NLP Project

Low-Resource Neural Machine Translation

AI VIET NAM
Nguyen Quoc Thai

Outline

- Introduction
- Pre-trained LMs: mBART50, mT5
- Back-Translation

Introduction

Translate a sentence $w^{(s)}$ in a **source language (input)** to a sentence $w^{(t)}$ in the **target language (output)**

Introduction

! Translate a sentence $w^{(s)}$ in a **source language (input)** to a sentence $w^{(t)}$ in the **target language (output)**

- Can be formulated as an optimization problem:

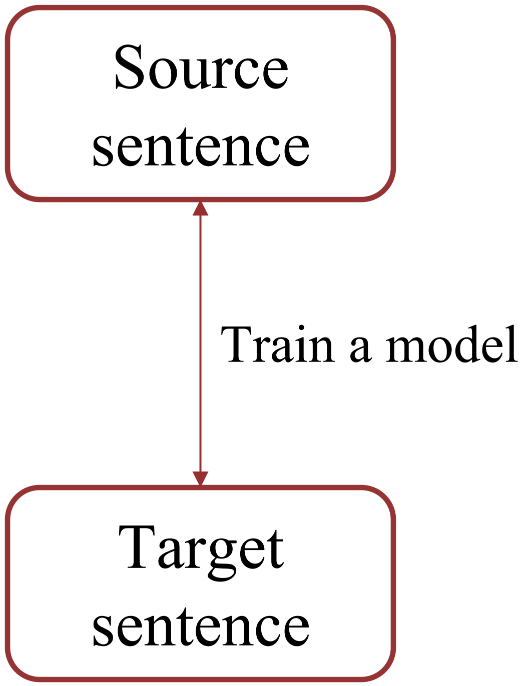
$$\hat{w}^{(\mathrm{t})} = \underset{w^{(t)}}{\operatorname{argmax}}\, \theta(w^{(s)}, w^{(t)})$$

Where θ is a scoring function over source and target sentences

- Requires two components:
  - Learning algorithm to compute parameters of θ
  - Decoding algorithm for computing the best translation $\hat{w}^{(\mathrm{t})}$

Introduction

Translate a sentence $w^{(s)}$ in a **source language (input)** to a sentence $w^{(t)}$ in the **target language (output)**



