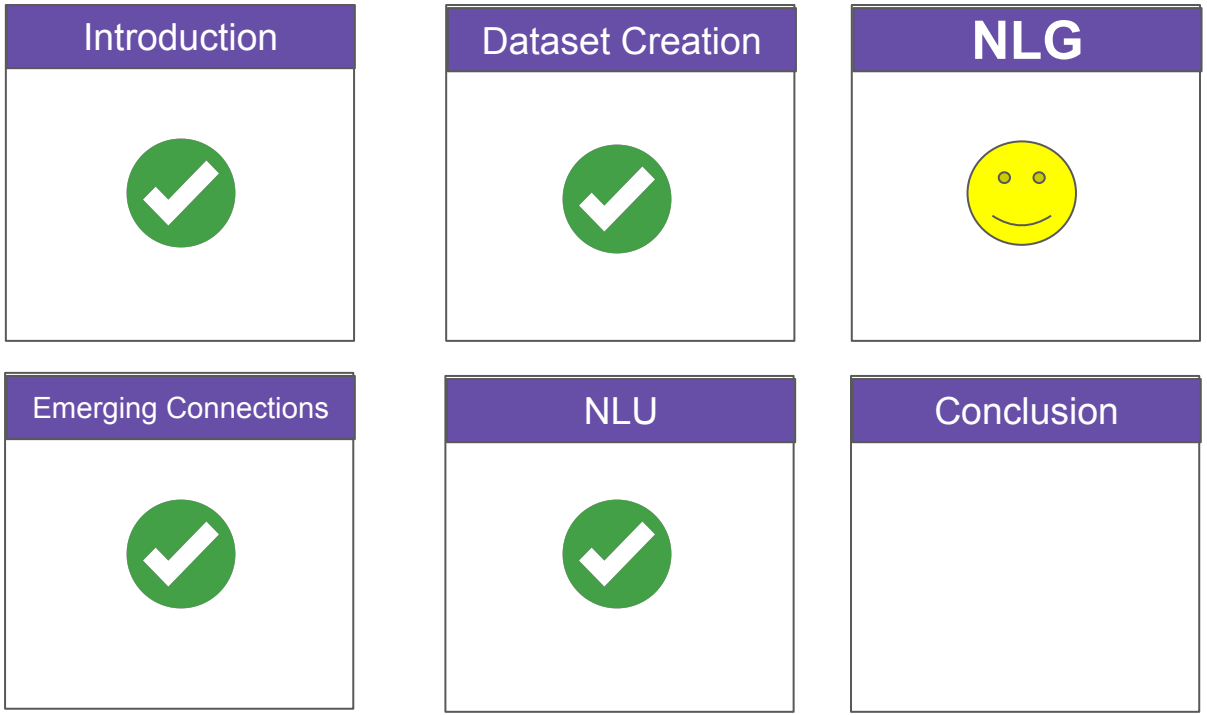
Connecting Ideas in ‘Lower-Resource’ Scenarios: NLP for National Varieties, Creoles and Other Low-resource Scenarios

Aditya Joshi, Diptesh Kanojia, Heather Lent, Hour Kaing, Haiyue Song

Tutorial Agenda



Three Parts in Module 5

- Natural Language Generation for **Low-Resource Languages**
- Natural Language Generation for **Dialects**
- Hands-on Session

Module 5 (Part 1):
Text Generation for Low-Resource Languages

We Will Introduce

Tasks:

- Machine translation
- Summarization, dialogue, and style transfer (in the dialect NLG part)

Challenges and solutions:

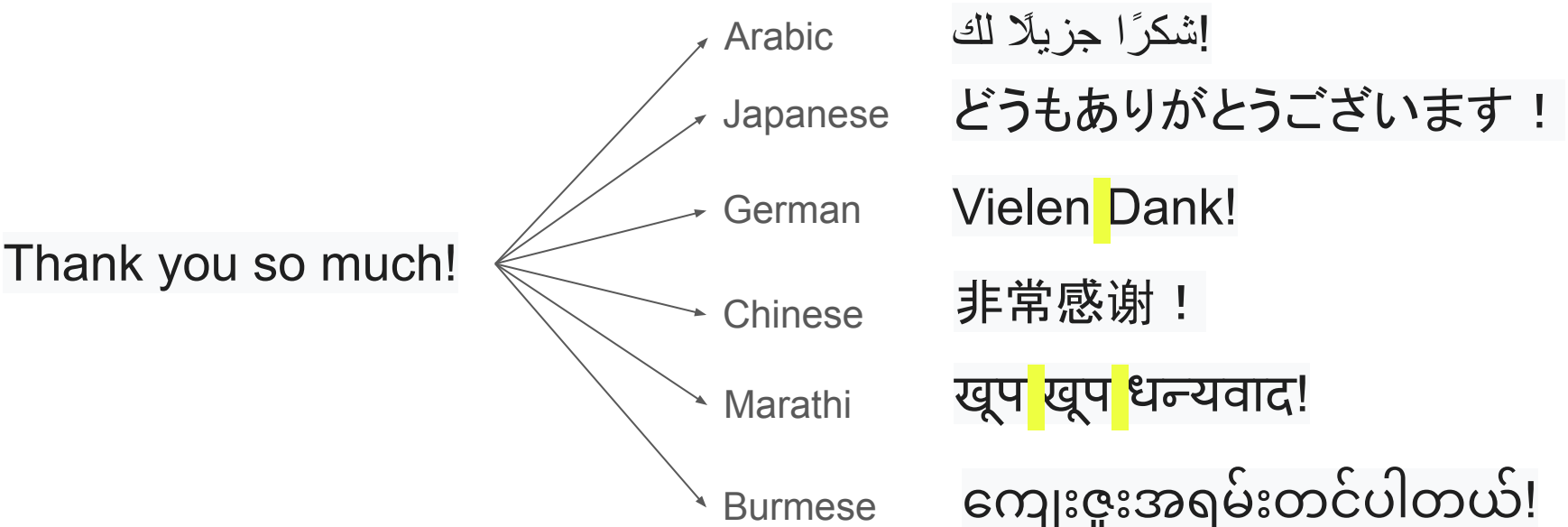
- Data scarcity -> Data augmentation
- Diverse scripts -> Script normalization

The performance of LLMs

Evaluation and Correction

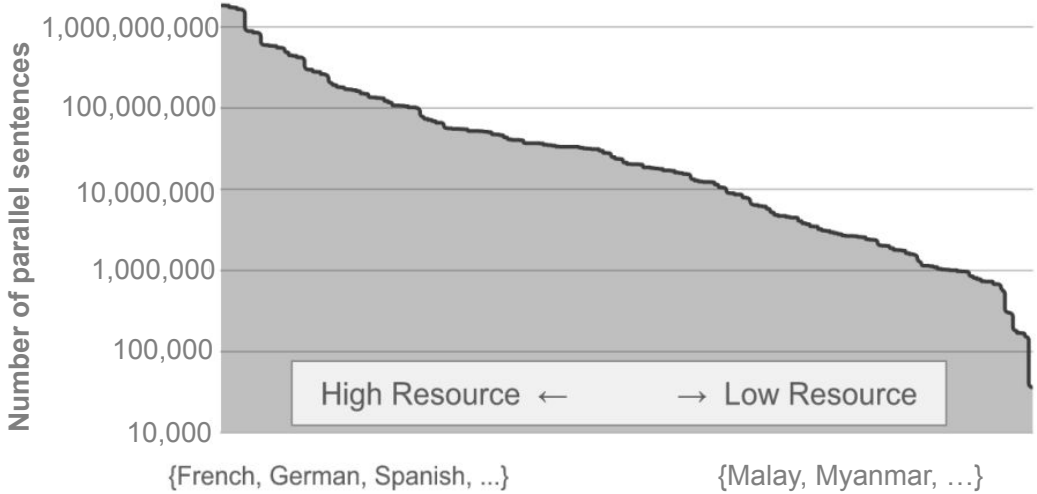
Machine Translation for Various (Low-Resource) Languages

- Different scripts, some with large number of distinct characters.
- No word boundaries.



Disparity Exists Among Languages: Data Size

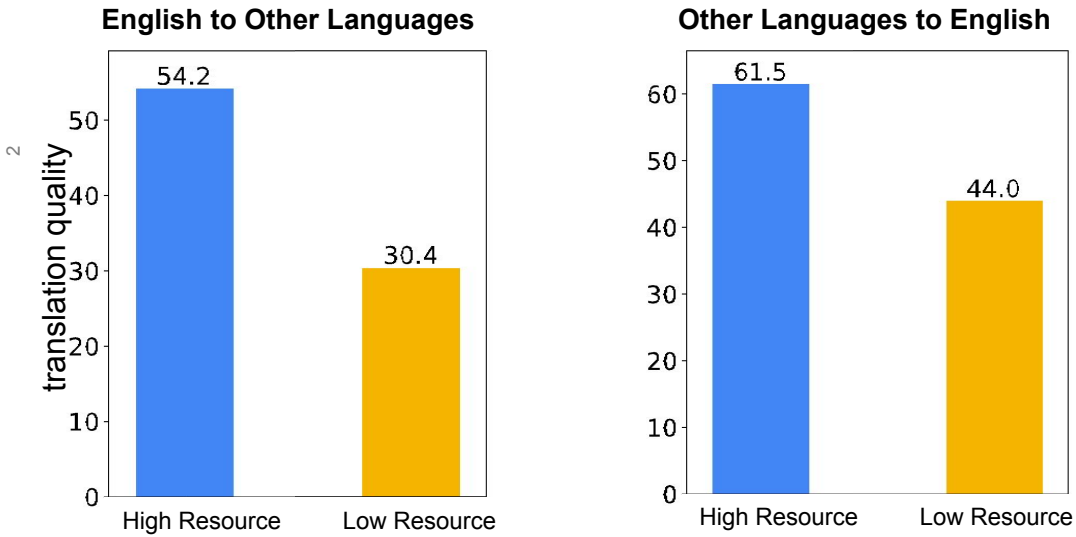
- Long-tail distribution of **parallel data** size over language pairs.
  - We consider the *Any*-English pairs.
  - The number of high-resource language pairs is small.
  - Most pairs in real-world are low-resource.



Google AI. 2019. [Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges]

Disparity Exists Among Languages: Performance

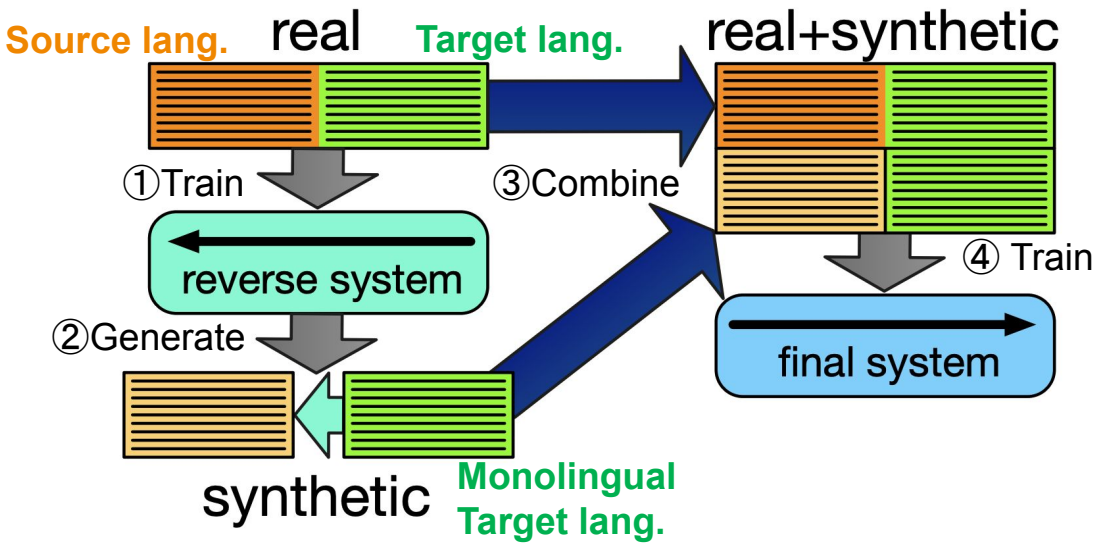
- High-resource languages like English-German benefit from MT advancements.
- **Low-resource** language pairs such as English-Malay with less training data (<100k **parallel** sentences[1]) did not achieve similar performance.



[1]Meta AI. 2023. Small Data, Big Impact: Leveraging Minimal Data for Effective Machine Translation
[2]measured by chrF++ score, the higher the better. Details in NLLB Team. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation

Back Translation as Data Augmentation

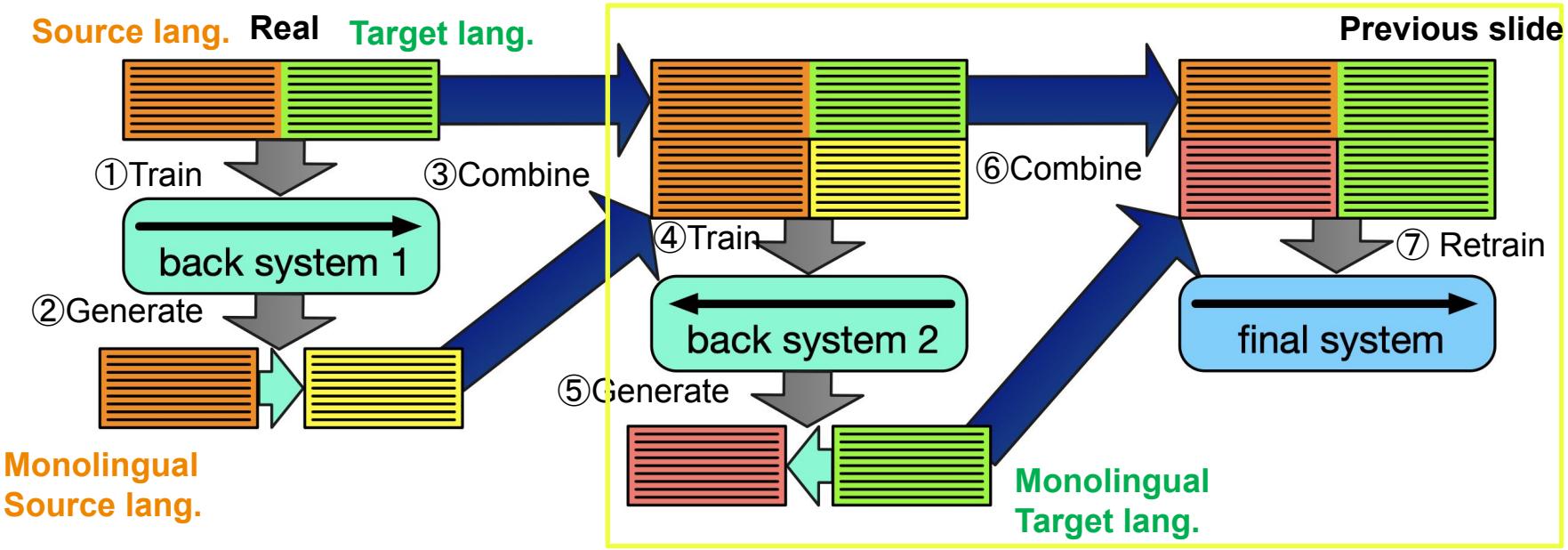
- Labelled data (e.g. parallel corpora) is expensive,
- How can we leverage **monolingual data**?



Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data.

Iterative Back Translation

- Motivation: a better back system will benefit the final system
- Idea: obtain a better back system 2 through back-back-translation system 1



Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative Back-Translation for Neural Machine Translation

More Synthetic Data by Sampling

- Motivation: generate **more synthetic source sentence** for a target sentence
- Method: use sampling instead of greedy/beam search, with quality control

	Log-likelihood	Synthetic Source Sentence
	-2.25	what should i do when i get injured or sick in japan ?
	-2.38	what should i do *if* i get injured or sick in japan ?
	-5.20	what should i do *if* i get injured or *illness* in japan ?
Filter out	-5.52	what should *we* do when *we* get injured or sick in japan ?
	-13.87	*if i get injured or a sickness in japan , what shall i do ?*
Target Sentence		日本で怪我や病気をしたときはどうすればいいのでしょうか？
Manual Back-Translation		what should i do when i get injured or sick in japan ?

Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita. 2018.
Enhancement of Encoder and Attention Using Target Monolingual Corpora in Neural Machine Translation.

How Well Does Back Translation Work?

- Back-translated sentences by different approaches:

	Perplexity
human data	75.34
beam	72.42
sampling	500.17
top10	87.15
beam+noise	2823.73

source	Diese gegenstzlichen Auffassungen von Fairness liegen nicht nur der politischen Debatte zugrunde.
reference	These competing principles of fairness underlie not only the political debate.
beam	These conflicting interpretations of fairness are not solely based on the political debate.
sample	*Mr President,* these contradictory interpretations of fairness are not based solely on the political debate.
top10	Those conflicting interpretations of fairness are not solely at the heart of the political debate.
beam+noise	conflicting BLANK interpretations BLANK are of not BLANK based on the political debate.

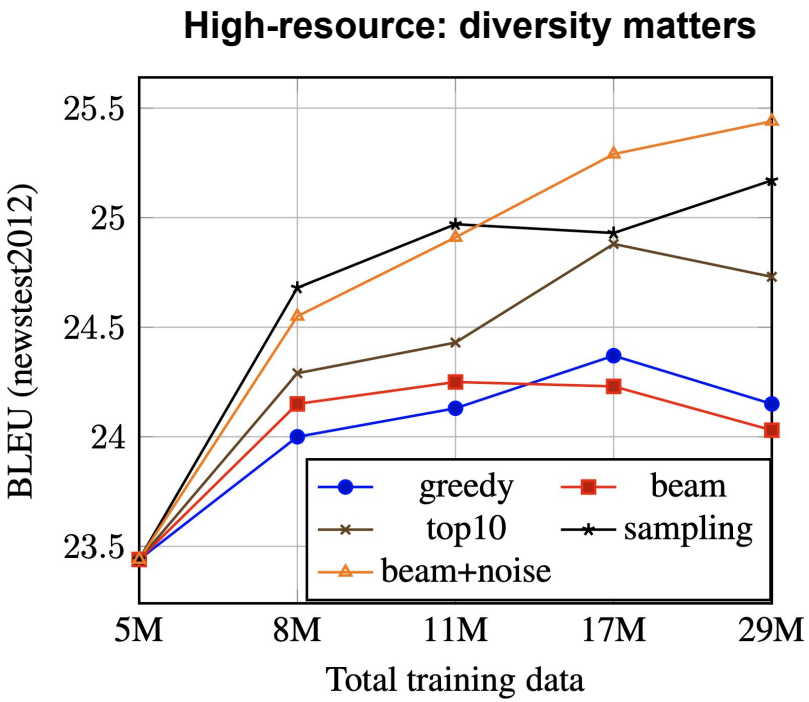
sample from top-10 words

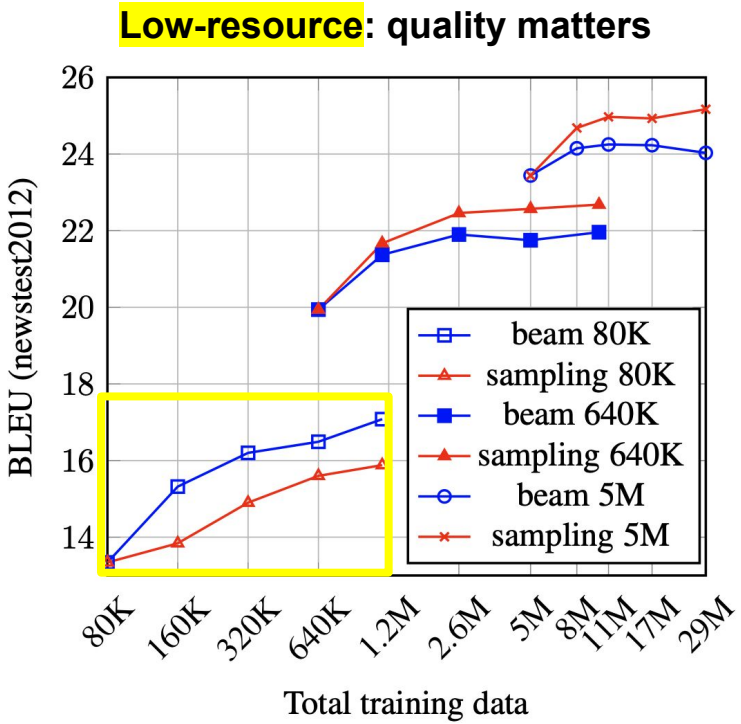
noise: delete/replace/swap

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale.

How Well Does Back Translation Work?

- Results

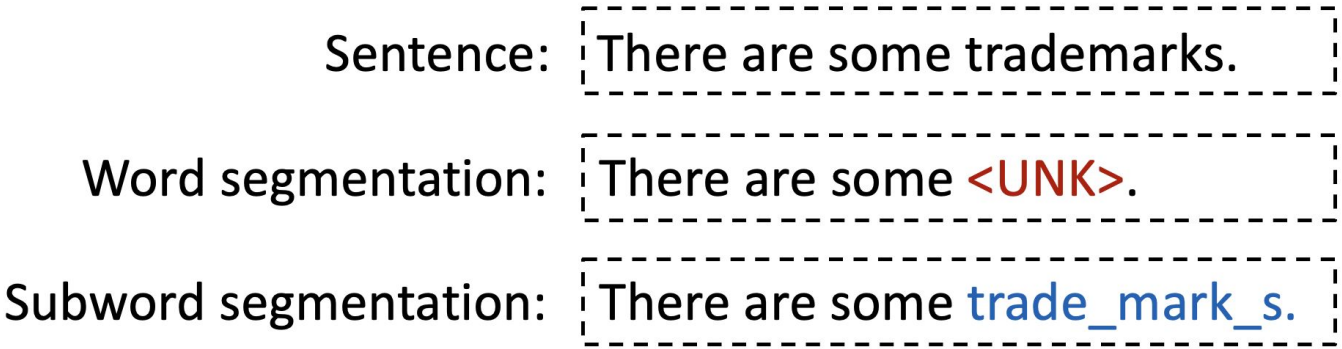




Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale.

Data Augmentation on Subwords (Background)

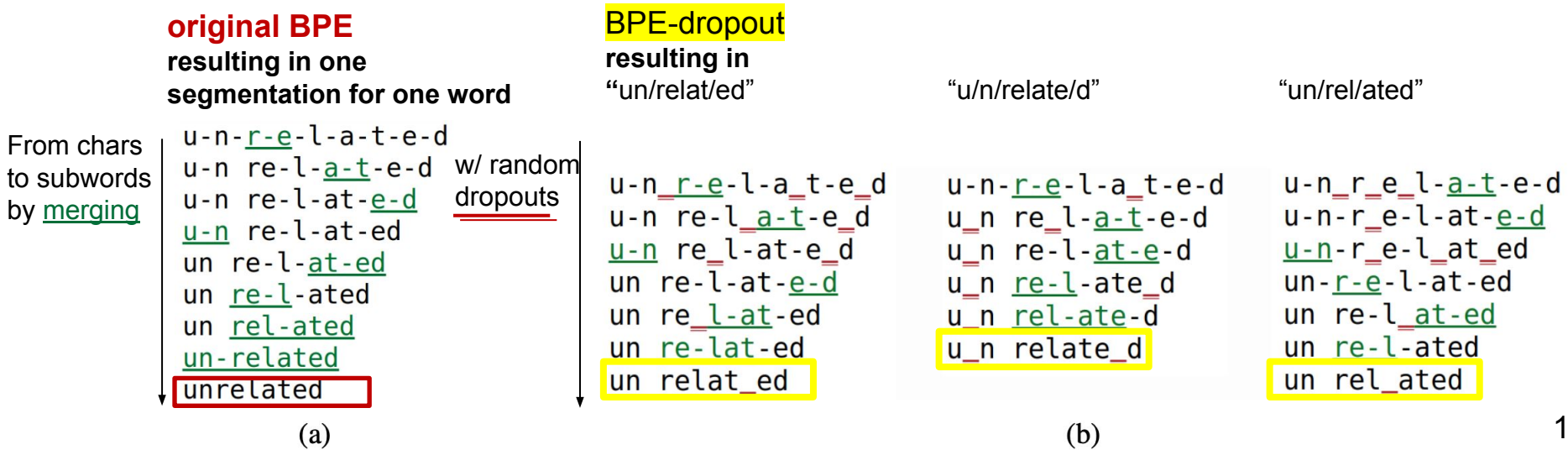
- Default NMT systems use subwords.
- Compared to *word*, subwords handles **unseen words** by segmenting them into **seen subwords** in the subword vocabulary.
  - The vocabulary size is finite (10k to 100k), but the number of words is infinite, causing the out-of-vocabulary (OOV) problem.



Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units
Taku Kudo, John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing

BPE-Dropout

- A dropout version of Byte-Pair-Encoding (BPE)
- There are multiple subword tokenizations for one sentence.
- Instead of training on one tokenization, leveraging all.



Subword Regularization

- A regularized version of SentencePiece.
- Instead of training on one tokenization, leveraging all!
- Simple but effective on low-resource scenarios.

Original SentencePiece

Regularizations

Subwords (_ means spaces)	Vocabulary id sequence
_Hell/o/_world	13586 137 255
_H/ello/_world	320 7363 255
_He/llo/_world	579 10115 255
_/He/l/l/o/_world	7 18085 356 356 137 255
_H/el/l/o/_/world	320 585 356 137 7 12295

Table 1: Multiple subword sequences encoding the same sentence "Hello World"

Domain (size)	Corpus	Language pair	Baseline (BPE)	Proposed (SR)
Web (5k)	IWSLT15	en → vi	13.86	17.36*
		vi → en	7.83	11.69*
		en → zh	9.71	13.85*
		zh → en	5.93	8.13*
	IWSLT17	en → fr	16.09	20.04*
		fr → en	14.77	19.99*
	WMT14	en → de	22.71	26.02*
		de → en	26.42	29.63*
		en → cs	19.53	21.41*
		cs → en	25.94	27.86*
Patent (2k)	WMT14	en → de	15.63	25.76*
		de → en	22.74	32.66*
		en → cs	16.70	19.38*
		cs → en	23.20	25.30*

Huge improvement!

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates.

Character Decomposition as Data Augmentation

- Characters in languages such as **Chinese**, Japanese, Korean may contain **sub-characters**.

Character Decomposition

森 → 木 / 木 / 木
forest — tree — tree — tree

驰 → 马/也
run — horse

林 → 木 / 木
woods — tree — tree

明 → 日/月
bright — sun — moon

website for character decomposition: https://www.cns11643.gov.tw/searchQ.jsp?ID=0

Character Decomposition: Method

- Replacing characters with ideograph sequences in the training data.

	Language	Word
Method in this paper*	JP-character	風 景
	JP-ideograph	几一虫 日亠口小_1
	JP-stroke	丿乁一丨冂一丨ノ丶 丨冂一一丶一丨フ一亅ノ丶_1
Method in this paper	CN-character	风 景
	CN-ideograph	几乂 日亠口小_1
	CN-stroke	丿乁ノ丶 丨冂一一丶一丨フ一亅ノ丶_1
	EN	landscape

Longtu Zhang and Mamoru Komachi. 2018. Neural Machine Translation of Logographic Language Using Sub-character Level Information.

