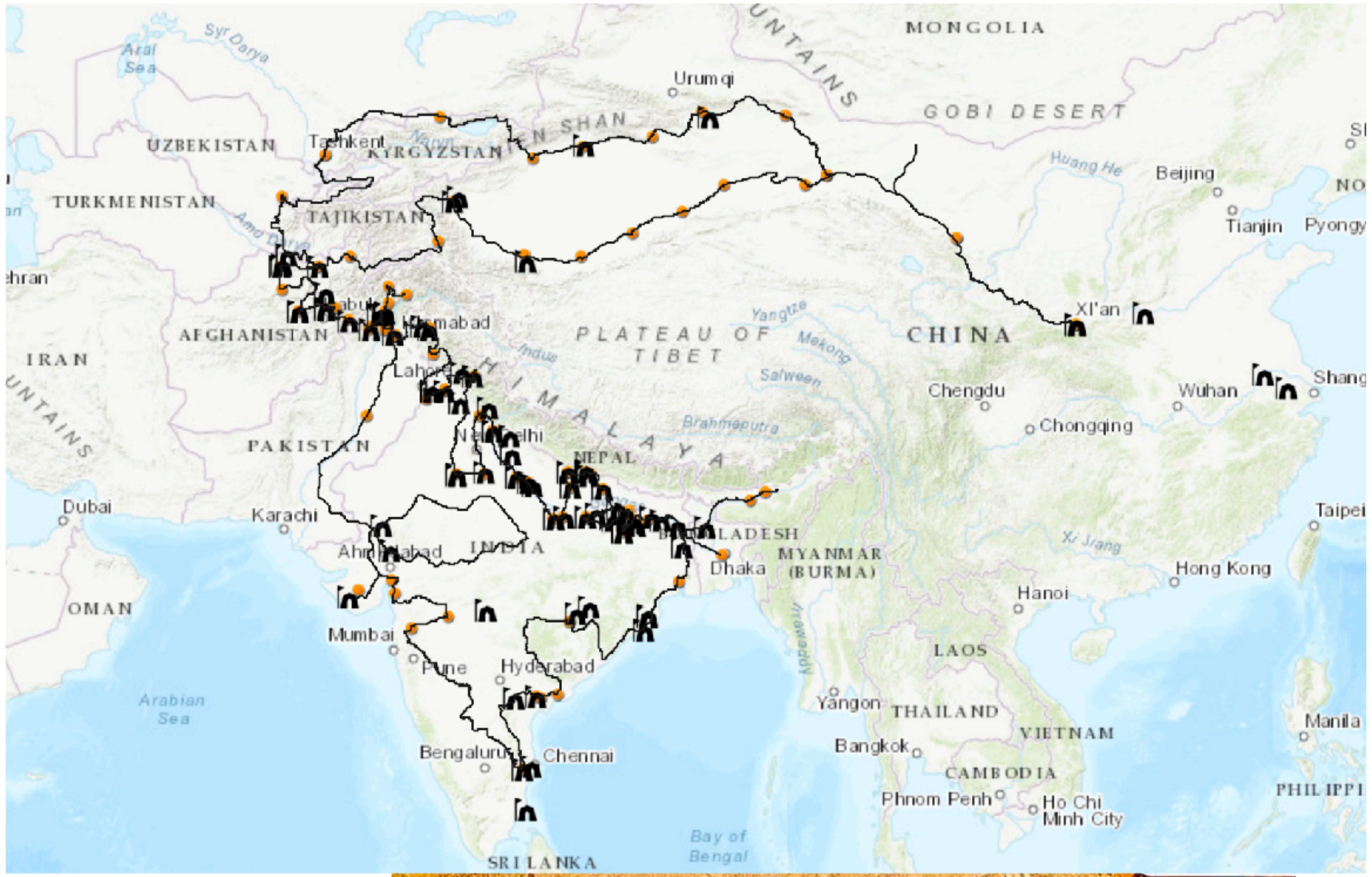
Breaking the Language Barrier with Neural Machine Translation

Lei Li
University of California Santa Barbara
leili@cs.ucsb.edu
10/12/2022

Once upon a time …

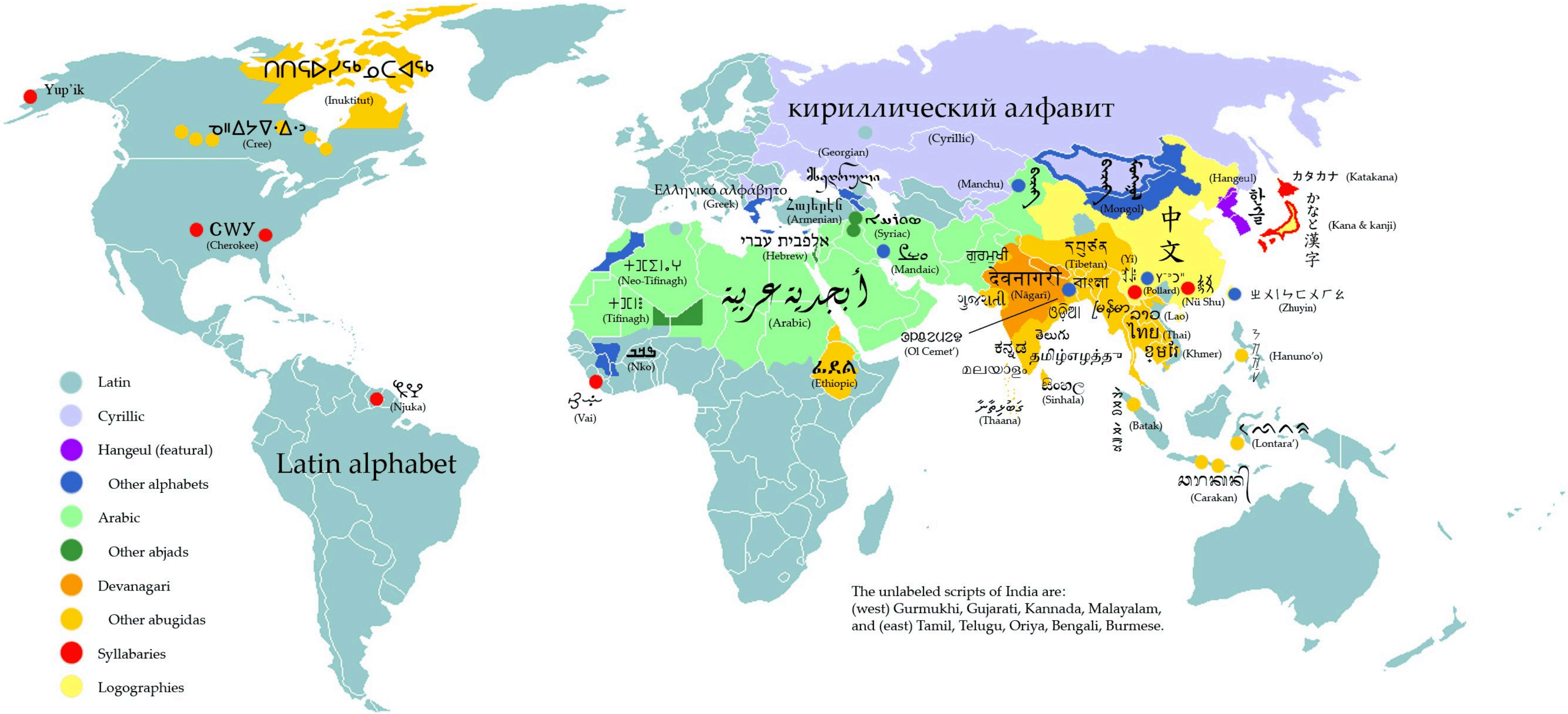
- Septuagint, translated from Hebrew Bible to Greek, mid 3rd century BCE
- Translating Buddhist texts written in Sanskrit to Chinese
  – Kumārajīva (कुमारजीव), 344-413 CE, translated 35-74 books
  – Xuanzang 602-664 CE, travel from Ancient China to India in 17 years, translated 75 books from Sanskrit to Chinese



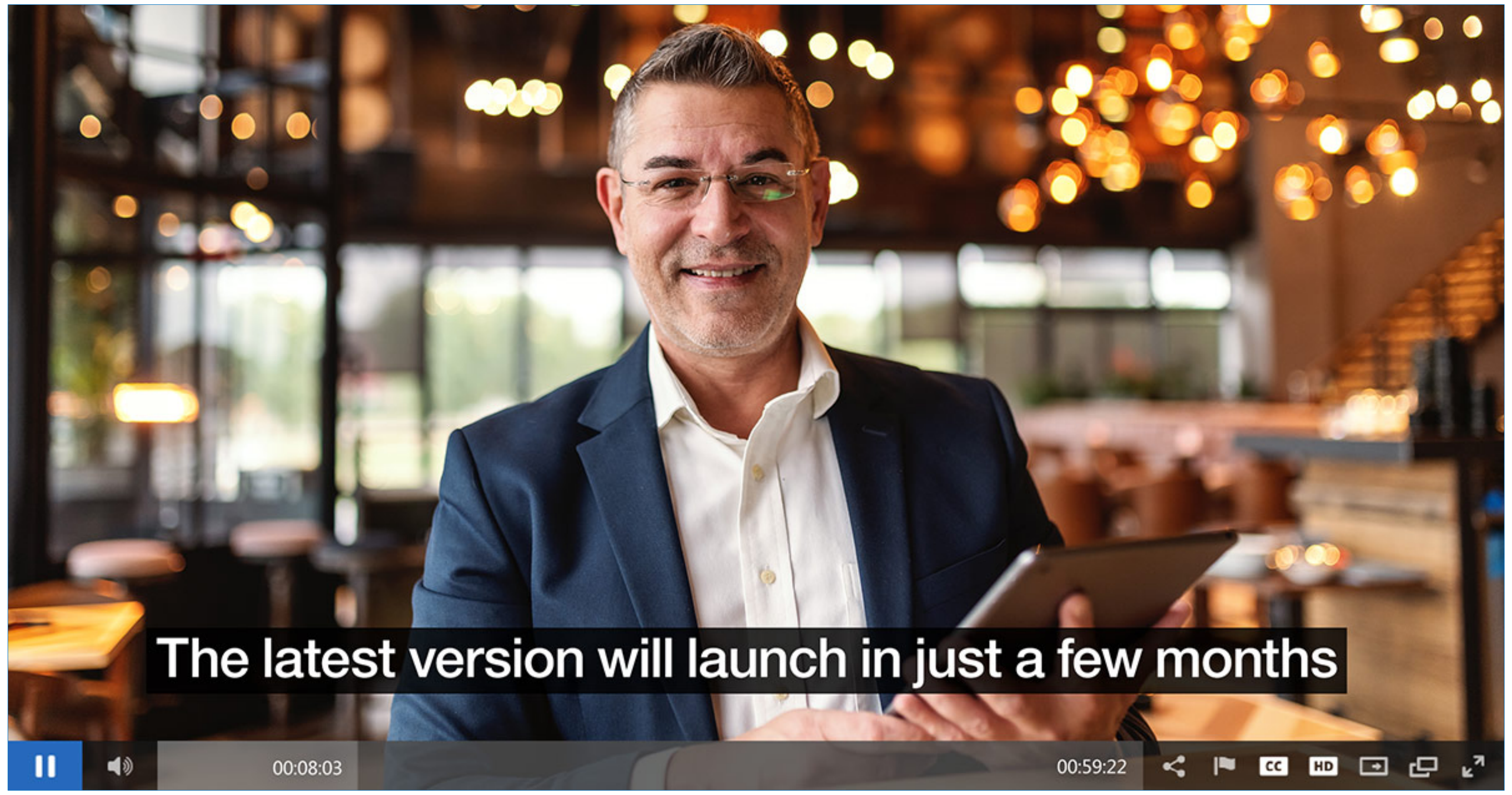
Xuanzang travelling, Dunhuang mural, China

7000 languages around the world

How to communicate efficiently across languages? Machine Translation



Cross Language Barrier with Machine Translation



Foreign Media

Global Conferences

Tourism

International Trade

When you really need Machine Translation

- Rimi Natsukawa live streaming on Tiktok July, 2021


INA 0 5
CHN 0 10
TOKYO 2020
TOKYO 2020
BWF
5-10
5-10
TOKYO 2020
TOKYO 2020

Machine Translation has increased international trade by over 10%

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 65, No. 12, December 2019, pp. 5449–5460
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Does Machine Translation Affect International Trade? Evidence from a Large Digital Platform

Erik Brynjolfsson,[a] Xiang Hui,[b] Meng Liu[b]

[a] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [b] Marketing, Olin School of Business, Washington University in St. Louis, St. Louis, Missouri 63130

Contact: erikb@mit.edu, http://orcid.org/0000-0002-8031-6990 (EB); hui@wustl.edu, http://orcid.org/0000-0001-7595-3461 (XH); mengl@wustl.edu, http://orcid.org/0000-0002-5512-7952 (ML)

Received: April 18, 2019
Revised: April 18, 2019
Accepted: April 18, 2019
Published Online in Articles in Advance: September 3, 2019

https://doi.org/10.1287/mnsc.2019.3388

Abstract. Artificial intelligence (AI) is surpassing human performance in a growing number of domains. However, there is limited evidence of its economic effects. Using data from a digital platform, we study a key application of AI: machine translation. We find that the introduction of a new machine translation system has significantly increased international trade on this platform, increasing exports by 10.9%. Furthermore, heterogeneous treatment effects are consistent with a substantial reduction in translation costs. Our results provide causal evidence that language barriers significantly hinder trade and that AI has already begun to improve economic efficiency in at least one domain.

History: Accepted by Joshua Gans, business strategy.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2019.3388.

Keywords: artificial intelligence • international trade • machine translation • machine learning • digital platforms

Equivalent to make the world smaller than 26%

study on ebay

Machine Translation

Translating information from one language to another

I bought a sweet persimmon in the store

⇩

Ich kaufte eine süße Persimone im laden

Types of Machine Translation

- Translating information from one language to another
- Media:
  - (Text) Machine Translation
  - Speech Translation: Speech-to-Text or Speech-to-speech translation
  - Visually Machine Translation: Text translation with additional image
- Genre:
  - Sentence level MT
  - Document level MT
  - Dialog Translation
- Number of Languages:
  - Bilingual
  - Multilingual

Why automatic Machine Translation?

- Too expensive to hire human translator
  - e.g. touring, shopping, restaurant eating in a foreign country
- Too much effort for human to translate massive text
  - can tolerate imprecise translation
- Need instantaneous translation
  - e.g. in international conference

A Brief History of Machine Translation

Commercial Machine Translation

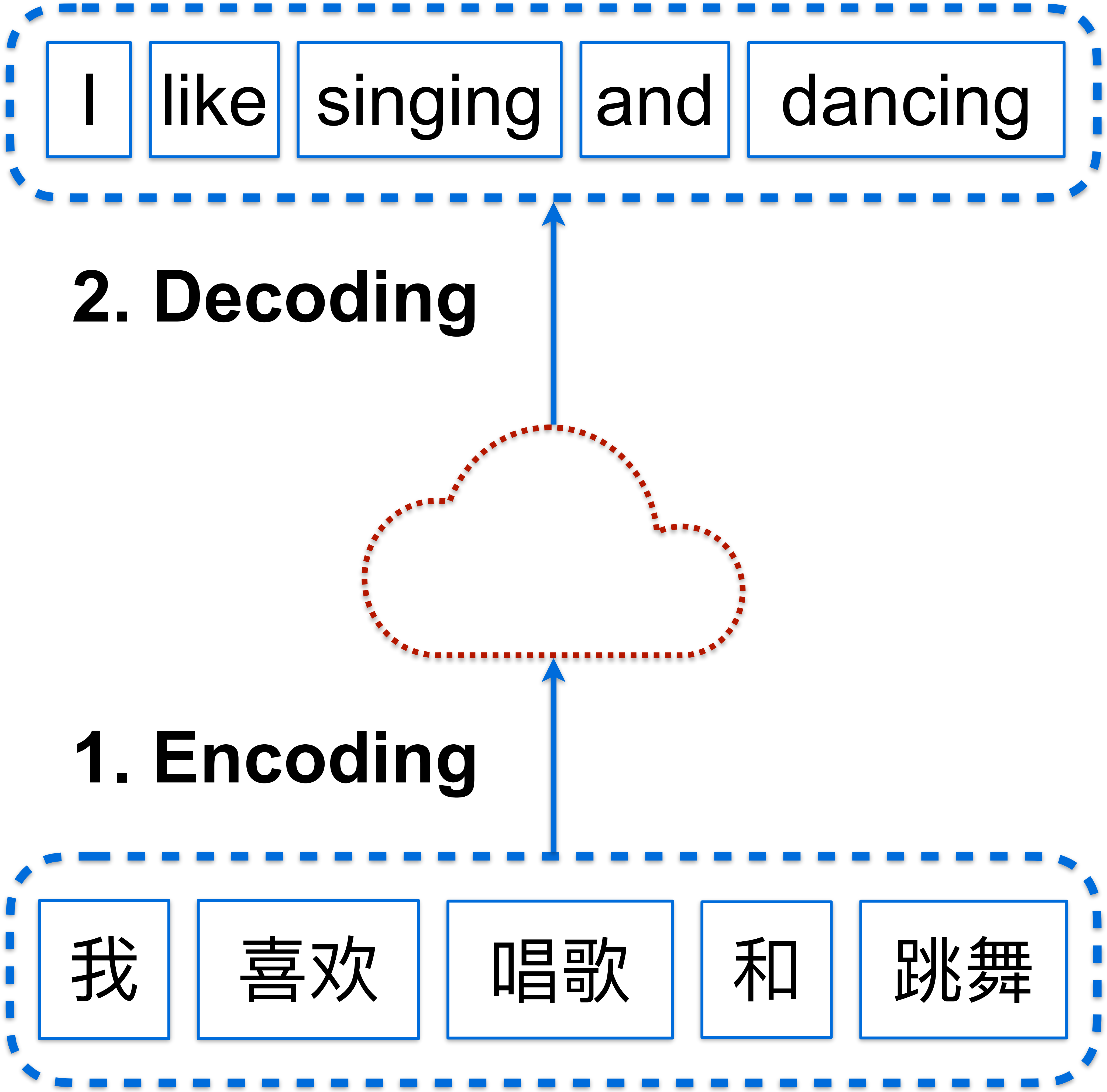
- Google translate: 109 languages, separate app, support text/ document translation, image translation, and speech translation
- Microsoft translate: 87 languages for text
- Baidu translate: 200+ languages
- ByteDance VolcTrans: 104 languages
- DeepL: good at European languages
- Youdao Translate: integrated with its own dictionary app
- Tencent Translate: native in wechat, and separate app
- NiuTrans: specialized in Chinese to many languages

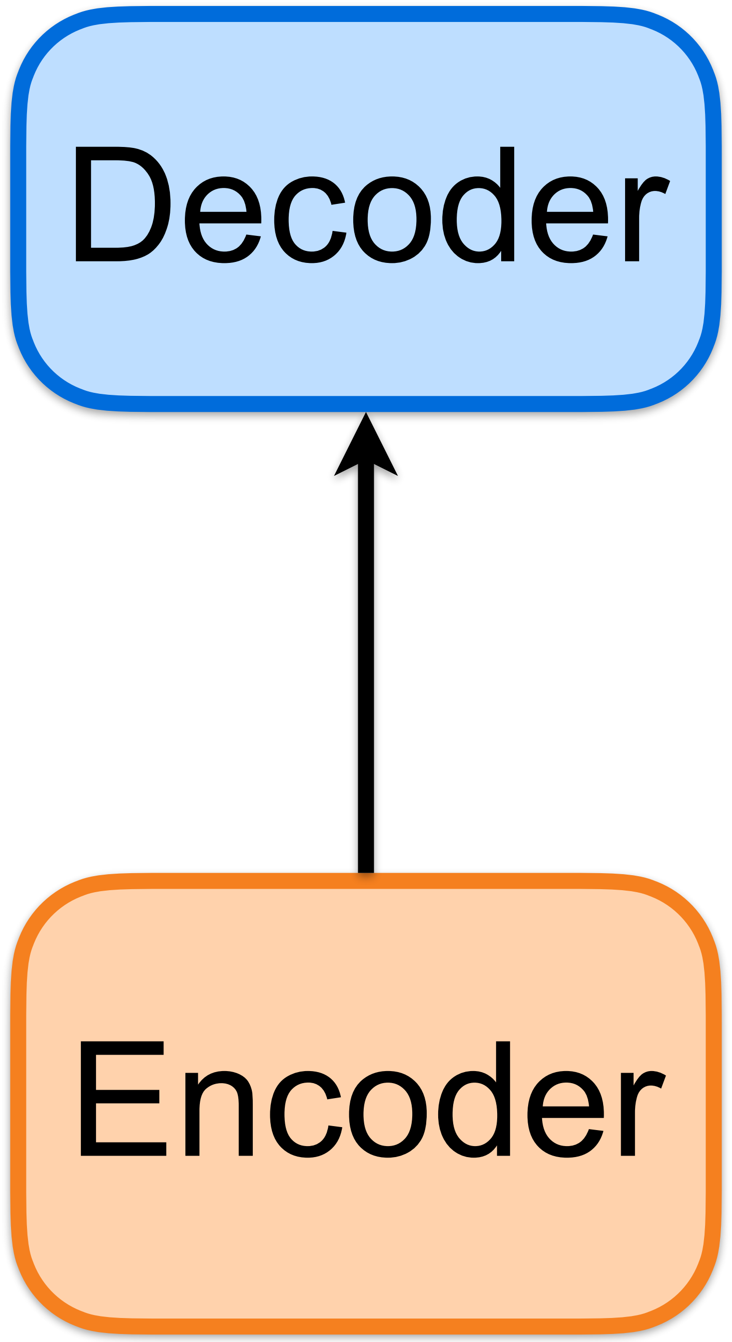
Outline

- Basics of Neural Machine Translation
  – Model, Data, Training, Low-resource
- Why is MT still hard?
- Multilingual MT
  – Contrastive Multilingual Training with Randomly Aligned Substitution (mRASP2)
  – Learning language-specific sub-network (LaSS)
  – Counter Interference Adapter (CIAT)
  – Graformer: Grafting Pre-trained Language Models
- Speech-to-Text Translation
  – Offline End-to-end ST: ConST, STEMM, Chimera, LUT, CosTT
  – Simultaneous Interpretation (Streaming ST)