Introduction

Low-resource Machine Translation

Source sentence

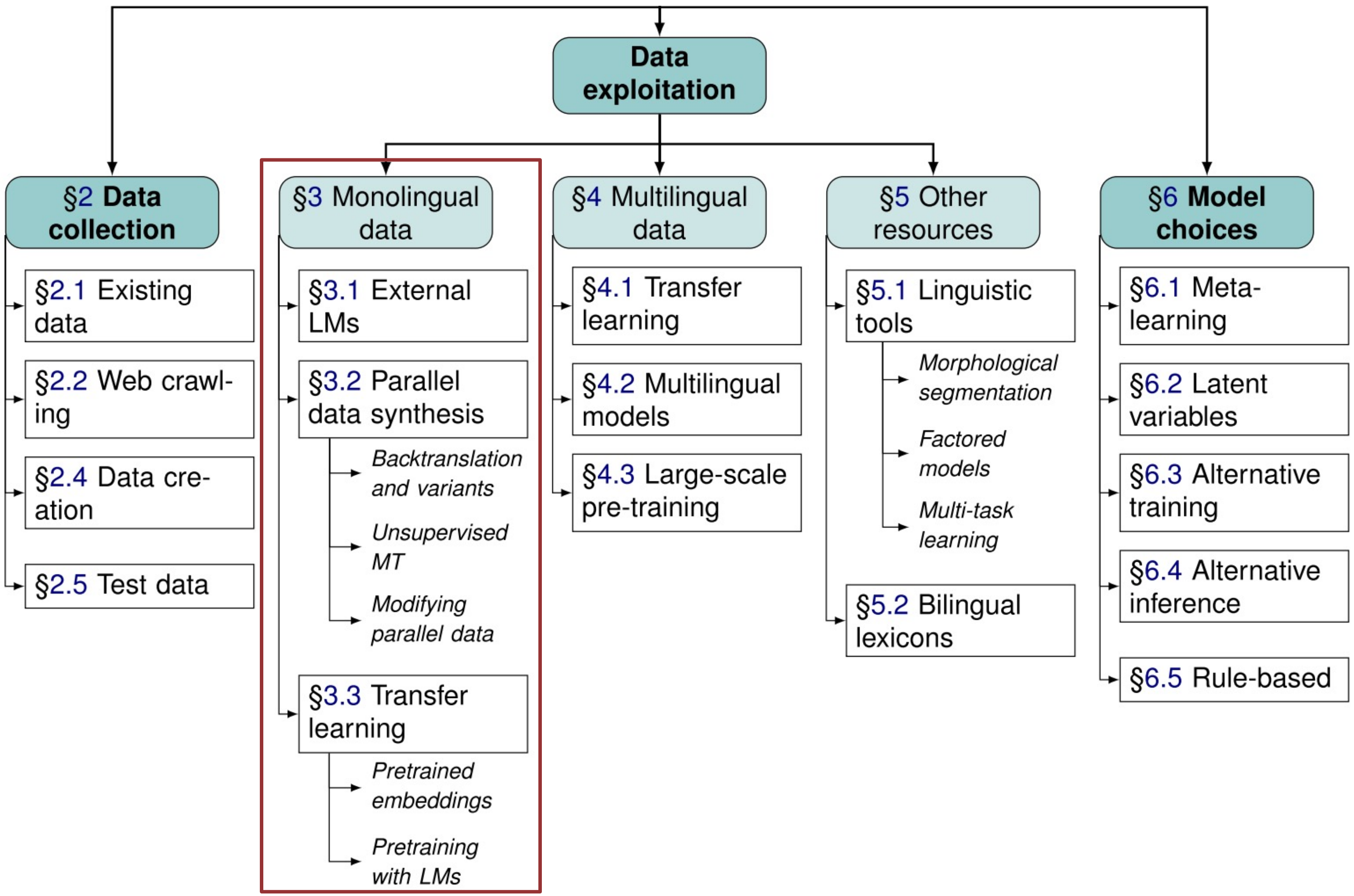
Train a model

Target sentence

Language Pair	Parallel Sentence
En-De	800M
En-Ko	500M
En-Vi	0.17M
De-Vi	0.05M

Introduction

Low-resource Machine Translation



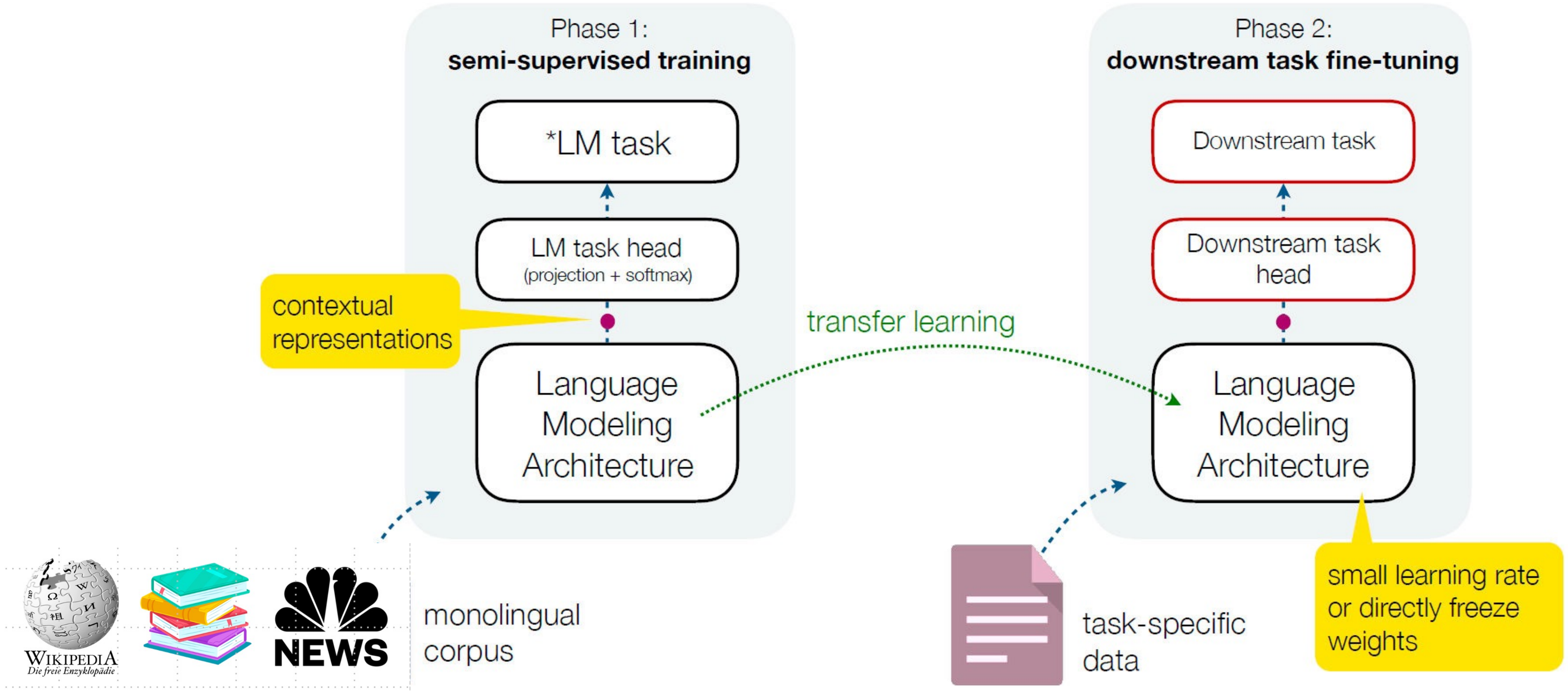
[Source]

Outline

- Introduction
- Pre-trained LMs: mBART50, mT5
- Back-Translation

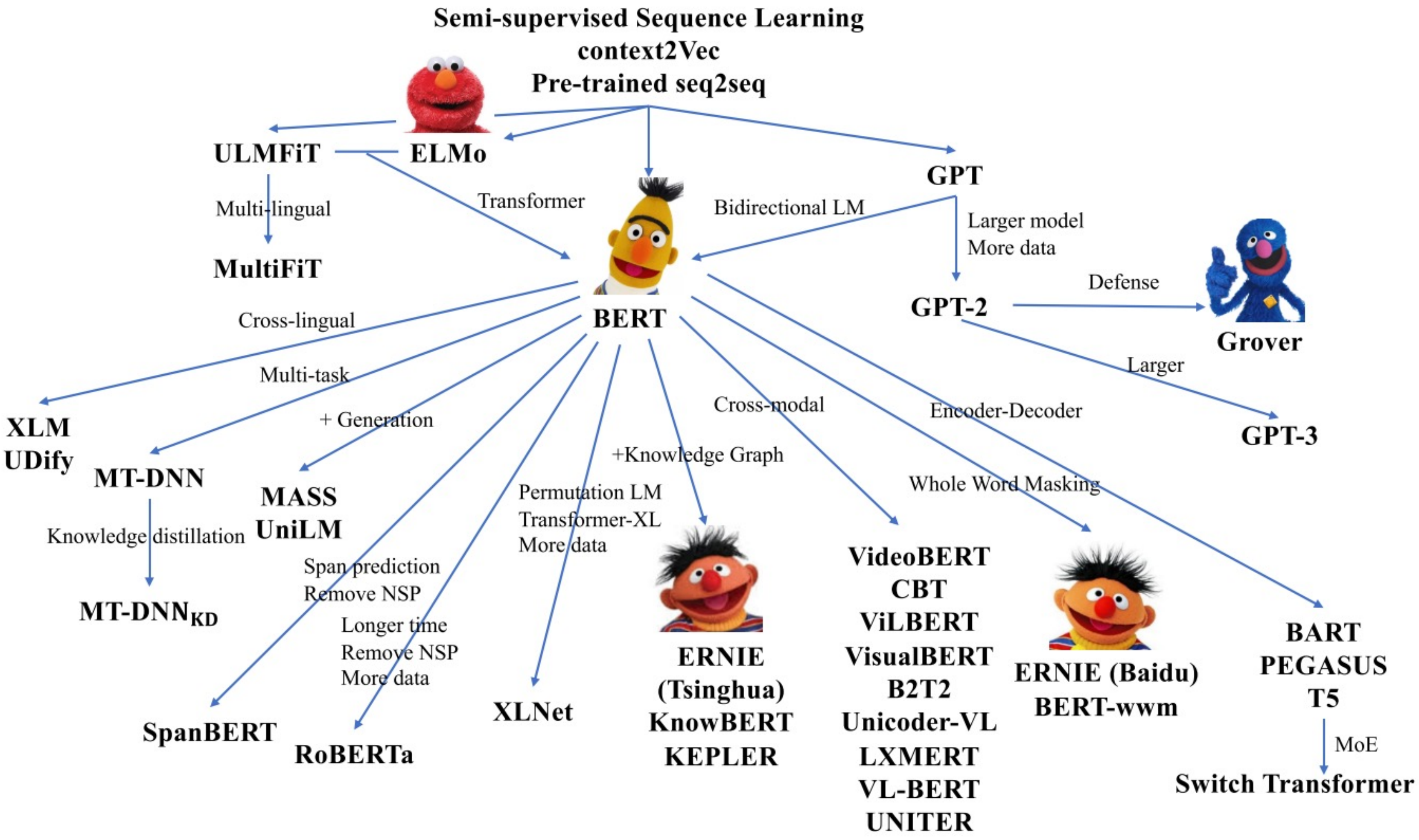
Pre-trained LMs

Pre-trained LMs



Pre-trained LMs

Pre-trained LMs

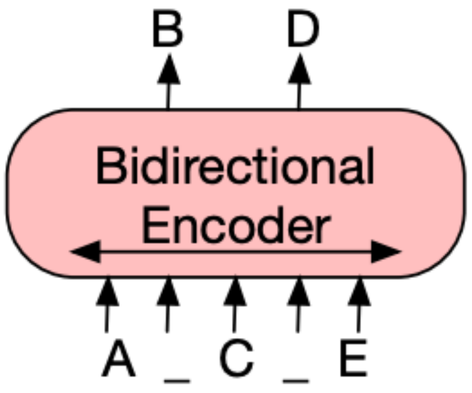


Source

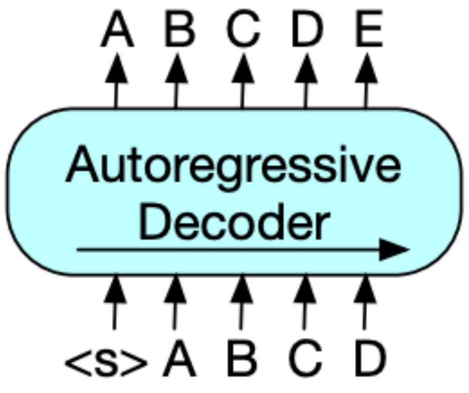
Pre-trained LMs

Pre-trained LMs

- BERT and GPT: a great catalyst for NLU
- But, less successful for sequence-to-sequence tasks: machine translation, text summarization,…



Missing tokens are predicted independently, soBERT cannot easily be used for generation

