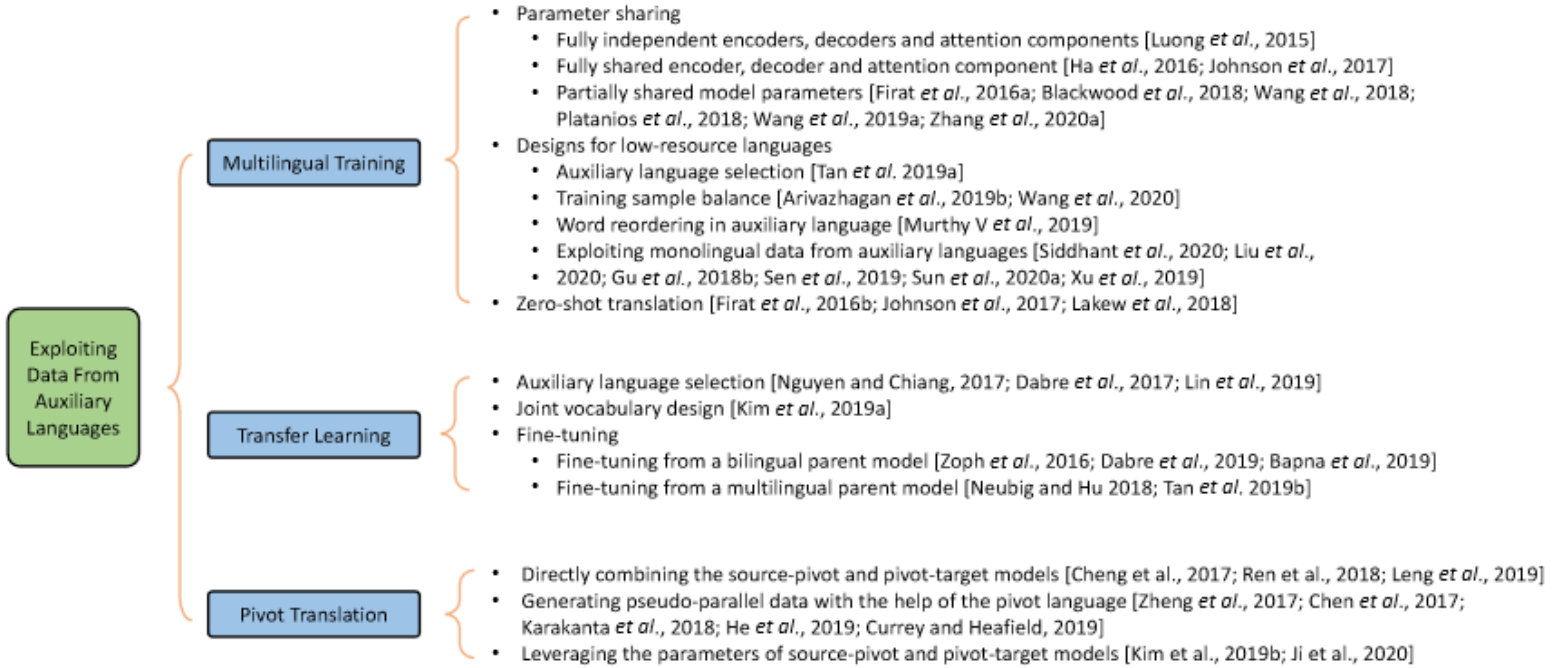
    低资源语言对与其他语言对在一个模型中联合训练

  - 迁移学习

    首先在包含丰富资源语言对的NMT模型训练，然后在低资源语言对上微调

  - 级联翻译

    选择一种或多种中转语言作为源语言和目标语言之间的桥梁，利用src-pivot和pivot-tgt数据来帮助src-tgt翻译



利用辅助语言数据预训练多语言模型应用在低资源NMT:

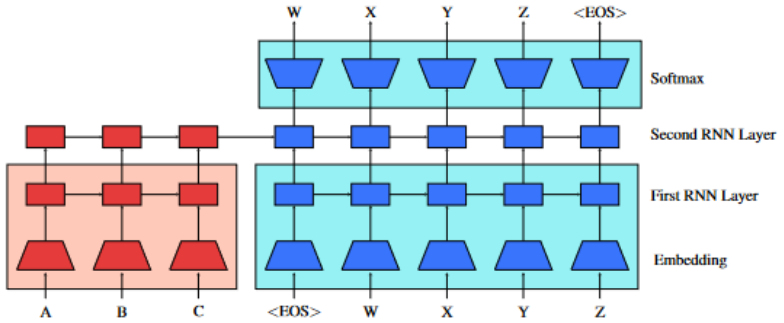
- 多语言训练模型三个优势:
  - 参数共享：与训练多个单独模型比，参数共享在单模型中训练多语言对可以降低训练和维护成本，且可从多种语言中集体学习知识，以帮助低资源语言。
    - 所有编码器、解码器和注意力组件在不同语言之间独立
    - 完全共享的编码器、解码器和注意力组件（其中在src语句中添加特定语言的token，以指定翻译tgt语言）
    - 部分共享模型参数
  - 低资源语言设计：低资源语言对通过联合训练受益于相关的丰富资源语言对
    - 辅助语言选择
    - 平衡训练样本比例
    - 利用辅助语言中的单语数据
      - 回译
      - 跨语言预训练模型 🚩：XLM [10] -> XLM-R [11] -> mBART [12] -> LaBSE [13] -> m2m-100 [14] -> deltaLM [15]
      - 元学习
  - Zero-shot翻译：多语言NMT提供了在训练期间不可见的语言对上进行翻译的可能性

[10] Cross-lingual language model pretraining. arXiv 19' Facebook AI Lample et.al
[11] Unsupervised cross-lingual representation learning at scale. ACL 20' Facebook AI Goyal et.al
[12] Multilingual denoising pre-training for neural machine translation. ACL 20 ' Facebook AI Liu et.al
[13] Language-agnostic bert sentence embedding. arXiv 20' Google AI Feng et.al
[14] Beyond English-Centric Multilingual Machine Translation. arXiv'20 Facebook AI Fan et.al
[15] DeltaLM: Encoder-Decoder Pre-training for Language Generation and Translation by Augmenting Pretrained Multilingual Encoders. arXiv'21 Microsoft et.al

利用额外单语数据预训练语言模型应用在提升NMT

Unsupervised Pretraining for Sequence to Sequence Learning. ENMLP 17' Google Ramachandran et.al

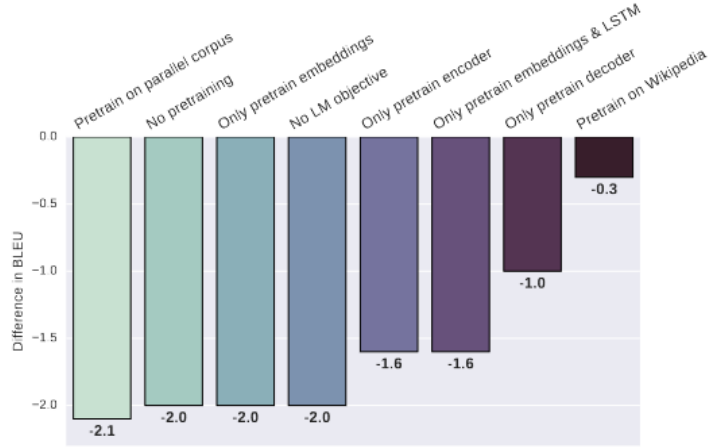
简介：使用两种预训练语言模型的权重分别初始化seq2seq的encoder和decoder，然后使用标签句对进行微调，联合训练seq2seq目标和语言建模目标（避免在小的标签数据集上微调导致灾难性遗忘，使得语言建模任务性能下降），在WMT'14/15 en-de ↑ 1.3BLEU。



System	ensemble?	BLEU newstest2014	BLEU newstest2015
Phrase Based MT (Williams et al., 2016)	-	21.9	23.7
Supervised NMT (Jean et al., 2015)	single	-	22.4
Edit Distance Transducer NMT (Stahlberg et al., 2016)	single	21.7	24.1
Edit Distance Transducer NMT (Stahlberg et al., 2016)	ensemble 8	22.9	25.7
Backtranslation (Sennrich et al., 2015a)	single	22.7	25.7
Backtranslation (Sennrich et al., 2015a)	ensemble 4	23.8	26.5
Backtranslation (Sennrich et al., 2015a)	ensemble 12	**24.7**	27.6
No pretraining	single	21.3	24.3
Pretrained seq2seq	single	**24.0**	**27.0**
Pretrained seq2seq	ensemble 5	**24.7**	**28.1**

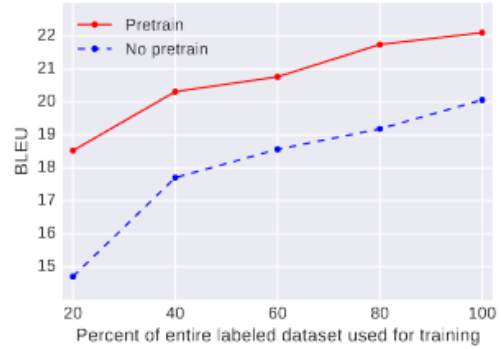
对于机器翻译,本文评估WMT en-de任务：

- 训练集：使用了WMT14训练数据集，使用语言检测系统来过滤，最后共约400万个训练样本。
- 分词：使用subword进行89500次合并操作，最终词汇量约为90000。
- 验证集：newstest2012和newstest2013合并，使用multi-BLEU.perl对分词翻译使用区分大小写的BLEU进行验证集评估。
- 测试集：newstest2014和newstest2015合并，使用mteval-v13a.pl对不分词翻译使用区分大小写的BLEU进行测试集评估。
- 单语预训练语料库：新闻抓取的英语和德语语料库，每个语料库都有超过10亿个token



结论：

- 仅预训练解码器比仅预训练编码器好
- 编码器和解码器都预训练，会比单独产生更多增益
- 预训练softmax层是重要的
- 语言建模目标是一个强大的正则化项（对模型进行预训练而不用LM目标的BLEU下降与不预训练而使用LM目标一样糟？）
- 使用在并行语料库的src和tgt部分进行预训练LMs，初始化模型，基本无作用，性能下降与不预训练基本一样。
- 在和测试集领域不同的大型非新闻维基百科语料库上进行预训练时，性能仍然很强。
- 随着有标签的数据集变小，预训练模型的退化程度降低。



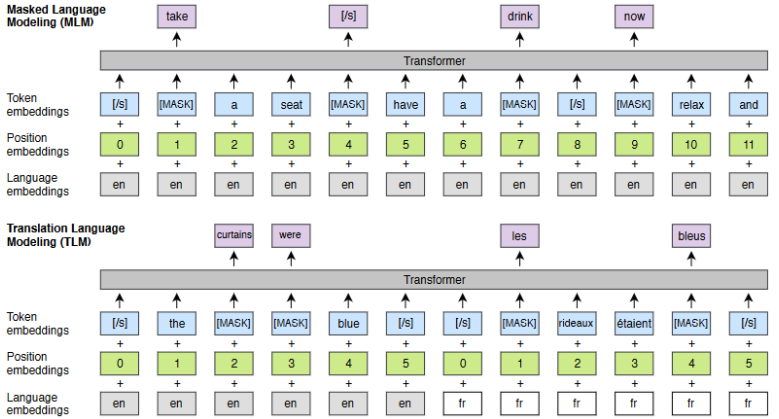
XLM：Cross-lingual language model pretraining. arXiv 19’ Facebook AI Lample et.al

模型：

首次将生成性预训练扩展到多语言，提出两种跨语言模型学习方法：

1）无监督，仅依赖单语数据（CTM+MLM）

2）有监督，利用平行语料（MLM+TLM）



注：

[1] CLM（Causal Language Modeling），即传统 LM 训练任务，已知全部上文序列，预测下一个词。XLM 根据不同的下游任务，使用了不同的预训练目标。CLM 只用于纯语言模型任务测 ppl。对于 Transformer 结构，可以控制 attention-mask 来保证只有上文可见。

[2] MLM（Masked Language Modeling），继承自 Bert，增加了对低频词的抽样概率补偿，以避免学习不充分。

[3] TLM（Translation Language Modeling），给平行语料设计的训练目标。借助译文句子，预测原文中被 Masked 的词。反之亦然。

