多语言预训练模型技术与应用
Multilingual Pre-trained Language Models — Technologies and Applications

Qun Liu (刘群)

Huawei Noah's Ark Lab

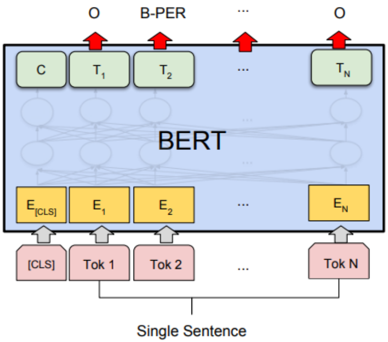
全国人工智能技术大会
多语种智能信息处理专业论坛
杭州，2021-06-06

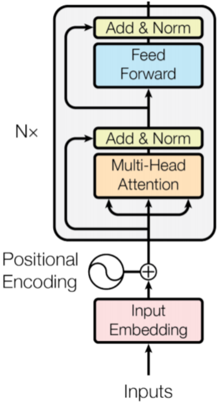
HUAWEI

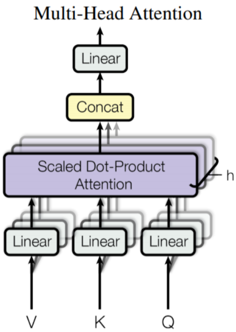
Content

Content

Pre-trained Language Models







Vaswani et al., Attention is All You Need, NIPS 2017: 5998-6008

Devlin, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805v2, 2018

Multi-Lingual BERT (mBERT)

Models

There are two multilingual models currently available. We do not plan to release more single-language models, but we may release `BERT-Large` versions of these two in the future:

- `BERT-Base, Multilingual Cased (New, recommended)` : 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- `BERT-Base, Multilingual Uncased (Orig, not recommended)` : 102 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- `BERT-Base, Chinese` : Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters

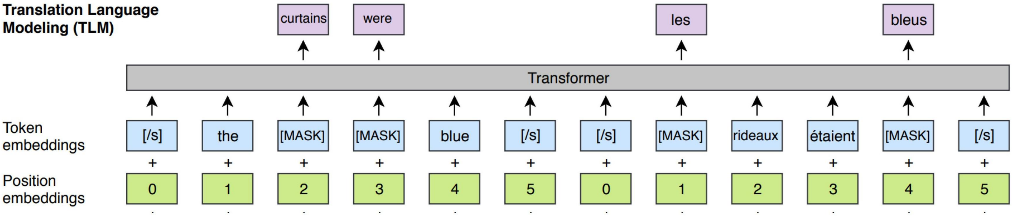
Data Source and Sampling

The languages chosen were the top 100 languages with the largest Wikipedias. The entire Wikipedia dump for each language (excluding user and talk pages) was taken as the training data for each language

HUAWEI

Cross-lingual Language Model (XLM)

Multilingual MLM is unsupervised, but we leverage parallel data with TLM:



Translation language modeling (TLM) pretraining

Lample and Conneau, Cross-lingual Language Model Pretraining, arXiv:1901.07291, 2019

Evaluation of Multilingual PLMs on XNLI

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
mBERT	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
XLM (MLM+TLM)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
XLM-R	CC	1	100	**88.8**	**83.6**	**84.2**	**82.7**	**82.3**	**83.1**	**80.1**	**79.0**	**78.8**	**79.7**	**78.6**	**80.2**	**75.8**	**72.0**	**71.7**	**80.1**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	CC	1	1	91.3	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
XLM (MLM)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
XLM (MLM+TLM)	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
XLM (MLM)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R	CC	1	100	**88.7**	**85.2**	**85.6**	**84.6**	**83.6**	**85.5**	**82.4**	**81.6**	**80.9**	**83.4**	**80.9**	**83.3**	**79.8**	**75.9**	**74.3**	**82.4**