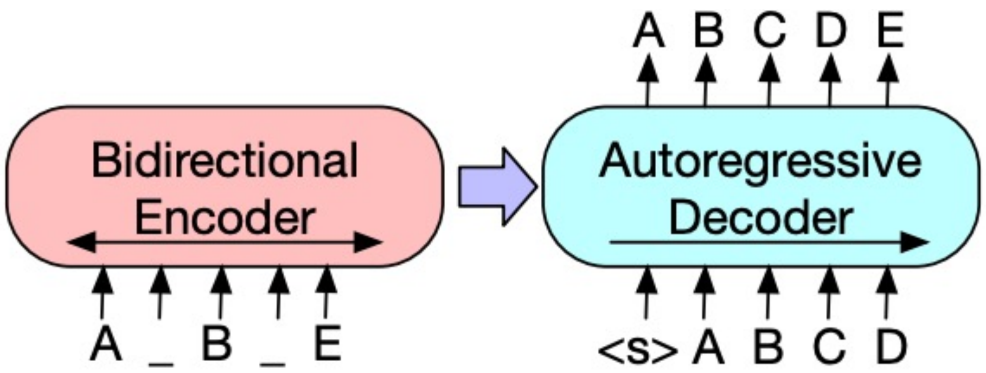


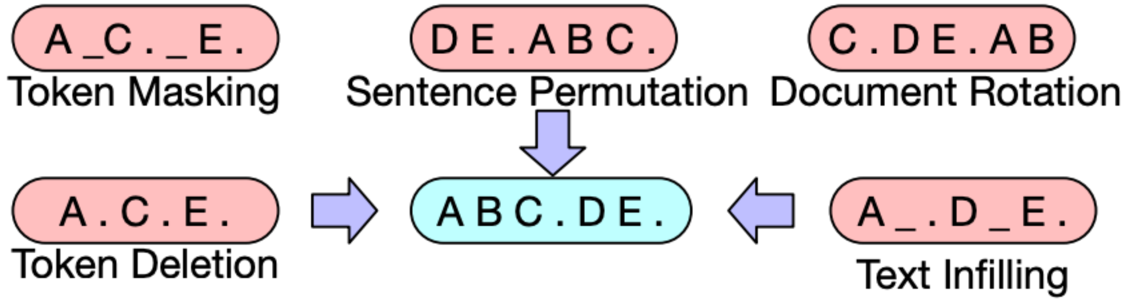
Tokens can only condition on leftward context, so it cannot learn bidirectional interactions

Pre-trained LMs

BART

- BART (Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation and Comprehension.

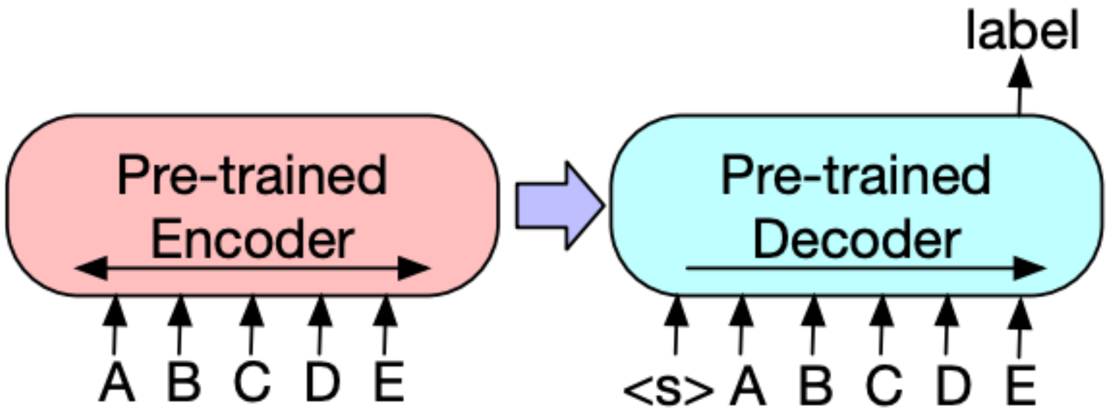




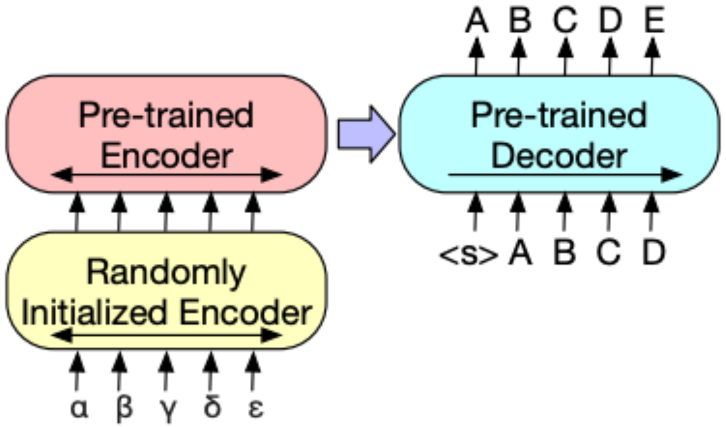
Pre-trained LMs

BART

- BART (Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation and Comprehension.
- Fine-Tuning



Classification Task



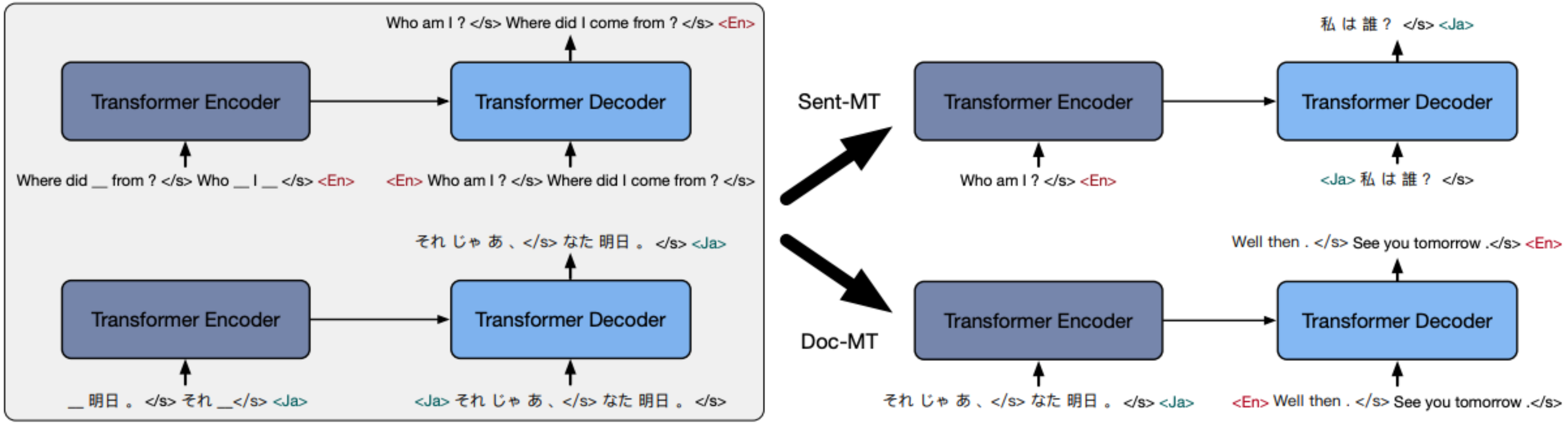
Machine Translation Task

AI VIET NAM
@aivietnam.edu.vn

Pre-trained LMs

BART

➢ mBART: Multilingual Denoising Pre-Training



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

Pre-trained LMs

BART

- mBART50: Multilingual Translation with Extensible Multilingual Pretraining

Data size	Languages
10M+	German, Czech, French, Japanese, Spanish, Russian, Polish, Chinese
1M - 10M	Finnish, Latvian, Lithuanian, Hindi, Estonian
100k to 1M	Tamil, Romanian, Pashto, Sinhala, Malayalam, Dutch, Nepali, Italian, Arabic, Korean, Hebrew, Turkish, Khmer, Farsi, Vietnamese, Croatian, Ukrainian
10K to 100K	Thai, Indonesian, Swedish, Portuguese, Xhosa, Afrikaans, Kazakh, Urdu, Macedonian, Telugu, Slovenian, Burmese, Georgia
10K-	Marathi, Gujarati, Mongolian, Azerbaijani, Bengali

Pre-trained LMs

BART

- mBART50
- PhoMT Dataset

Model	# Params	Pretrained	Finetuned Dataset	Finetuned # pairs	En-Vi	Vi-En
M2M100	1.2B	-	CCMatrix + CCAligned	7.5B	35.83	31.15
Google Translate	-	-	-		39.86	35.76
Bing Translator	-	-	-		40.37	35.74
Transformer-base	65M	-	PhoMT	3M	42.12	37.19
Transformer-big	213M	-	PhoMT	3M	42.94	37.83
mBART†	448M	CC25	PhoMT	3M	43.46	39.78
EnViT5-base	275M	CC100	MTet	4.2M	43.87	39.57
			MTet + PhoMT	6.2M	**45.47**	**40.57**