Character Decomposition: Results

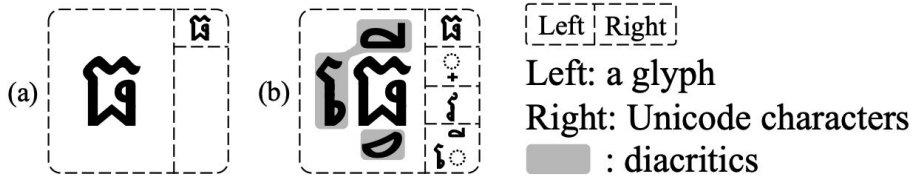
- Best performance achieved by character decomposition compared to word/character/stroke

English-Chinese NMT		BLEU
EN_word	CN_word	11.8
EN_word	CN_character	10.3
EN_word	CN_ideograph	**14.6***
EN_word	CN_stroke	14.1*
Chinese-English NMT		**BLEU**
CN_word	EN_word	14.7
CN_character	EN_word	14.5
CN_ideograph	EN_word	**15.6***
CN_stroke	EN_word	15.5*

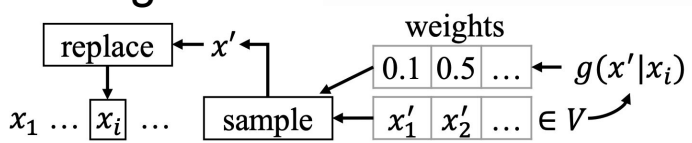
Longtu Zhang and Mamoru Komachi. 2018. Neural Machine Translation of Logographic Language Using Sub-character Level Information.

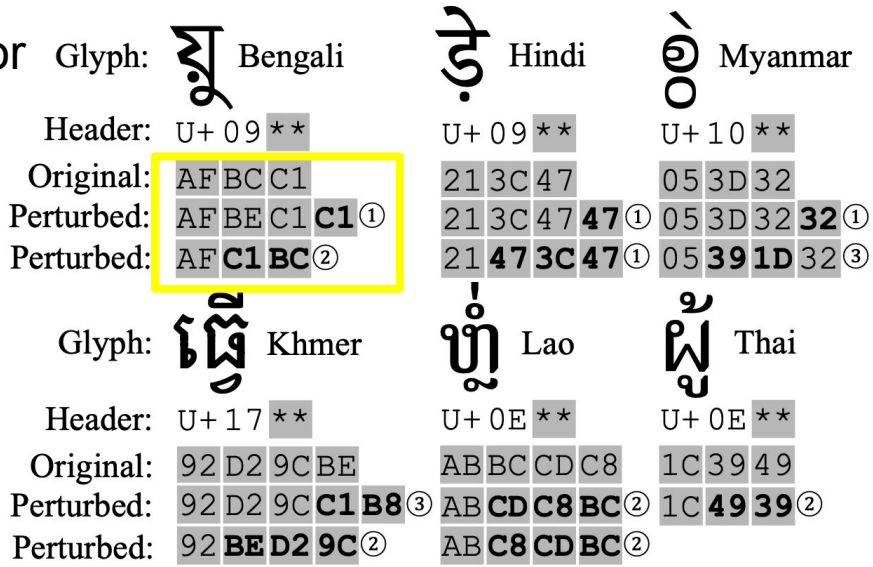
Glyph Perturbation: Data Augmentation for Abugidas

- Examples of Khmer glyphs
  - (a) is a glyph without diacritics and (b) with diacritics.



- There are multiple homoglyphs for one character, caused by
  ① repetition
  ② permutation, and
  ③ decomposition
- Generate homoglyphs as data augmentation.

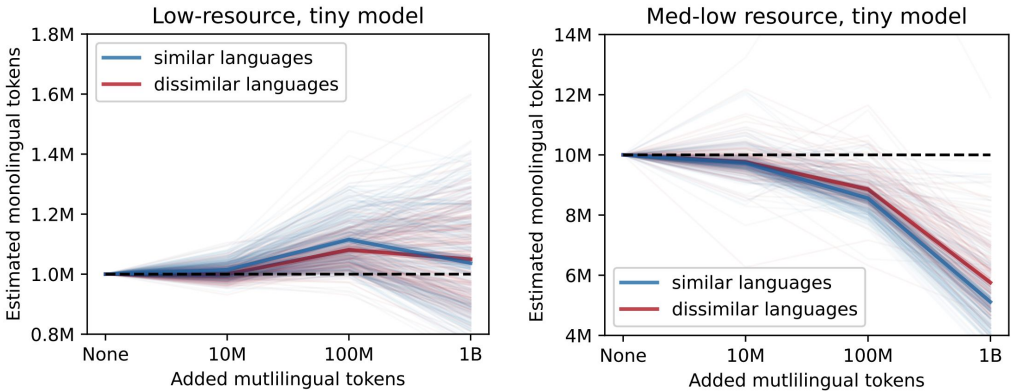


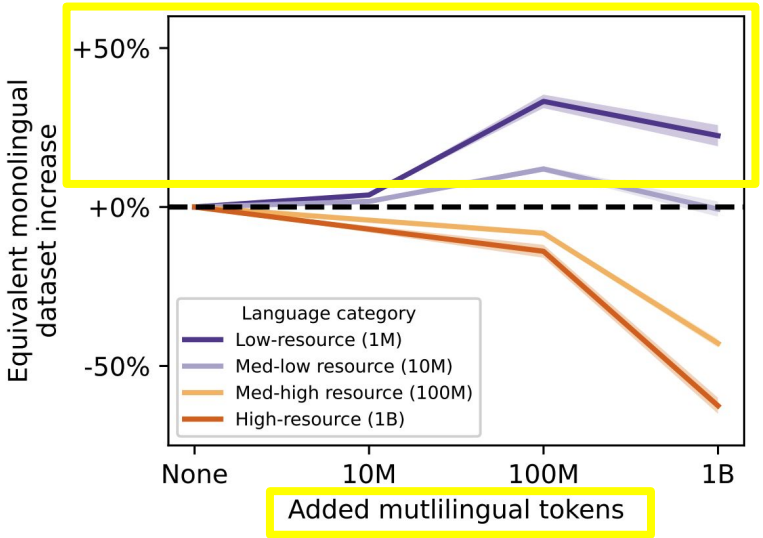


Hour Kaing, Chenchen Ding, Hideki Tanaka, and Masao Utiyama. 2024. Robust Neural Machine Translation for Abugidas by Glyph Perturbation.

Does **Multilingual Data** Help?

- More multilingual data improves the downstream tasks performance of low-resource languages
- The **more similar** with high-resource languages, **the better**
- At the same time, it hurts the performance of higher-resource languages





Tyler A. Chang et al. 2024. [When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages]

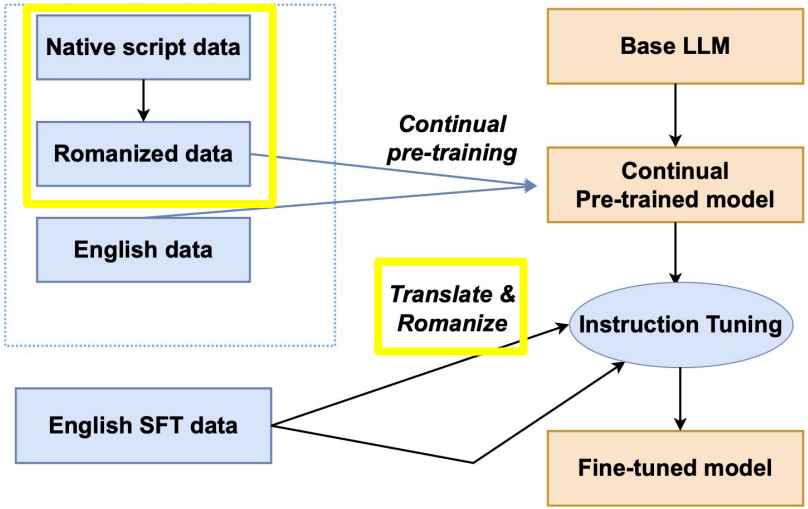
Handle Diverse Scripts

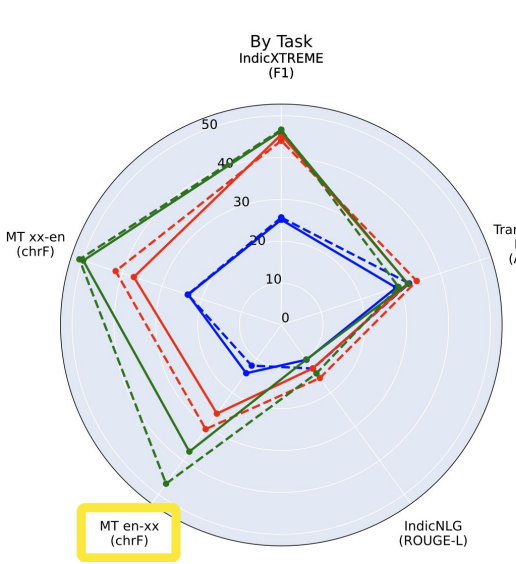
- High-resource languages: vocabulary extension is possible for languages such as Chinese[1]
- Low-resource languages: increase similarity with higher-resource languages?
  - Map them to Roman script
  - Map data in high-resource language to the target language, especially when
    - They in the same language family (Hindi and Gujarati)
    - They are related/sharing some scripts (Japanese and Chinese)

[1]Cui et al. 2024. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca

Romanization

- Extending Large Language Models (LLMs) to non- English languages that use non-Roman scripts.





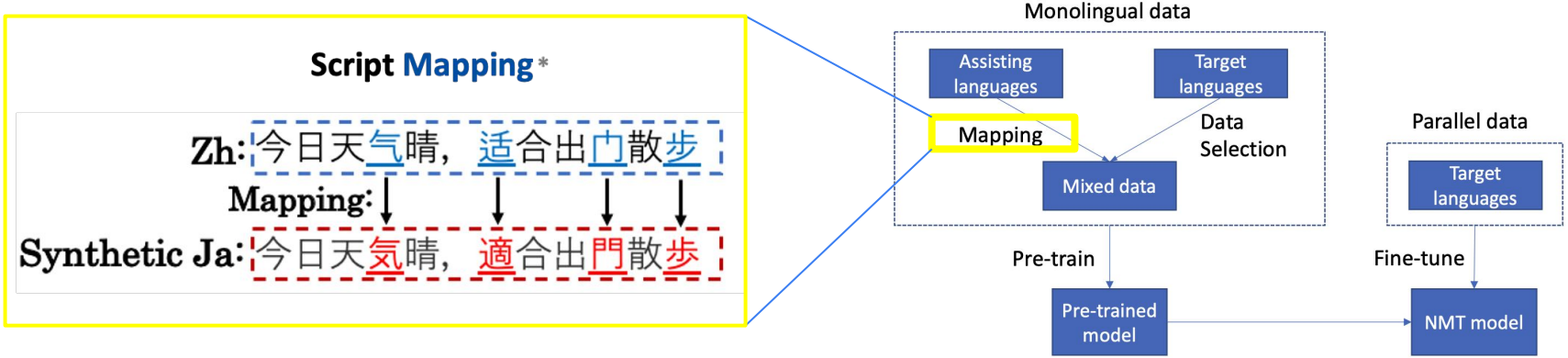
RomanSetu

Languages:
Hindi, Marathi,
Gujarati,
Tamil and Malayalam

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models via Romanization.

Script Normalization

- From one language to another **related** language
  - Map Chinese data to **Synthetic Japanese** data, which is then used for pre-training.
  - By script mapping, where there is a *character-to-character dictionary.*



Haiyue Song, et al. 2020. [Pre-training via Leveraging Assisting Languages for Neural Machine Translation](#).

LLMs for Low-Resource Machine Translation

- WMT23 Findings:
  - Although GPT4 excelled in some areas
  - It struggled with other aspects such as specific domains
  - It ranked lower than encoder-decoder systems when translating from English into low-resource languages (e.g. Czech and Russian)

English→Czech		
Rank	Ave.	System
1	85.4	Human-refA
2	84.1	ONLINE-W
3-5	81.8	GPT4-5shot
3-4	80.4	CUNI-GA
5-8	80.3	ONLINE-A
5-8	79.4	CUNI-DocTransformer
4-7	78.8	ONLINE-B
8-14	78.6	NLLB_MBR_BLEU
6-11	78.4	GTCOM_DLUT
8-12	77.4	CUNI-Transformer
10-14	76.8	NLLB_Greedy
9-14	75.7	ONLINE-M
10-15	75.2	ONLINE-G
13-15	75.0	ONLINE-Y
8-15	75.0	Lan-BridgeMT
16	74.1	LanguageX

Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here But Not Quite There Yet. 2023.

LLMs for Low-Resource Machine Translation

- WMT24 Findings:
  - Three top-ranked systems use LLMs (Qwen, TowerLLM, Claude3.5)
  - Best system on 10 language pairs (Zhang, 24):
    - Continue pretraining using the open-source LLMs
    - Leverage open LLMs to generate **synthetic data**
    - **Supervised fine-tuning** on synthetic & real data
  - Human references are in the winning cluster in 7 out of 11 language pairs.

Findings of the WMT24 General Machine Translation Shared Task: The LLM Era is Here but MT is Not Solved Yet. 2024.
Zhang et al. IOL Research Machine Translation Systems for WMT24 General Machine Translation Shared Task. 2024

Prompting LLMs for Low-Resource Machine Translation

- Zero-shot prompting performance varies greatly across templates.

ID	Template (in English)	English		German		Chinese	
		w/o	w/	w/o	w/	w/o	w/
A	`[src]: [input] ◇ [tgt]:`	**38.78**	**31.17**	-26.15	-16.48	**14.82**	**-1.08**
B	`[input] ◇ [tgt]:`	-88.62	-85.35	-135.97	-99.65	-66.55	-85.84
C	`[input] ◇` Translate to `[tgt]:`	-87.63	-68.75	-106.30	-73.23	-63.38	-70.91
D	`[input] ◇` Translate from `[src]` to `[tgt]:`	-113.80	-89.16	-153.80	-130.65	-76.79	-67.71
E	`[src]: [input] ◇` Translate to `[tgt]:`	20.81	16.69	**-24.33**	**-5.68**	-8.61	-30.38
F	`[src]: [input] ◇` Translate from `[src]` to `[tgt]:`	-27.14	-6.88	-34.36	-9.22	-32.22	-44.95

Table 1: COMET scores averaged over 6 language pairs for *zero-shot* prompting with different templates and different template languages on Wiki Ablation sets. *w/* and *w/o* denote whether adding line breaks into the template or not; ◇ indicates the position of the line break. `[src]` and `[tgt]` denote source and target test language name, respectively, and `[input]` denotes the test input; all of them are placeholders. *English, German* and *Chinese* indicate template languages. Best results are shown in **bold**.