Encoder-Decoder Framework

Translation as an encoding-decoding problem





A generic formulation

ImageCaption

Text-to-Image Generation

ASR (speech-to-text)

MT (text-to-text)

Mathematical Formulation of MT

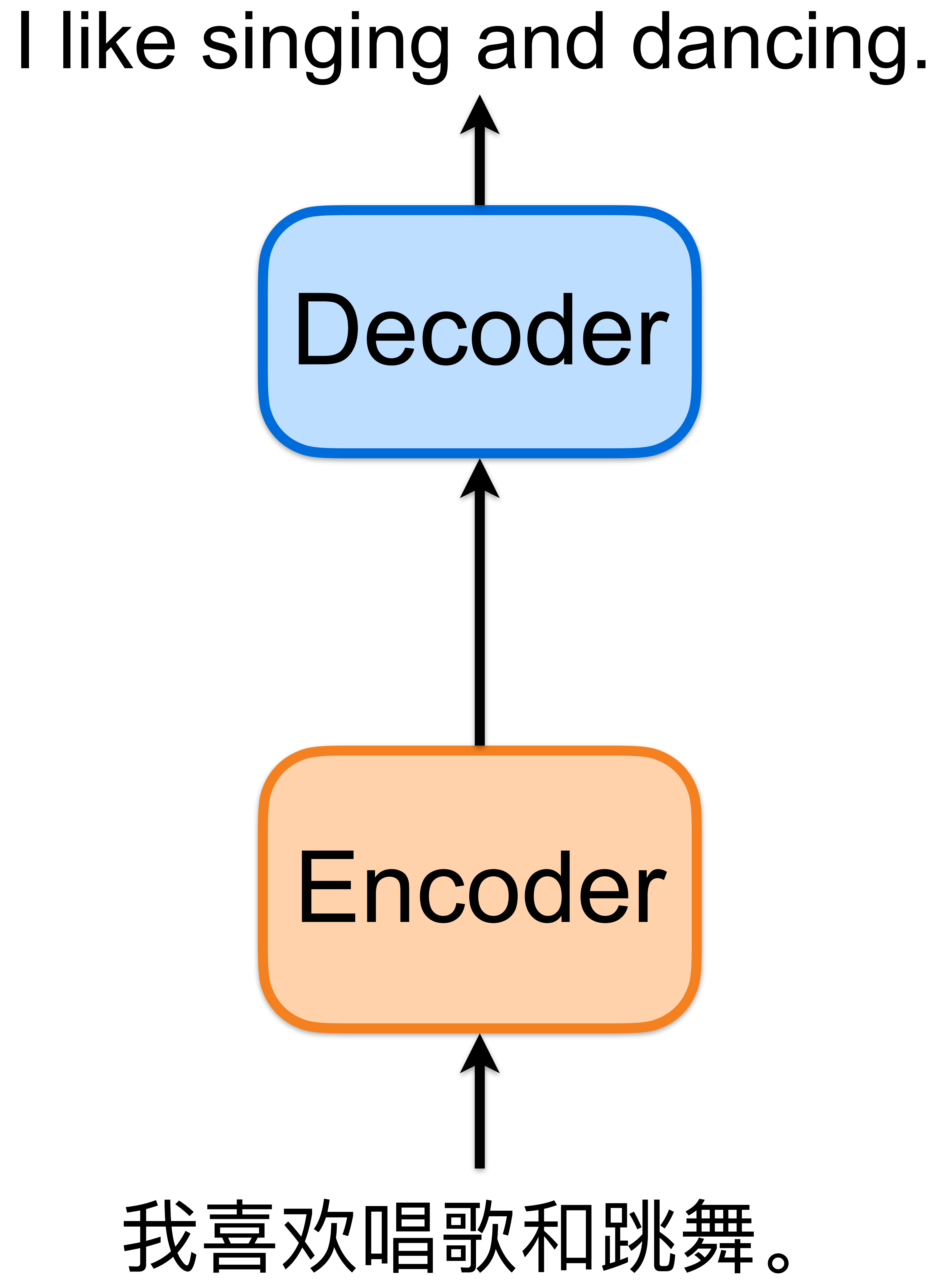
- MT model as a function mapping from source sequence to target sequence

$$P(Y|X;\theta) = \prod P(y_t|y_{<t}, x; \theta)$$

$$P(y_t|y_{<t}, x; \theta) = f_\theta(x_{1\ldots k}, y_{1\ldots t-1})$$

- Training: finding the optimal model parameter θ
- Inference: decode the best target text given an input

$$Y^\star = \underset{Y}{\operatorname{argmax}}\, P(Y|X;\theta)$$



Neural MT Models

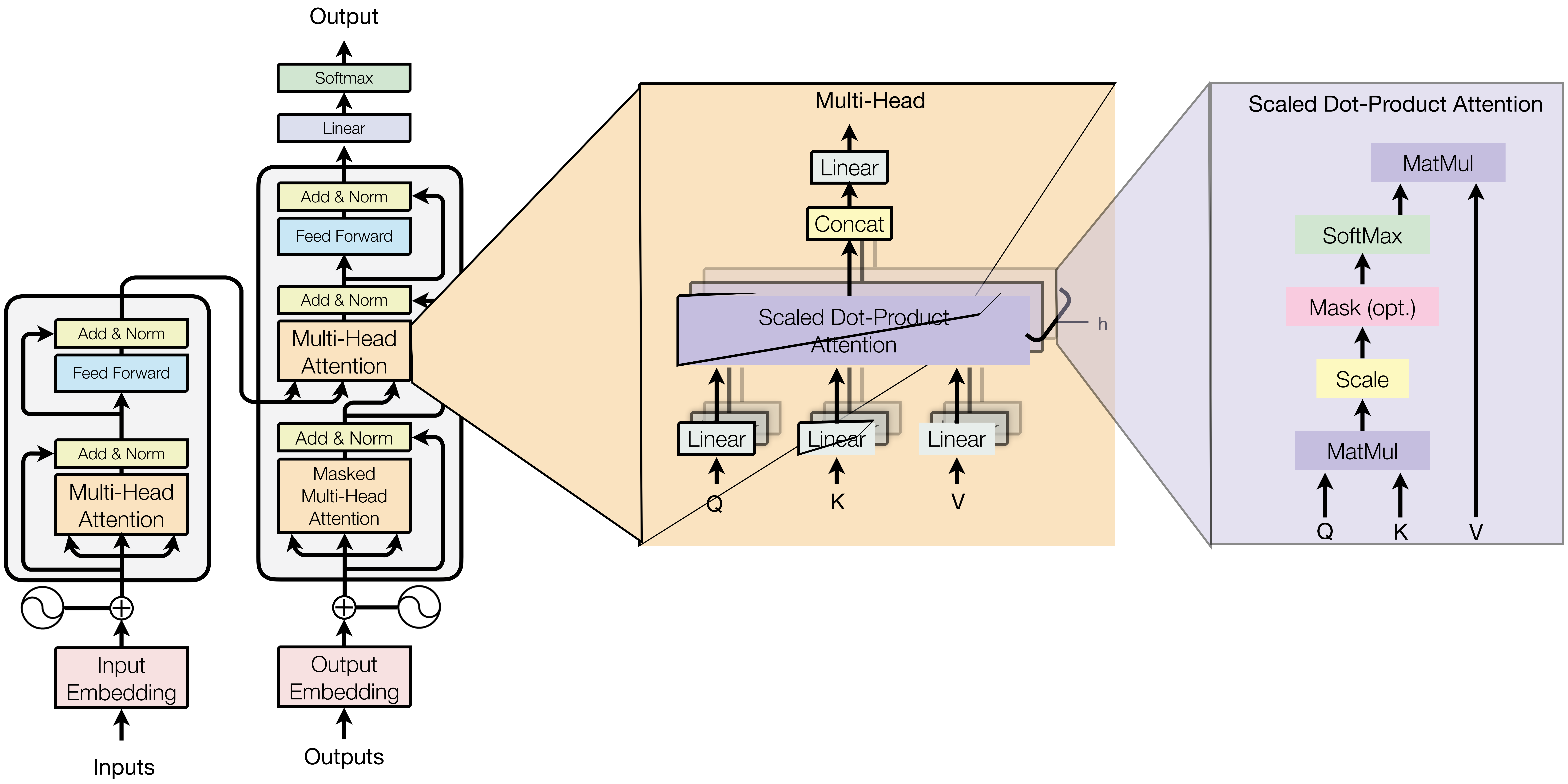
- Transformer: the most popular model for MT since 2017
  - use attention+FFN, many variations
- Sequence-to-sequence (seq2seq): using multiple layers of (bidirectional) LSTM/GRU as the encoder and decoder, 2014
- CNN MT: using convolutional neural networks at encoder/decoder

Transformer

Vaswani et al. Attention is All You Need. 2017

Multi-head Attention Layer (MHA)

- C layers of encoder (=6)
- D layers of decoder (=6)



How does Transformer Translate?

Translation Performance on WMT14

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Demo

- translate.volcengine.com

Volctrans

Why is MT challenging?

Why is MT challenging?

- Polysemy

He deposited money in a bank account with a high interest rate.

Sitting on the bank of the Mississippi, a passing ship piqued his interest.

- New entity names
  - COVID-19
- Complex structure
- Ellipsis (i.e. omission)

New Terms

周四经济数据面，美国劳工部报告称，截至8月28日当周美国首次申请失业救济人数为34万，降至2020年美国新冠疫情危机爆发以来的最低点。市场预计该数字为34.5万。

Google Translation (2021.9.1)

On Thursday’s economic data, the U.S. Department of Labor reported that as of August 28, the number of people applying for unemployment benefits for the first time was 340,000, which dropped to the lowest point since the outbreak of the new crown crisis in the United States in 2020. The market expects the number to be 345,000.

VolcTrans (2021.9.1)

On Thursday's economic data, the U.S. Labor Department reported that the number of first-time jobless claims in the United States for the week ending August 28 was 340 thousand, falling to the lowest level since the COVID-19 Epide COVID-19 epidemic crisis broke out in the United States in 2020. The market expects the number to be 345 thousand.

New Terms

周四经济数据面，美国劳工部报告称，截至8月28日当周美国首次申请失业救济人数为34万，降至2020年美国新冠疫情危机爆发以来的最低点。市场预计该数字为34.5万。

Bing Translation (2021.9.1)

On Thursday, the *Labor Department reported that 340,000 people applied for * unemployment benefits for the week ended Aug. 28, the lowest level since the * crisis began in 2020. The market expects the figure to be 345,000.

DeepL (2021.9.1)

On Thursday's economic data front, the U.S. Labor Department reported that the number of first-time U.S. jobless claims for the week ended Aug. 28 was 340,000, falling to the lowest point since the outbreak of the new U.S. crown epidemic crisis in 2020. The market expected the figure to be 345,000.

Complex sentences

周四美股成交额冠军苹果(153.65, 1.14, 0.75%)公司收高0.75%，报153.65美元，创历史收盘新高，成交108.9亿美元，市值逼近2.54万亿美元。

Bing Translation (2021.9.1)

U.S. stock market champion Apple Inc (153.65, 1.14, 0.75 percent) closed up 0.75 percent at $153.65 on Thursday, a record closing high of $10.89 billion, giving it a market capitalization of nearly $2.54 trillion.

DeepL (2021.9.1)

Thursday's U.S. stock turnover leader Apple (153.65, 1.14, 0.75%) closed 0.75% higher at $153.65, an all-time closing high, with $10.89 billion traded and a market cap approaching $2.54 trillion.

他的爷爷和奶奶没见过他的姥姥和姥爷。

Google Translate: His grandpa and grandma have never met his grandma and grandpa.

Correct: His father’s parents never met his mother’s.

- Acronym and incorrect word segmentation

一些立陶宛人士表示，中立关系恶化，影响最大的当属立陶宛的出口企业。

Google Translate: Some Lithuanians said that the deterioration of Sino-Lithuanian relations has affected Lithuanian export companies the most.

Bing Translate: Some Lithuanians say the deterioration in neutral relations has affected Lithuania's exporters the most.

Culture and Slang

这个人很牛

MT1/MT3: This person is very cattle.

MT2: This man is a cow.

MT4: This guy's good.

MT0: This guy is awesome.

Robustness

– variation of auxiliary function words or symbols

这个人很牛
MT1: This person is very cattle.
MT3: This person is very cattle.
MT0: This guy is awesome.

这个人很牛。
MT1: This person is very bullish.
MT3: This man is very good.
MT4: This guy is good.
MT0: This guy is very good.

这个人非常牛。
MT1: This person is very cattle.
MT3: This person is very cattle.
MT0: This guy is awesome.

这个人很牛!
MT1: This person is very cow!
MT3: This man is very good.
MT4: This man is good!
MT0: This guy is awesome!

Robustness

乔丹最早周日伤愈复出

MT0: Jordan came back from his first injury on Sunday.

MT1: Jordan first recovered from injury on Sunday

乔丹最早周日伤愈复出。

MT0: Jordan came back from injury on Sunday.

MT1: Jordan returned from injury on Sunday.

Reference: Jordan may return from injury as early as this Sunday.

MT: From fluency to nativeness

No, Scarlett, the seeds of greatness were never in me.

MT1: 不，思嘉，伟大的种子永远不会在我身上。

MT0: 不，思嘉，伟大的种子从来就不存在。

Ref: 不，斯佳丽，我根本就不是当大人物的料。

(Average) Human Level Translation

You say that you love rain, but you ope
n your umbrella when it rains.
You say that you love the sun, but you f
ind a shadow spot when the sun shine
s.
You say that you love the wind, but you
close your windows when wind blows.
This is why I am afraid, you say that yo
u love me too.

MT: 你说你喜欢雨，但雨下的时候你打开雨伞。

你说你爱太阳，但当太阳照耀时，你发现了一个阴影斑点。

你说你喜欢风，但是当风吹起的时候你会关上窗户。

这就是为什么我害怕，你说你也爱我。

Expert Level Translation

诗经体：

子言慕雨，启伞避之。子言好阳，寻荫拒之。子言喜风，阖户离之。子言偕老，吾所畏之。

离骚版:

君乐雨兮启伞枝，君乐昼兮林蔽日，君乐风兮栏帐起，君乐吾兮吾心噬。

七律：

江南三月雨微茫，罗伞叠烟湿幽香。夏日微醺正可人，却傍佳木趁荫凉。霜风清和更初霁，轻蹙蛾眉锁朱窗。怜卿一片相思意，犹恐流年拆鸳鸯。

网络咆哮体：

你有本事爱雨天，你有本事别打伞啊！你有本事爱阳光，你有本事别乘凉啊！！你有本事爱吹风，你有本事别关窗啊！！！你有本事说爱我，你有本事捡肥皂啊！！！！

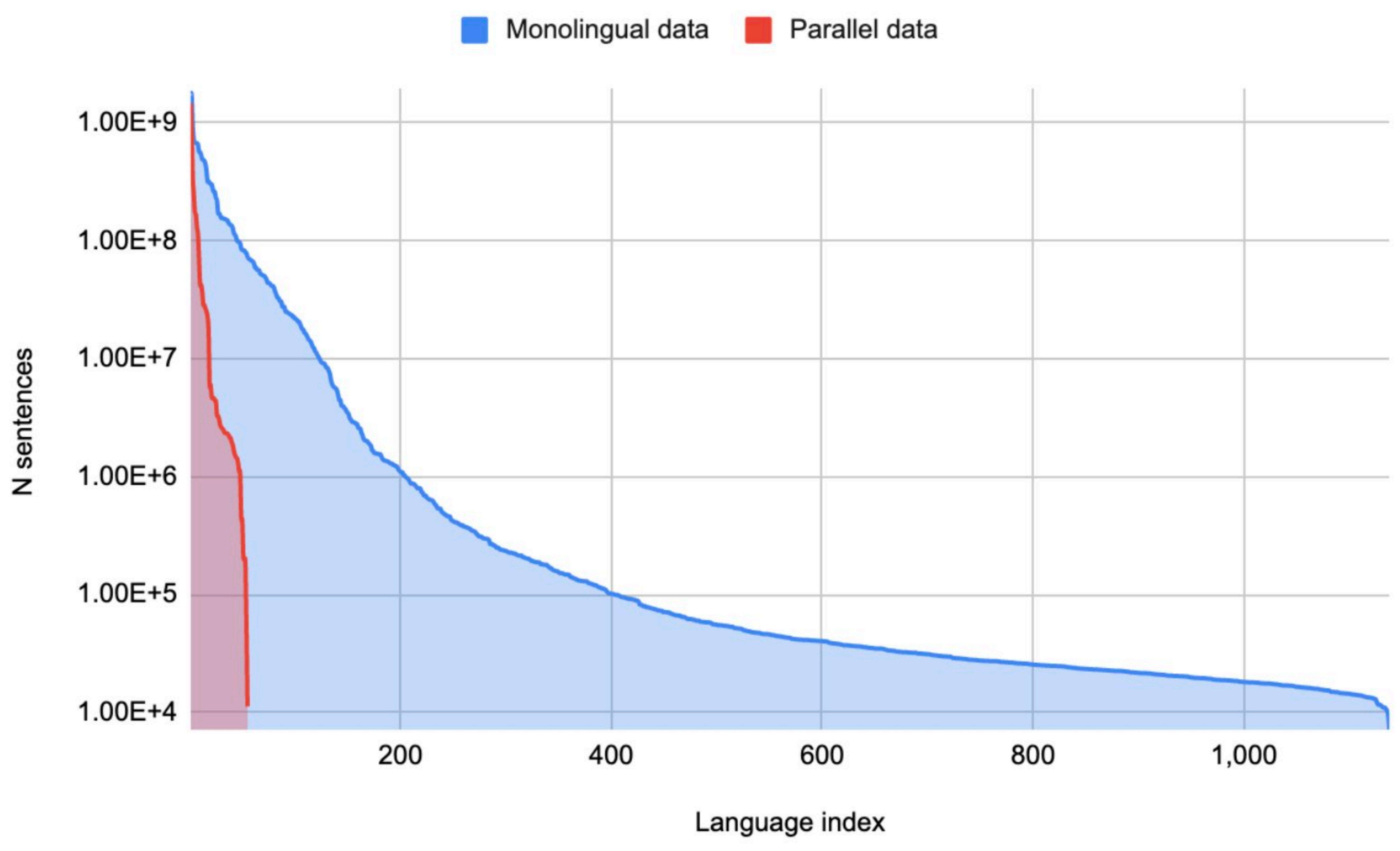
Multilingual Machine Translation

Multilingual Neural Machine Translation

- Bilingual NMT: one model for each translation direction
- Multilingual NMT: Develop one model to translate between all language pairs.
- Why? Motivation
  - Potential better performance: Languages with rich resource could benefit those with low resource
  - Economic: only one model deployment versus of many deployments. Simpler workload and job management and scheduling.
    - vs Bilingual models: Many languages would have much few requests but still need to occupy the servers.

Imbalanced Data across Languages

- NMT requires large amount of parallel bilingual data
- Parallel data, However, very expensive/ non-trivial to obtain
  - Low resource language pairs (e.g., English-to-Tamil)
  - Low resource domains (e.g., social network)
  - but additional monolingual data on source side and/or target side. can we do reasonably well?
- Rich resource setting: in addition to parallel data (>10 millions), much larger monolingual data, can we further improve?