Pre-trained LMs

BART

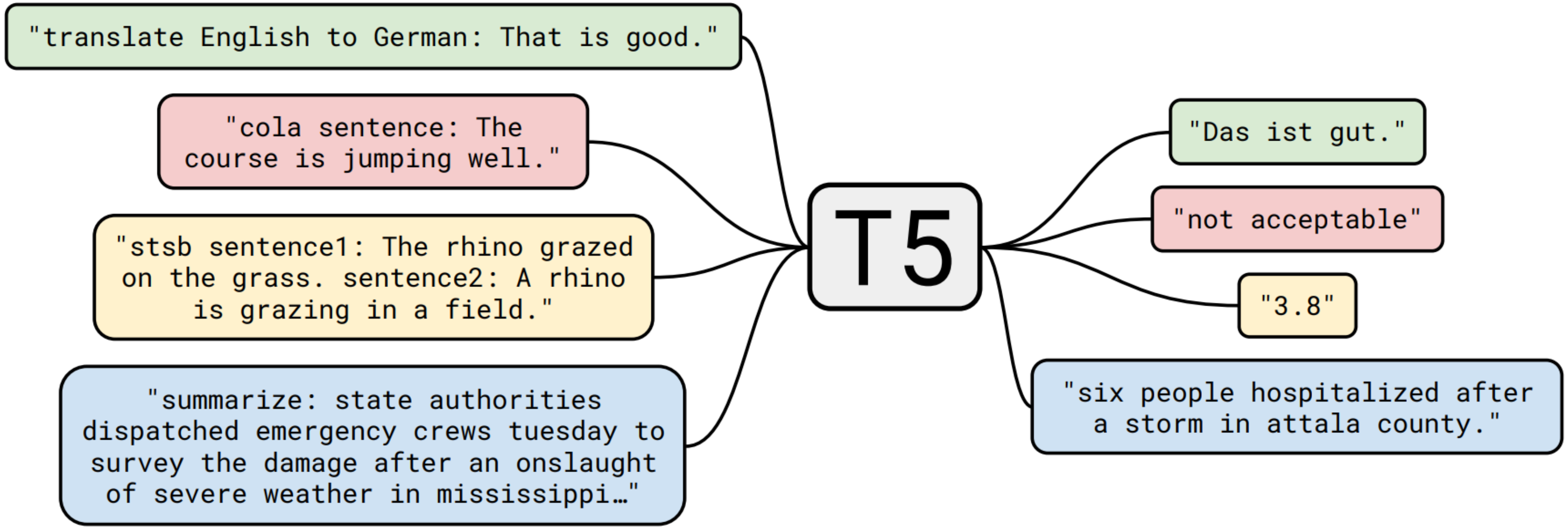
- mBART50
- PhoMT Dataset

Model	Validation set				Test set					
	En-to-Vi		Vi-to-En		En-to-Vi			Vi-to-En		
	TER↓	BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	Human↑	TER↓	BLEU↑	Human↑
Google Translate	45.86	40.10	44.69	36.89	46.52	39.86	23/100	45.86	35.76	10/100
Bing Translator	45.36	40.82	45.32	36.61	46.04	40.37	14/100	46.09	35.74	15/100
Transformer-base	42.77	43.01	43.42	38.26	43.79	42.12	13/100	44.28	37.19	13/100
Transformer-big	42.13	43.75	43.08	39.04	43.04	42.94	18/100	44.06	37.83	28/100
mBART	**41.56**	**44.32**	**41.44**	**40.88**	**42.57**	**43.46**	**32/100**	**42.54**	**39.78**	**34/100**

Pre-trained LMs

T5

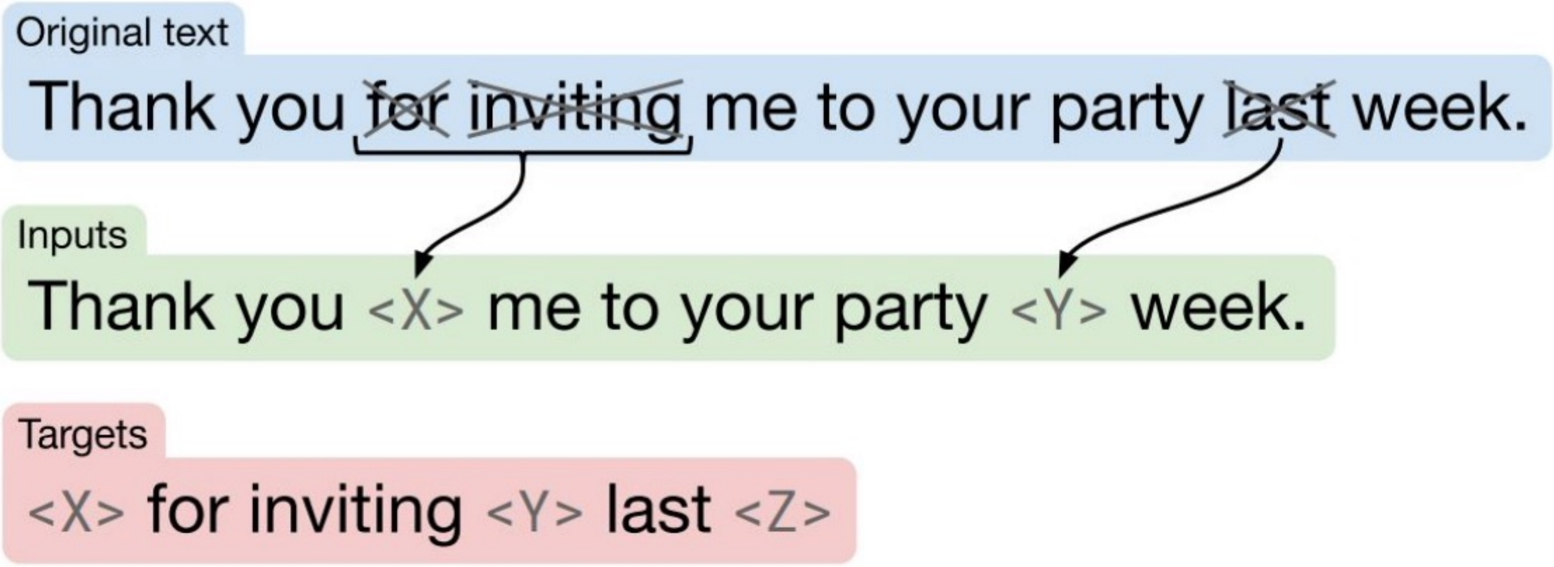
- T5 (Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer)
- Every task, one format!
- ["Task-specific prefix]: [Input text]" => "[Output text]"