Zhang et al. Prompting Large Language Model for Machine Translation: A Case Study. 2023.

Prompting LLMs for Low-Resource Machine Translation

- More prompt examples improves translation significantly.

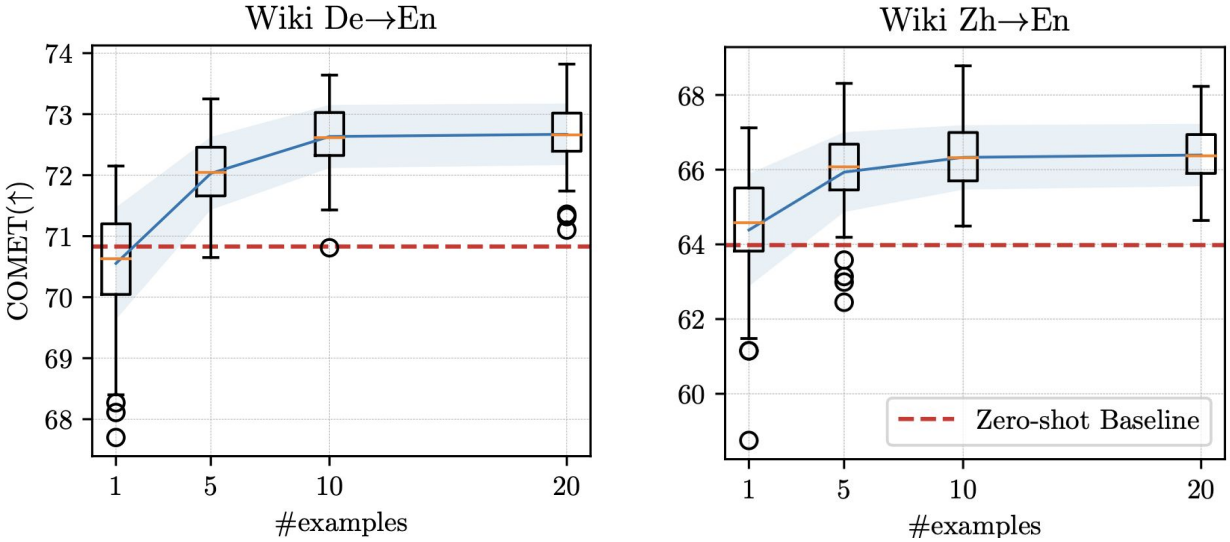
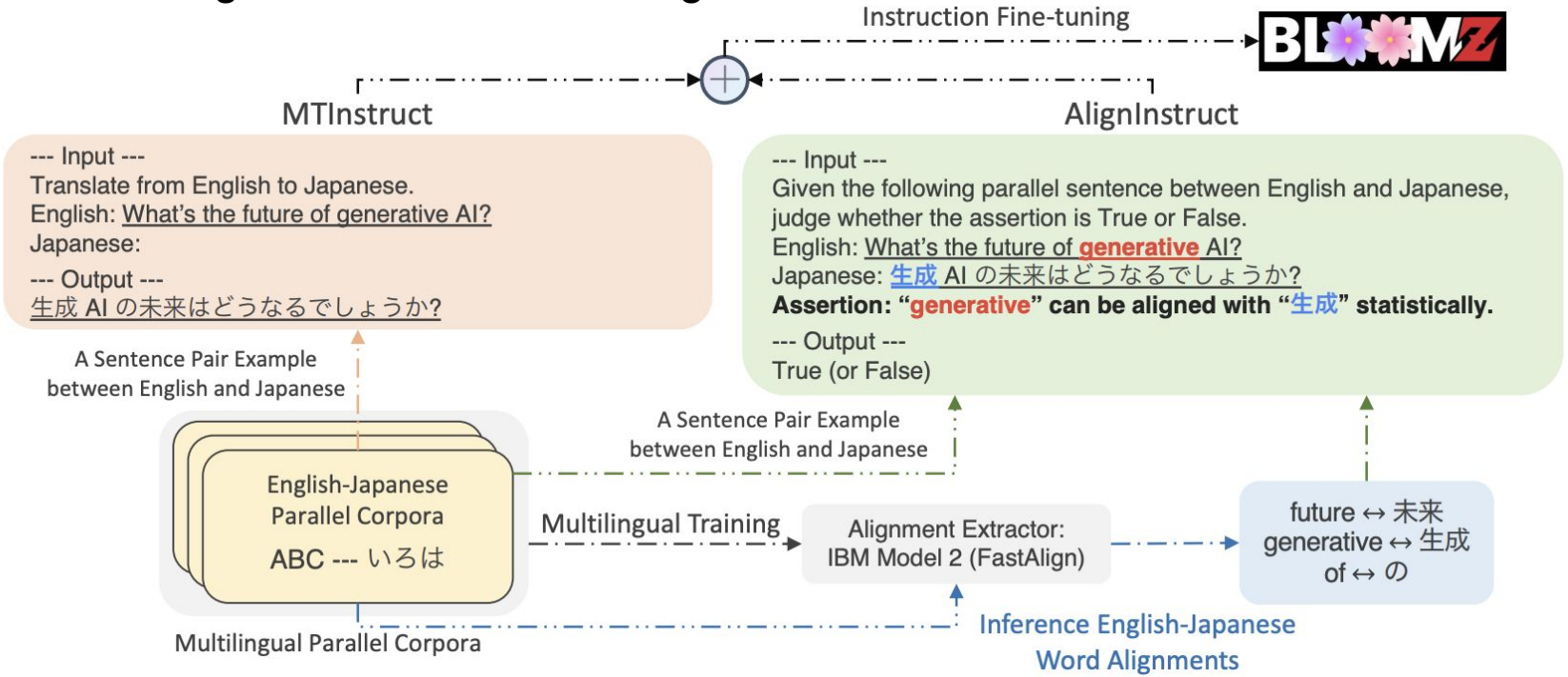


Figure 1: COMET scores for *few-shot* prompting as a function of the number of prompt examples ($K = 1, 5, 10, 20$) on Wiki Ablation sets. For each setup, we randomly sample 100 times from the example pool and show the performance distribution via box plots. Dashed red line denotes the zero-shot baseline; blue curve and shadow area denote the mean and standard deviation.

Zhang et al. Prompting Large Language Model for Machine Translation: A Case Study. 2023.

Instruction-Tuning for Low-Resource Machine Translation

- Fine-tuning with MT and Word alignment tasks.



Mao et al. Tuning LLMs with Contrastive Alignment Instructions for Machine Translation in Unseen, Low-resource Languages. 2024.

Evaluation Metrics (Background)

- BLEU is a metrics showing N-gram overlap between MT output and reference
  - Compute the **precision of *n*-gram** where *n* varies from 1 to 4

$$\text{BLEU} = \left(\prod_{n=1}^{4} \text{precision}_n \right)^{\frac{1}{4}}$$

- Problem
  - Cannot (directly) apply to languages without word boundary
  - It only captures surface similarity, but does not capture semantic similarity

Robust Metrics for Low-Resource Languages

- ChrF: no word tokenizer needed

$$\text{CHRF}\beta = (1+\beta^2)\frac{\text{CHRP}\cdot\text{CHRR}}{\beta^2\cdot\text{CHRP}+\text{CHRR}}$$

where CHRP and CHRR stand for character *n*-gram precision and recall arithmetically averaged over all *n*-grams

- ChrF++ considers both **character *n*-gram** (1 to 4) and **word *n*-gram** (1 to 2)

$$\text{CHRF++}_\beta = \frac{1}{N}\sum_{n=1}^{N}\left[(1+\beta^2)\frac{\text{Prec}_n\cdot\text{Rec}_n}{\beta^2\,\text{Prec}_n+\text{Rec}_n}\right]$$

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation.

Robust Metrics for Low-Resource Languages

- COMET: a multilingual quality estimation metrics
  - Training data: human judgment of machine translation outputs, including Direct Assessment and Multidimensional Quality Metrics in WMT (represented by scores or rankings)
  - Task: quality score prediction and ranking

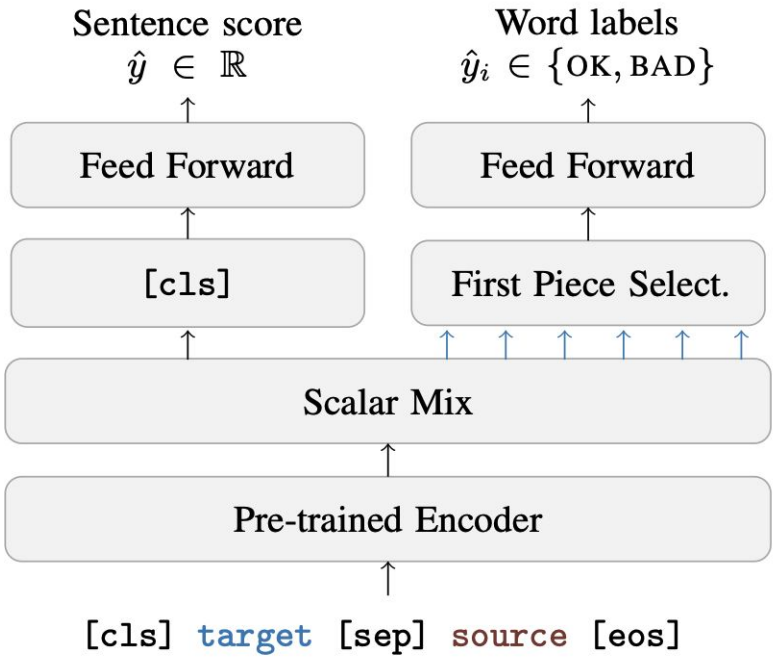
Metric	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
BLEU	0.364	0.248	0.395	0.463	0.363	0.333	0.469	0.235
CHRF	0.444	0.321	0.518	0.548	0.510	0.438	0.548	0.241
YISI-1	0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355
BERTSCORE (default)	0.500	0.363	0.527	0.568	0.540	0.464	0.585	0.356
BERTSCORE (xlmr-base)	0.503	0.369	0.553	0.584	0.536	0.514	0.599	0.317
COMET-HTER	0.524	0.383	0.560	0.552	0.508	0.577	0.539	0.380
COMET-MQM	0.537	0.398	0.567	0.564	0.534	0.574	**0.615**	0.378
COMET-RANK	**0.603**	**0.427**	**0.664**	**0.611**	**0.693**	**0.665**	0.580	**0.449**

Ricardo Rei, et al. 2020. COMET: A Neural Framework for MT Evaluation

Correlation with human judgements

Robust Metrics for Low-Resource Languages

- CometKiwi: A multi-task version of **Comet**
- Task: quality prediction
  - Training data: quality estimation labeled data from Direct Assessment and Multidimensional Quality Metrics in WMT
  - + Multi-task learning: both **sentence-leven and word-level quality estimation** task
  - + Few-shot language adaptation: using half of the dev set (500 out of 1k) for 5 lang pairs
  - + Model Ensemble of 6 models in total, including InfoXLM, RemBERT, XML-R with different settings



Ricardo Rei *et al.* 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task.

Robust Metrics for Low-Resource Languages

- CometKiwi shows a high correlation of human evaluation (DA).
- Improvements brought by multi-task learning and few-shot language adaptation.

	Direct Assessment												
Encoder	**km-en**	**ps-en**	**en-ja**	**en-cs**	**en-mr**	**ru-en**	**ro-en**	**en-zh**	**en-de**	**et-en**	**si-en**	**ne-en**	**avg.**
Baseline (Zerva et al., 2021)													
XLM-R	0.615	0.601	0.295	0.535	0.419	0.703	0.828	0.513	0.500	0.806	0.565	0.793	0.598
Pretrained models													
InfoXLM	0.619	0.603	0.328	0.510	0.462	0.731	0.829	0.554	0.516	0.803	0.561	0.777	0.608
RemBERT	0.600	0.621	0.338	0.525	0.447	0.680	0.818	0.487	0.491	0.810	0.525	0.747	0.591
XLM-R	0.610	0.579	0.325	0.503	0.405	0.715	0.832	0.541	0.514	0.782	0.540	0.740	0.591
Sentence-level only													
XLM-R	0.628	0.591	0.350	0.531	0.551	0.761	0.859	0.577	0.568	0.800	0.565	0.796	0.631
InfoXLM	0.629	0.623	0.348	0.515	0.574	0.747	0.858	0.586	0.551	0.828	0.568	0.790	0.635
RemBERT	0.634	0.631	0.346	0.570	0.564	0.754	0.862	0.534	0.531	0.822	0.550	0.782	0.632
Few-shot Language Adaptation													
XLM-R	0.650	0.619	0.352	0.551	0.546	0.753	0.852	0.571	0.554	0.813	0.562	0.798	0.635
InfoXLM	0.641	0.650	0.367	0.549	0.549	0.751	0.855	0.591	0.565	0.824	0.563	0.803	0.642
RemBERT	0.625	0.641	0.367	0.568	0.563	0.756	0.857	0.540	0.527	0.824	0.568	0.796	0.636
Sentence + word-level training													
InfoXLM	0.617	0.586	0.344	0.532	0.572	0.761	0.865	0.586	0.579	0.829	0.576	0.804	0.637
RemBERT	0.634	0.628	0.356	0.564	0.571	0.762	0.860	0.541	0.553	0.826	0.564	0.799	0.638
Few-shot Language Adaptation													
InfoXLM	0.643	0.632	0.335	0.557	0.560	0.766	0.860	0.575	0.582	0.833	0.578	0.809	0.644
RemBERT	0.644	0.645	0.356	0.567	0.568	0.759	0.856	0.545	0.552	0.835	0.561	0.804	0.641
Final Ensemble													
Ensemble 6x	**0.664**	**0.669**	**0.380**	**0.591**	**0.593**	**0.782**	**0.871**	**0.597**	**0.593**	**0.845**	**0.588**	**0.820**	**0.666**

Ricardo Rei *et al.* 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task.