https://peltarion.com/blog/data-science/a-deep-dive-into-multilingual-nlp-models

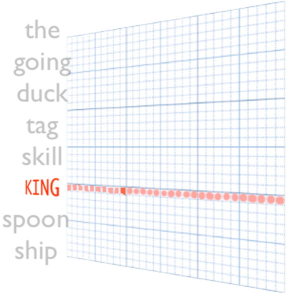
HUAWEI

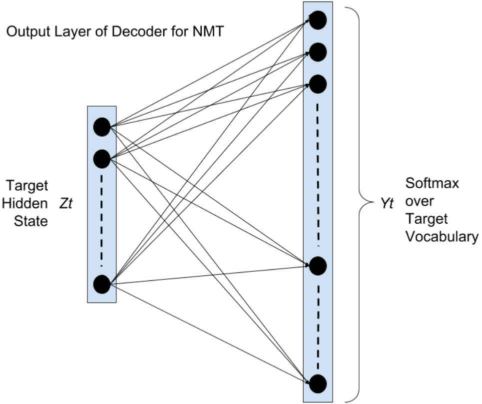
In this talk, we will briefly introduce our work on:

- Training Multilingual PLMs with Byte-Level Subwords
  - A tokenization technique for multilingual PLMs
  - Junqiu Wei, Qun Liu, Yinpeng Guo, Xin Jiang, arXiv:2101.09469 [cs.CL]
- Zero-Shot Paraphrase Generation with Multilingual PLMs
  - Using multilingual PLMs in zero-shot paraphrasing
  - Yinpeng Guo, Yi Liao, Xin Jiang, Qing Zhang, Yibo Zhang, Qun Liu, arXiv:1911.03597 [cs.CL]
- Two Parents, One Child: Dual Transfer for Low-Resource NMT
  - Using multilingual PLMs for low-resource machine translation
  - Meng Zhang, Liangyou Li and Qun Liu, accepted by Findings of ACL-IJCNLP 2021

Content

Word Embeddings in Neural NLP





- In neural NLP, we represent words in a fixed-size vocabulary, either in encoder and in decoder.

HUAWEI

Subword Level Tokenization: BPE or WordPiece

- Although word embedding is successful in neural NLP, the fixed size of vocabulary results in OOV problem.
- Subword level tokenization techniques, like Bype-Pair Encoding (BPE) or WordPiece, solve the OOV problem well.

Merge ops	Byte-pair encoded text
5000	豊 田 駅 (と よ だ え き) は 、 東京都 日 野 市 豊 田 四 丁目 にある
10000	豊 田 駅 (と よ だ えき) は 、 東京都 日 野市 豊 田 四 丁目にある
25000	豊 田駅 (とよ だ えき) は 、 東京都 日 野市 豊田 四 丁目にある
50000	豊 田駅 (とよ だ えき) は 、 東京都 日 野市 豊田 四丁目にある
Tokenized	豊田 駅 （ と よ だ え き ） は 、 東京 都 日野 市 豊田 四 丁目 に ある
10000	豐 田 站 是 東 日本 旅 客 鐵 道 (JR 東 日本) 中央 本 線 的 鐵路 車站
25000	豐田 站是 東日本旅客鐵道 (JR 東日本) 中央 本 線的鐵路車站
50000	豐田 站是 東日本旅客鐵道 (JR 東日本) 中央 本線的鐵路車站
Tokenized	豐田站 是 東日本 旅客 鐵道 （ JR 東日本 ） 中央本線 的 鐵路車站
1000	to y od a ▁station is ▁a ▁r ail way ▁station ▁on ▁the ▁ch ū ō ▁main ▁l ine
3000	to y od a ▁station ▁is ▁a ▁railway ▁station ▁on ▁the ▁ch ū ō ▁main ▁line
10000	toy oda ▁station ▁is ▁a ▁railway ▁station ▁on ▁the ▁ch ū ō ▁main ▁line
50000	toy oda ▁station ▁is ▁a ▁railway ▁station ▁on ▁the ▁chū ō ▁main ▁line
100000	toy oda ▁station ▁is ▁a ▁railway ▁station ▁on ▁the ▁chūō ▁main ▁line
Tokenized	toyoda station is a railway station on the chūō main line