Pre-trained LMs

T5

- Baseline Objective

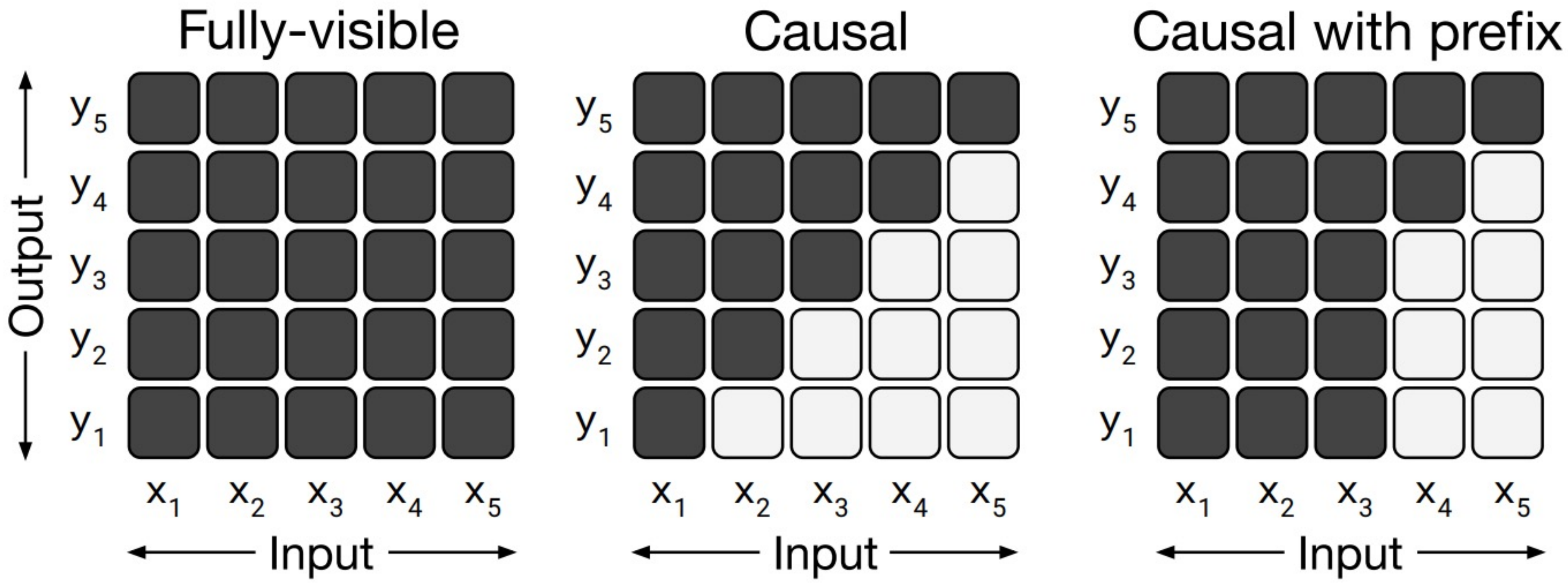


AI VIET NAM
@aivietnam.edu.vn

Pre-trained LMs

T5

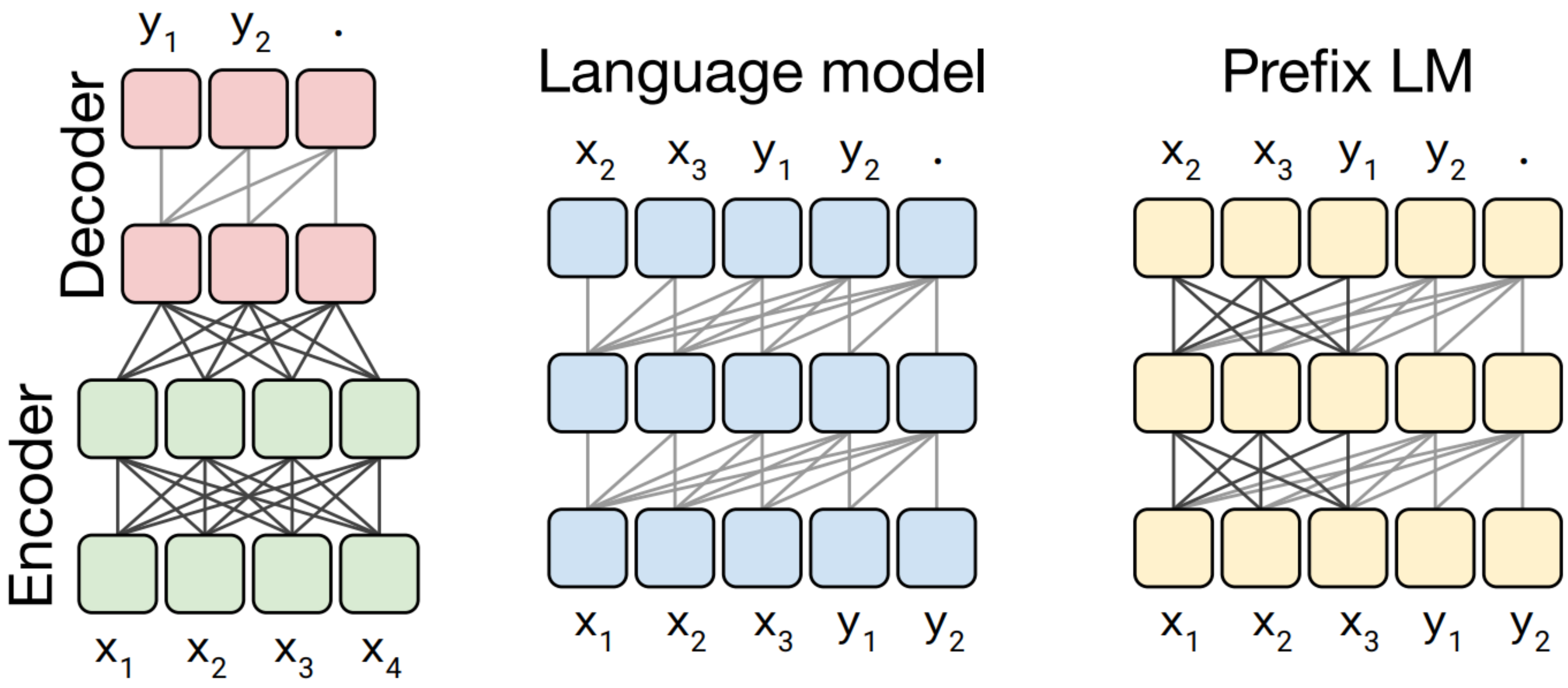
➤ Different Attention Mask Patterns



Pre-trained LMs

T5

➢ Transformer Architecture Variants

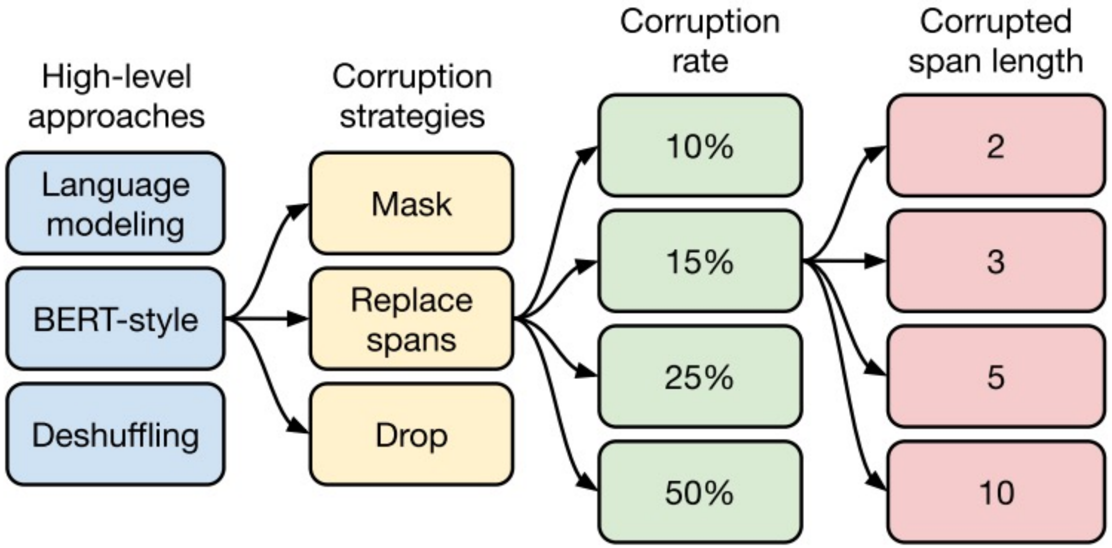


Pre-trained LMs

T5

➢ Different Unsupervised Objectives

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style Devlin et al. (2018)	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*