MT Evaluation: Low-Resource Scenario

MT evaluation using predictive measures - QE

TransQuest-, and COMET-based encoders **work well for English on the target side.**

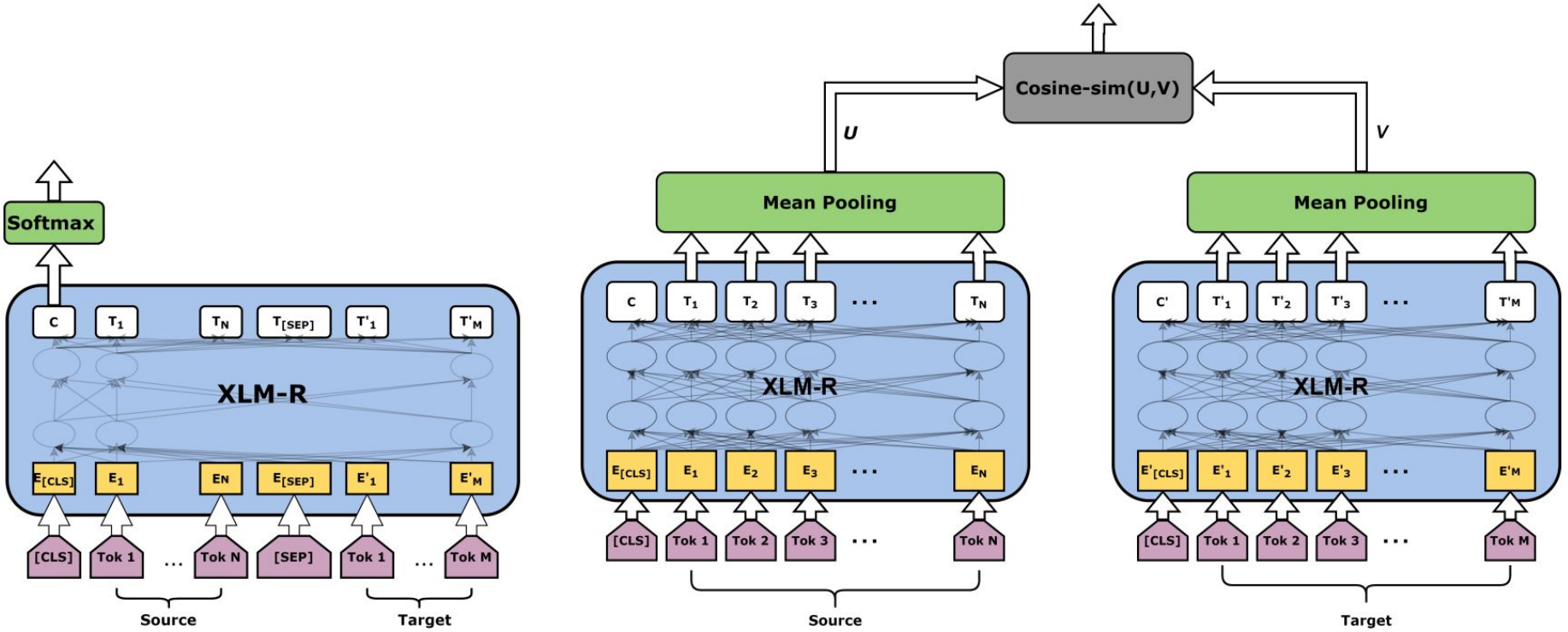
Despite 21k training instances (En-Mr), compared to 7-8k for most other high-resource languages.

LP	TransQuest	COMET
EN-DE	0.3811	0.3579
EN-MR	0.2489	0.5135
EN-ZH	0.6360	0.5410
ET-EN	0.8148	0.7018
NE-EN	0.8034	0.6393
RO-EN	0.8739	0.7699
RU-EN	0.8252	0.6482
SI-EN	0.7233	0.5874

What do Large Language Models Need for Machine Translation Evaluation?

LP	T1		T2		T3		T4		T5		T6	
	ρ	D	ρ	D	ρ	D	ρ	**D**	ρ	**D**	ρ	D
						OpenChat3.5						
EN-DE	0.2258	1	0.2209	2	0.2849	0	0.2599	0	0.2812	0	**0.2960**	0
EN-MR	0.2295	3	0.3110	9	0.3546	0	0.3347	0	0.3565	0	0.3446	0
EN-ZH	0.2722	0	0.2603	4	0.3995	0	0.3002	0	0.3333	0	0.3635	0
ET-EN	0.5402	0	0.5798	2	**0.6980**	0	0.5879	0	0.6700	0	0.6925	0
NE-EN	0.3784	9	0.4855	25	0.5937	0	0.5008	0	0.5832	0	**0.6073**	0
RO-EN	0.4712	2	0.5669	25	0.7294	0	0.6900	0	0.7096	0	**0.7385**	0
RU-EN	0.5714	0	0.5320	13	**0.6066**	0	0.5494	0	0.5322	0	0.5938	0
SI-EN	0.4120	4	0.4201	7	**0.6034**	0	0.4364	0	0.5990	0	0.5963	0

TransQuest - A Scalable, Adaptable Framework for MT Evaluation

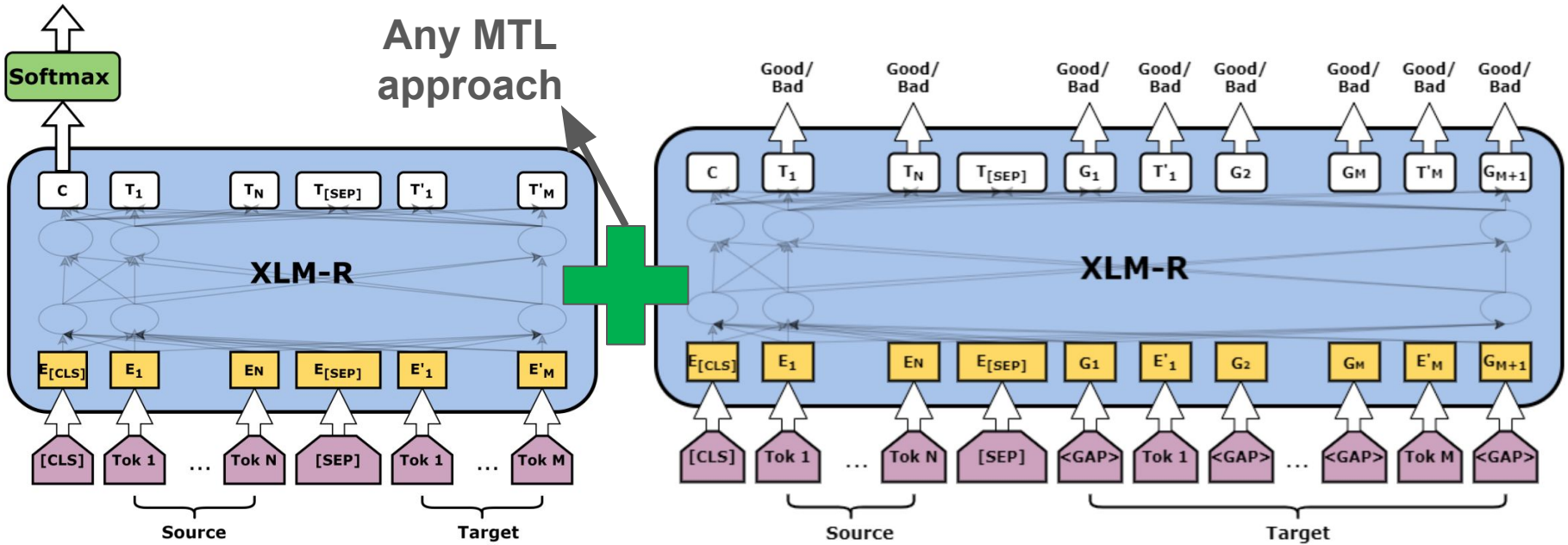


(a) *MTransQuest* architecture (b) *STransQuest* Architecture

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation Quality Estimation with Cross-lingual Transformers

MT Evaluation: Multi-tasking for QE

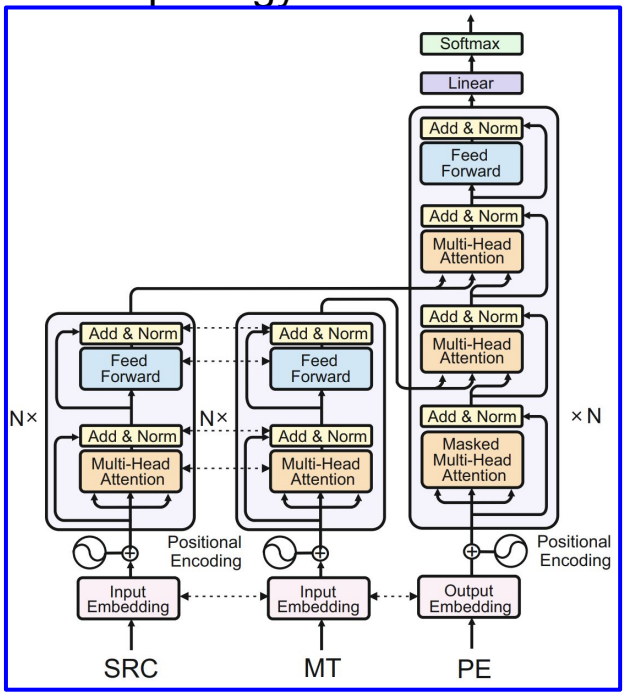
Multitask learning based framework for QE [based on TransQuest]



Sourabh Deoghare, Paramveer Choudhary, Diptesh Kanojia, Tharindu Ranasinghe, Pushpak Bhattacharyya, and Constantin Orăsan. 2023. A Multi-task Learning Framework for Quality Estimation.

Automatic Post-Editing (APE)

MT output **can be problematic** given **domain-specificity**, **low-resourceness**, or challenging morphology.



PE is the task of **minimally** editing MT output for accurate semantic transfer.

APE allows us to build systems which can mimic this behaviour, and **produce corrected translations**.

System — En-Mr APE task	TER ↓	BLEU ↑
Do Nothing (Baseline)	22.93	64.51
+ CTS-based Training and External MT	20.08	67.39
+ LaBSE-based Data Filtering and in-domain training data	19.73	67.86
+ Phrase-level APE triplets	19.39	68.35
+ Sentence-level QE	**19.01**	**68.87**

IIT Bombay's WMT22 Automatic Post-Editing Shared Task Submission [Winning System, APE 2022 Shared Task

MS-UEdin Submission to the WMT2018 APE Shared Task: Dual-Source Transformer for Automatic Post-Editing

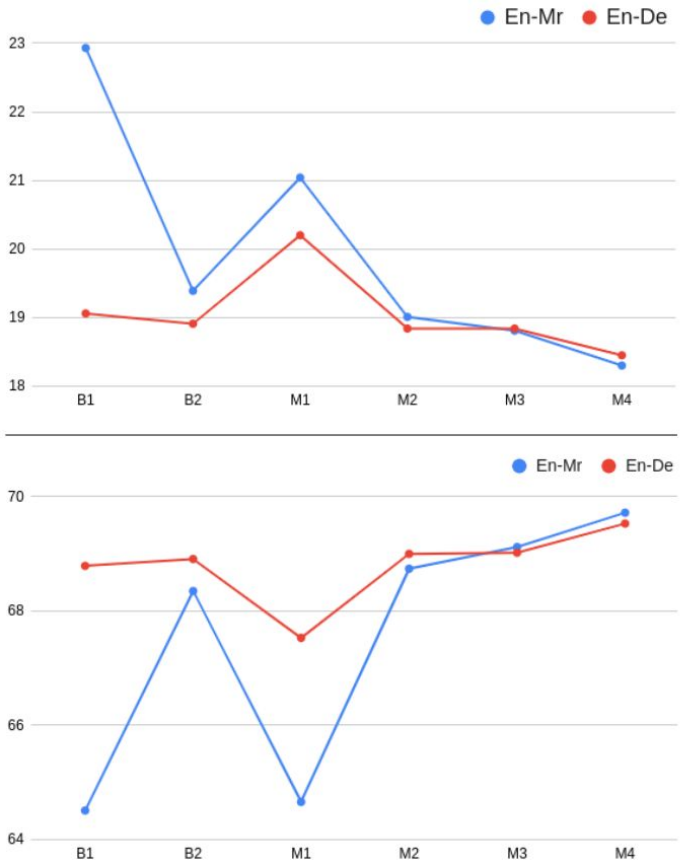
Unified Evaluation and Correction

Merging QE and APE - Sentence-level + Word-level + APE, for context-aware unified evaluation and correction.

Progressively Integrating QE with APE

- ➔ QE as APE Activator
- ➔ QE as MT/APE Selector
- ➔ QE as APE Guide
- ➔ **Joint Training over QE and APE**
  - ◆ **Linear Scalarization (LS-MTL) vs. Nash-MTL**

$$L_{LS-MTL} = L_{sent} + L_{word} + L_{APE}$$



Sourabh Deoghare, et al. 2023.
Quality Estimation-Assisted Automatic Post-Editing.