Heinzerling & Strube, BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages, arXiv:1710.02187

While BPE or WordPiece Encounters Large Alphabet Languages

- However, BPE or WordPiece techniques face problems when the alphabet of the languages are very large, especially in multilingual scenario which including languages like Chinese, Japanese and Korean (CJK):
  - The vocabulary have to include a large number of CJK characters, while most of these CJK characters have very low frequencies.
  - This makes the efficiency of the use vocabulary space very low.
  - Even so, OOV problem still exists for some characters, for example, "詒" in "章詒和" is not included in mBERT.

Subword Frequencies in mBERT Vocabulary

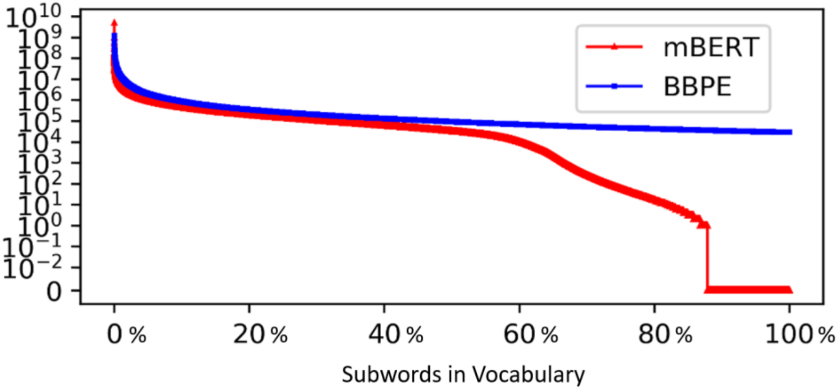


Figure 1: Subwords Frequencies in Google Multilingual Vocabulary (mBERT) and Our Byte-Level Vocabulary (BBPE)

Idea: Further Segment Rare Characters into Bytes with UTF-8

- To solve this problem, the idea is to further rare characters to small pieces, rather than keep all these rare characters in the BPE vocabulary;
- To achieve this, a Byte-Level BPE (BBPE) is proposed:
  - Texts are represented in UTF-8 strings, where each character is encoded in 1-4 bytes;
  - The BPE algorithm is conducted on UTF-8 strings, using bytes as the basic units.

Code point <-> UTF-8 conversion

First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
U+0000	U+007F	0xxxxxxx			
U+0080	U+07FF	110xxxxx	10xxxxxx		
U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
U+10000	[nb 2]U+10FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Byte-Level BPE (BBPE)

- BBPE has been adopted by GPT-2 and RoBERTa;
- Wang et al.(2019) analysis the benefits of using BBPE in neural machine translation;
- Our work:
  - analysis the benefits of using BBPE in multilingual pre-trained language models;
  - impletement an BBPE tool as a part of our NEZHA model.

BBPE tokenization: Examples

Original Text (English)	Anarchism is a political philocophy that advocates self-governed societies based on voluntary institutions .
Our BBPE tokenization	Ana ##r ##chis ##m is a political philocophy that advoc ##ates self - govern ##ed societies based on volunt ##ary institutions .
mBERT WordPiece tokenization	Ana ##rch ##ism is a political philocophy that advocate ##s self - governed societies based on vol## untary institutions .