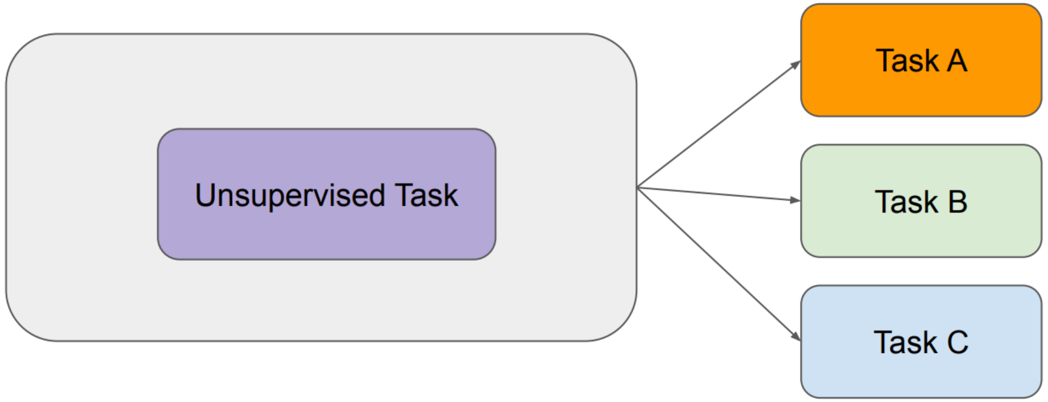


Pre-trained LMs

T5

- Multi-task Learning

Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	**40.13**	**28.04**

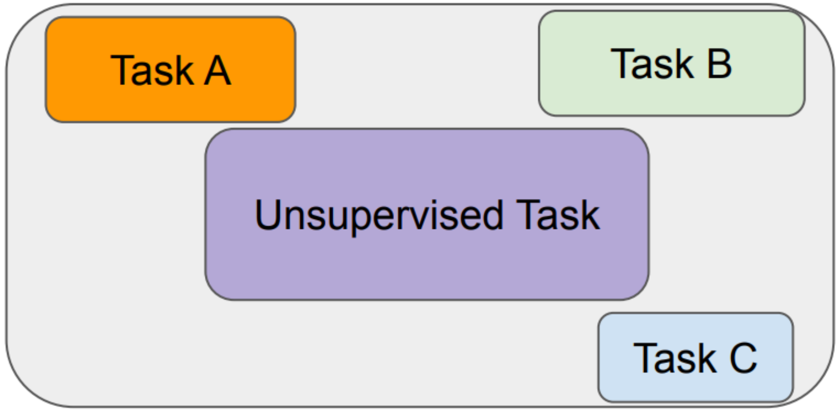


Pre-trained LMs

T5

- Multi-task Learning

Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	**40.13**	**28.04**

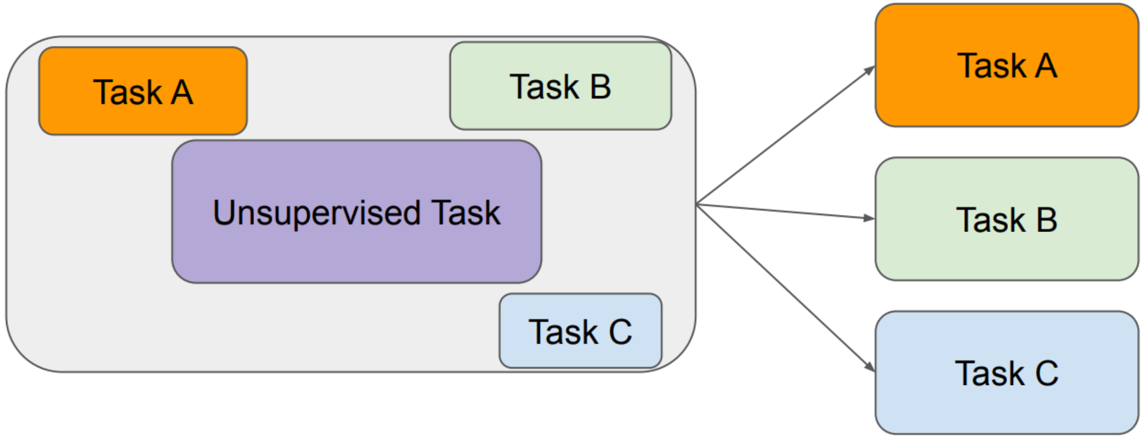


Pre-trained LMs

T5

- Multi-task Learning

Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	**40.13**	**28.04**

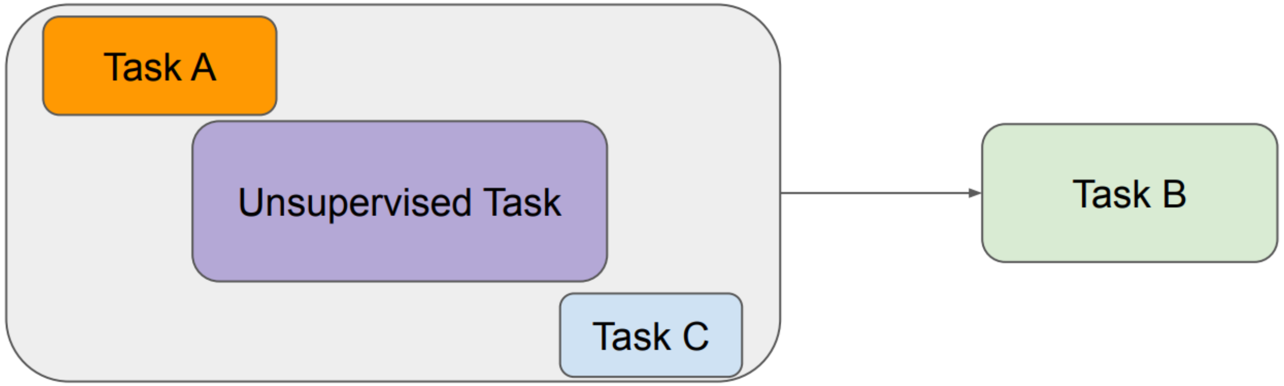


Pre-trained LMs

T5

- Multi-task Learning

Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	**40.13**	**28.04**

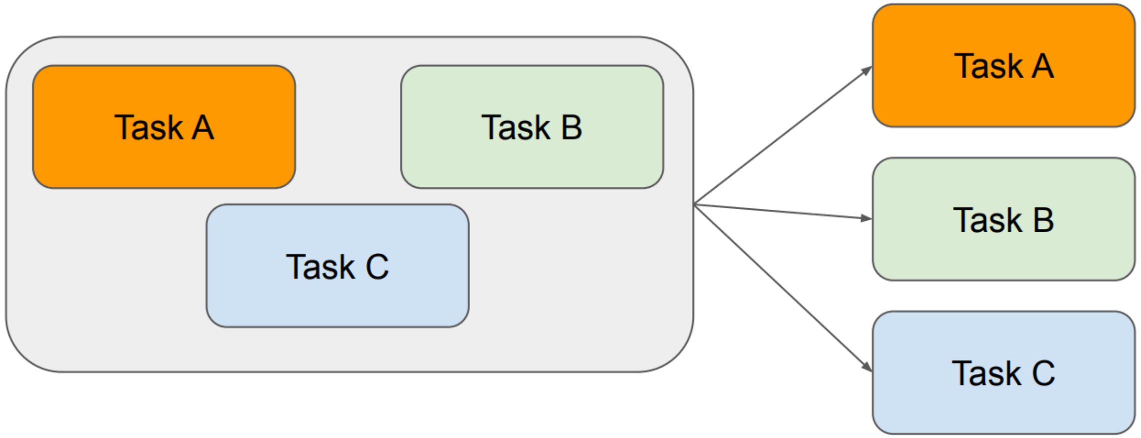


Pre-trained LMs

T5

➢ Multi-task Learning

Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	**40.13**	**28.04**



Pre-trained LMs

T5

- mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer

Model	Architecture	Parameters	# languages	Data source
mBERT (Devlin, 2018)	Encoder-only	180M	104	Wikipedia
XLM (Conneau and Lample, 2019)	Encoder-only	570M	100	Wikipedia
XLM-R (Conneau et al., 2020)	Encoder-only	270M – 550M	100	Common Crawl (CCNet)
mBART (Lewis et al., 2020b)	Encoder-decoder	680M	25	Common Crawl (CC25)
MARGE (Lewis et al., 2020a)	Encoder-decoder	960M	26	Wikipedia or CC-News
mT5 (ours)	Encoder-decoder	300M – 13B	101	Common Crawl (mC4)