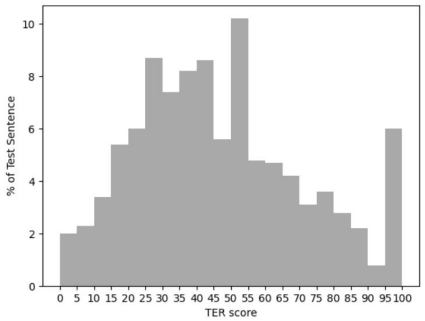
Unified Evaluation and Correction [QE Shared Task @ WMT24]

Merging QE and APE shared tasks

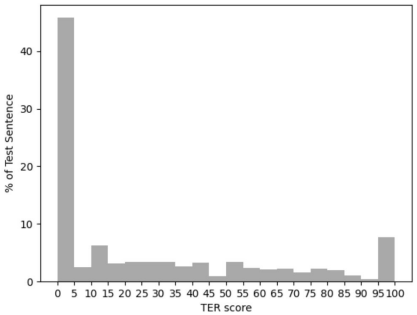
Subtask 3: QE-informed APE

Systems perform aggressive post-editing; top submission modifying 96.5% translations, most modifications lead to improvement in quality with a precision of 84.56%.

		TER	BLEU	CHRF	COMET
En-Hi	IT-Unbabel	**27.08**	58.38	73.45	0.8646
	HW-TSC_yjwsss	**30.37**	54.50	71.06	0.8514
	HW-TSC_zhaoxf4	**31.32**	52.74	69.83	0.8517
	BASELINE (MT)	46.36	39.28	59.48	0.8084
En-Ta	HW-TSC	24.24	69.64	82.36	0.9186
	IT-Unbabel	24.54	70.05	82.30	0.9163
	BASELINE (MT)	24.71	70.16	81.80	0.9137



En-Hi Distribution



En-Ta Distribution

➔ Final QE-driven selection to choose- original MT output *vs.* generated APE hypothesis.

➔ HW-TSC exploited QE information only for final selection step, while IT-Unbabel **integrated the two technologies more tightly** by **generating APE outputs with an LLM informed by free-text explanations for translation errors.**

Summary

- Back-translation
  - Greedy search/beam search to generate synthetic parallel data
  - Sampling with quality control
  - Iterative back-translation
- Diverse tokenizations as data augmentation
  - BPE-dropout
  - SentencePiece regularization
  - Character decomposition for Chinese/Japanese/Abugidas
- Script mapping
  - Romanization
  - Mapping to related languages
- LLMs for low-resource machine translation
- MT Evaluation and Correction

Future Direction

- More data
  - More effective, automatic data collection/data cleaning
  - Synthetic data creation
  - Mining more supervision signal from existing data
- Effective vocabulary extension
- Evaluation metrics for low-resource language pairs
- Leveraging LLMs and NMT
  - How to combine them?
    - Direct fine-tuning LLMs on MT data
    - LLMs provide synthetic data -> fine-tuning LLM/NMT models
    - Does the encoder-decoder architecture still necessary?
- Unifying evaluation and correction
  - Emerging as a unified paradigm (QE+APE subtask, WMT24)

Module 5 (Part 2):
NLG tasks for Dialects

We will introduce

NLG tasks for dialects

- Machine Translation
- Summarization
- Dialogue systems
- Evaluation

Hands-on session

NLG tasks for dialects: Machine Translation

– Commercial systems have a good performance (understanding) on a **standard dialect (high resource)** but worse on **its variants (low resource)**.

– E.g., Swiss German and Italian Variants

Standard Italian Variant:		
Source:	*Hanno rubato il quadro*	
GTranslate:	They stole the painting	✓
Alassio Variant:		
Source:	*I han rubbau u quaddru*	
GTranslate:	I han rubbau u quaddru	✗

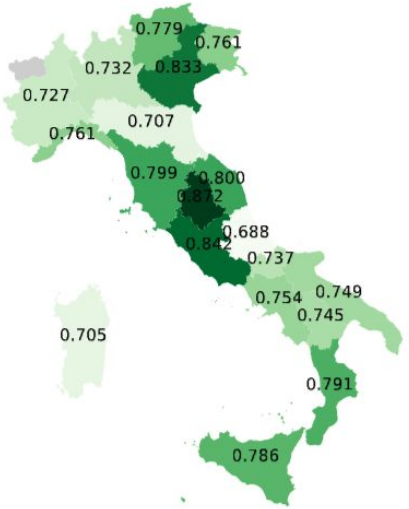


Figure 2: Map of Italy with COMET scores per region.

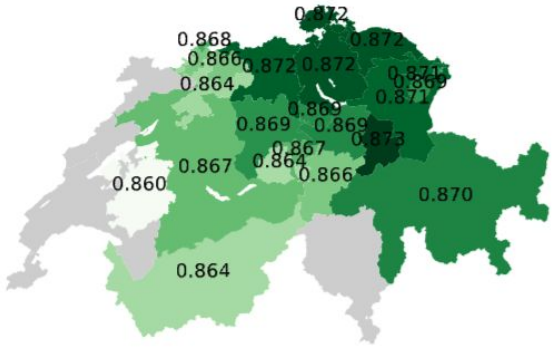


Figure 3: Map of Switzerland with COMET scores for different regions.

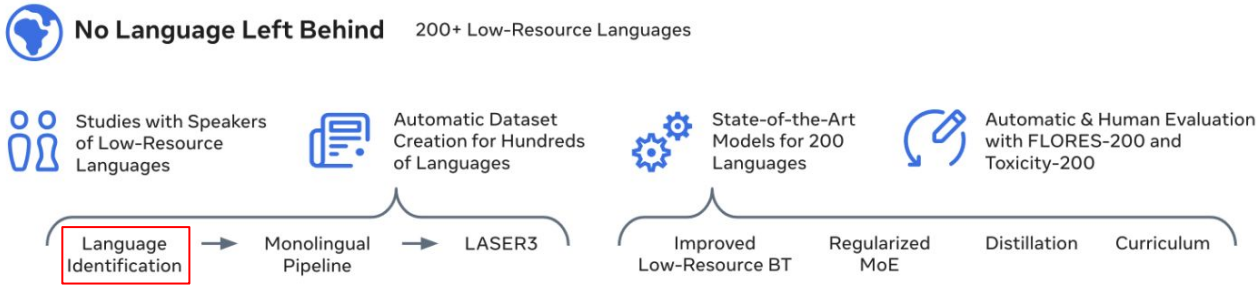
Md Mahfuz Ibn Alam et al., 2024. CODET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation.

Generation Capability: Language Coverage

– There are ~7000 languages
    – Including dialects?
– Current models could cover up to 1000 languages.
    – mT5: 101 languages
    – NLLB: 200 languages
    – Google translate: 249 [Wikipedia] and towards 1500+ [Bapna+23]
    – etc

Linting Xue, et al., 2021. mT5: A massively multilingual pre-trained text-to-text transformer
Ankur Bapna, et al., 2023. Building Machine Translation Systems for the Next Thousand Languages.
NLLB Team, et al., 2022. No Language Left Behind: Scaling Human-Centered Machine Translation

Generation Capability: Increasing Language Diversity



LID	#Lang
FastText	217
CLD3 [Bapna+23]	1,745

– Potential mixture of dialects in datasets

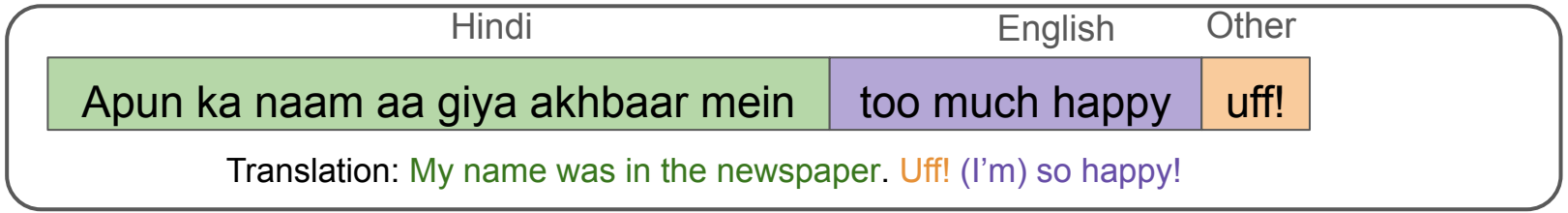
– Increase the number of languages/dialects of the identification model (LID)

– Need diverse dialectal datasets to train diverse LID

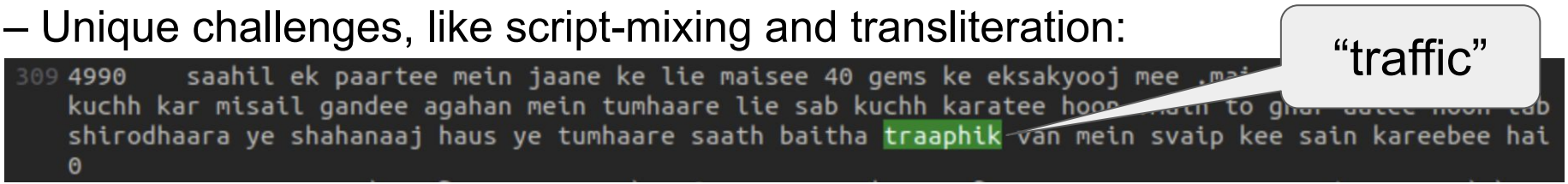
NLLB Team, et al., 2022. No Language Left Behind: Scaling Human-Centered Machine Translation
Ankur Bapna, et al., 2023. Building Machine Translation Systems for the Next Thousand Languages.

Connecting Ideas: Similar Challenges for LID of code-mixing

– Most LID is doc/sent level, precluding code-mixing, which needs token-level LID (Caswell et al., 2020)



– Unique challenges, like script-mixing and transliteration:

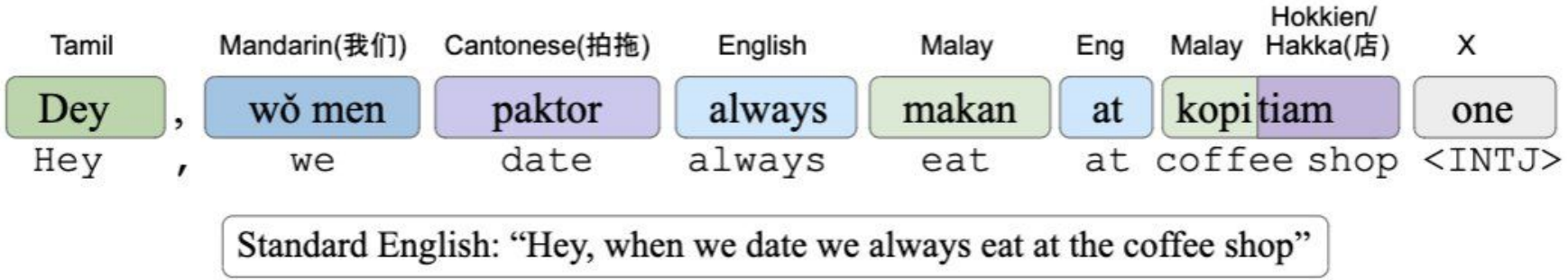


– **GLUE-CoS** (Khanuja et al., 2020) has LID for Spanish-English, Hindi-English

– **LINCE** (Aguilar et al., 2020) includes LID for Spanish-English, Hindi-English, Nepali-English, Arabic-Arabic* (Modern standard* and Egyptian dialects)

Connecting Ideas: Similar Challenges for LID of Creoles

– LID is also a problem for Creoles with highly multilingual vocabulary, such as **Nigerian Pidgin** or **Singlish** (Lent et al., 2021):



Not to be confused with Singapore English or Simlish.

Singlish (a portmanteau of *Singapore* and *English*), formally known as **Colloquial Singaporean English**, is an English-based creole language originating in Singapore.[1][2][3] Singlish arose out of a situation of prolonged language contact between speakers of many different Asian languages in Singapore, such as Malay, Cantonese, Hokkien, Mandarin, Teochew, and Tamil.[4] The term *Singlish* was first recorded in the early 1970s.[5]

Challenge of LID: Dialect Similarity

– Surface form's similarity of dialects.
  – E.g., Chinese and Japanese

– Dialectal datasets are crucial for model to distinguish dialects.

English	Being a professor is my lifelong wish.	Don't strew things all over the ground.
Mandarin	做教授是我一生的願望。	東西不要撒得滿地都是
Hokkien	做教授是我一生的願望。	物件毋通掖甲一四界

n - / so re ma de / o - ni wa
(んー / それまで / おーにわ)

Japanese Standard Dialect

u n / so re ma de / a i o i ni ha
(うん / それまで / あいおいには)

Hyogo region

Yes, until then, in Aioi ...

Sin-En Lu et al., 2022. Exploring Methods for Building Dialects-Mandarin Code-Mixing Corpora: A Case Study in Taiwanese Hokkien.
Kaori Abe et al., 2018. Multi-dialect Neural Machine Translation and Dialectometry.

Dialectal Dataset within Languages

– 48 Japanese dialects [Abe+18]

– 63 Vietnamese dialects [Dinh+24]

– 10 Thai dialects [Suwanbandit+23]

– 9 Chinese dialects [Tang+24]

– 8 Arabic dialects [Talafha+24]

– 7 Swiss German dialects [Plüss+23]

transcripted from speech

Kaori Abe et al., 2018. Multi-dialect Neural Machine Translation and Dialectometry.
Nguyen Van Dinh, et al., 2024. Multi-Dialect Vietnamese: Task, Dataset, Baseline Models and Challenges.
Zhiyuan Tang, et al., 2024. KeSpeech: An Open Source Speech Dataset of Mandarin and Its Eight Subdialects
Michel Plüss, et al., 2023. STT4SG-350: A Speech Corpus for All Swiss German Dialect Regions.
Artit Suwanbandit, et al., 2023. Thai-Dialect: Low Resource Thai Dialectal Speech to Text Corpora.
Bashar Talafha, et al., 2024. Casablanca: Data and Models for Multidialectal Arabic Speech Recognition.

Dialectal Dataset from Speech

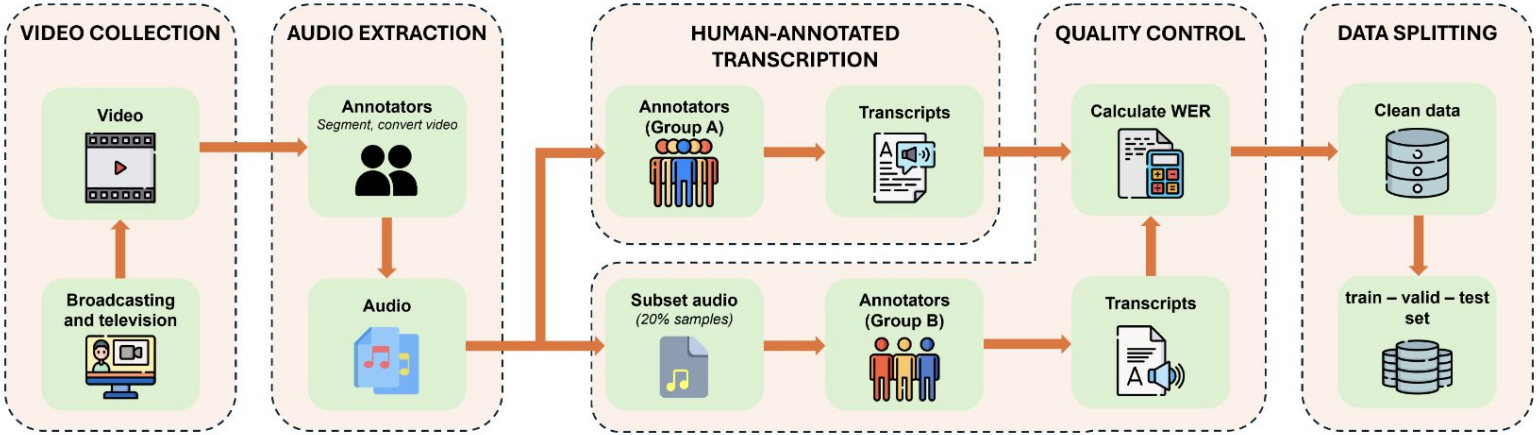


Figure 1: Data Collection Pipeline for the ViMD Dataset.

Nguyen Van Dinh, et al., 2024. Multi-Dialect Vietnamese: Task, Dataset, Baseline Models and Challenges

More Dialectal Datasets

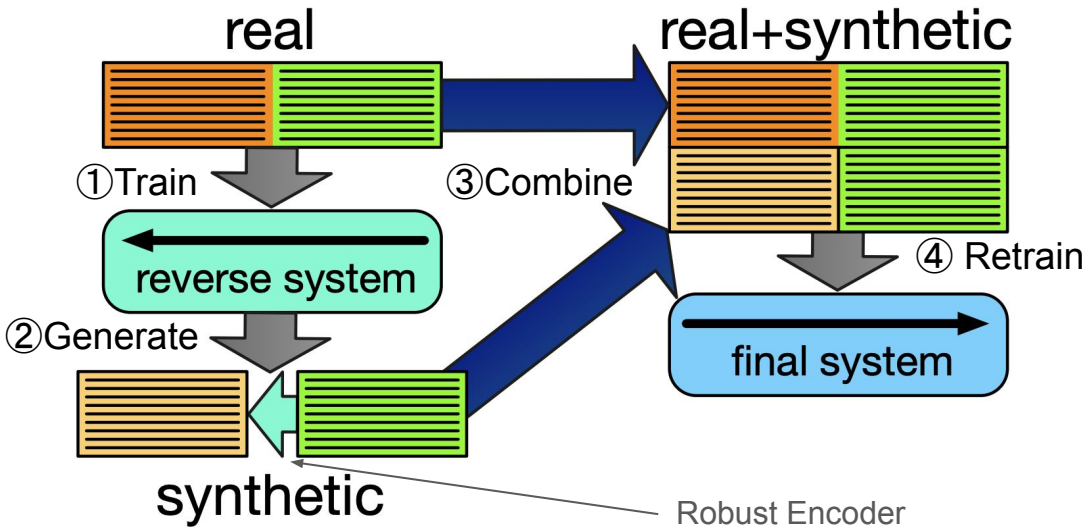
	Languages						Innovation				Problem/Area						
	English	Chinese	Arabic	German	Indic Languages	Other	Dataset	Method/Model	Evaluation/Metric	Benchmark	Dialect Classification	Sentiment Analysis	Machine Translation	Morphology/Parsing	Conversational AI	Summarisation	Speech/Visual
[Hassan et al. 2017]	✓					✓	✓	✓					✓				
[Artemova and Plank 2023]				✓			✓		✓				✓				
[Ahia et al. 2024]						✓	✓			✓			✓			✓
[Dabre et al. 2024]					✓		✓			✓			✓				
[Olabisi et al. 2022]	✓						✓									✓	
[Estival et al. 2014]	✓						✓										✓
[Talafha et al. 2024]			✓				✓			✓	✓						✓
[Dinh et al. 2024]						✓	✓				✓						✓
[Zhan et al. 2023]		✓					✓		✓	✓					✓		
[Zhan et al. 2024]		✓					✓		✓	✓					✓		
[Artemova et al. 2024]				✓			✓		✓	✓					✓		

Table 2. State of NLP research on Dialects.

Aditya Joshi, et al., 2024. Natural Language Processing for Dialects of a Language: A Survey

Having Monolingual Data for Dialects

– Let's say a base model's encoder can understand well the standard dialect
  – standard ~ its variants.
– Back translation (BT)
– Iterative BT
– Robust encoder
  – Subword regularization
  – Input perturbation
  – Code-mixing



Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data.
Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative Back-Translation for Neural Machine Translation

Robust Encoder for Related Languages

– Adding noises improve zero-shot cross-lingual transfer between related languages

– This should be applied for dialects

Languages	Baseline	BPE-Dropout	Noise	BPE-Drop-out+Noise
DE→GSW	73.14	76.48	77.11	**78.13**
FI→OLO	69.32	69.66	**73.03**	71.76
FI→KRL	72.44	76.35	**79.18**	78.57
SV→FO	84.76	86.20	**87.63**	87.31
IS→FO	85.94	86.80	87.43	**87.46**
FR→OFR	63.42	66.65	66.73	**67.27**
DE→FO	81.74	81.34	81.38	**82.27**
DE→OLO	**52.63**	52.09	51.10	49.26
DE→KRL	**57.51**	57.47	55.71	53.37
DE→OFR	**44.08**	39.17	38.32	40.03
FR→OLO	56.49	56.72	**58.59**	56.64
FR→KRL	59.46	62.27	**64.52**	64.15
FR→FO	81.13	82.09	81.81	**82.62**

Table 2: Zero-shot POS tagging accuracy of different strategies for several languages (TRAIN→TEST). The training and test languages are closely related in the upper but not in the lower part of the table as indicated by the colors (Finnic, West Germanic, North Germanic, and Western Romance language branches.) Noise consistently adds additional accuracy points beyond BPE-dropout performance increase.

Noëmi Aepli et al., 2022. Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise

Robust Encoder for Extremely Low-Resource Languages

– “All the proposed noise augmentation models outperform vanilla NMT and all baseline models that utilize lexical similarity (i.e., OBPE, BPE-Dropout, and SDE).”

– “introducing character span noise after segmentation provides a statistically significant improvement over baselines”

Models	Indo-Aryan								Romance		Malay-Polynesian		Average
	Gom	Bho	Hne	San	Npi	Mai	Mag	Awa	Cat	Glg	Jav	Sun	
BPE*	26.75	39.75	46.57	27.97	30.84	39.79	48.08	46.28	33.32	53.75	31.44	32.21	38.06
WordDropout	27.01	39.57	46.19	28.13	31.91	40.31	47.37	46.48	34.20	52.21	32.03	32.52	38.16
SubwordDropout	27.91	40.11	46.26	29.46	32.56	40.99	47.91	47.43	35.09	52.28	33.38	33.47	38.90
WordSwitchOut	25.17	38.81	45.87	26.21	29.95	39.69	47.53	44.54	32.98	51.81	31.84	32.49	37.24
SubwordSwitchOut	26.08	38.84	45.84	28.19	30.81	40.19	47.28	45.93	33.26	53.71	31.24	32.06	37.78
OBPE	27.90	40.57	47.46	28.52	31.99	40.71	49.10	47.16	32.33	52.77	29.98	30.88	38.28
SDE	28.01	40.91	47.88	28.66	32.03	40.82	48.96	47.30	33.72	53.95	31.84	31.24	38.77
BPE-Dropout*	28.65	40.84	46.58	28.80	31.88	40.79	47.86	47.32	34.56	55.83	32.01	32.97	39.00
unigram char-noise**	28.85	42.53	49.35	29.80	34.61	42.67	50.97	49.43	43.16	54.81	35.42	36.69	41.52
BPE → SpanNoise*** (*ours*)	28.66	41.94	49.48	30.49	35.66	44.75	50.55	49.21	43.11	54.89	36.12	37.11	40.16
CHARSPAN (*ours*)	29.71	43.75	51.69	**31.40**	36.52	45.84	51.90	50.55	43.51	55.46	36.24	37.31	42.82
CHARSPAN + BPE-Dropout (*ours*)	**29.91**	**44.02**	**51.86**	30.88	**37.15**	**46.52**	**52.99**	**51.34**	**44.93**	**55.87**	**36.97**	**38.09**	**43.37**

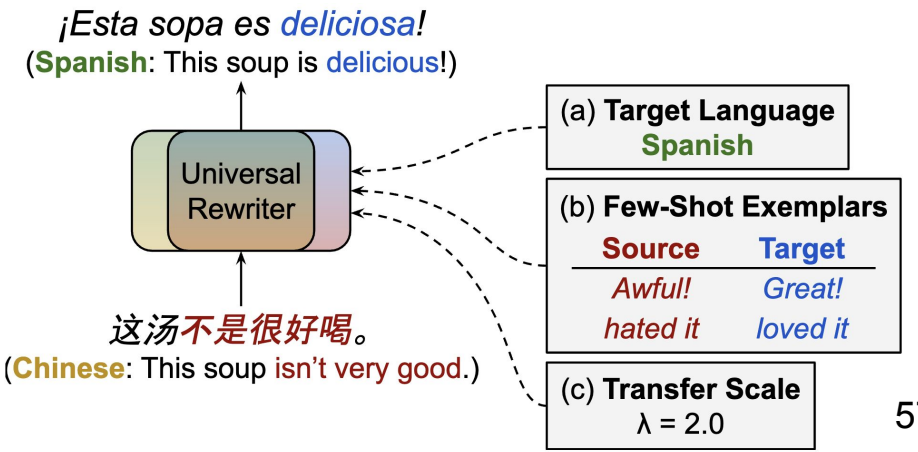
Kaushal Kumar Maurya et al., 2024. CharSpan: Utilizing Lexical Similarity to Enable Zero-Shot Machine Translation for Extremely Low-resource Languages

Few-Shot Example Generation: Rewriter

– (a) The target language is signaled by a unique language code

– (b) Other attributes are controlled through few-shot examplars, which may leverage distinct languages from the input and target.

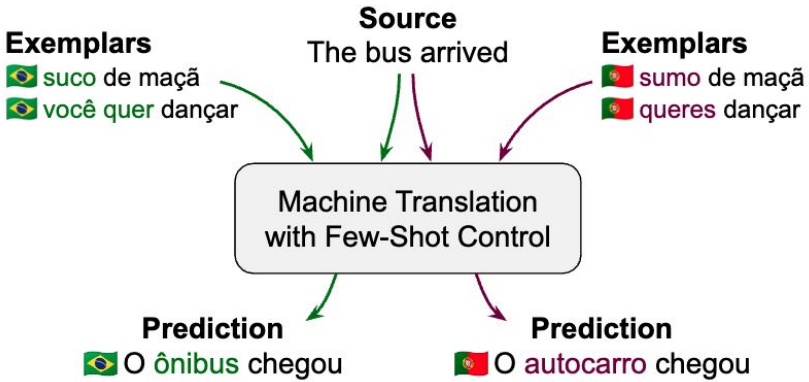
– (c) Transfer scale λ modulates the strength of the attribute transfer.



Xavier Garcia, et al., 2021. Towards Universality in Multilingual Text Rewriting.

Few-Shot Example Generation: Few-Shot Prompt

– PaLM 540B shows impressive few-shot region control
– But there is still significant room for improvement.
– None of the models match human performance



Model	pt	zh
Gold	98.6	94.4
UR	50.4	50.6
M4-UR	51.2	50.9
M4-Prompts	66.7	50.0
M4-Prompts FT	66.7	51.0
PaLM 8B	85.0	69.0
PaLM 62B	90.4	70.8
PaLM 540B	**93.2**	**83.6**
Google Translate	50.0	50.0

Table 5: Lexical accuracy on FRMT test. PaLM outperforms other approaches, while region-agnostic models like Google Translate are guaranteed 50%.

Parker Riley, et al., 2023. FRMT: A Benchmark for Few-Shot Region-Aware Machine Translation.

Having Dialect-Standard Pairs

– **Dialectal Normalization**

– Granularity: subwords vs characters [Kuparinen+23]

  – “the subwords improve the performance on the large corpora …”

  – “... but worsen it on the small corpora …”

– Context: full sentence vs sliding window [Kuparinen+23]

  – sliding window is superior in most case.

Olli Kuparinen et al., 2023. Dialect-to-Standard Normalization: A Large-Scale Multilingual Evaluation

LLM Prompting for Normalization

– LLM on dialect normalization is still limited for low-resource languages such as Vietnamese.

	Correctness	Fluency	Style
ChatGPT (Zero-shot)	5%	37%	54%
ChatGPT (Few-shot)	9%	39%	58%
BARTpho (Fine-tuned)	82%	86%	95%

Table 6: Human analysis of ChatGPT and BARTpho's output qualities for central-to-northern dialect transfer.