Original Text (Chinese)	兰叶春葳蕤，桂华秋皎洁。
Our BBPE tokenization	兰 叶 春 0xE891 ##0xB3 0xE895 ##0xA4 ， 桂 华 秋 0xE79A ##0x8E 洁 。
mBERT WordPiece tokenization	兰 叶 春 葳 [UNK] ， 桂 华 秋 皎 洁 。

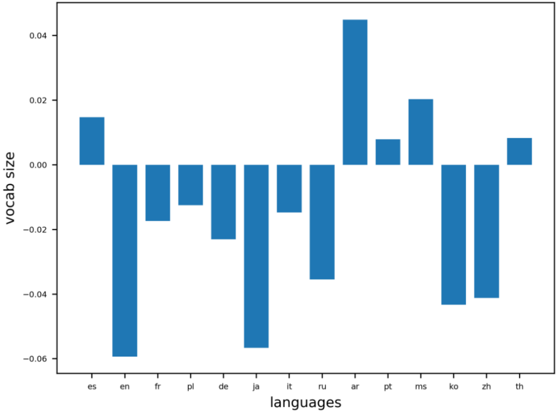
BBPE vs WordPiece: Experimental Results on XNLI

Table 2: Results on *XNLI*

Model	En	Ar	De	Es	Fr	Ru	Th	Avg.	Vocab
BERT (google)	82.4	72.0	76.1	78.4	76.9	74.5	67.1	75.3	120k
NEZHA	83.0	75.9	79.1	81.1	79.3	76.3	68.8	77.6	120k
NEZHA (BPE)	81.2	72.7	77.0	78.8	79.0	73.8	67.3	77.0	100k
NEZHA (BBPE)	83.2	73.9	79.8	81.5	80.8	77.1	71.5	78.3	100k
NEZHA (BBPE)	82.6	74.1	78.3	80.5	79.6	76.1	69.6	77.3	50k

BERT(Google) and NEZHA use Character-Level WordPiece tokenizer.

BBPE vs WordPiece: Distribution of Entries in Vocabulary



- The entries of CJK langauges are much reduced because the rare characters are no longer kept in the vocabulary.
- The entries of Spanish, Arabic, Malay and Thai are increased significantly.

Note: Shared entries for multiple languages are counted multiple times.

HUAWEI

Content

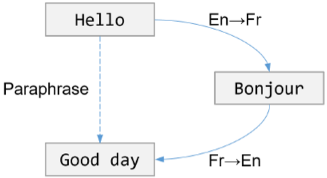
Content

Why zero-shot paraphrase generation?

- Application of paraphrase generation
  - Response diversification in dialogue system
  - Query reformulation in information retrieval
  - Data augmentation
- Conventional methods
  - Based on expensive paraphrase corpora (Prakash et al., 2016; Ziqiang Cao, 2017; Ankush Gupta, 2018; Zichao Li, 2018, 2019)
- Zero-shot paraphrase generation?
  - No paraphrase corpora
  - Translation corpora existed (also synonymous sentences)
  - Pivot-based / round-trip translation

Round-trip translation

- Strengths
  - Leverage tremendous translation corpora
  - Able to do zero-shot paraphrasing



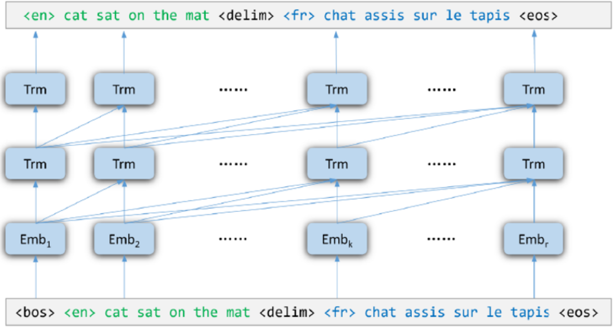
- Weakness
  - Semantic drift
    - pivoted through finite intermedia outputs
    - can hardly explore all paths of paraphrasing
  - Quality determined by MT systems
    - Not easy to be optimized end-to-end for paraphrasing
  - Not efficient
    - Two-pass translation for inference