Pre-trained LMs

Pre-trained LMs

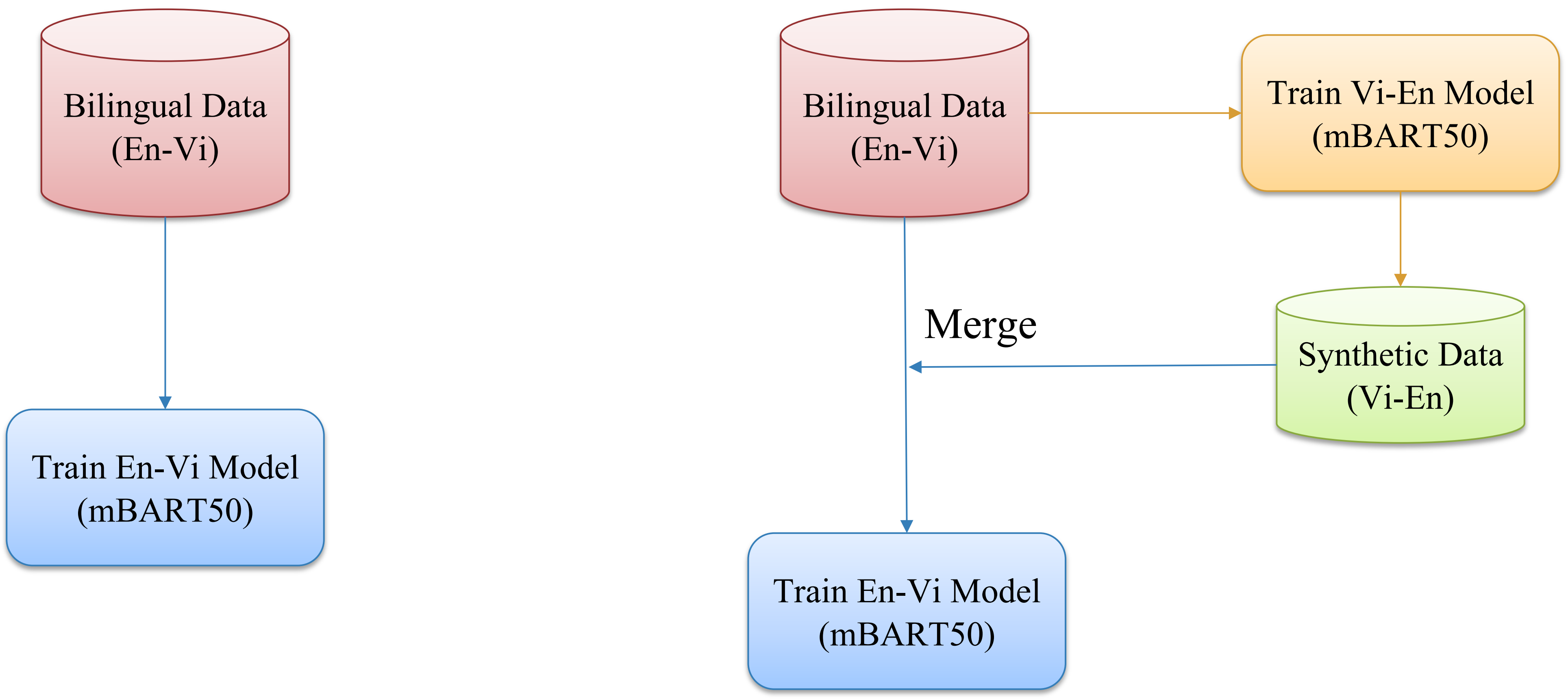
```
# MBart50TokenizerFast.from_pretrained(model_name,
      src_lang="en_XX",tgt_lang = "vi_VN")
model_name = "facebook/mbart-large-50-many-to-many-mmt"
tokenizer = MBart50TokenizerFast.from_pretrained(model_name)
model = AutoModelForSeq2SeqLM.from_pretrained(model_name)


# prefix: translate English to Vietnamese
model_name = ”google/mt5-base"
tokenizer = T5TokenizerFast.from_pretrained(model_name)
model = AutoModelForSeq2SeqLM.from_pretrained(model_name)
```