Thang Le et al., 2023. A Parallel Corpus for Vietnamese Central-Northern Dialect Text Transfer.

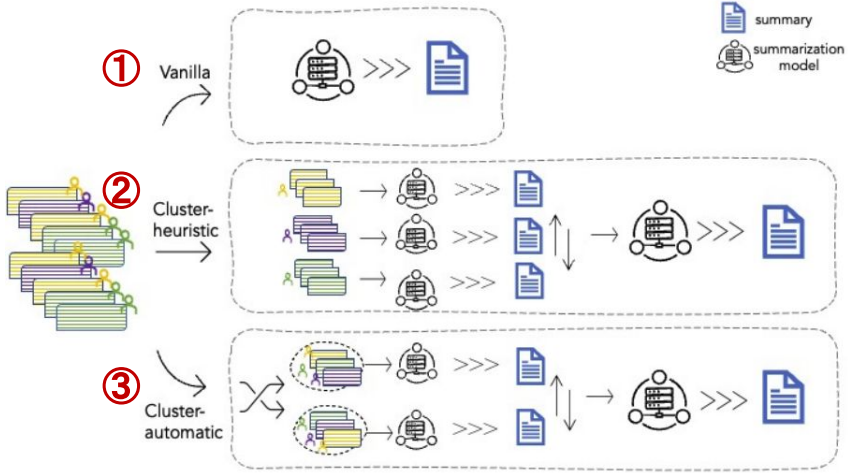
```
Here are a few examples of parallel
Vietnamese central-northern dialect text
utterance pairs.

Central Text: {Central-style input 1}
Northern Text: {Northern-style output 1}

Central Text: {Central-style input 2}
Northern Text: {Northern-style output 2}

...

Central Text: {Central-style input 5}
Northern Text: {Northern-style output 5}

Convert the following Vietnamese central
dialect text utterance into the northern
dialect. Explain the difference and how
you do it.
Central Text: {Central-style test input}
Northern Text:
```

NLG tasks for dialects: Summarization

Multi-document summarization: Dialect-aware clusters can bring better representativeness from texts written in dialects of a language

Input Documents (Tweets)
G3: If Lakers play like that every game no chance for anyone else in the NBA #Nc
G2: The Fan Fictions and Imagines other Beliebers come up with , OMFG I LOVE
G2: Remember when they the NBA play-offs were boring and disappointing-not s
G1: NBATV shittin on u niggas.lol. Melo and Monroe doing work doe.
G3: Going to take in some NBA tonight. Pacers/Knicks. NBA games are much bet
G2: Just fast forward to the trophy presentation and the sucking off of LeBron by y
G1: No point guard in the NBA can hold Russell Westbrook
G1: The Black Mamba 81 point showcase in showing now on NBATV !!
G2: NBA reffs.. Responsible for half of Miami heats points since 2010. -.-t #refsfc
G3: And te NBA was in a lockout last year too.....NOBODY complained about th
...
...
...
G3: Is it possible that boozer is the worst player in the NBA? #GETOUT
G3: NBA rule change I'm shocked has never happened: An assist if the player you
G3: If this was an NBA game, Michael Carrera would be walking away from it wit
G1: Lakers??? Is that even a NBA team? I thought they were D-League.....

Tweet for the same topic in three dialects (G1, G2, G3)



Methods

Olubusayo Olabisi, et al. 2022. Analyzing the Dialect Diversity in Multi-document Summaries

NLG tasks for dialects: Dialogue systems

- Models for dialectal dialogue lag behind models for standard languages.
- Rule-based
  - Elmadany et al. (2018); Ali and Habash (2016) employ rule-based systems (Artificial Intelligence Markup Language) to manage dialectal variation in simple Egyptian Arabic chit-chat dialogue systems

U: Good evening (lit. evening of goodness)

B: Good evening (lit: evening of light), your good name?

U: My name is Ichiro

B: Hello Ichiro. Are you a man or a woman?

U: Man

B: Nice meeting you Mr. Ichiro

U: How are you?

B: Good, *Alhamdulillah* (thanks to God), and you?

A Elmadany, et al.. Arsas: An arabic speech-act and sentiment corpus of tweets. 2018.
Dana Abu Ali and Nizar Habash. Botta: An arabic dialect chatbot. COLING 2016.

NLG tasks for dialects: Dialogue systems

- Transformer models show decreased performance when handling dialect inputs.
  - Artemova et al (2024) manually evaluate perturbation rules that transform German sentences into colloquial forms and use them to synthesize test sets.
  - Experiments across six different transformers for **user intent** and **slot detection** show increased performance.

Category	Perturbation	Example: Before → After
Noun Phrase		
	`possession_von`	des Baums → von dem Baum
	von construction instead of genitive	*the*.GEN *tree's* → *of the*.DAT *tree*
	`possession_pron`	Kafkas Werke → Kafka seine Werke
	Dative with poss. pron. instead of genitive	*Kafka's works* → *Kafka*.DAT *his works*

		Intact		Individual Perturbations	
		Intent Acc	Slot F_1	Δ Intent Acc	Δ Slot F_1
xSID	mBERT	76.36	70.57	0.40	2.32
	XLM-R	90.20	76.23	0.31	2.70
	RemBERT	91.08	79.44	0.34	2.78
	mDeBERTa	94.88	82.62	0.24	2.69
	DistilmBERT	71.04	66.62	0.43	2.17
	mMiniLM	72.16	69.29	0.34	2.25

Acc/F1 Lost

Ekaterina Artemova, et al. Exploring the Robustness of Task-oriented Dialogue Systems for Colloquial German Varieties. EACL 2024.

Dialectal text transfer

- Problem:
  - Models perform worse on translation and text-image retrieval, when inputs are in dialects
  - This is due to significant vocabulary differences
- Solution:
  - Construct a parallel corpus
    - First collect conversations in central dialect, then convert them into northern dialect
  - Train mBART and Vietnamese BARTPho models to transfer (translate) other dialects to standard dialect
  - The Northern dialect style generated by BARTPho improved Vietnamese-English translation for Google Translate.

	Vietnamese Input Text		Gold Translation
	răng tự nhiên ngá cực kỳ luôn [Central Dialect]	sao tự nhiên ngứa cực kỳ luôn [Northern Dialect]	
Google Translate	natural teeth are very yawn	Why is it so itchy all of a sudden?	Oh I feel so itchy
Yandex Translate	natural teeth are extremely toothed	why does it naturally itch extremely well	
ChatGPT	My natural teeth are extremely sharp	why does it suddenly itch so much all the time	

Dialect	#Samples	#Avg. syll.	#Avg. word
Central	3761	10.88	10.35
Northern	3761	10.97	10.13

Text Style	Content	Fluency
Northern (Gold)	4.04	4.24
Northern (BARTpho)	3.96	4.21
Central	1.72	2.6

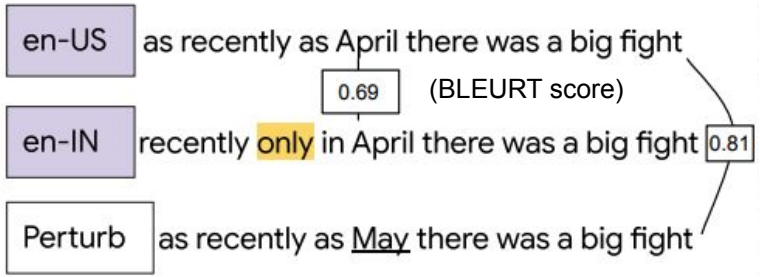
Thang Le and Anh Luu. 2023. A Parallel Corpus for Vietnamese Central-Northern Dialect Text Transfer.

Dialect-robust evaluation for NLG tasks

- Current NLG metrics is not ***dialect robust***
- This study proposes a **suite** to measure the dialect robustness of existing metrics
  - Dialects of Mandarin, English and Portuguese
  - Metrics: BLEURT, COMET, …
  - Win/Loss analysis

	EN	PT	ZH
All	148	2616	2227
Replace	96	962	866
Insert	89	550	528
Delete	63	693	614
AGG.	115	1415	1252

Number of evaluation examples in the suite.



Dialect edited sentence **should show higher similarity** than semantics edited sentence

Jiao Sun, et al. Dialect-robust Evaluation of Generated Text. In ACL 2023.

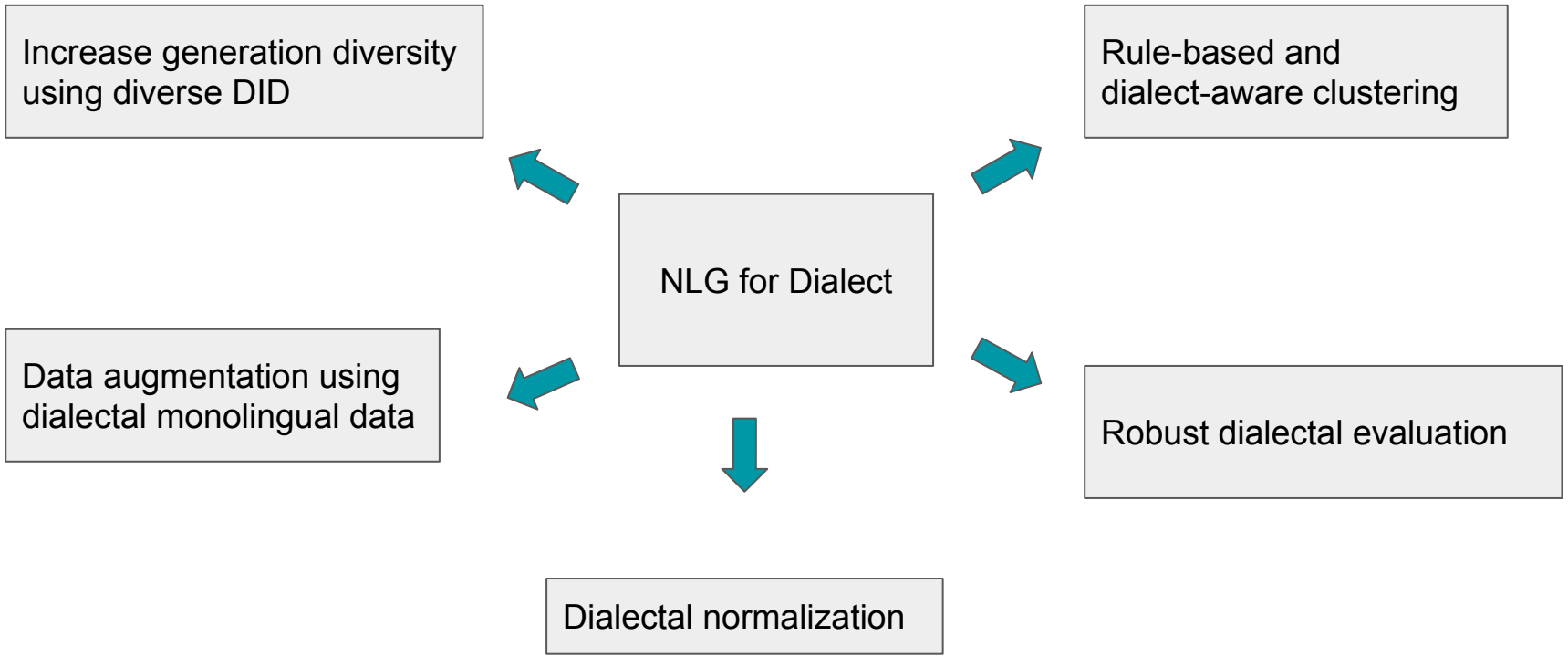
Dialect-robust evaluation for NLG tasks

- Current NLG metrics is not ***dialect robust***
- Proposed NANO
  - Distill regional and language information into metrics, using mT5 model in this study
    - Through binary classification **pre-training**
      - Input:
        - candidate: {sentence} language: {language_tag}
      - Output:
        - 0 or 1 indicating whether the sentence belongs to the language tag
    - Fine-tuning on WMT shared task to make mT5 model a quality estimation model.
  - Improves dialect robustness while maintaining performance on standard QE benchmark.

	Learned			*Lexical*		mT5_{base}		mT5_{XL}		mT5_{XXL}	
	BLEURT	PRISM	YISI	BLEU	CHRF	-NANO	+NANO	-NANO	♦ +NANO	-NANO	🏆 +NANO
EN	0.53	0.51	0.53	0.49	0.46	0.50	0.50	0.55	0.54	0.57	0.57
PT	**0.59**	0.53	0.36	0.35	0.35	0.39	0.44	**0.57**	**0.65**	**0.82**	**0.81**
ZH	**0.59**	0.47	0.46	0.35	0.36	0.46	0.45	0.51	**0.59**	**0.74**	**0.74**

Jiao Sun, et al. Dialect-robust Evaluation of Generated Text. In ACL 2023.

Summary


Increase generation diversity using diverse DID
Rule-based and dialect-aware clustering
NLG for Dialect
Data augmentation using dialectal monolingual data
Robust dialectal evaluation
Dialectal normalization

Future directions

- Data
  - How to identify/mine dialectal data from raw data more effectively?
  - Synthetic dialectal data creation by rules/LLMs/text style transfer?
- Application
  - Dialect/low-resource dialogue systems are in demand
- Evaluation
  - Evaluation of outputs by dialectal system is still an open question
    - How to capture the **semantic similarity** and **style difference** at the same time?

Hand-on session (Module 5 Part 3)

- English→Hindi machine translation
- Train from scratch using fairseq
  - fairseq.ipynb
- Zero/few-shot prompting of Airavata LLM for translation
  - prompt.ipynb
- HuggingFace models fine-tuning for translation
  - https://huggingface.co/docs/transformers/v4.17.0/en/tasks/translation
- Fine-tuning IndicBART for script conversion
  - https://colab.research.google.com/drive/13Gj7bAhR2HIdgSXEzp8fu4xwqrKsEYaa?usp=sharing

Tutorial Agenda