HUAWEI

Content

Multilingual and cross-lingual language modeling

- Transformer-based Language Model
  - Directly learns paraphrasing distribution, without intermedia translations
  - Single-step end-to-end training
  - Shared parameters across languages
- Cross-lingual language modeling
  - Input concatenation of translation sentence pairs
- Multilingual language modeling
  - Input monolingual sentences from different languages $\rightarrow$ to improve fluency



Language embeddings and denoising auto-encoder

- Language embeddings
  - To guide the generation language
  - Add language-specific embedding to each word embedding

```
<bos> <en> cat sat on the mat <delim> <fr> chat assis sur le tapis <eos>
```

- Denosing auto-encoder
  - To improve generation robustness
  - In source sentences
    - Deletion: randomly delete 1% tokens

      cat <u>sat</u> on the mat → cat on the mat

    - Insertion: insert a random token in 1% random positions

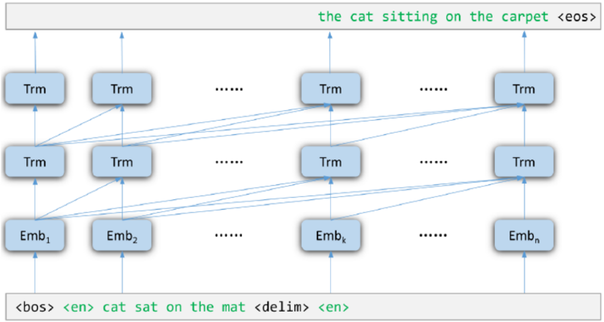
      cat sat on the mat → cat sat on <u>red</u> the mat

    - Reordering: randomly swap 1% tokens

      <u>cat</u> sat on the <u>mat</u> → <u>mat</u> sat on the <u>cat</u>

Zero-shot paraphrasing

- Input:
  - Source language identifier & embeddings
  - Source sentence
  - Target language identifier & embeddings
    - Generated paraphrase sentence
- Output:
  - Input concatenation of translation sentence pairs



Content

Experimental settings

- Datasets
  - MultiUN (Andreas Eisele, 2010)
  - OpenSubtitles (Pierre Lison, 2016)

Table 1: Statistics of training data (#sentences).

	En↔Es	En↔Ru	En↔Zh	Es↔Ru	Es↔Zh	Ru↔Zh
OpenSubtitles	11.7M	11.7M	11.2M	10.5M	8.5M	9.6M
MultiUN	11.4M	11.7M	9.6M	10.6M	9.8M	9.6M
Total	23.1M	23.4M	20.8M	21.1M	18.3M	19.2M

- Basic configurations
  - 12 layers of Transformer blocks, 12 attention heads
  - 768 embedding dimensions, 768 hidden dimensions, 3072 FFN projection dimensions

Evaluation

- Automatic evaluation
  - Consider Relevance and Diversity simultaneously
  - Semantic relevance
    - Cosine similarity between sentential representations

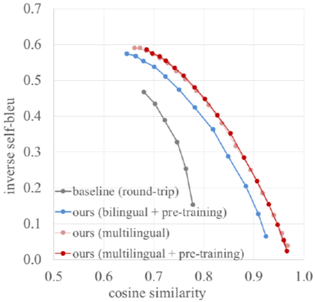
    → Glove-840B word embeddings (Jeffrey Pennington, 2014)

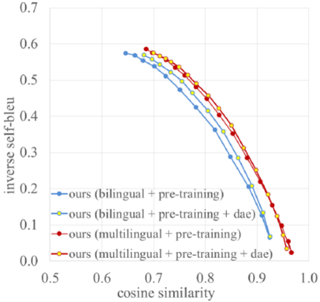
    → Vector Extrema sentential representations (Chia-Wei Liu, 2016)
- Human evaluation

Content

Result

- Upper-right the better
  - Round-trip translation
  - Bilingual: Cross-lingual language modeling w/ only one language pair
  - Multilingual: Cross-lingual language modeling w/ multiple language pairs
  - Pre-training: Monolingual language modeling w/ multiple languages





HUAWEI

Content

Zero-Shot Paraphrase Generation with Multilingual PLMs

- Background
- Method
- Experiments
- Results
- Summary

Summary

- The proposed method performs better than round-trip baseline
- More language pairs benefits the performance
- Denoising auto-encoder further boost the performance
- Multilingual language model pre-training improves fluency

Table 3: Human evaluation results.