Outline

- Introduction
- Pre-trained LMs: mBART50, mT5
- Back-Translation

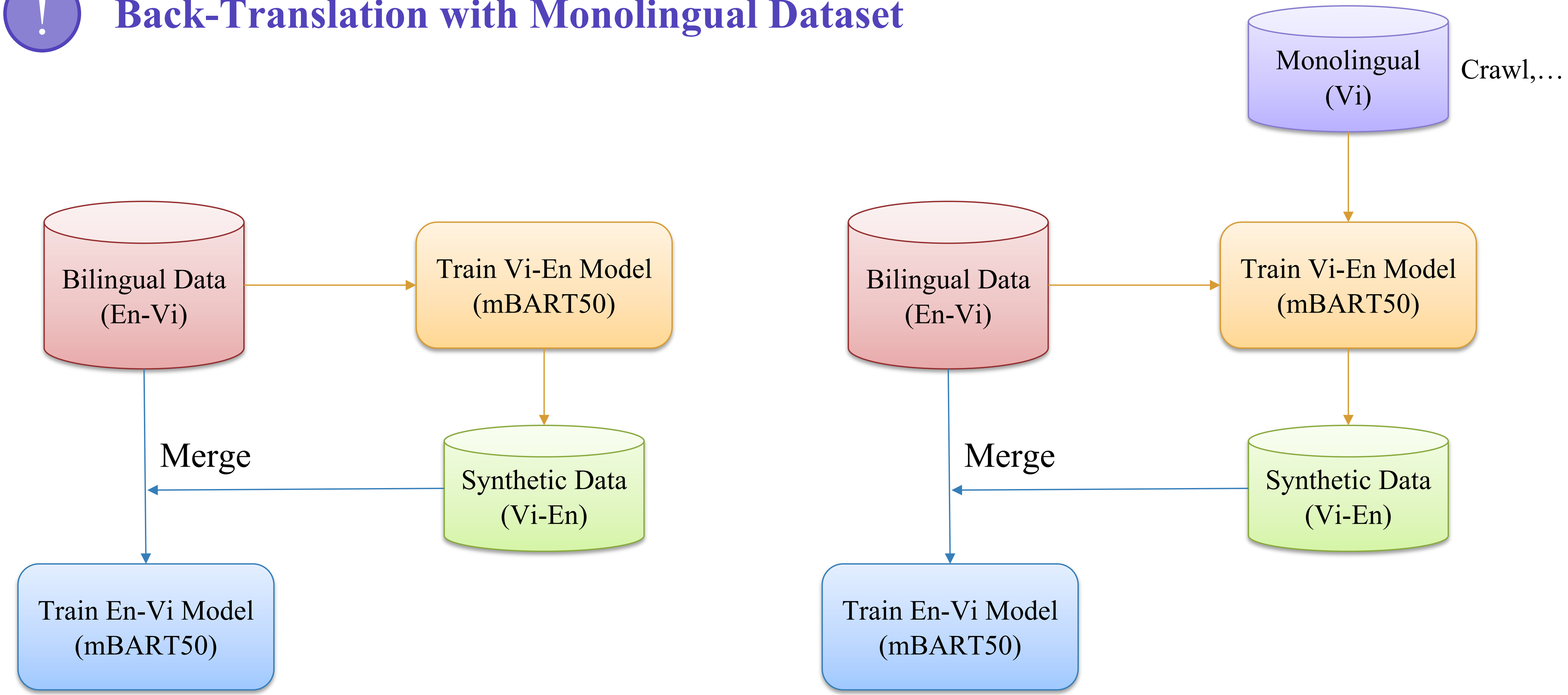
Back-Translation

Back-Translation Technique



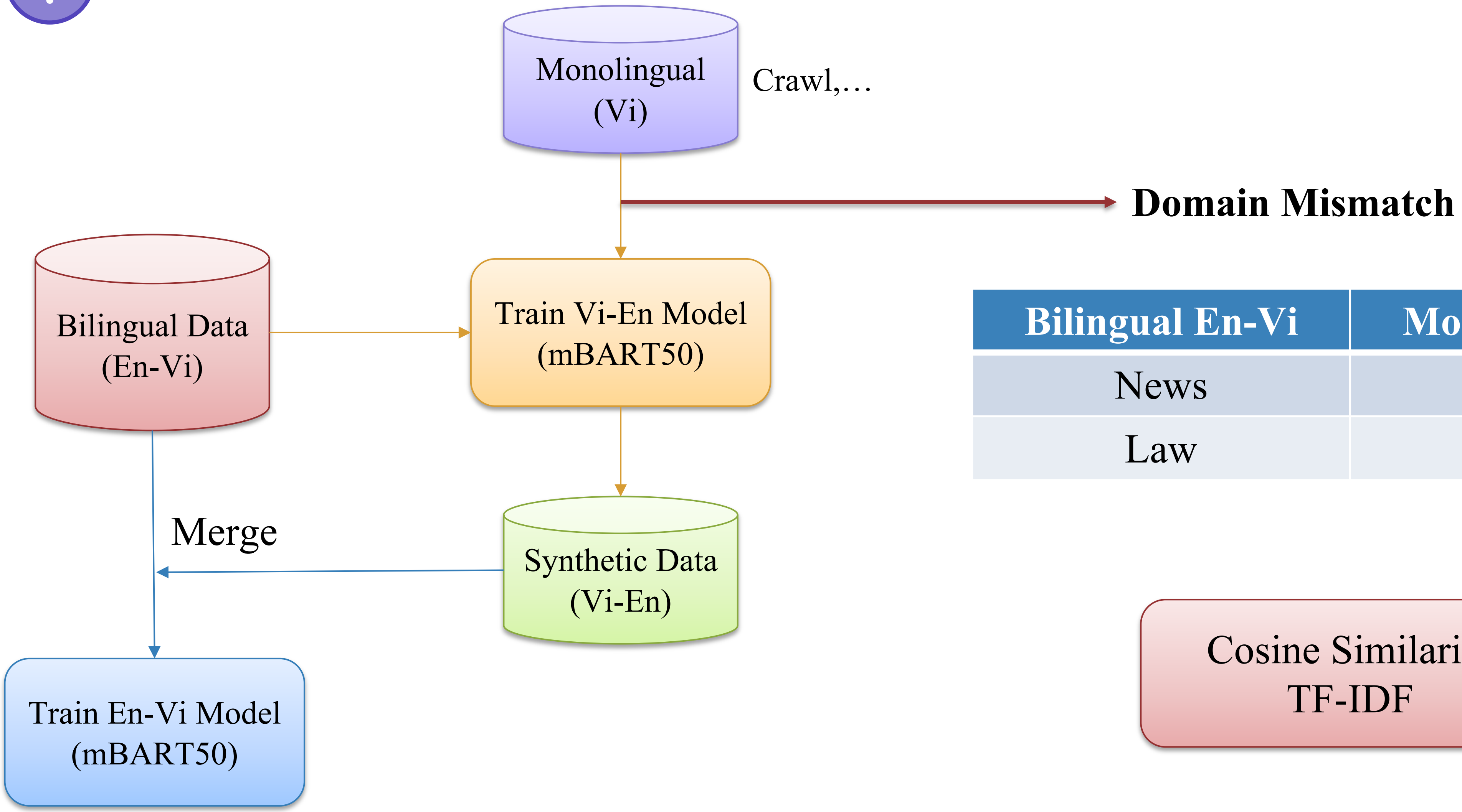
Back-Translation

Back-Translation with Monolingual Dataset



Back-Translation

Data Selection



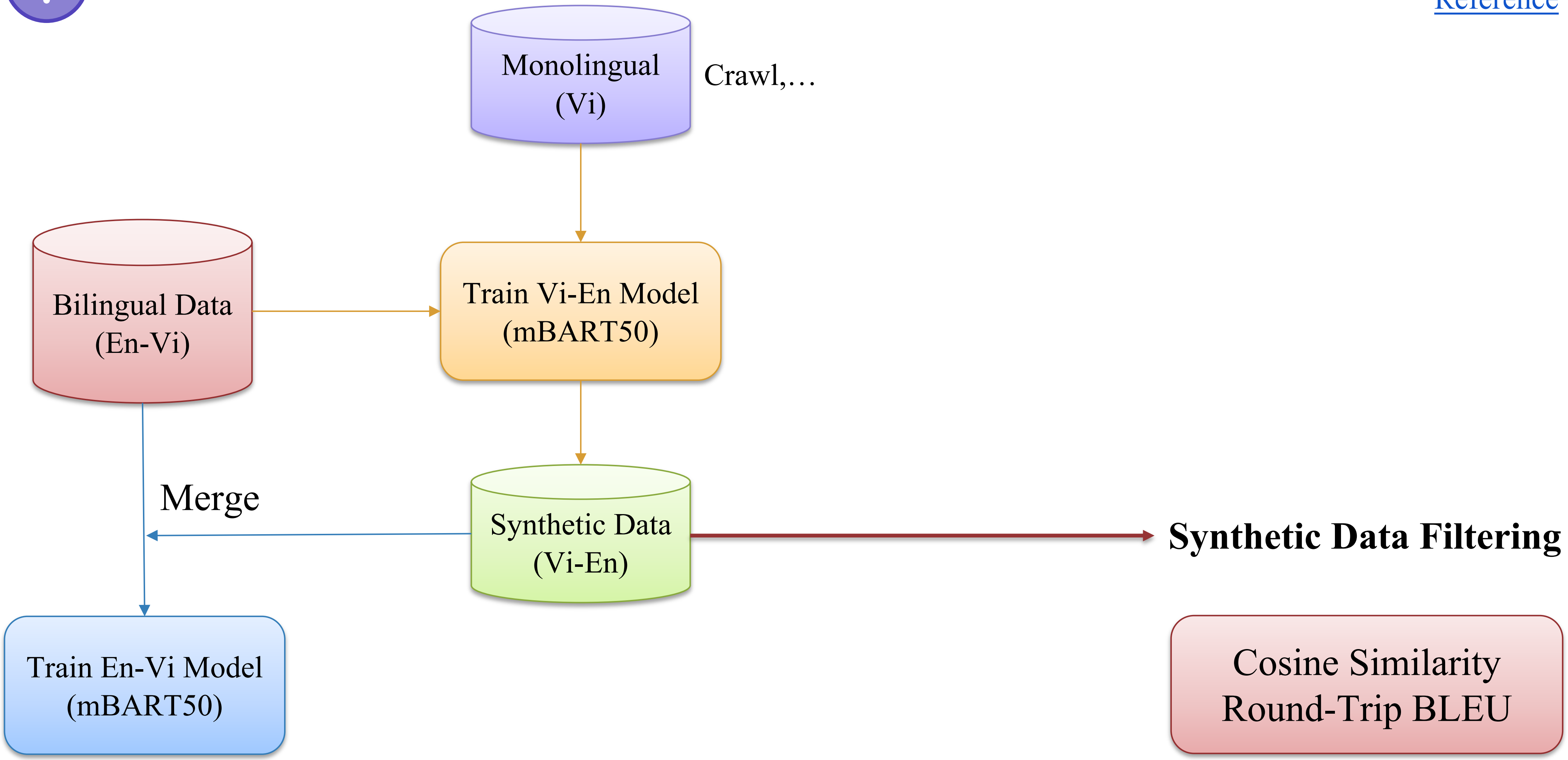
Bilingual En-Vi	Monolingual Vi
News	Medical
Law	News

Cosine Similarity
TF-IDF

Back-Translation

Data Selection

Reference



Back-Translation

Experiment

- **Dataset: IWSLT'15 English-Vietnamese**

Training: 133 317 Validation: 1 553 Test: 1 269

Experiment	Model	ScareBLEU
#1	Standard Transformer (Greedy Search)	24.66
#2	BERT-to-BERT (Greedy Search)	25.41
#3	BERT-to-GPT2 (Greedy Search)	23.56
#4	mBART50	34.87
#5	Back-Translation (Monolingual)	35.22

Thanks!

Any questions?