[4] 混合学习 BPE 构建共享词表,对小语种进行了采样补偿，避免因为不占优而被切得过碎。

数据：

无监督单语数据： 用WikiExtractor从Wikipedia dumps中抽取原始句子

有监督平行数据：只使用涉及到英文的平行语料

高资源： MultiIUN——法语、西班牙语、俄语、阿拉伯语、汉语 / IIT Bombay corpus——印度语

低资源： OPUS website Tiedemann (2012)——EUbookshop corpus for German, Greek and Bulgarian, OpenSubtitles 2018 for Turkish, Vietnamese and Thai, Tanzil for both Urdu and Swahili and GlobalVoices for Swahili

分词：对于中文、日语和泰语，分别采用[18]分词器、Kytea分词器和PyThaiNLP分词器。其他统一采用Moses分词器。BPE：FastBPE。

总结：

在操作层面上，XLM 先是混合所有数据建立共享 BPE 词表，然后加入 Language Embedding，根据下游任务选择使用 CLM、MLM 和 TLM 三个目标进行训练。以下是应用：

A. 无监督机器翻译：弃用 MUSE 中使用 fastText 初始化 embedding 的方法，转而用 XLM 初始化整个 encoder 和 decoder，分为随机初始化/CLM/MLM共3*3=9种初始化编码和解码器的方法+1种只初始化嵌入层的方法。

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

Pretraining	-	CLM	MLM
Sennrich et al. (2016)	33.9	-	-
ro → en	28.4	31.5	35.3
ro ↔ en	28.5	31.5	35.6
ro ↔ en + BT	34.4	37.0	**38.5**

Training languages	Nepali perplexity
Nepali	157.2
Nepali + English	140.1
Nepali + Hindi	115.6
Nepali + English + Hindi	**109.3**

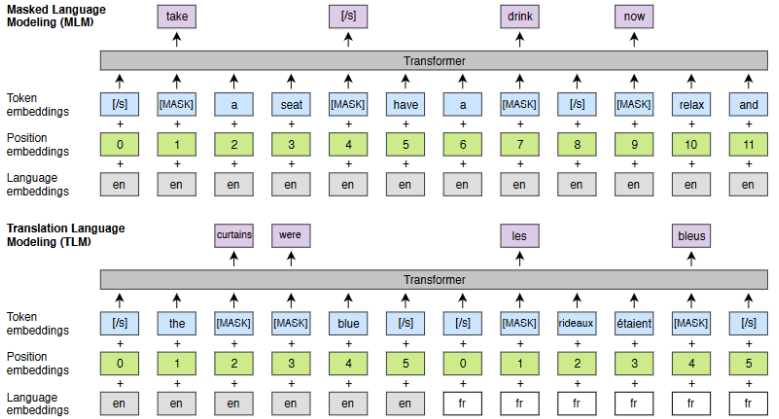
[18] Optimizing chinese word segmentation for machine translation performance. SML workshop’08 Chang et.al

Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. TACL 19’ Google Sascha et.al

简介:

本文证明了预训练checkpoint用于序列生成的有效性。研究了各种模型(BERT/GPT-2/RoBERTa/随机初始化)来初始化encoder和decoder的性能

- BERT checkpoint:
  - 分词：wordpiece
  - BERT-Base-Cased / BERT-Base-Uncased / BERT-Base Multilingual Cased



数据:

无监督单语数据： 用WikiExtractor从Wikipedia dumps中抽取原始句子

有监督平行数据：只使用涉及到英文的平行语料

高资源： MultilUN——法语、西班牙语、俄语、阿拉伯语、汉语 / IIT Bombay corpus——印度语

低资源： OPUS website Tiedemann (2012)——EUbookshop corpus for German, Greek and Bulgarian, OpenSubtitles 2018 for Turkish, Vietnamese and Thai, Tanzil for both Urdu and Swahili and GlobalVoices for Swahili

分词：对于中文、日语和泰语，分别采用[18]分词器、Kytea分词器和PyThaiNLP分词器。其他统一采用Moses分词器。BPE：FastBPE。

总结:

在操作层面上，XLM 先是混合所有数据建立共享 BPE 词表，然后加入 Language Embedding，根据下游任务选择使用 CLM、MLM 和 TLM 三个目标进行训练。以下是应用:

A. 无监督机器翻译：弃用 MUSE 中使用 fastText 初始化 embedding 的方法，转而用 XLM 初始化整个 encoder 和 decoder，分为随机初始化/CLM/MLM共3*3=9种初始化编码和解码器的方法+1种只初始化嵌入层的方法。

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

Pretraining	-	CLM	MLM
Sennrich et al. (2016)	33.9	-	-
ro → en	28.4	31.5	35.3
ro ↔ en	28.5	31.5	35.6
ro ↔ en + BT	34.4	37.0	**38.5**

Training languages	Nepali perplexity
Nepali	157.2
Nepali + English	140.1
Nepali + Hindi	115.6
Nepali + English + Hindi	**109.3**

代码：https://github.com/google-research/google-re search/tree/master/bertseq2seq

利用额外辅助语言数据预训练多语言模型应用在提升NMT

mBART：Multilingual Denoising Pre-training for Neural Machine Translation. ACL 20’ Facebook AI Liu et.al

数据：爬取的25种语言CC-25M，遵循XLM使用重采样平衡每种语言。

预处理：sentence piece model，对包括25万个子词标记的完整CC数据tokenize

模型：遵循BART，不同的是BART只针对英语进行预训练，
mBART系统研究了预训练对不同语言集的影响。

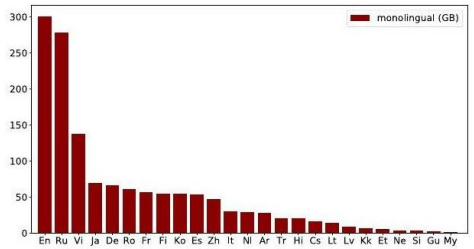
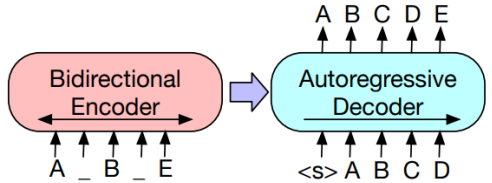
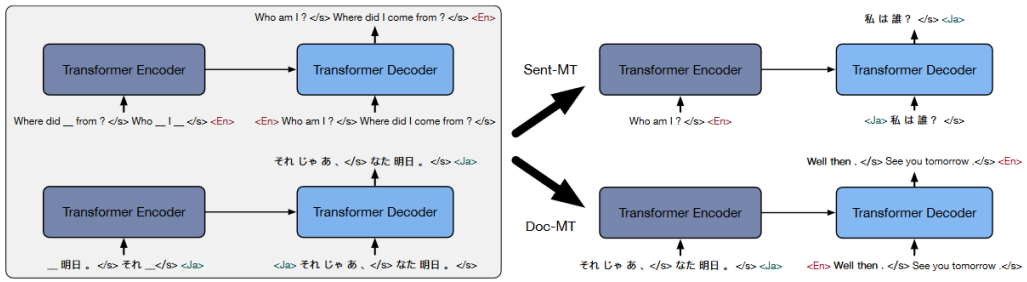


Figure 1: Sizes of the CC25 Corpus. A list of 25 languages ranked with monolingual corpus size.



预训练的一系列模型：

- mBART25——对所有25种语言进行模型预训练。
- mBART06——在六种欧洲语言的子集上预训练一个模型。Ro、It、Cs、Fr、Es和En（探索预训练对相似语言的影响）
- mBART02——预训练双语模型，使用英语和另一种语言进行四种语言对。En-De, En-Ro, En-It.
- BART-En/Ro——只在En和Ro语料库上训练单语言BART模型。



Multilingual Denoising Pre-Training (mBART)　　Fine-tuning on Machine Translation

结论：

对于有监督的句级MT，mBART 初始化在低/中资源对（<10M 双文本对）上带来了显著的收益（高达 12 个 BLEU 点）。这些结果在反译(BT)的情况下得到进一步改善，在WMT16英语-罗马尼亚语和FloRes测试集上创造了新的先进水平。（缺点：高资源无提升，甚至掉点）

代码： https://github.com/facebookresearch/fairseq/blob/main/examples/mbart/README.md

[18] Optimizing chinese word segmentation for machine translation performance. SML workshop’08 Chang et.al