Model	Relevance	Fluency	Agreement
Round-trip	2.72	3.61	0.36
Multilingual (ours)	**3.43**	**3.75**	0.35

Table 2: Log-probabilities of the generated sentences. $\surd$ and $\times$ symbols denote learning with or without pre-training respectively, **bold** font denotes greater values.

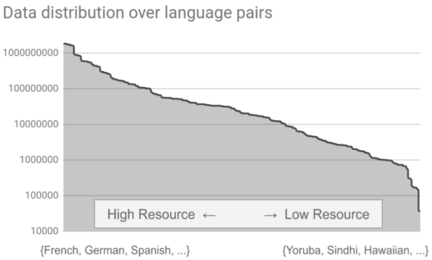
Model	Sampling	Pre-Training	Log-Prob
Multilingual	greedy, temp=1	$\surd$	**-0.1427**
		$\times$	-0.1428
	top-3, temp=1	$\surd$	**-0.1425**
		$\times$	-0.1448
	top-3, temp=1.5	$\surd$	**-0.1420**
		$\times$	-0.1425
Bilingual	greedy, temp=1	$\surd$	**-0.1472**
		$\times$	-0.1484
	top-3, temp=1	$\surd$	**-0.1487**
		$\times$	-0.1502
	top-3, temp=1.5	$\surd$	**-0.1461**
		$\times$	-0.1506

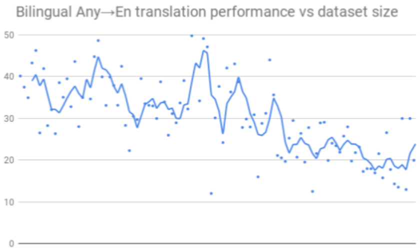
Content

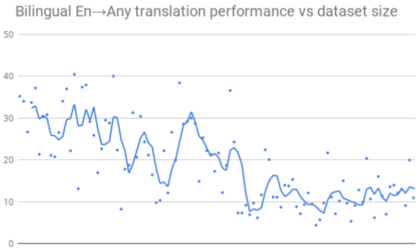
Content

Background

- Neural machine translation has been quite successful in high-resource conditions
- But still suffers in low-resource settings







HUAWEI

Prior work

- Low-resource machine translation commonly uses auxiliary data

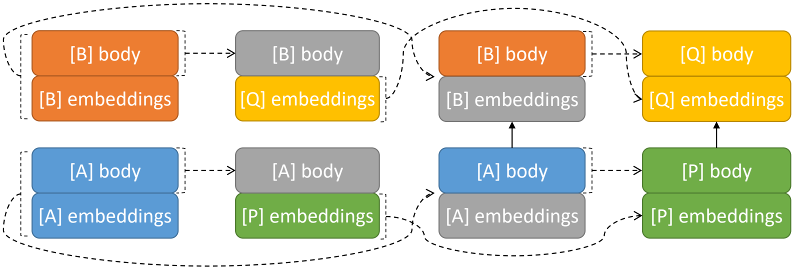
- Using parallel data of high-resource languages: transfer learning
  - (Zoph et al., 2016)
  - (Kim et al., 2019)
- Using monolingual data
  - Back-translation (BT) (Sennrich et al., 2016)
  - Pretrained language model (PLM) (Rothe et al., 2020)
- Multilingual machine translation
  - Multilingual machine translation commonly shares vocabulary, which makes it difficult to extend to new languages (Kocmi and Bojar, 2018)* (asterisk indicates such limitation)