[Credit: Isaac Caswell, 2022]

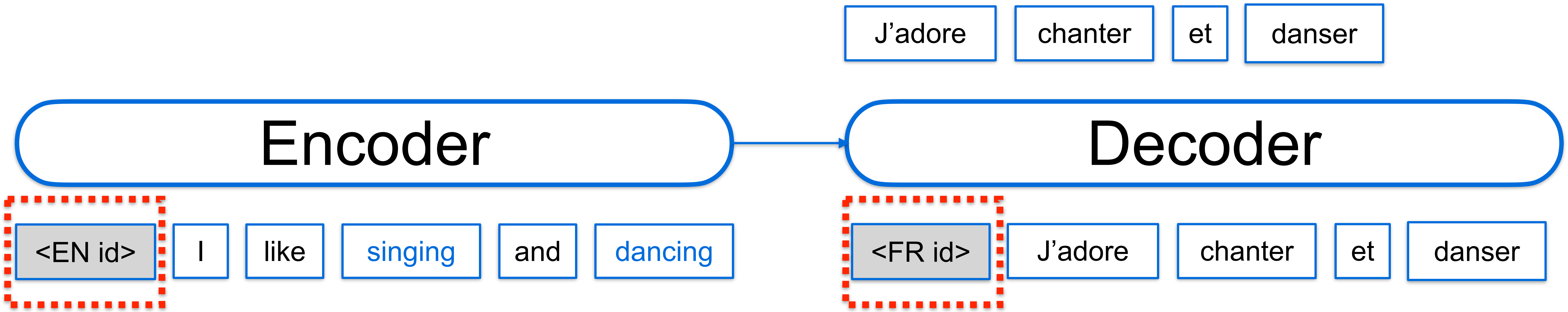
Types of MNMT

- Many-to-one:
  - Many source language to a target language
  - Usually the target is English
- One-to-Many:
  - One source language to many target languages
  - Usually the source is English
- Many-to-many
  - Many source language to many target languages
  - Should include non-English pairs (often low-resource or zero-resource setting)
  - very challenging if Non-english directions have little data!

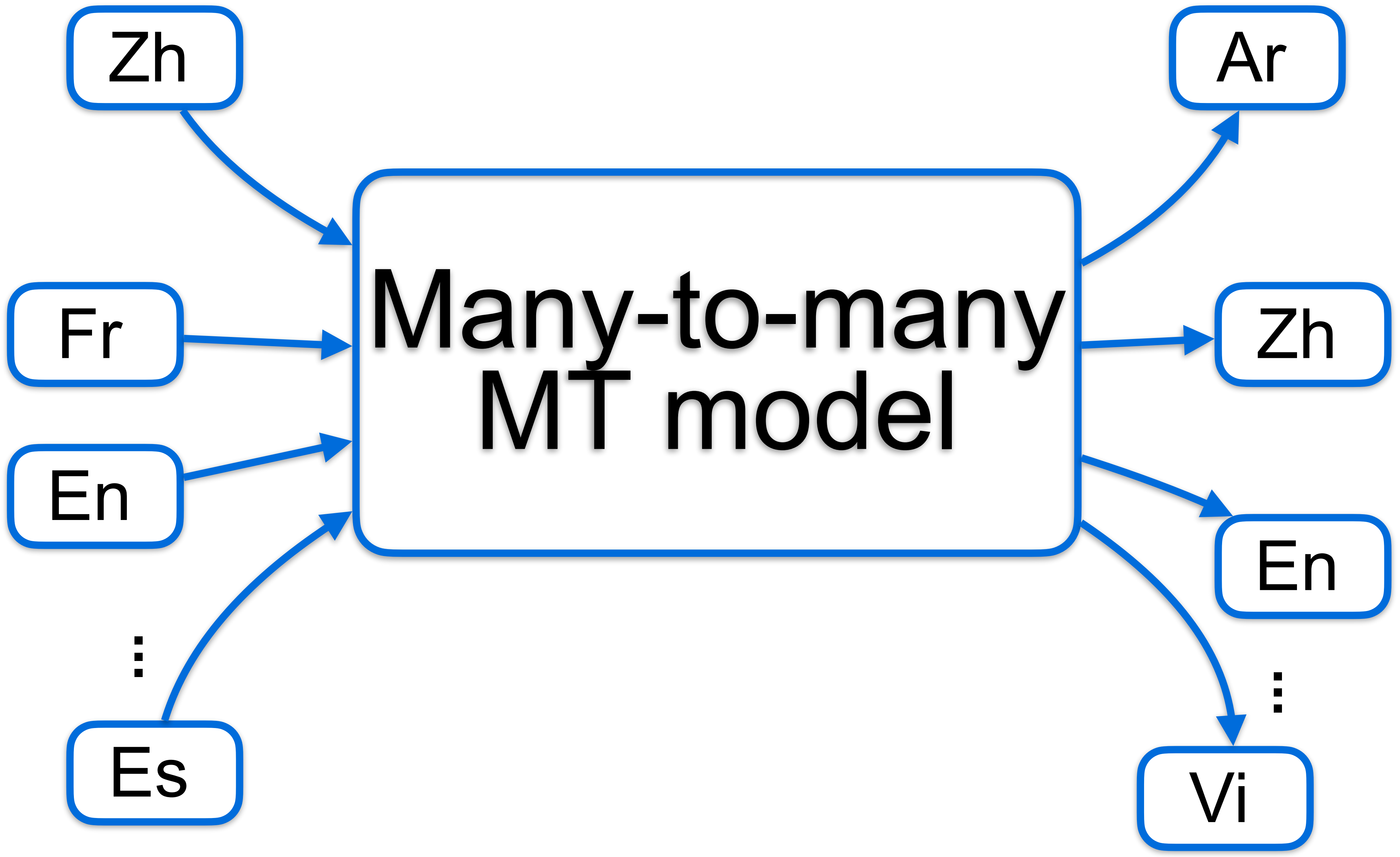
MNMT at Testing Time

- Supervised:
  - Testing language pairs (usually English-centric) appeared during training
- Zero-shot (Exotic/unseen pair)
  - Both the testing source language and target language appeared in the training, but the source-target pair never appeared in the training
  - Training on En-De, En-Fr, testing on De-Fr
- Unsupervised
  - Exotic source/target
    - Testing source/target language with no parallel sentence in the training. (but with Monolingual)
    - Training on En-De, En-Fr, En-Zh, and Japanese monolingual text, then testing on Ja-De
  - Exotic/Unseen full (most challenging)
    - Neither the source language nor the target language for testing occur in the training

Single Model for Multilingual MT



- One model can translate between many languages.
- Language Tag is used to indicate the source and target language.
- Vocabulary is built jointly



Johnson et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. 2017
Arivazhagan et al. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. 2019

Google's MNMT: Success and Limitation

- Training 12 language pairs together
- A single model (LSTM seq2seq) with comparable performance as individual bilingual models 😁
- But only one direction is better, many are noticeably worse than bilingual 😭

Table 4: Large-scale experiments: BLEU scores for singl language pair and multilingual models.

Model	Single	Multi	Multi	Multi	Multi
#nodes	1024	1024	1280	1536	1792
#params	3B	255M	367M	499M	650M
En→Ja	23.66	21.10	21.17	21.72	21.70
En→Ko	19.75	18.41	18.36	18.30	18.28
Ja→En	23.41	21.62	22.03	22.51	23.18
Ko→En	25.42	22.87	23.46	24.00	24.67
En→Es	34.50	34.25	34.40	34.77	34.70
En→Pt	38.40	37.35	37.42	37.80	37.92
Es→En	38.00	36.04	36.50	37.26	37.45
Pt→En	44.40	42.53	42.82	43.64	43.87
En→De	26.43	23.15	23.77	23.63	24.01
En→Fr	35.37	34.00	34.19	34.91	34.81
De→En	31.77	31.17	31.65	32.24	32.32
Fr→En	36.47	34.40	34.56	35.35	35.52
ave diff	-	-1.72	-1.43	-0.95	-0.76
vs single	-	-5.6%	-4.7%	-3.1%	-2.5%

Johnson et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. 2017

Multilingual Transformer: works but …

- Data: 25 billion sentence pairs in 103 languages
- Model: mTransformer with 375million params (larger than Transformer-big)

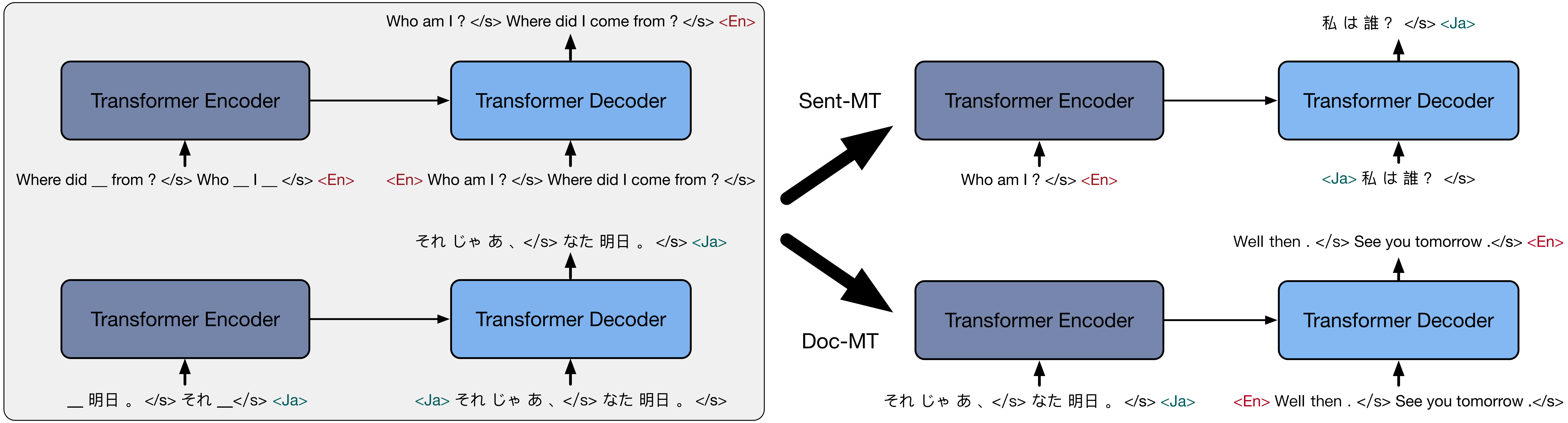
En→Any	High 25	Med. 52	Low 25
Bilingual	29.34	17.50	11.72
All→All	28.03	16.91	12.75
En→Any	28.75	17.32	12.98

Any→En	High 25	Med. 52	Low 25
Bilingual	37.61	31.41	21.63
All→All	33.85	30.25	26.96
Any→En	36.61	33.66	30.56

Observation: MNMT is good for low-resource, but bad for high/med-resource

Arivazhagan et al. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. 2019

Pre-training Fine-tuning Paradigm for MNMT



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

- Multilingual denoising pre-training (25 languages)
  - Sentence permutation
  - Word-span masking
- Fine-tuning on MT with special language id

Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

mBART: Multilingual Denoising Pre-training

Instead of a single model. Pre-train & fine-tuning

Languages	En-Gu		En-Kk		En-Vi		En-Tr		En-Ja		En-Ko	
Data Source	WMT19		WMT19		IWSLT15		WMT17		IWSLT17		IWSLT17	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	**0.3**	**0.1**	**7.4**	**2.5**	**36.1**	**35.4**	**22.5**	**17.8**	**19.1**	**19.4**	**24.6**	**22.6**

Languages	En-Nl		En-Ar		En-It		En-My		En-Ne		En-Ro	
Data Source	IWSLT17		IWSLT17		IWSLT17		WAT19		FLoRes		WMT16	
Size	237K		250K		250K		259K		564K		608K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	34.6	29.3	27.5	16.9	31.7	28.0	23.3	34.9	7.6	4.3	34.0	34.3
mBART25	**43.3**	**34.8**	**37.6**	**21.6**	**39.8**	**34.0**	**28.3**	**36.9**	**14.5**	**7.4**	**37.8**	**37.7**

Languages	En-Si		En-Hi		En-Et		En-Lt		En-Fi		En-Lv	
Data Source	FLoRes		ITTB		WMT18		WMT19		WMT17		WMT17	
Size	647K		1.56M		1.94M		2.11M		2.66M		4.50M	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	7.2	1.2	10.9	14.2	22.6	17.9	18.1	12.1	21.8	20.2	15.6	12.9
mBART25	**13.7**	**3.3**	**23.5**	**20.8**	**27.8**	**21.4**	**22.4**	**15.3**	**28.5**	**22.4**	**19.3**	**15.9**

Low resource: more than 6 BLEU. But fails in extremely low-resource setting

Medium resource: more than 3 BLEU improvement

Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

mBART on Rich-resource translation

Languages	Cs	Es	Zh	De	Ru	Fr
Size	11M	15M	25M	28M	29M	41M
Random	16.5	33.2	**35.0**	**30.9**	**31.5**	**41.4**
mBART25	**18.0**	**34.0**	33.3	30.5	31.3	41.0

- Pre-training slightly hurts performance when >25M parallel sentence are available.
- When a significant amount of bi-text data is given, supervised training are supposed to wash out the pre-trained weights completely.

Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

Summary of Challenges for MNMT

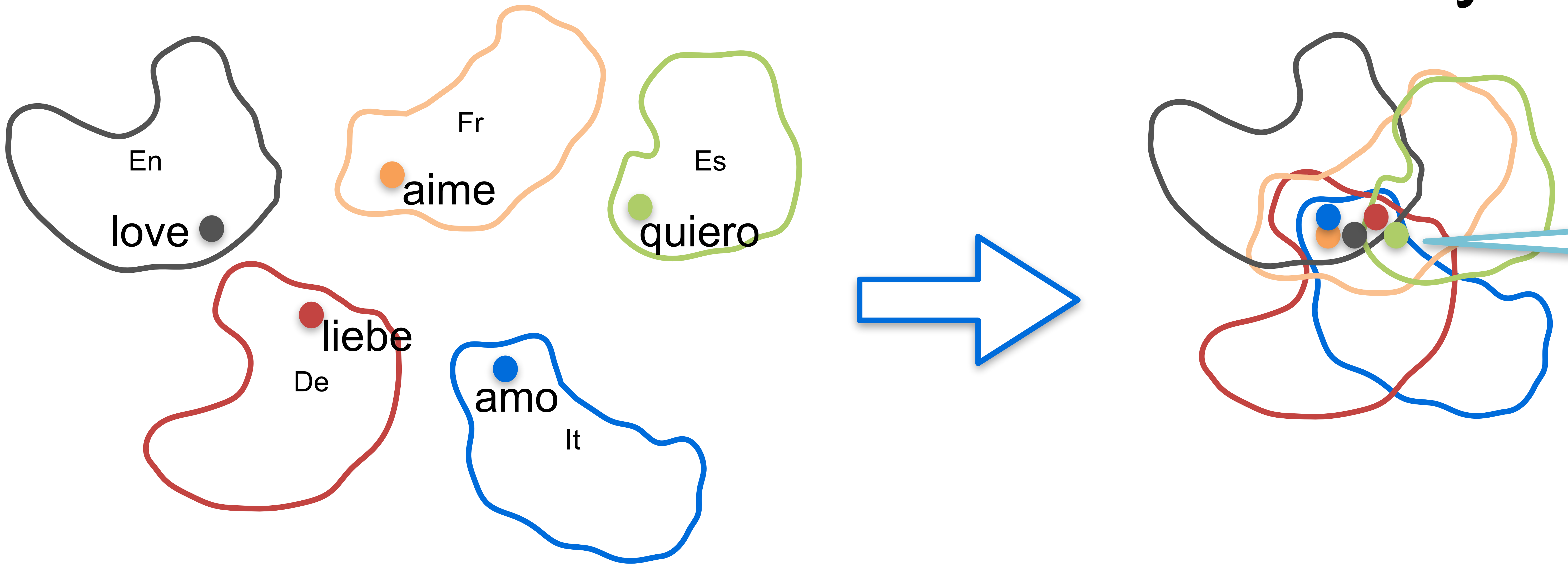
- Unified MNMT model has *inferior* performance than bilingual models
- Limited performance on zero-shot directions
- Possible causes:
  - highly imbalanced parallel data
  - parameter interference
  - insufficient use of monolingual data

build a single unified Multilingual MT models with superior performance on all language directions

Aligning Semantic Representations across Languages

- Key idea:
  1. Words in difference languages with the same meaning should have the same embedding
     - –but the training objective does not necessarily encourage that!

ideally

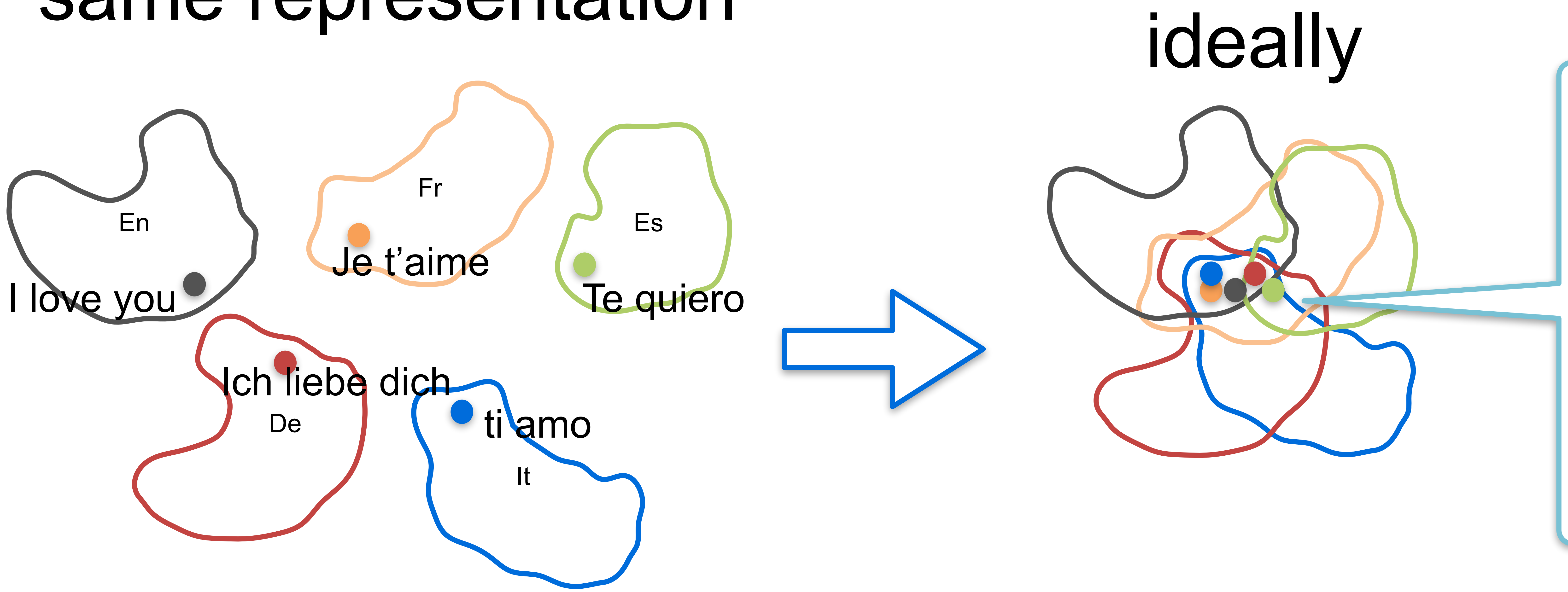


<En> I love you.
<Fr> Je t'aime.
<De> Ich liebe dich.
<Es> Te quiero.
<It> ti amo.

Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]
Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

Aligning Semantic Representations across Languages

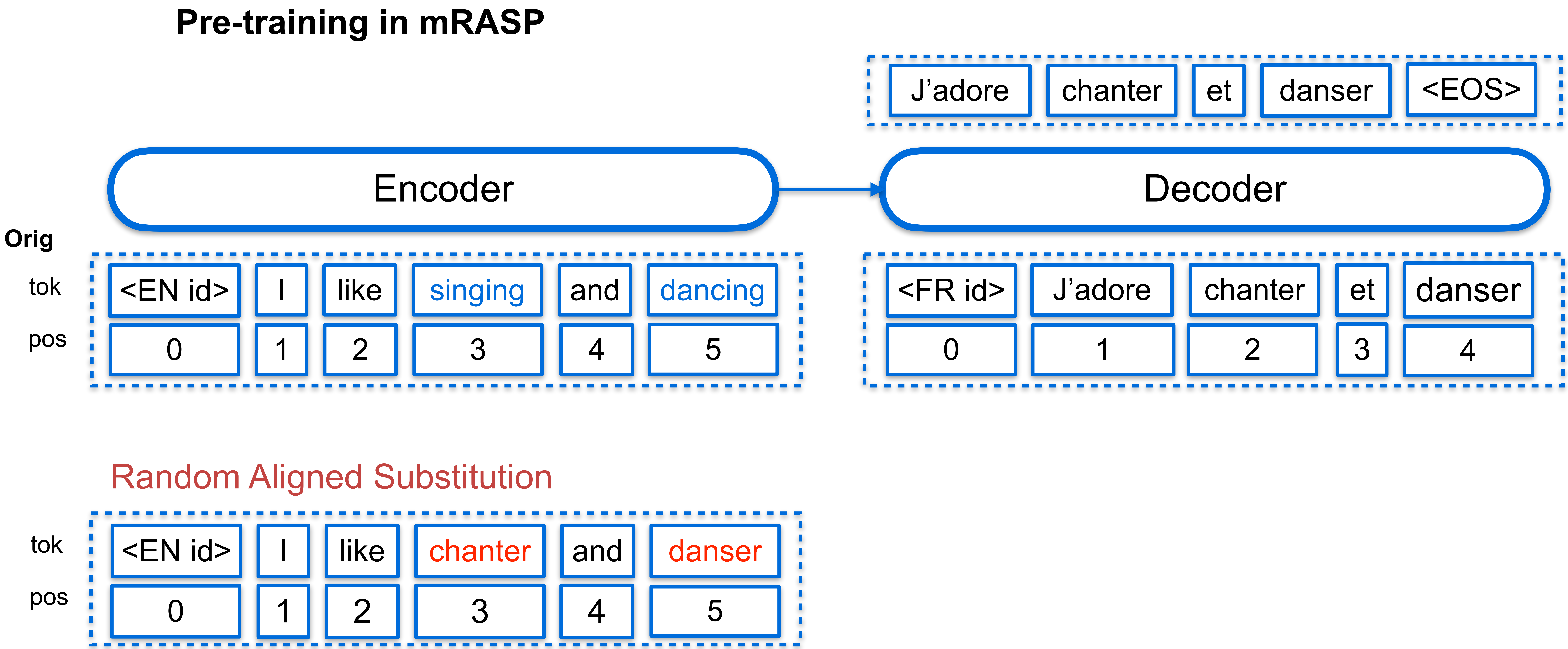
- Key idea:
  1. Words in difference languages with the same meaning should have the same embedding
  2. Parallel sentences in difference languages should have the same representation



<En> I love you.
<Fr> Je t'aime.
<De> Ich liebe dich.
<Es> Te quiero.
<It> ti amo.

Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]
Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

Idea 1: Training with RAS augmented samples



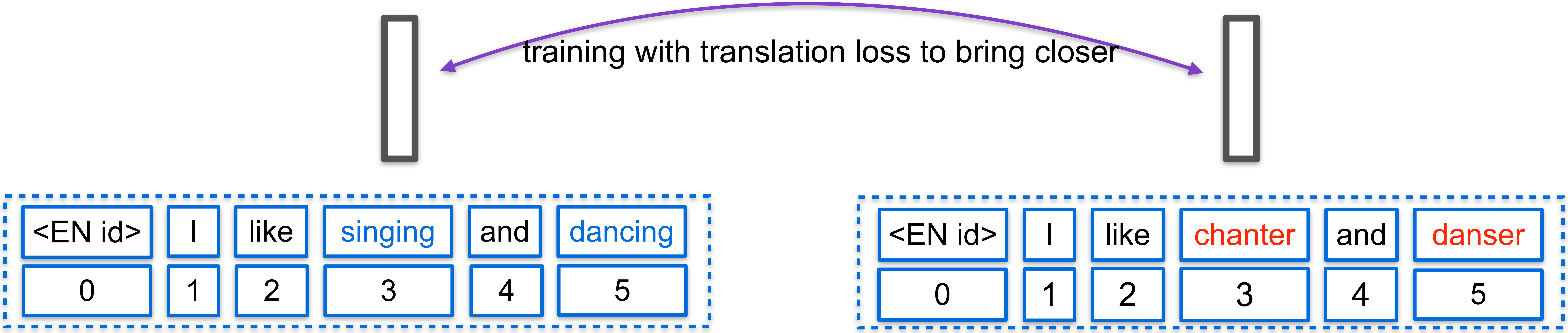
- Randomly replace a source word to its synonym in different language.

Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP: Bringing synonym representations closer

RAS: for each source sentence, randomly pick tokens, substitute with synonyms in other languages.

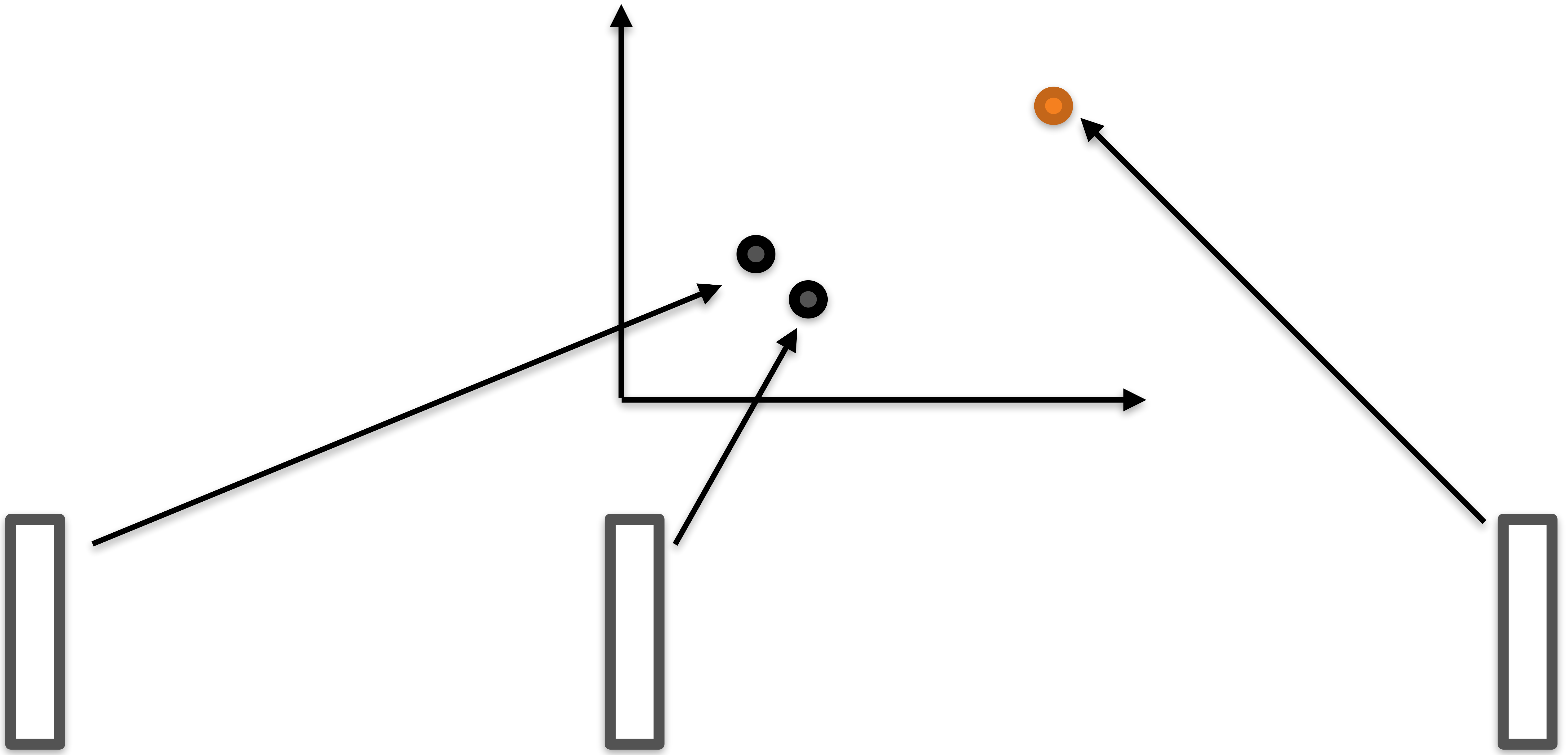
pair with original target and train in normal translation objective (cross-entropy)



$$\mathcal{L}^{pre} = \sum_{i,j \in \mathcal{E}} \mathbb{E}_{(\mathbf{x}^i, \mathbf{x}^j) \sim \mathcal{D}_{i,j}} \left[-\log P_\theta \left(\mathbf{x}^i \mid C\left(\mathbf{x}^j\right) \right) \right]$$

Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

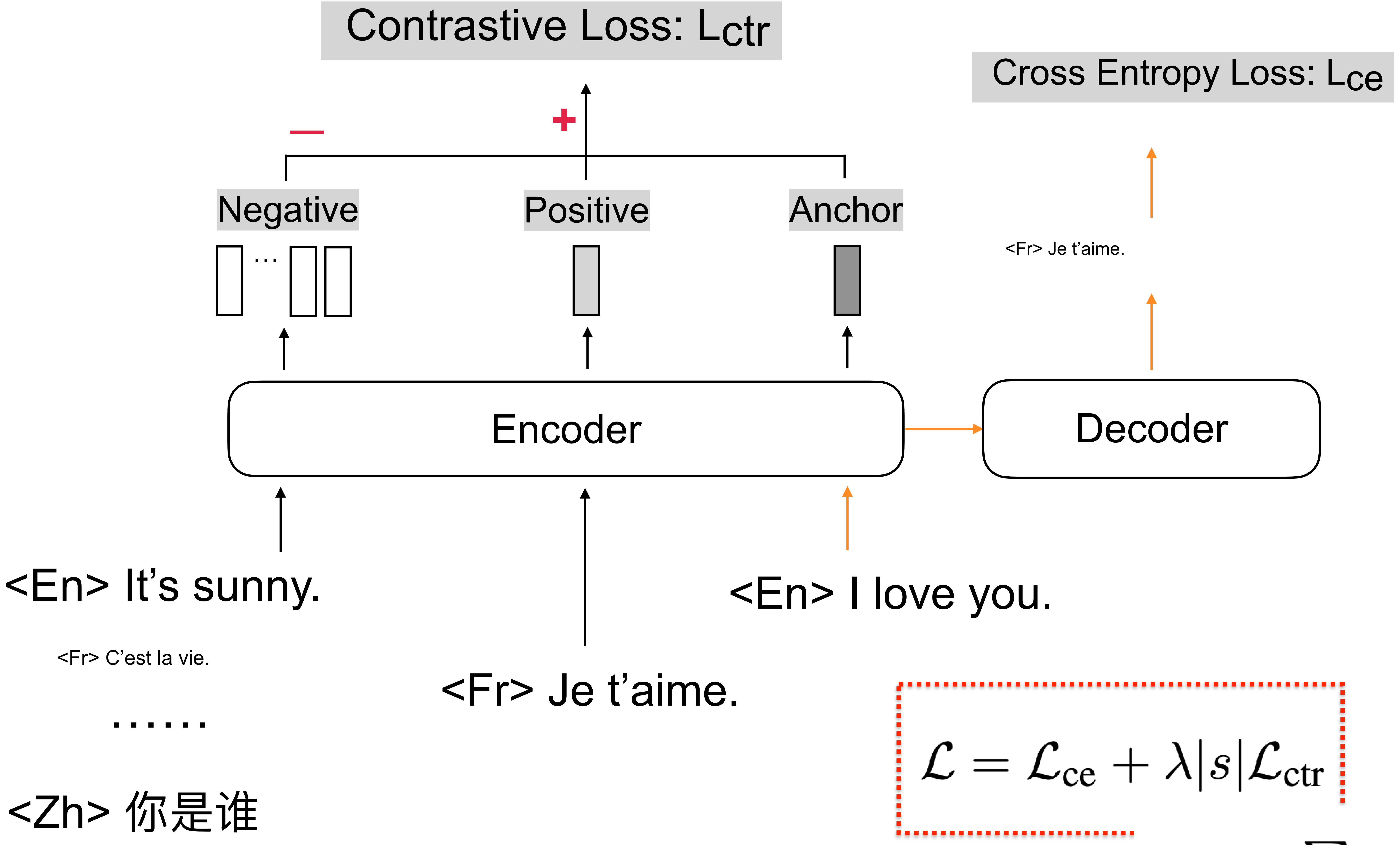
Idea 2: Bring parallel sentence representations closer



<En> I love you.　<Fr> Je t'aime.　<En> It's sunny.

Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

mRASP2: Contrastive Learning to bring sentence representations closer



$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda |s| \mathcal{L}_{\text{ctr}}$$

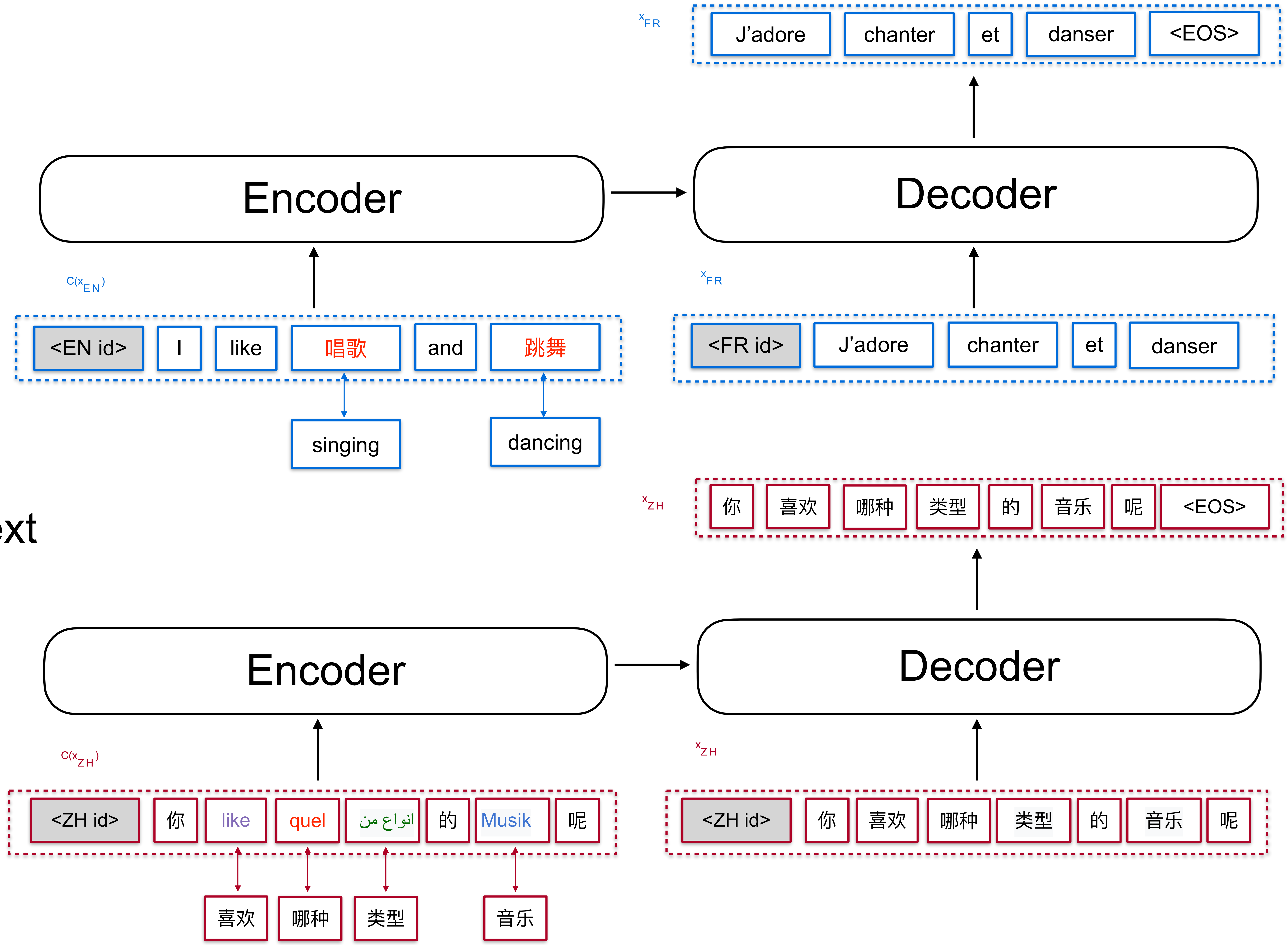
$$\mathcal{L}_{\text{ctr}} = -\sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}} \log \frac{e^{\text{sim}^{+}(\mathcal{R}(\mathbf{x}^i), \mathcal{R}(\mathbf{x}^j))/\tau}}{\sum_{\mathbf{y}^j} e^{\text{sim}^{-}(\mathcal{R}(\mathbf{x}^i), \mathcal{R}(\mathbf{y}^j))/\tau}}$$

Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

Idea 3: Integrating monolingual data in a unified training framework

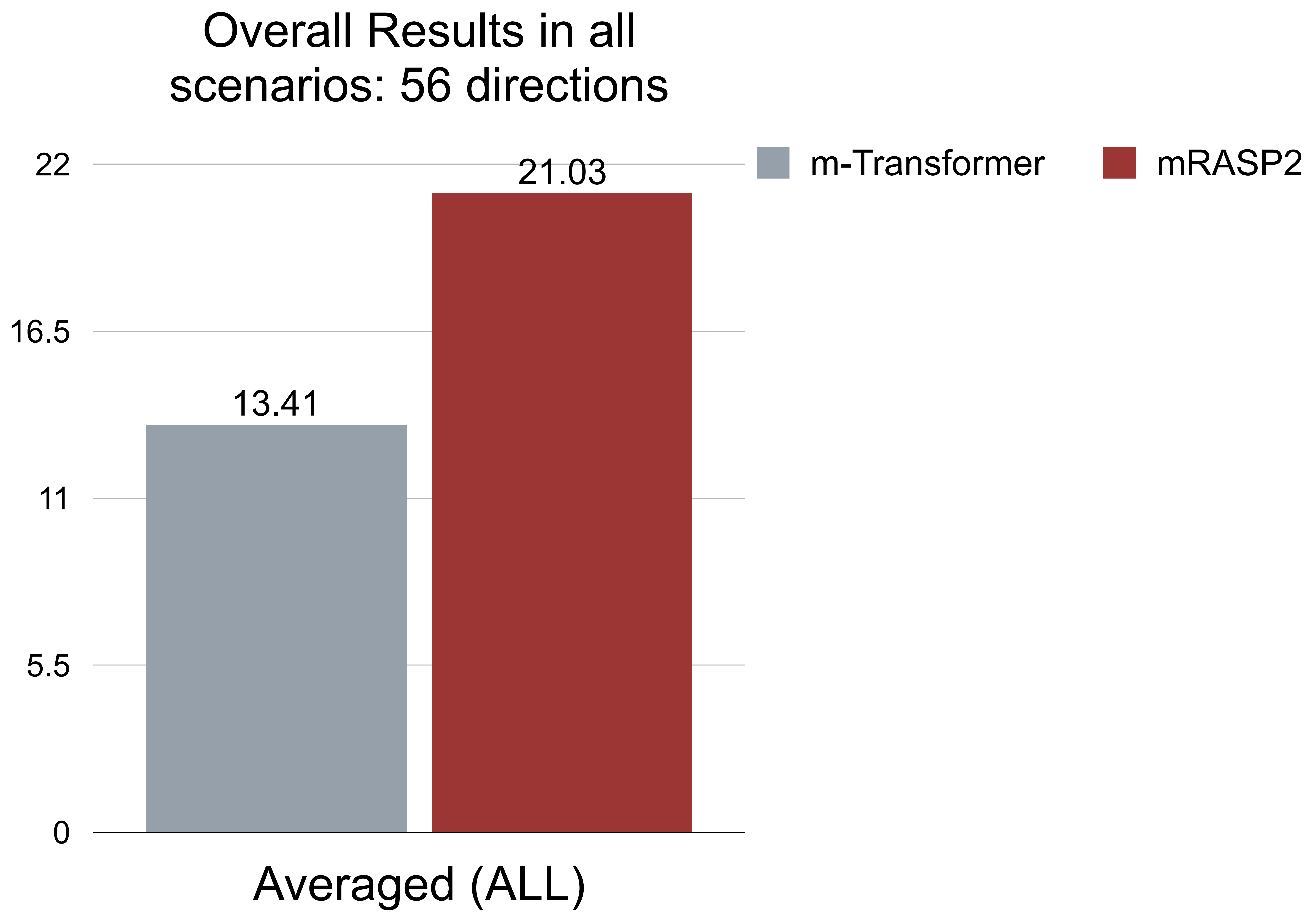
- Parallel text

- Monolingual text

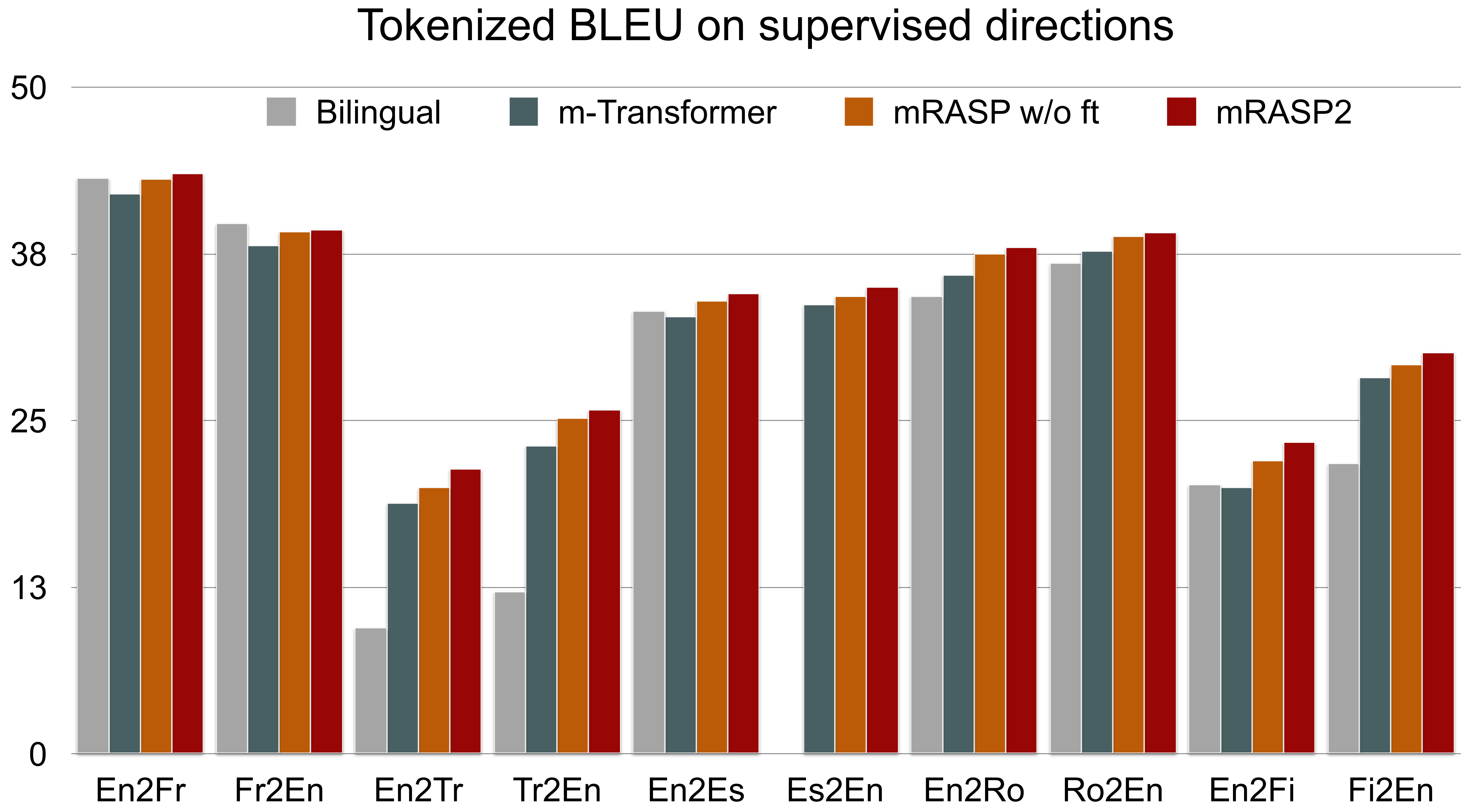


Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

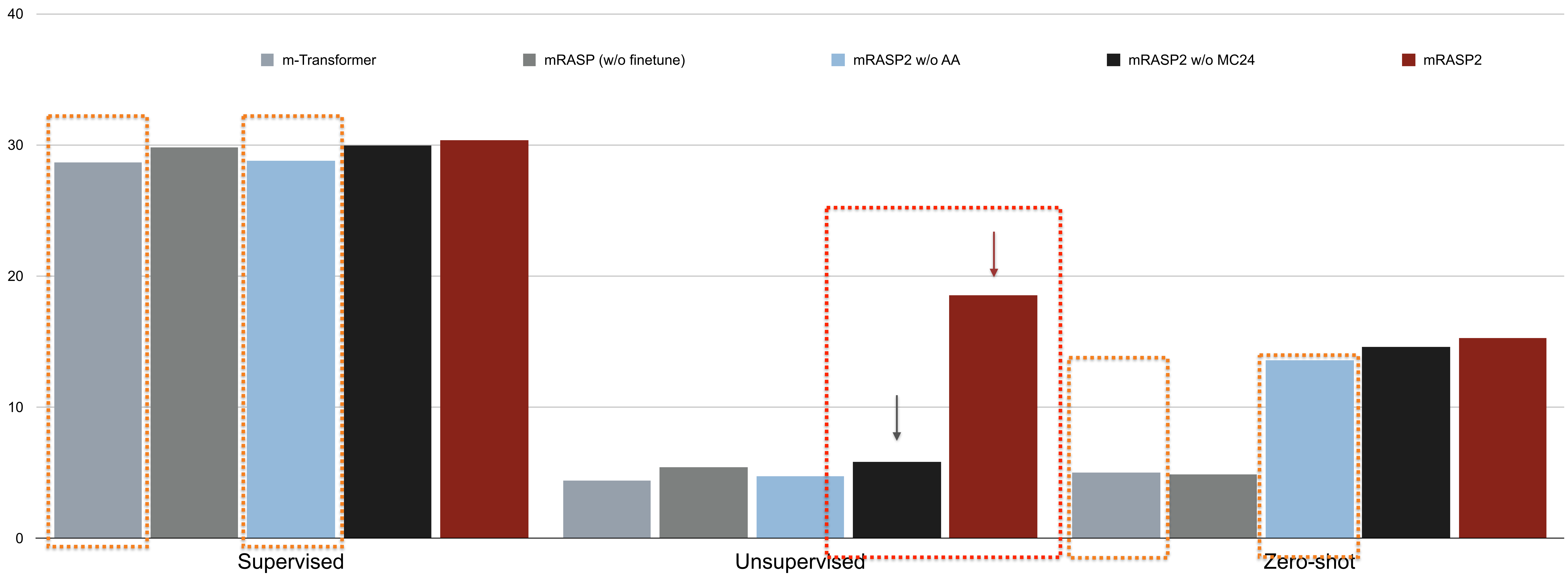
mRASP2: a single MNMT model (no fine-tuning)



mRASP2: Comparable or Better Performance on Supervised Directions

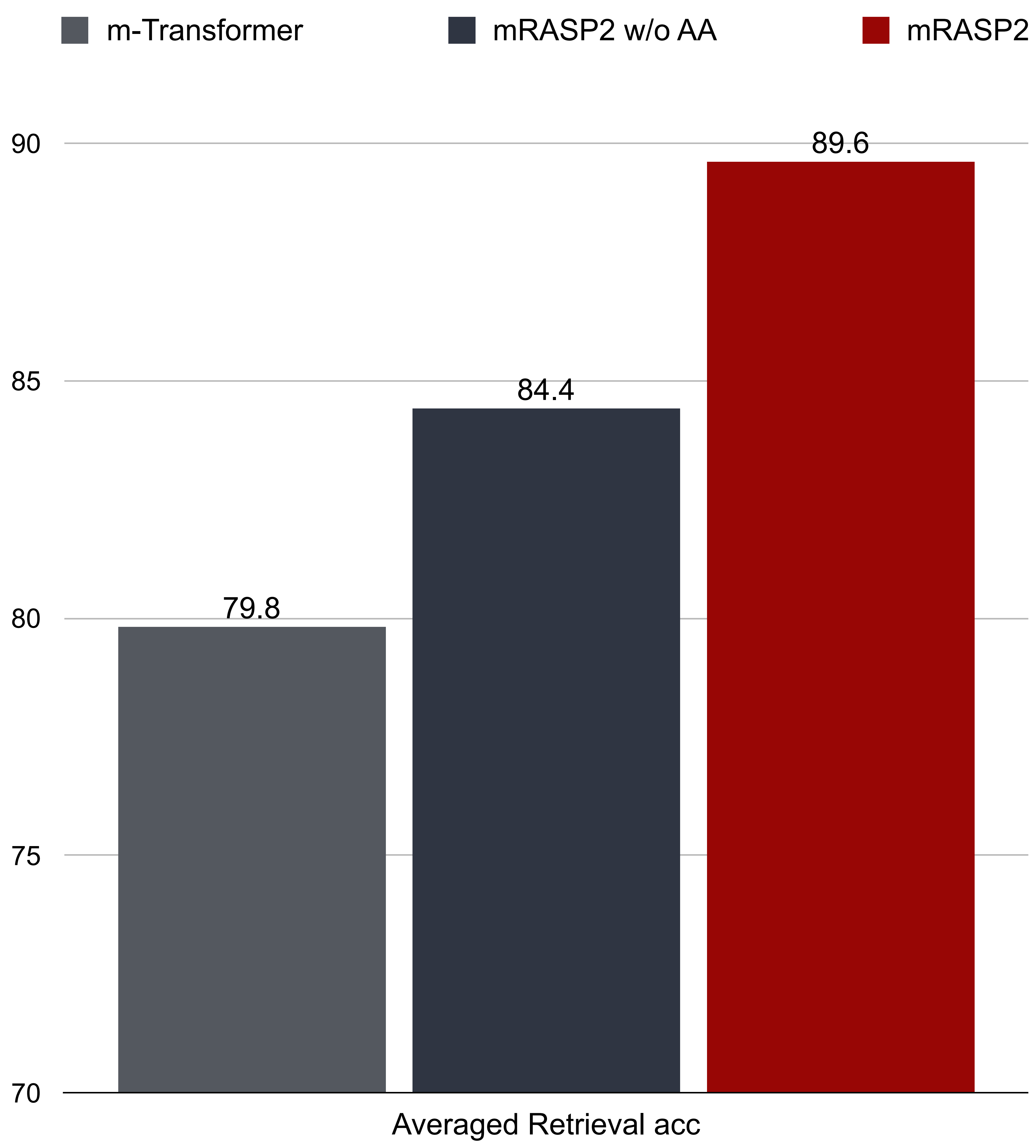


Contrastive Learning effectively improves zero-shot translation without hurting supervised translation performance



Monolingual Corpus mainly contributes to unsupervised translation

Better Semantic Alignment: Sentence Retrieval

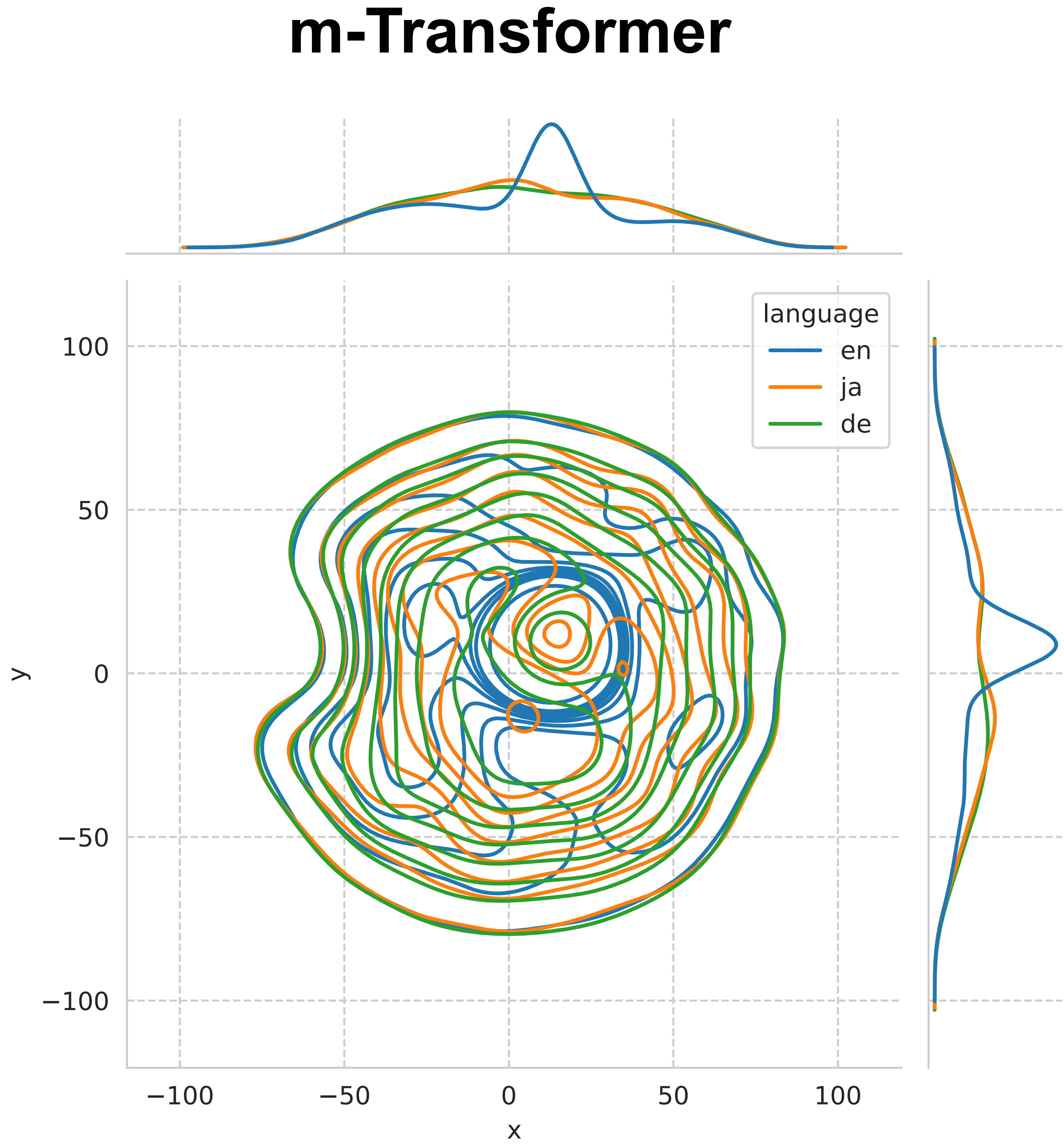


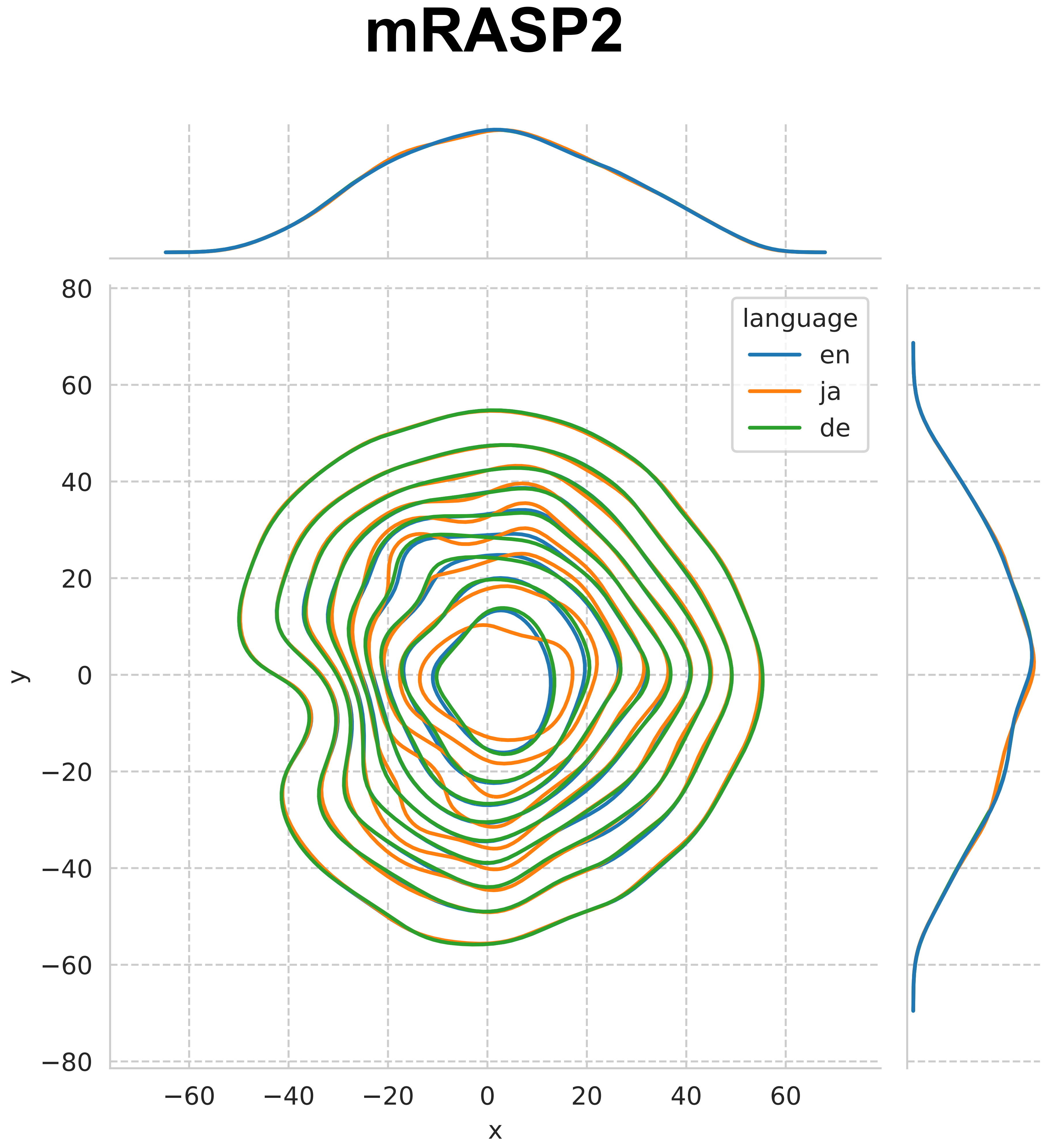
15-way parallel test set(Ted-M): 2284 samples

Contrastive Learning and Randomly Aligned Substitution both contribute to the improvement on sentence retrieval

mRASP2 produces Better Semantic Alignment

Visualization of Sentence Representation

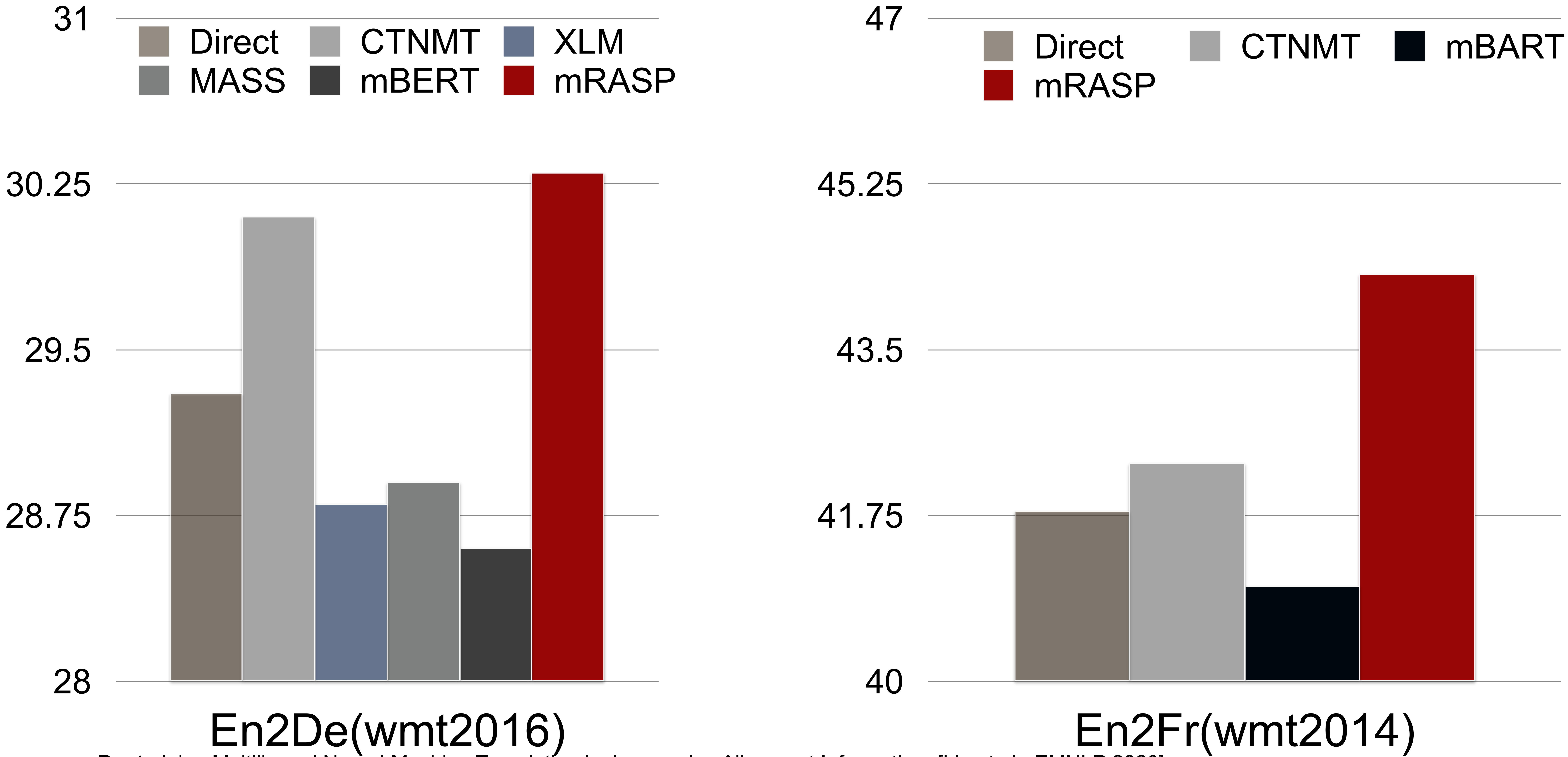




Better Alignment of En, Ja, De Representations !!

mRASP Fine-tunes better: Rich resource works

- En->Fr +1.1BLEU.



Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP: Unseen languages

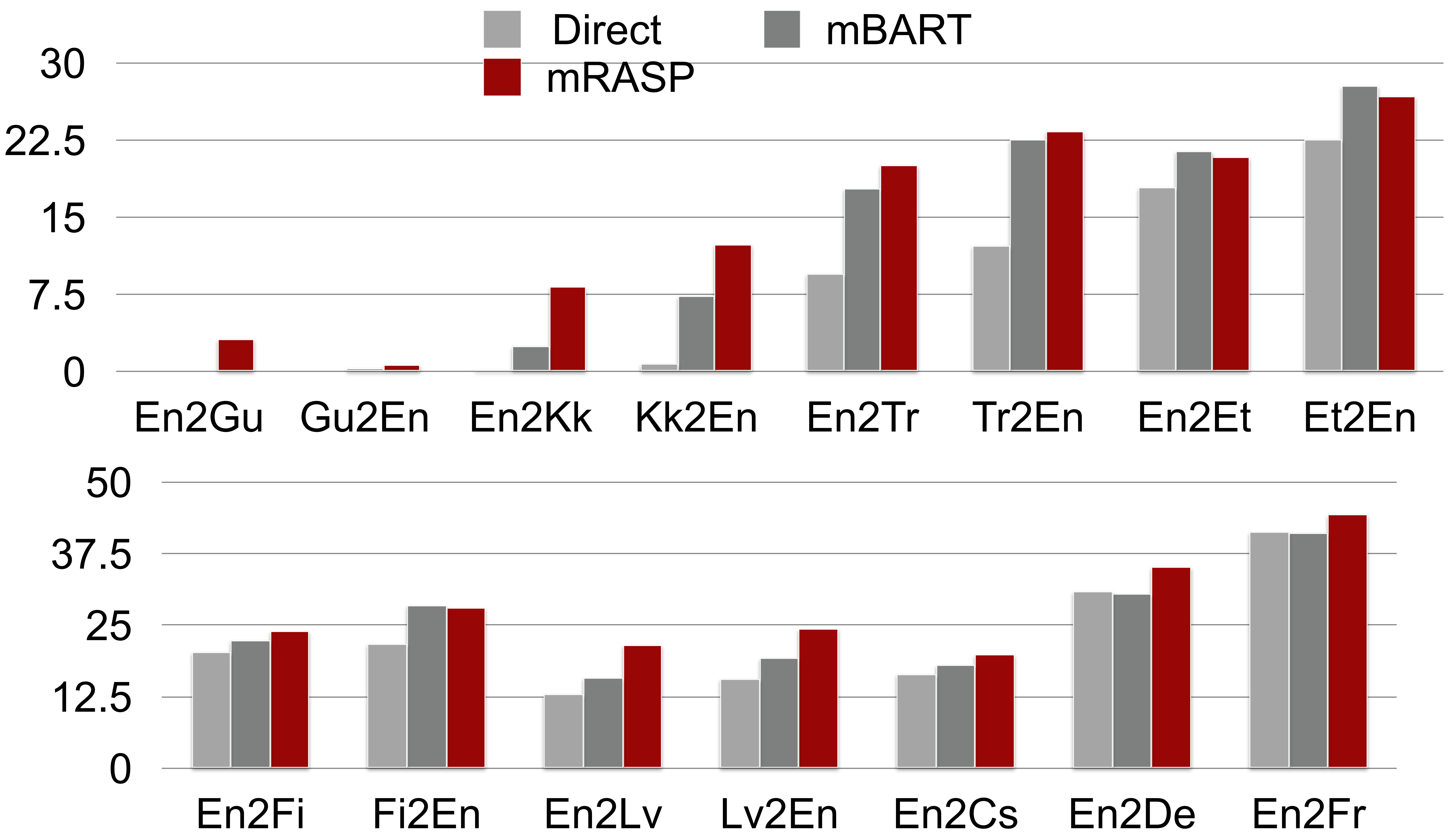
- mRASP generalizes on all exotic scenarios.

		Fr-Zh(20K)		De-Fr(9M)	
		—>	<—	—>	<—
Exotic Pair	Direct	0.7	3	23.5	21.2
	mRAS	25.8	26.7	29.9	23.4
		Nl-Pt(12K)		Da-El(1.2M)	
		—>	<—	—>	<—
Exotic Full	Direct	0.0	0.0	14.1	16.9
	mRAS	14.1	13.2	17.6	19.9
		En-Mr(11K)		En-Gl(1.2M)	
		—>	<—	—>	<—
Exotic Source/Target	Direct	6.4	6.8	8.9	12.8
	mRAS	22.7	22.9	32.1	38.1
		En-Eu(726k)		En-Sl(2M)	
		—>	<—	—>	<—
	Direct	7.1	10.9	24.2	28.2
	mRAS	19.1	28.4	27.6	29.5

12k: **Direct** not work **VS** **mRASP** achieves 10+ BLEU!!

mRASP: Compare with other methods

- mRASP outperforms mBART for all but two language pairs.



Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

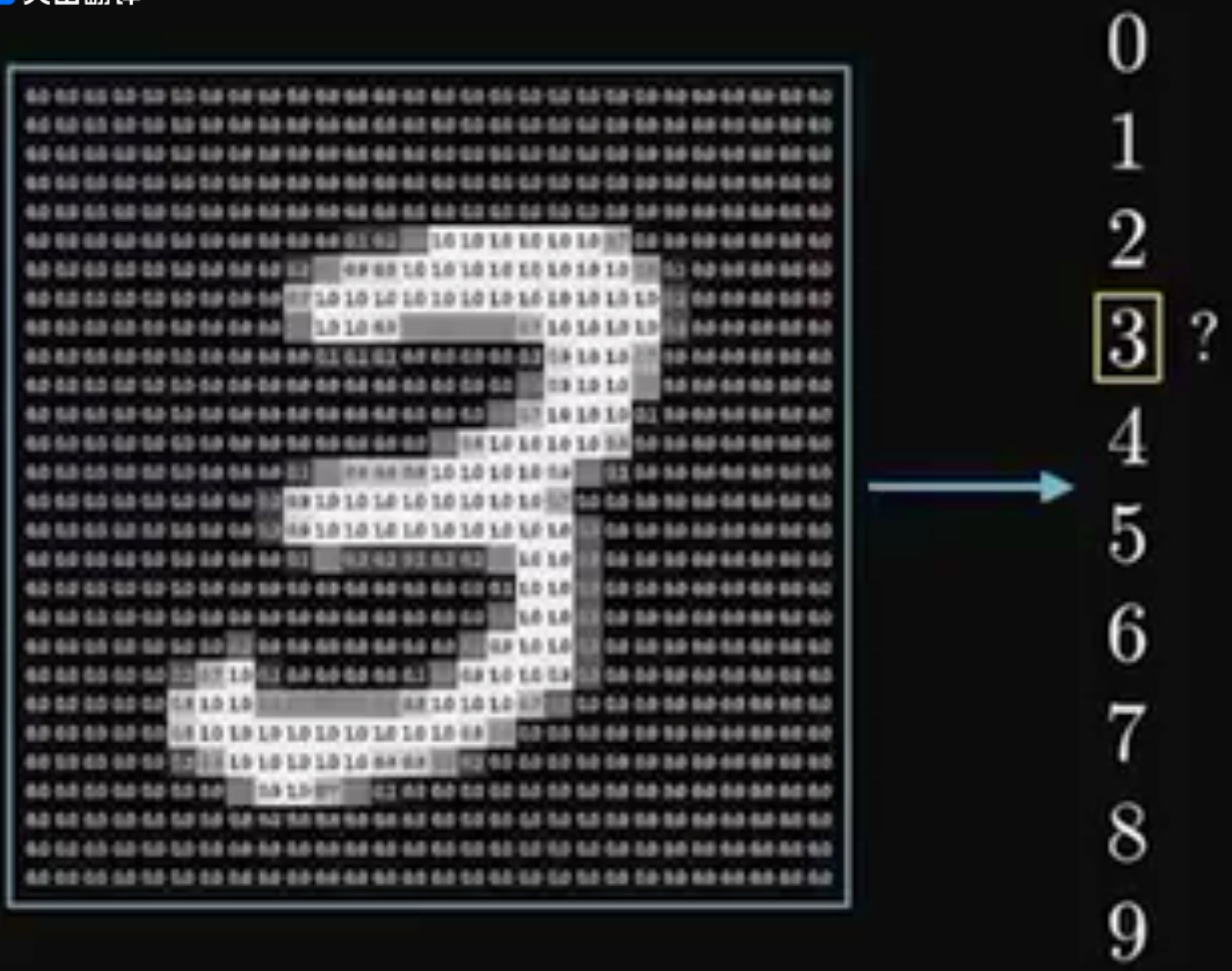
Speech Translation

Speech-to-Text Translation(ST)

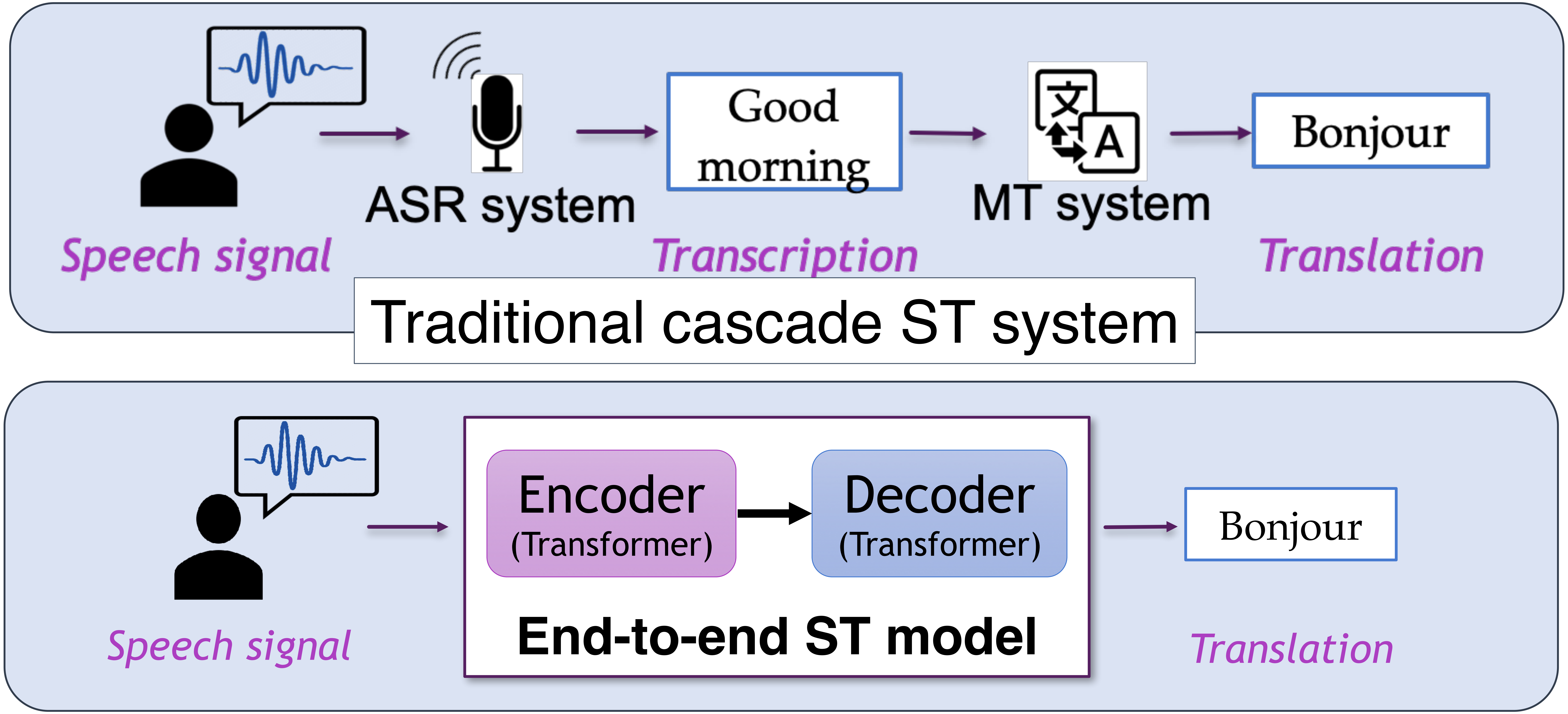
- source language ***speech(audio)*** ⟶ target lang ***text***

Application Type
• (Non-streaming) ST e.g. video translation • Streaming ST e.g. realtime conference translation

System
• Cascaded ST • End-to-end ST


0
1
2
3 ?
4
5
6
7
8
9

End-to-end model: makes ST easier

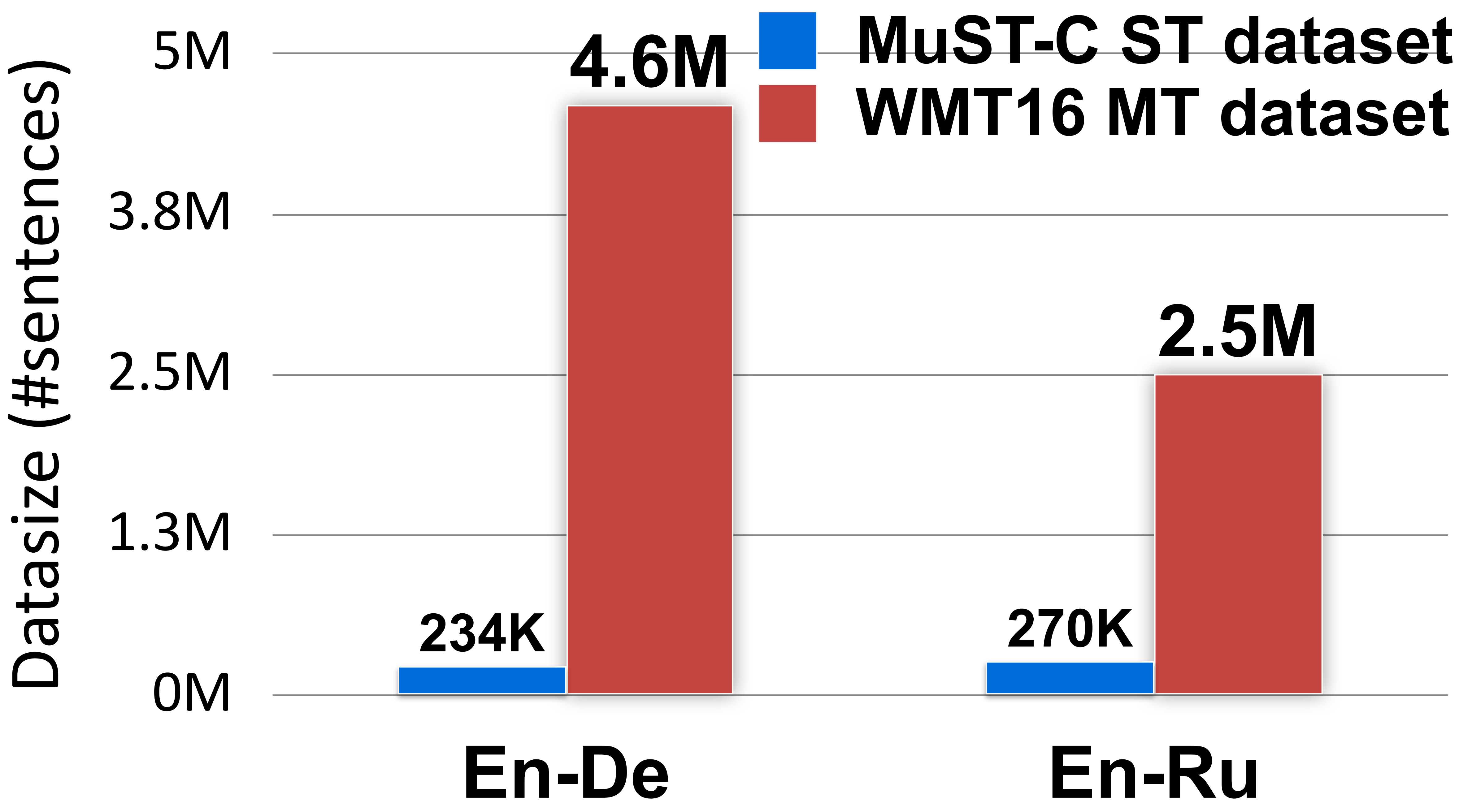


* Pictures are from our previous video talk at InterSpeech 2021.

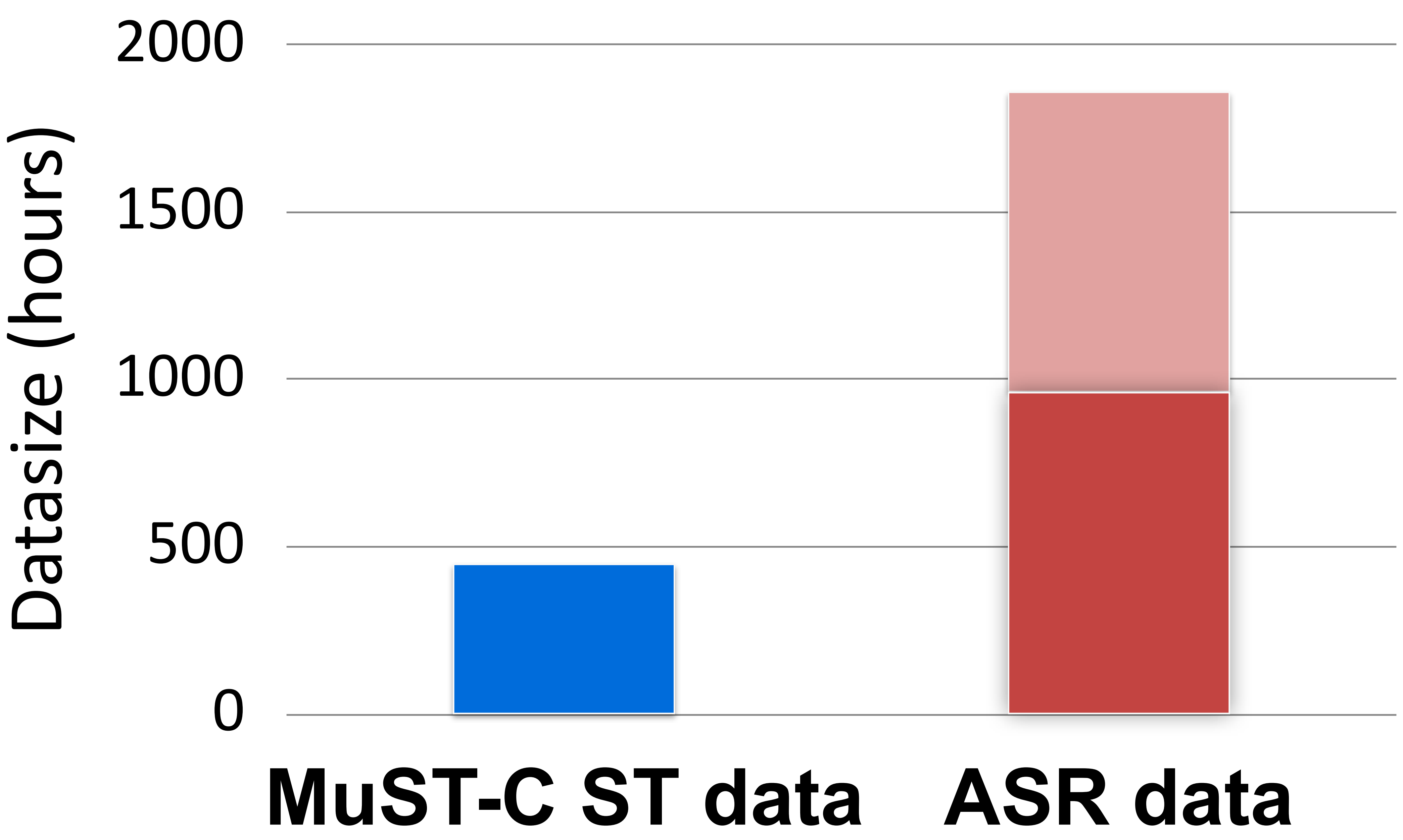
Challenge

- Data scarcity - lack of large parallel audio-translation corpus
- Modality Disparity between speech and text

Dataset size (Text)
ST vs MT



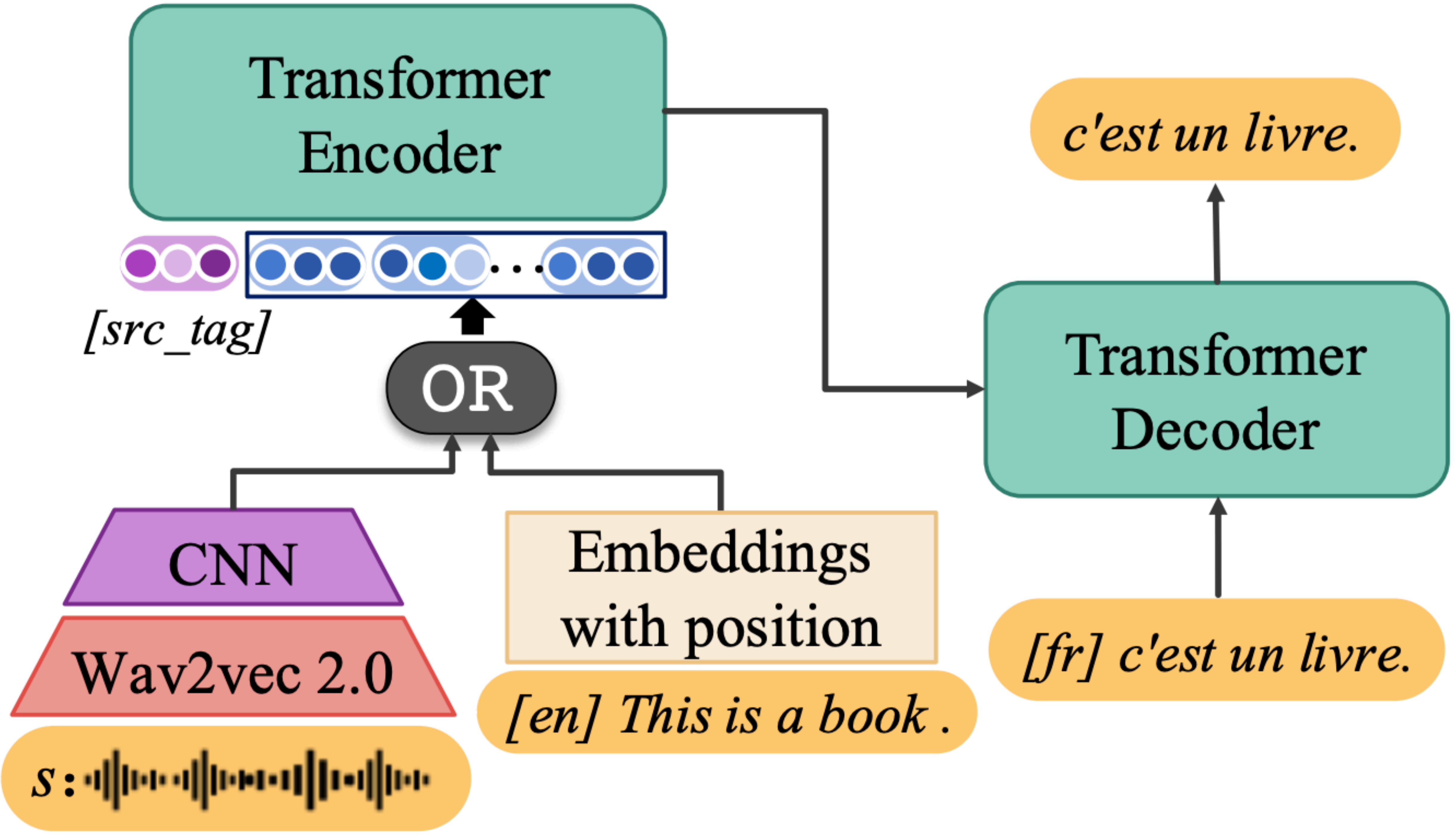
Dataset size
ST vs ASR



Multi-task learning leads to better ST

- To joint train
  ST, ASR and MT tasks.

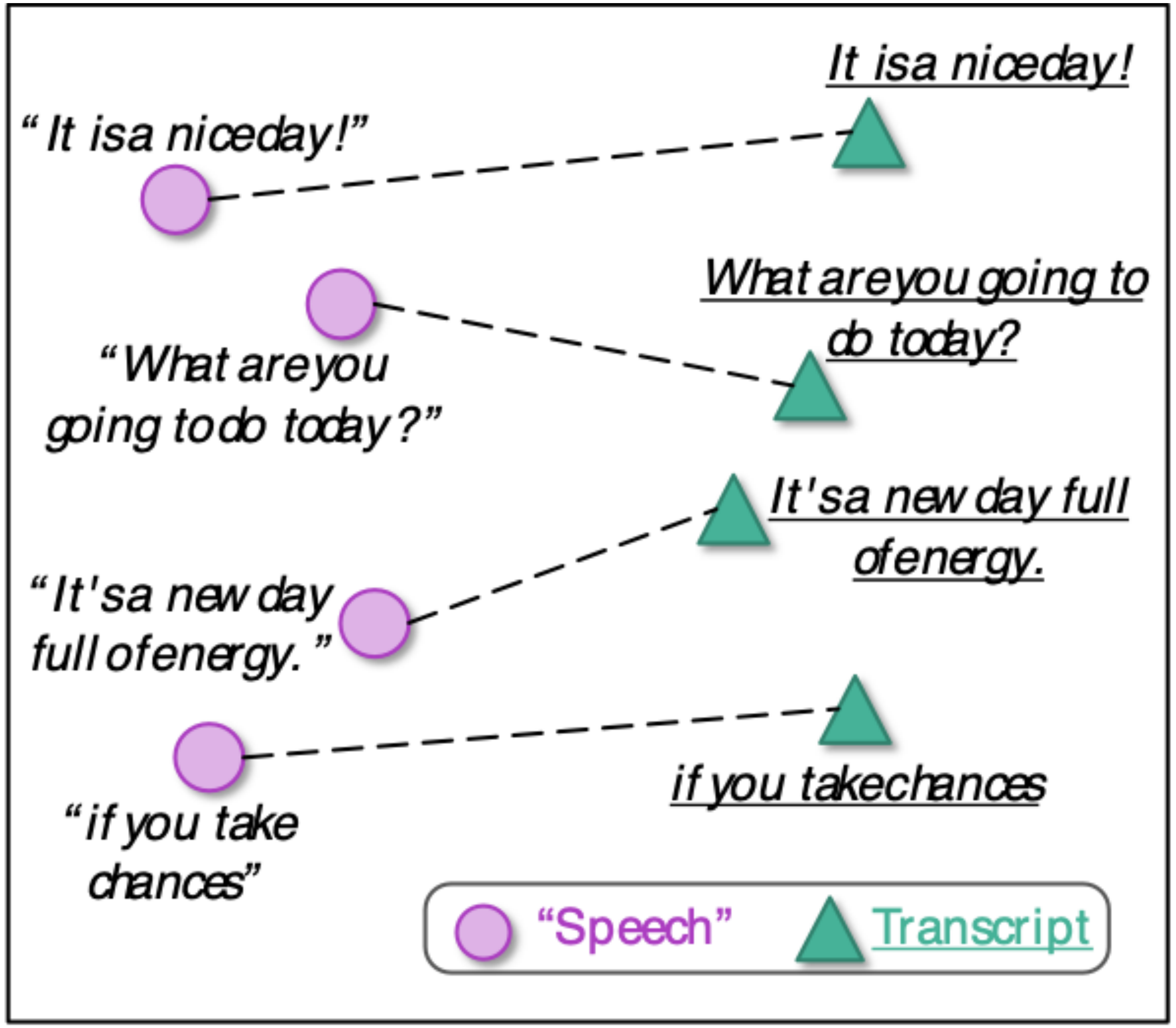
- **Advantages**:
  – Better generalization
  – Utilizing large-scale extra



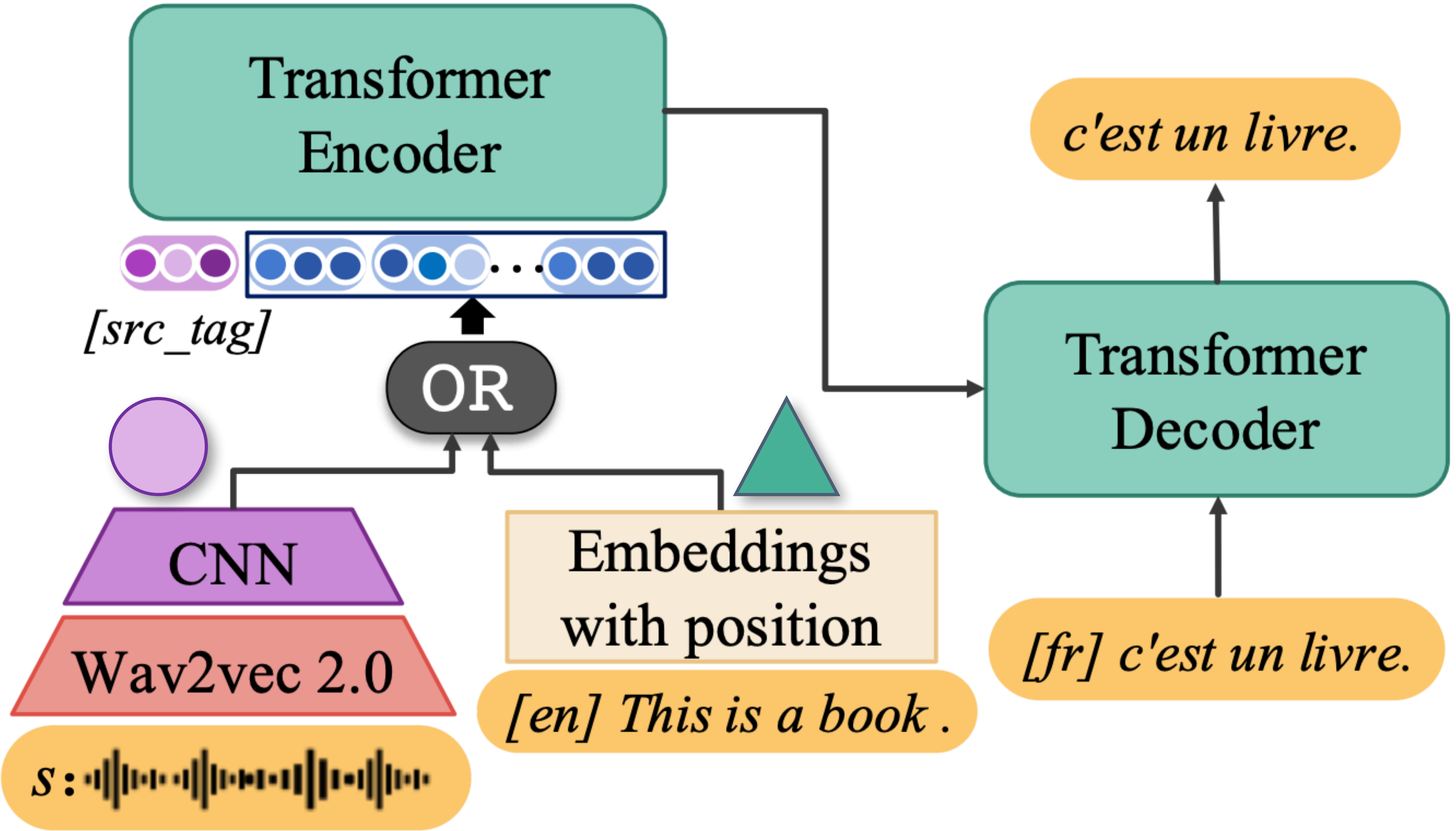
XSTNet (Ye et al., 2021[1])

[1] Rong Ye, Mingxuan Wang, and Lei Li. XSTNet: End-to-end Speech Translation via Cross-modal Progressive Training. InterSpeech 2021.

Representation Perspective: Modality Gap Exists!

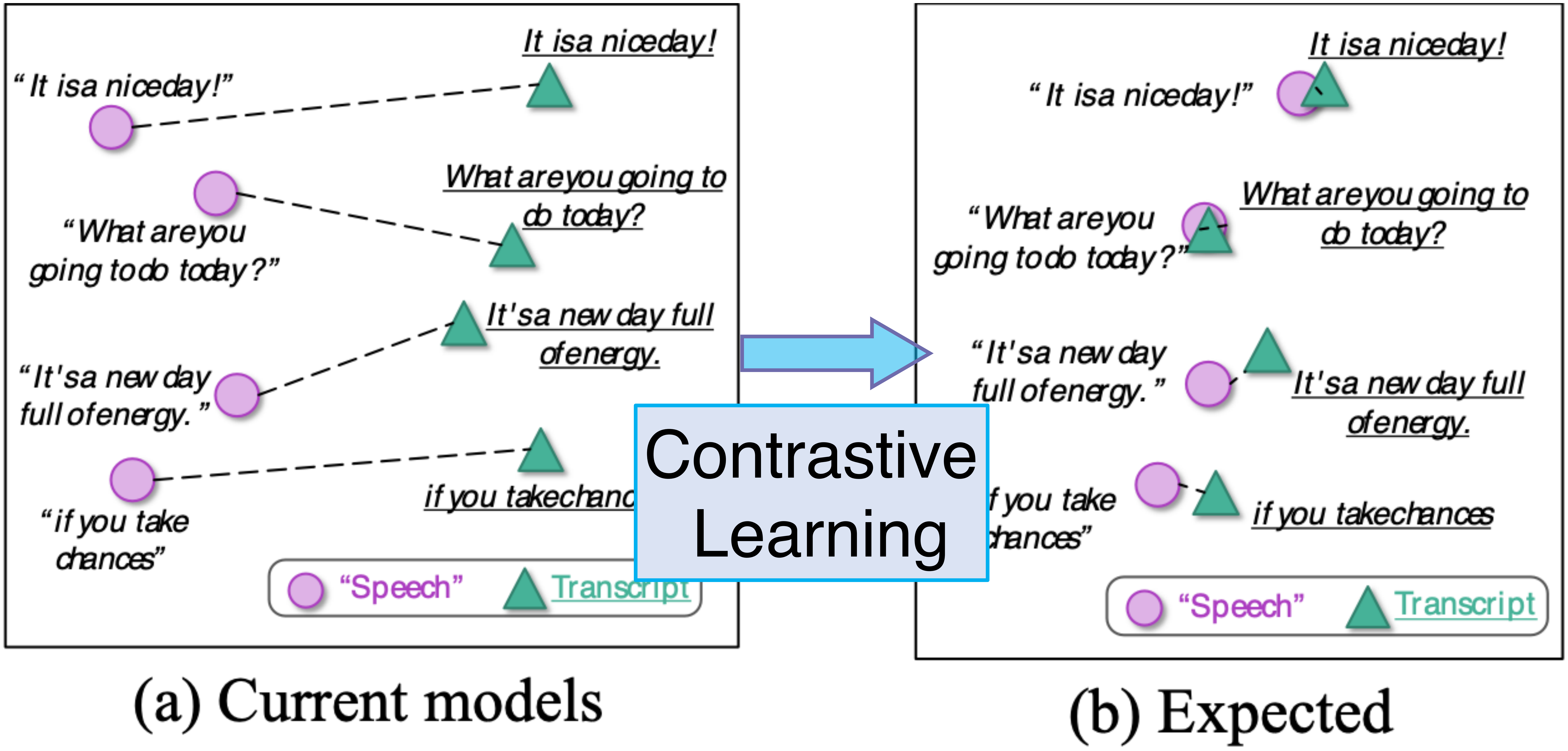


XSTNet (Ye et al., 2021[1])

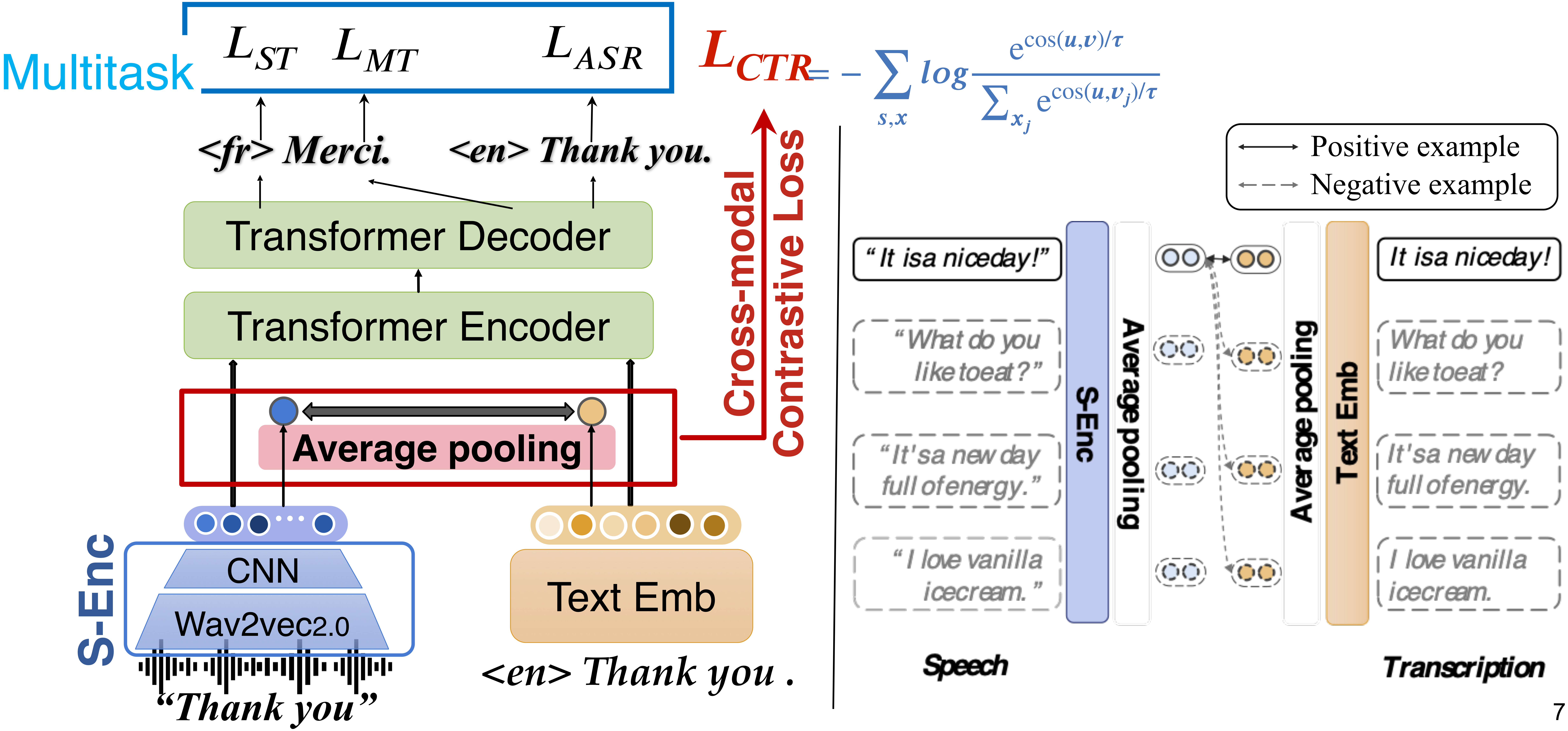


[1] Rong Ye, Mingxuan Wang, and Lei Li. XSTNet: End-to-end Speech Translation via Cross-modal Progressive Training. InterSpeech 2021.

Text and speech with same meaning should be similar in representation!



Contrastive Learning (ConST)



$$L_{CTR} = -\sum_{s,x} \log \frac{e^{\cos(u,v)/\tau}}{\sum_{x_j} e^{\cos(u,v_j)/\tau}}$$

Experimental Setups

- **Datasets**
  - All 8 directions of **MuST-C** benchmark
  - MT datasets for pretraining

En→	ST (MuST-C) hours	ST (MuST-C) #sents	MT name	MT #sents
De	408	234K	WMT16	4.6M
Fr	492	280K	WMT14	40.8M
Ru	489	270K	WMT16	2.5M
Es	504	270K	WMT13	15.2M
Ro	432	240K	WMT16	0.6M
It	465	258K	OPUS100	1.0M
Pt	385	211K	OPUS100	1.0M
Nl	442	253K	OPUS100	1.0M

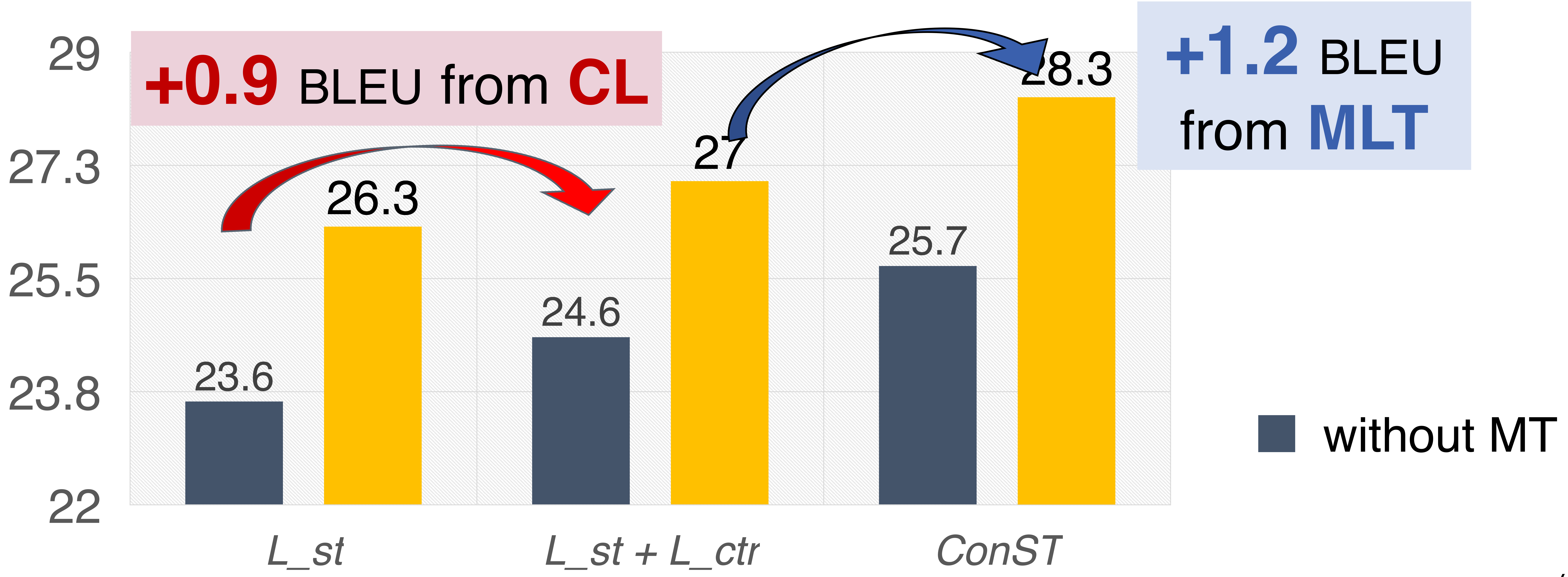
- **Settings**
  - without external MT data
  - with external MT data

- **Baseline**
  - W2v2-Transformer
  - XSTNet (Ye et. al.)[1]

[1] Rong Ye, Mingxuan Wang, and Lei Li. XSTNet: End-to-end Speech Translation via Cross-modal Progressive Training. InterSpeech 2021.

Both **Multi-task** and **Contrastive** Learning are important!

$$\mathcal{L} = \boxed{\mathcal{L}_{\mathrm{ST}} + \mathcal{L}_{\mathrm{ASR}} + \mathcal{L}_{\mathrm{MT}}} + \lambda \mathcal{L}_{\mathrm{CTR}}$$



Contrastive Learning improves ST

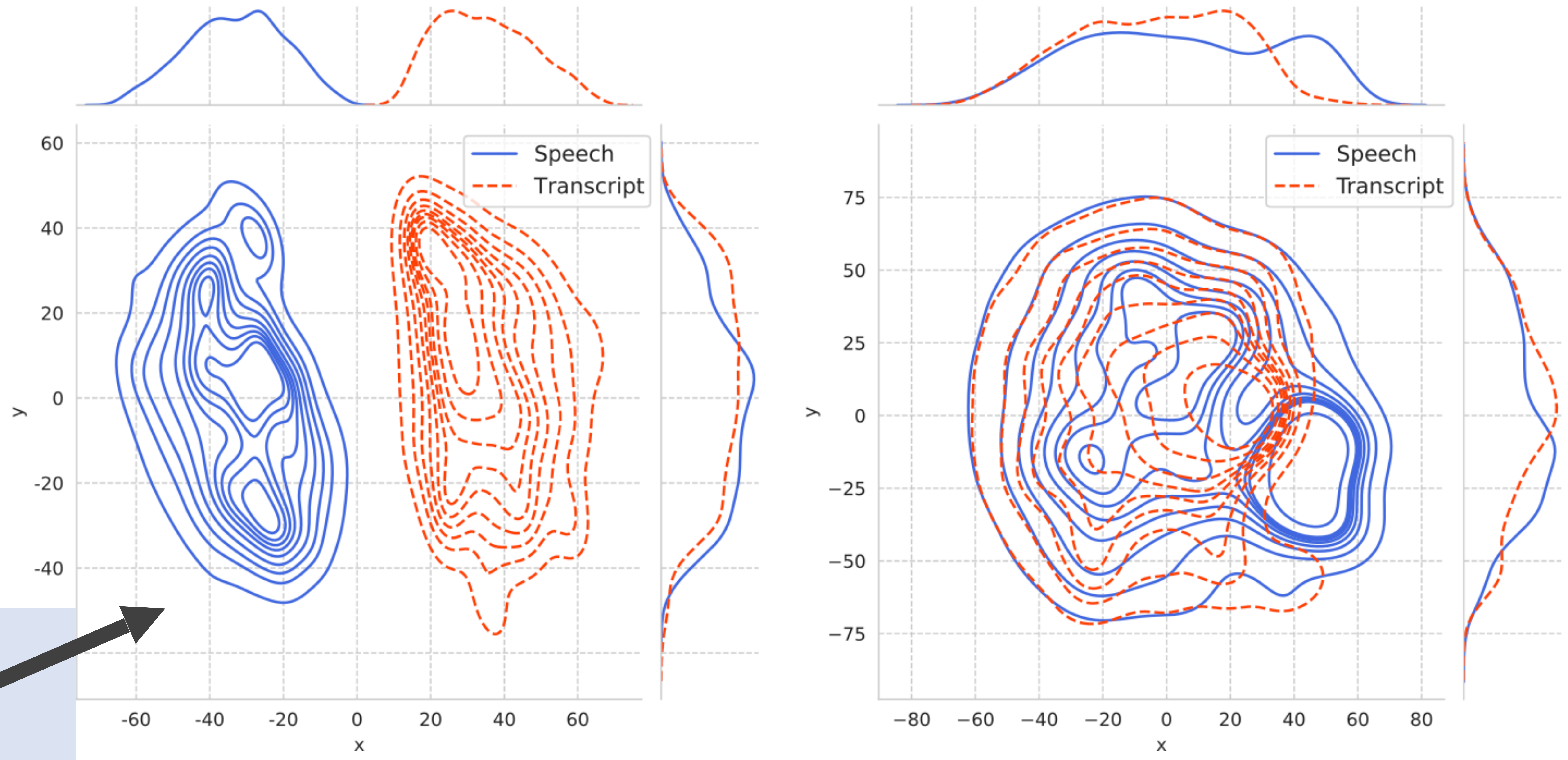
Models	External Data				BLEU								
	Speech	Text	ASR	MT	De	Es	Fr	It	Nl	Pt	Ro	Ru	Avg.
w/o external MT data													
Fairseq ST (Wang et al., 2020a)	-	-	-	-	22.7	27.2	32.9	22.7	27.3	28.1	21.9	15.3	24.8
NeurST (Zhao et al., 2021a)	-	-	-	-	22.8	27.4	33.3	22.9	27.2	28.7	22.2	15.1	24.9
Espnet ST (Inaguma et al., 2020)	-	-	-	-	22.9	28.0	32.8	23.8	27.4	28.0	21.9	15.6	25.1
Dual Decoder (Le et al., 2020)	-	-	-	-	23.6	28.1	33.5	24.2	27.6	30.0	22.9	15.2	25.6
W-Transf. (Ye et al., 2021)	✓	-	-	-	23.6	28.4	34.6	24.0	29.0	29.6	22.4	14.4	25.7
Speechformer (Papi et al., 2021)	-	-	-	-	23.6	28.5	-	-	27.7	-	-	-	-
LightweightAdaptor (Le et al., 2021)	-	-	-	-	24.7	28.7	35.0	25.0	28.8	31.1	23.8	16.4	26.6
Self-training (Pino et al., 2020)	✓	-	✓	-	25.2	-	34.5	-	-	-	-	-	-
SATE (Xu et al., 2021)	-	-	-	-	25.2	-	-	-	-	-	-	-	-
BiKD (Inaguma et al., 2021)	-	-	-	-	25.3	-	35.3	-	-	-	-	-	-
Mutual-learning (Zhao et al., 2021b)	-	-	-	-	-	28.7	36.3	-	-	-	-	-	-
XSTNet (Ye et al., 2021)	✓	-	-	-	25.5	29.6	36.0	25.5	30.0	31.3	**25.1**	16.9	27.5
ConST	✓	-	-	-	**25.7**	**30.4**	**36.8**	**26.3**	**30.6**	**32.0**	24.8	**17.3**	**28.0**
w/ external MT data													
Chimera (Han et al., 2021)	✓			✓	27.1†	30.6	35.6	25.0	29.2	30.2	24.0	17.4	27.4
XSTNet (Ye et al., 2021)	✓	-	-	✓	27.1	30.8	38.0	26.4	31.2	32.4	**25.7**	18.5	28.8
STEMM (Fang et al., 2022)	✓	-	-	✓	28.7	31.0	37.4	25.8	30.5	31.7	24.5	17.8	28.4
ConST	✓	-	-	✓	**28.3**	**32.0**	**38.3**	**27.2**	**31.7**	**33.1**	25.6	**18.9**	**29.4**

+ 0.5 BLEU

+ 0.6 BLEU

Visualization:CL draws the distance of two modalities!

XSTNet[1]:
(BLEU=27.1)



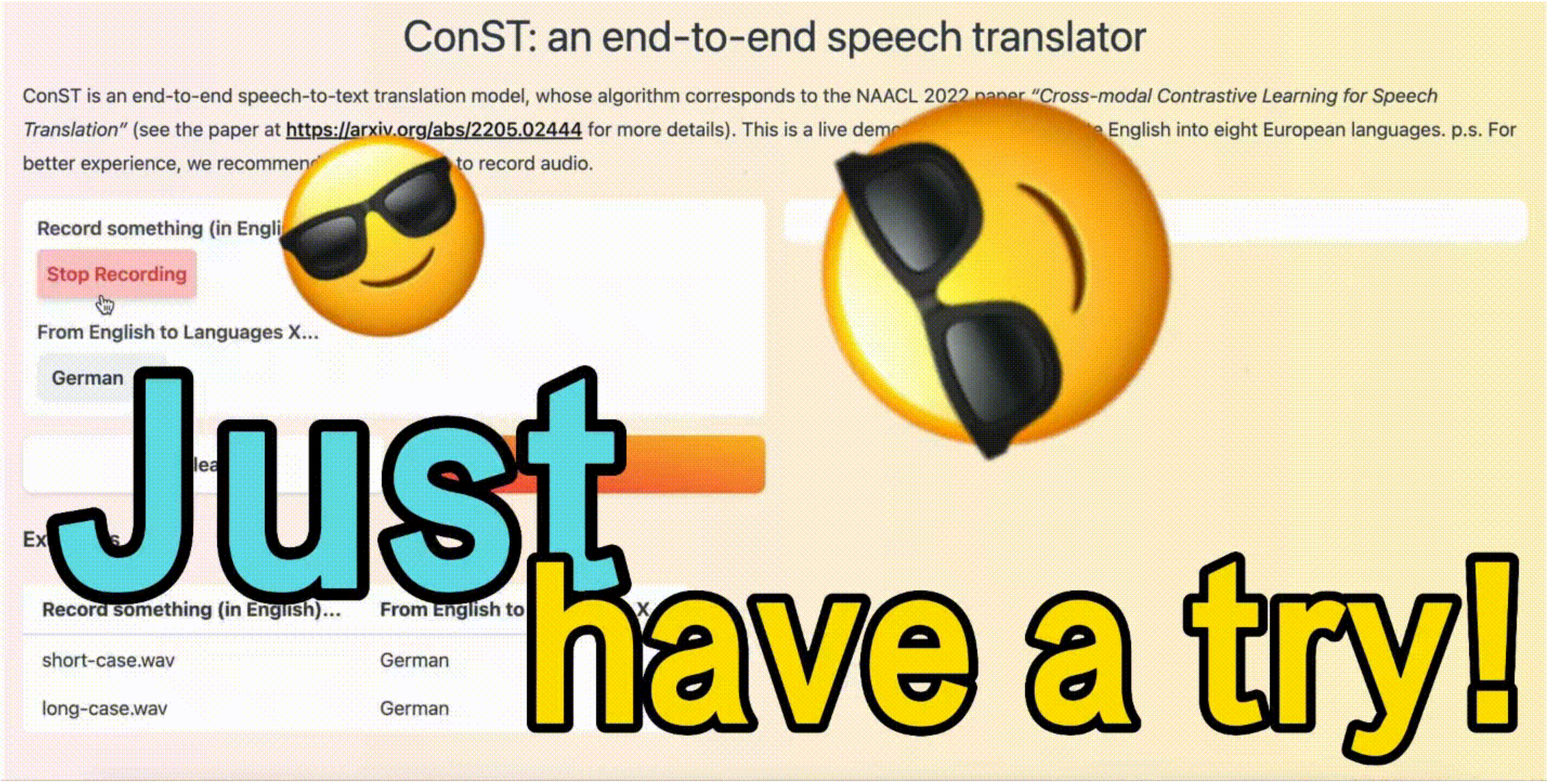
(a) w/o CTR loss (b) ConST

[1] Rong Ye, Mingxuan Wang, and Lei Li. XSTNet: End-to-end Speech Translation via Cross-modal Progressive Training. InterSpeech 2021.

Wanna have a try?

- https://huggingface.co/spaces/ReneeYe/ConST-speech2text-translator

*Best practice on ***Chrome***



MT works from my group

Machine Translation

VOLT
ACL 2021
best paper award

LaSS
ACL 2021

mRASP
EMNLP 2020
ACL 2021

GLAT
ACL 2021

REDER
NeurIPS 2021

MGNMT
ICLR 2020

Graformer
EMNLP-Findings 2021

KSTER
EMNLP 2021

CIAT
EMNLP-Findings 2021

NAT-theory
ICML 2022

switch-GLAT
ICLR 2022

Speech Translation

LUT
AAAI 2021

COSTT
AAAI 2021

Chimera
ACL-Findings 2021

XSTNet
InterSpeech 2021

STEMM
ACL 2022

MoSST
ACL 2022

ConST
NAACL 2022

Open Source Library

LightSeq
High performance sequence inference
https://github.com/bytedance/lightseq

NeurST
neural speech translation toolkit
https://github.com/bytedance/neurst

Summary and Takeaway

- Transformer is powerful MT model
- MT is still challenging
- Benefits of MNMT
  - boosting performance on low-resource
  - economic in training/deployment/maintenance
- Bringing representations of words/sentences closer across languages/modality proves beneficial
  - mRASP & mRASP2: augmenting data with randomly substitute of words from bilingual lexicon + monolingual reconstruction + contrastive learning
  - ConST: contrastive learning to bring speech and text representation closer

Resource

- Code:
  - mRASP https://github.com/PANXiao1994/mRASP2
  - ConST: https://github.com/ReneeYe/ConST
- Joint work with


Mingxuan Wang

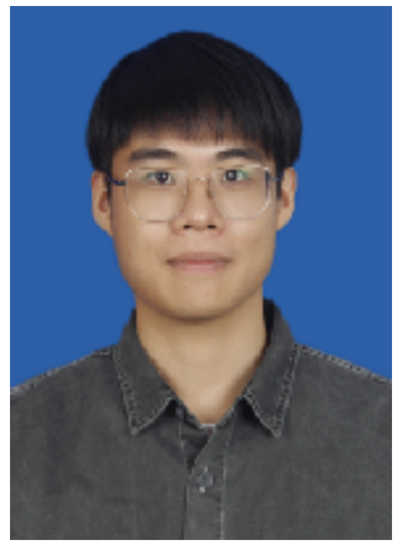

Rong Ye

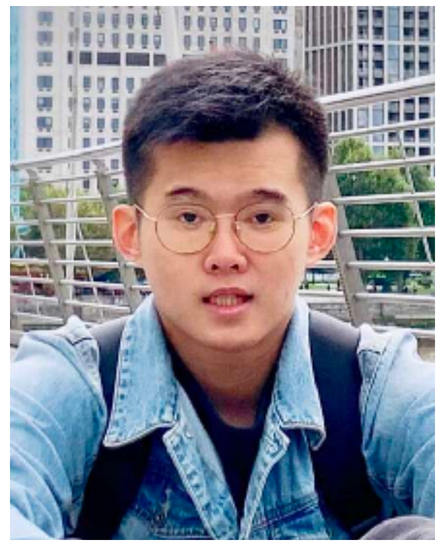

Xiao Pan


Qianqian Dong


Jingjing Xu


Yu Bao


Lihua Qian

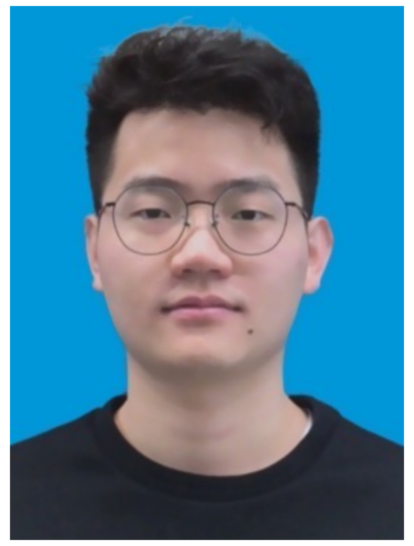

Zaixiang Zheng


Yaoming Zhu


Zewei Sun


Hao Zhou


Xiaohui Wang

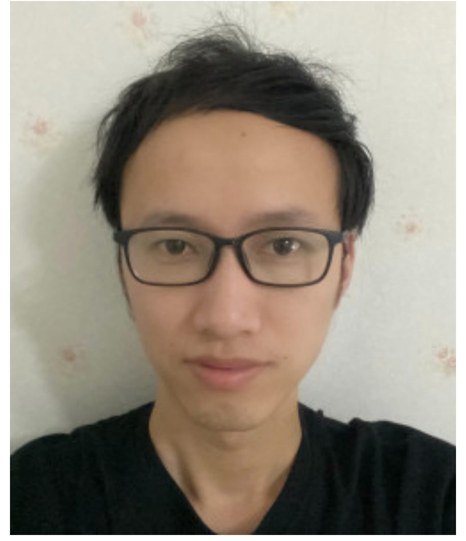

Zehui Lin


Ying Xiong


Liwei Wu

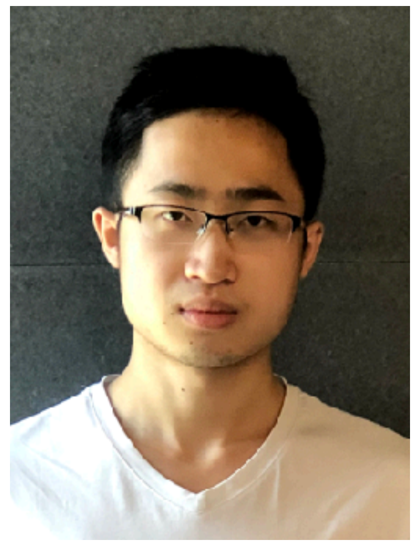

Chun Gan


Xian Qian


Yang Wei


Jiangtao Feng


Chenyang Huang


Chi Han


Chengqi Zhao