HUAWEI

Content

Approach

--> : initialization
Gray: frozen parameters
Color: trainable parameters



(1) Train PLM on mono-lingual data of A and B separately

(2) Train PLM on mono-lingual data of P and Q separately

(3) Train MT on A→B parallel data

(4) Finetune on P→Q parallel data

HUAWEI

Approach

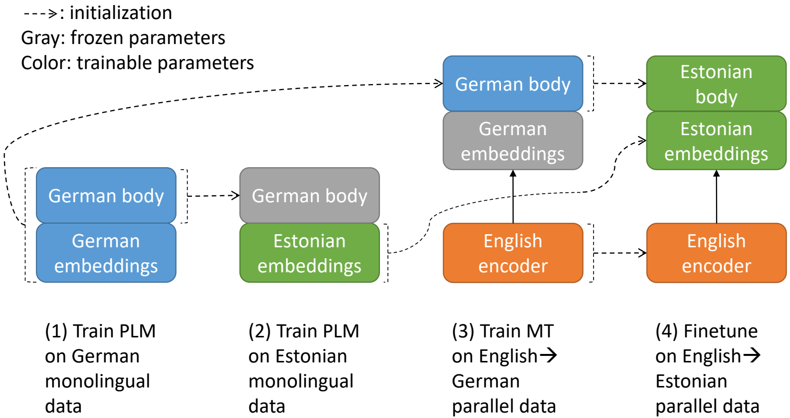
Consider transferring a high-resource A$\rightarrow$B MT model to a low-resource P$\rightarrow$Q

1. Train PLM_A and PLM_B on monolingual data of A and B separately
2. Train PLM_P and PLM_Q on monolingual data of P and Q as follows:
   - 2.1 Initialize PLM_P with PLM_A (except word embeddings); freeze parameters other than word embeddings
   - 2.2 Initialize PLM_Q with PLM_B (except word embeddings); freeze parameters other than word embeddings
3. Train $\text{MT}_{\text{A}\rightarrow\text{B}}$ on A$\rightarrow$B parallel data as follows:
   - 3.1 Initialize MT encoder with PLM_A, and decoder with PLM_B
   - 3.2 Freeze word embeddings during training
4. Replace word embeddings:
   - 4.1 Replace $\text{MT}_{\text{A}\rightarrow\text{B}}$ encoder word embeddings with those in PLM_P
   - 4.2 Replace $\text{MT}_{\text{A}\rightarrow\text{B}}$ decoder word embeddings with those in PLM_Q
5. Finetune on P$\rightarrow$Q parallel data to obtain $\text{MT}_{\text{P}\rightarrow\text{Q}}$

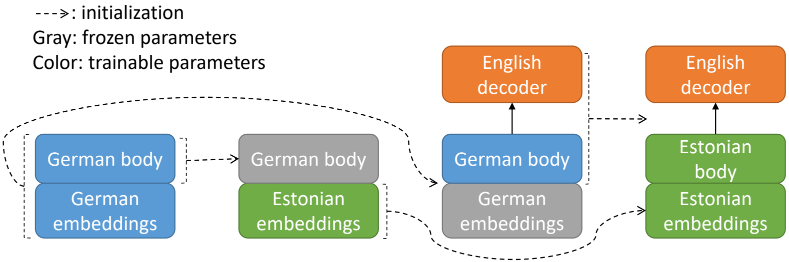
Approach

- If the high-resource and low-resource language pairs share the same target language or source language, then the corresponding PLM can be inherited
  - If B=Q, then PLM_Q is not needed; decoder word embeddings can be adjusted when training $MT_{A \to B}$; PLM_B may also be dispensed with and the decoder can be randomly initialized
  - If A=P, then PLM_P is not needed; encoder word embeddings can be adjusted when training $MT_{A \to B}$; PLM_A may also be dispensed with and the encoder can be randomly initialized
- Symbols may represent specific domains, extending to domain adaptation
  - A=src-lang-src-domain, B=tgt-lang-src-domain, P=src-lang-tgt-domain, Q=tgt-lang-tgt-domain
- The framework is applicable to various network architectures
  - For example, if a low-resource RNN-based NMT is desired, RNN-based PLMs and a high-resource RNN-based NMT can be prepared as parent models

Shared Source Transfer (A=P=en)



Shared Target Transfer (B=Q=en)



(1) Train PLM on German monolingual data

(2) Train PLM on Estonian monolingual data

(3) Train MT on German→English parallel data

(4) Finetune on Estonian→English parallel data

Content

Experiments

- Usage of auxiliary data

	High-resource language		Low-resource language	
	monolingual	parallel	monolingual	parallel
Baseline				✓
(Zoph et al., 2016)		✓		✓
(Kim et al., 2019)		✓	✓	✓
(Kocmi and Bojar, 2018)*		✓		✓
(Rothe et al., 2020)			✓	✓
Ours	✓	✓	✓	✓

HUAWEI

Content

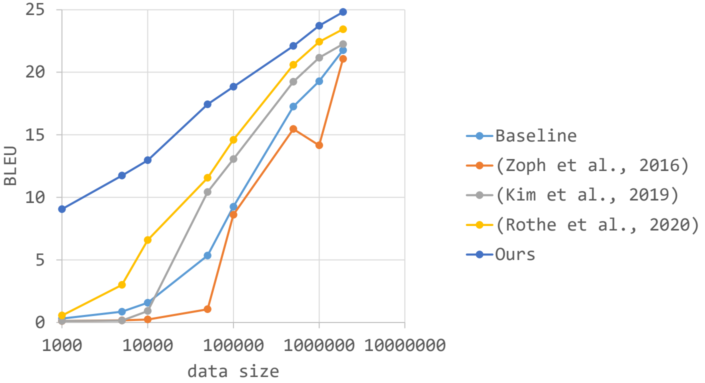
Results

Our approach significantly outperforms strong baselines:

	et-en BLEU
Baseline	21.76
(Zoph et al., 2016)	21.07
(Kim et al., 2019)	22.25
(Kocmi and Bojar, 2018)*	23.58
(Rothe et al., 2020)	23.44
Ours	24.81

Results

Ours performs reasonably well even with a very small amount of parallel data, alleviating the data issue for low-resource language pairs:



Results

Other translation directions:

	tr-en BLEU	en-et BLEU	en-tr BLEU	fr-es BLEU
Baseline	15.44	16.29	9.63	10.59
(Rothe et al., 2020)	19.73	17.36	11.78	18.26
Ours	21.12	19.41	13.18	22.28

Results

Our approach is complementary to back-translation:

	en-et BLEU
Baseline	16.29
Ours	19.41
Baseline + 4m BT	19.78
Ours + 4m BT	21.74
Baseline + 130m BT	20.52
Ours + 130m BT	22.23

Content

Summary

- Our approach to low-resource machine translation transfers knowledge from both pretrained language models and a high-resource neural machine translation model by freezing subsets of parameters during the transfer procedure.
- It significantly outperforms competitors, and possesses several features:
  - It performs reasonably well even with a very small amount of parallel data in the language pair of interest, alleviating the data issue for low-resource language pairs.
  - It is complementary to back-translation, a strong data augmentation approach.
  - It is agnostic to network architectures and thus applicable to any translation models.
  - It is widely applicable to low-resource languages and can be applied to domain adaptation.
  - The same high-resource NMT model can be used to transfer to future low-resource languages, saving computation.

Content

Introduction

Training Multilingual PLMs with Byte-Level Subwords

Zero-Shot Paraphrase Generation with Multilingual PLMs

Dual Transfer for Low-Resource Neural Machine Translation

Conclusion

Content

Thank you!

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI