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

UNSW SYDNEY

People-Centred AI
UNIVERSITY OF SURREY

AALBORG UNIVERSITY

NICT

Tutorial Agenda

- Introduction ✓
- Dataset Creation ✓
- NLG ✓
- Emerging Connections ✓
- NLU ✓
- Conclusion

Tutorial Recap

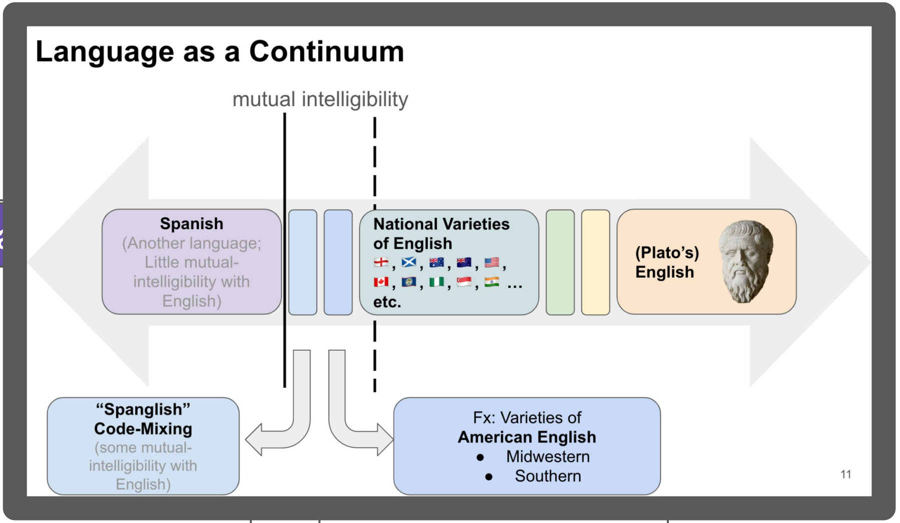
Introduction

✓

Emerging Connections

✓

Data



Language as a Continuum

mutual intelligibility

Spanish (Another language; Little mutual-intelligibility with English)

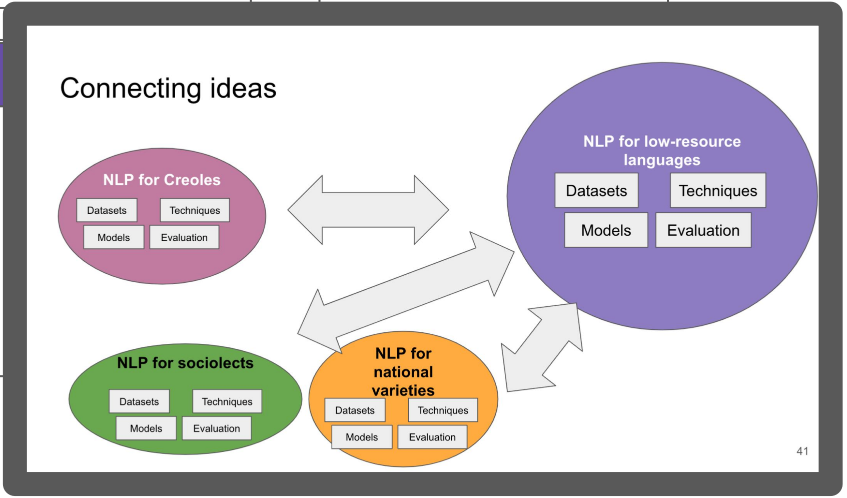
National Varieties of English 🏴󠁧󠁢󠁥󠁮󠁧󠁿, 🏴󠁧󠁢󠁳󠁣󠁴󠁿, 🇦🇺, 🇳🇿, 🇺🇸, 🇨🇦, 🇧🇿, 🇳🇬, 🇸🇬, 🇮🇳 ... **etc.**

(Plato's) English

"Spanglish" Code-Mixing (some mutual-intelligibility with English)

Fx: Varieties of **American English**
- Midwestern
- Southern

11



Connecting ideas

NLP for Creoles: Datasets, Techniques, Models, Evaluation

NLP for low-resource languages: Datasets, Techniques, Models, Evaluation

NLP for sociolects: Datasets, Techniques, Models, Evaluation

NLP for national varieties: Datasets, Techniques, Models, Evaluation

41

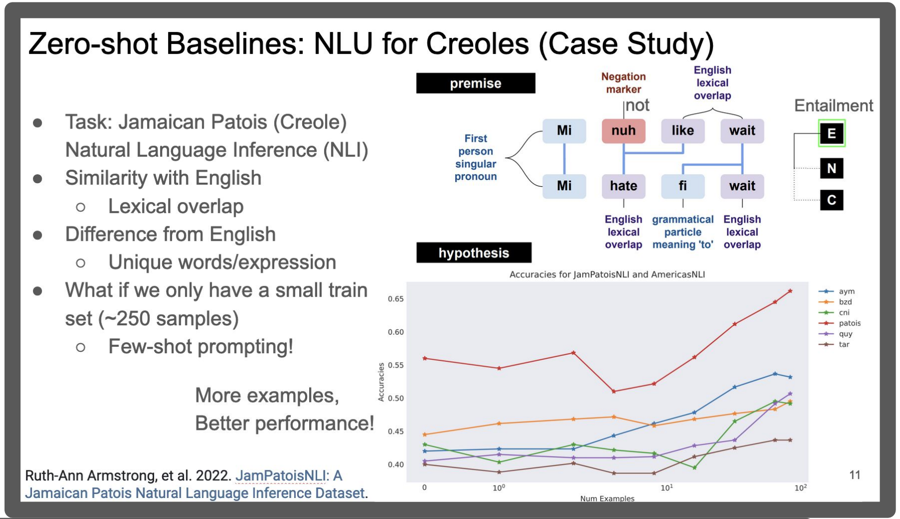
Tutorial Recap

Introduction

✓

Emerging Connections

✓



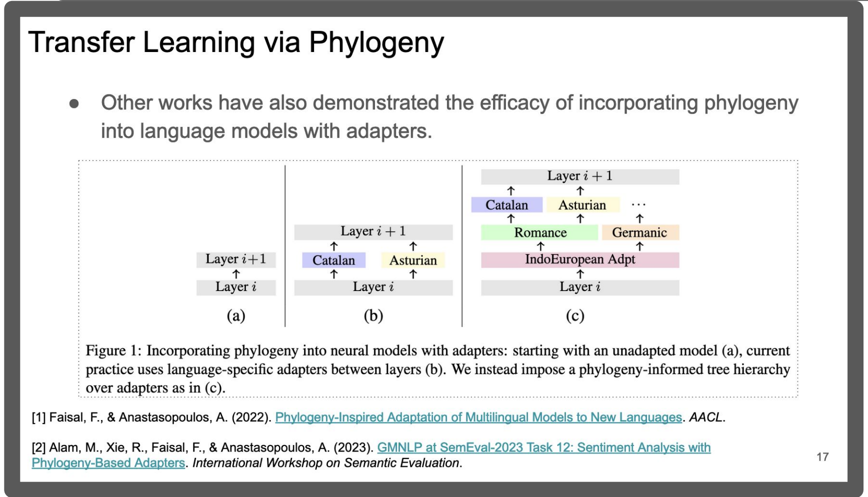
Zero-shot Baselines: NLU for Creoles (Case Study)

- Task: Jamaican Patois (Creole) Natural Language Inference (NLI)
- Similarity with English
  - Lexical overlap
- Difference from English
  - Unique words/expression
- What if we only have a small train set (~250 samples)
  - Few-shot prompting!

More examples, Better performance!

Ruth-Ann Armstrong, et al. 2022. JamPatoisNLI: A Jamaican Patois Natural Language Inference Dataset.

11



Transfer Learning via Phylogeny

- Other works have also demonstrated the efficacy of incorporating phylogeny into language models with adapters.

Figure 1: Incorporating phylogeny into neural models with adapters: starting with an unadapted model (a), current practice uses language-specific adapters between layers (b). We instead impose a phylogeny-informed tree hierarchy over adapters as in (c).

[1] Faisal, F., & Anastasopoulos, A. (2022). Phylogeny-Inspired Adaptation of Multilingual Models to New Languages. *AACL*.

[2] Alam, M., Xie, R., Faisal, F., & Anastasopoulos, A. (2023). GMNLP at SemEval-2023 Task 12: Sentiment Analysis with Phylogeny-Based Adapters. *International Workshop on Semantic Evaluation.*

17

Tutorial Recap

Introduction

✓

Dataset Creation

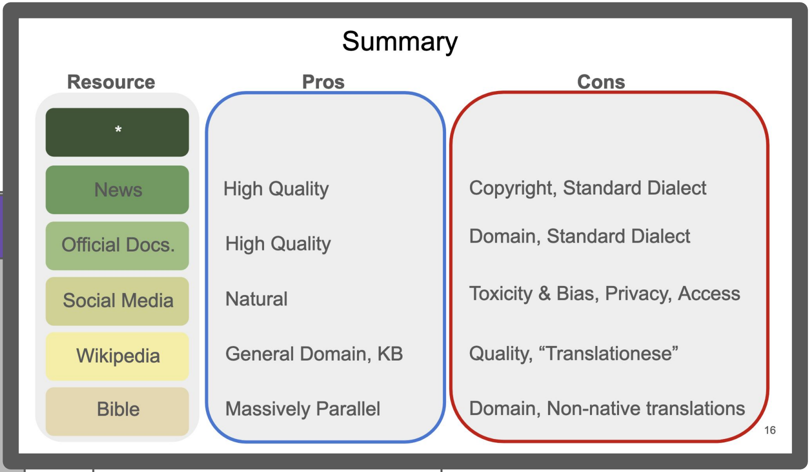
✓

Emerging Connections

✓

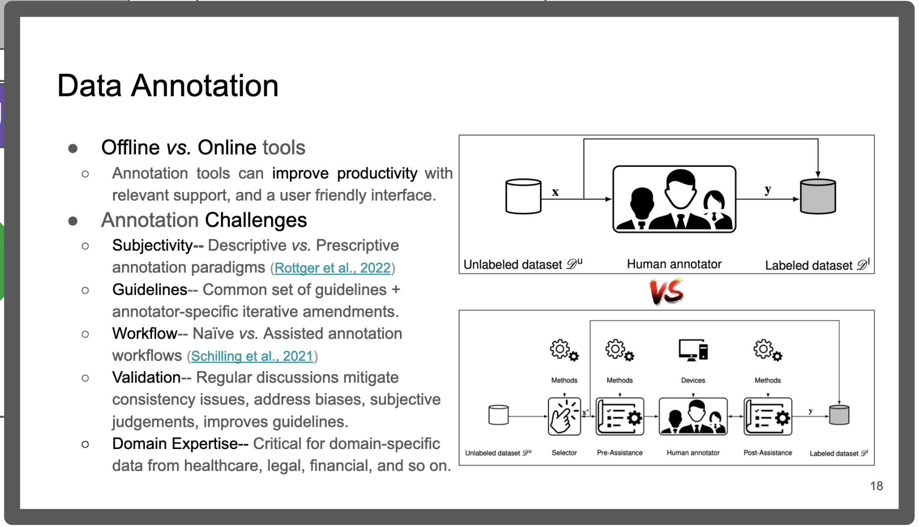
NLU

✓



Summary

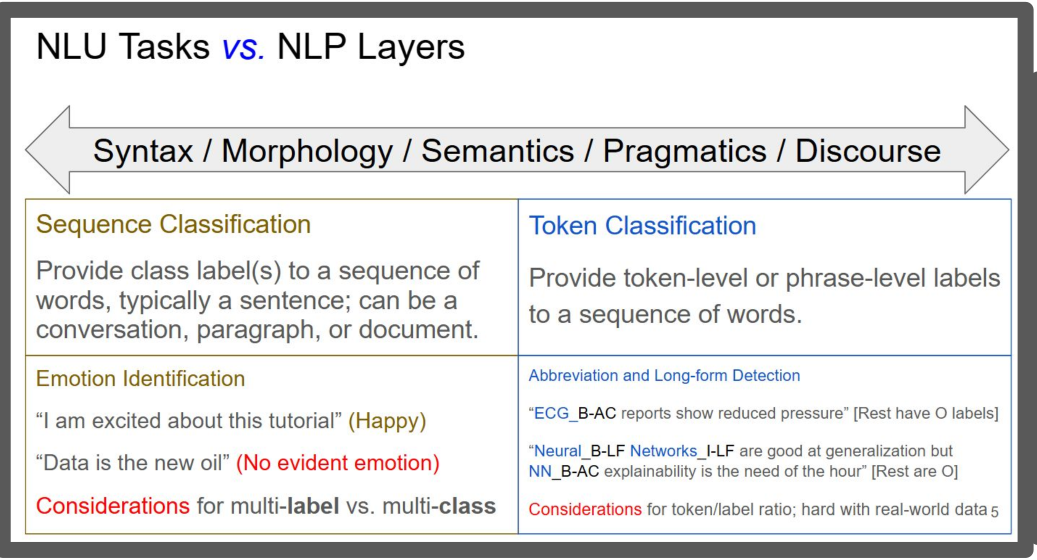
Resource	Pros	Cons
*		
News	High Quality	Copyright, Standard Dialect
Official Docs.	High Quality	Domain, Standard Dialect
Social Media	Natural	Toxicity & Bias, Privacy, Access
Wikipedia	General Domain, KB	Quality, “Translationese”
Bible	Massively Parallel	Domain, Non-native translations



Data Annotation

- Offline *vs.* Online tools
  - Annotation tools can **improve productivity** with relevant support, and a user friendly interface.
- Annotation Challenges
  - Subjectivity-- Descriptive *vs.* Prescriptive annotation paradigms (Rottger et al., 2022)
  - Guidelines-- Common set of guidelines + annotator-specific iterative amendments.
  - Workflow-- Naïve *vs.* Assisted annotation workflows (Schilling et al., 2021)
  - Validation-- Regular discussions mitigate consistency issues, address biases, subjective judgements, improves guidelines.
  - Domain Expertise-- Critical for domain-specific data from healthcare, legal, financial, and so on.

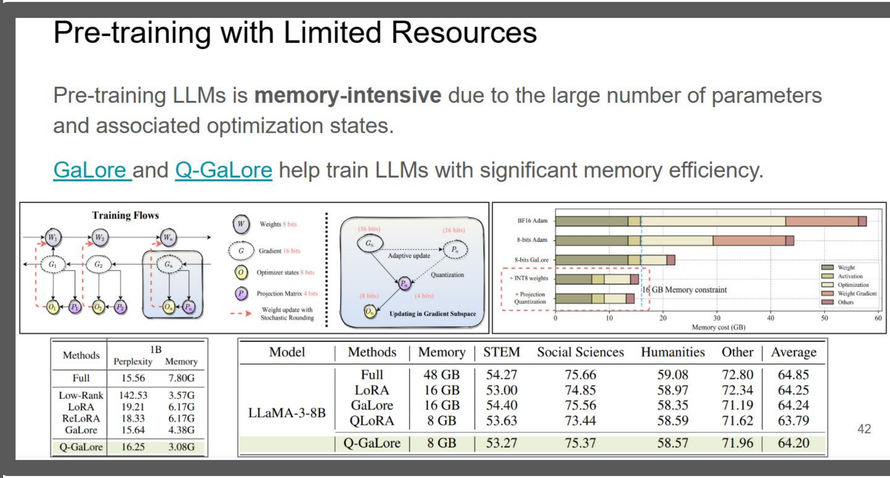
x
y
Unlabeled dataset $\mathscr{D}^u$
Human annotator
Labeled dataset $\mathscr{D}^l$
VS
Methods
Methods
Devices
Methods
x'
y
Unlabeled dataset $\mathscr{D}^u$
Selector
Pre-Assistance
Human annotator
Post-Assistance
Labeled dataset $\mathscr{D}^l$



NLU Tasks *vs.* NLP Layers

Syntax / Morphology / Semantics / Pragmatics / Discourse

Sequence Classification	Token Classification
Provide class label(s) to a sequence of words, typically a sentence; can be a conversation, paragraph, or document.	Provide token-level or phrase-level labels to a sequence of words.
Emotion Identification "I am excited about this tutorial" (Happy) "Data is the new oil" (No evident emotion) Considerations for multi-**label** vs. multi-**class**	Abbreviation and Long-form Detection "ECG_B-AC reports show reduced pressure" [Rest have O labels] "Neural_B-LF Networks_I-LF are good at generalization but NN_B-AC explainability is the need of the hour" [Rest are O] Considerations for token/label ratio; hard with real-world data 5



Pre-training with Limited Resources

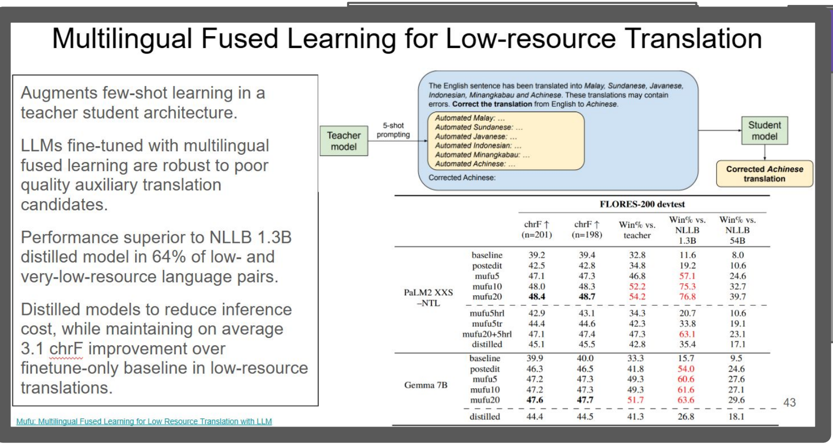
Pre-training LLMs is **memory-intensive** due to the large number of parameters and associated optimization states.

GaLore and Q-GaLore help train LLMs with significant memory efficiency.

Methods	1B Perplexity	1B Memory
Full	15.56	7.80G
Low-Rank	142.53	3.57G
LoRA	19.21	6.17G
ReLoRA	18.33	6.17G
GaLore	15.64	4.38G
Q-GaLore	16.25	3.08G

Model	Methods	Memory	STEM	Social Sciences	Humanities	Other	Average
LLaMA-3-8B	Full	48 GB	54.27	75.66	59.08	72.80	64.85
	LoRA	16 GB	53.00	74.85	58.97	72.34	64.25
	GaLore	16 GB	54.40	75.56	58.35	71.19	64.24
	QLoRA	8 GB	53.63	73.44	58.59	71.62	63.79
	Q-GaLore	8 GB	53.27	75.37	58.57	71.96	64.20

42



Multilingual Fused Learning for Low-resource Translation

Augments few-shot learning in a teacher student architecture.

LLMs fine-tuned with multilingual fused learning are robust to poor quality auxiliary translation candidates.

Performance superior to NLLB 1.3B distilled model in 64% of low- and very-low-resource language pairs.

Distilled models to reduce inference cost, while maintaining on average 3.1 chrF improvement over finetune-only baseline in low-resource translations.

The English sentence has been translated into *Malay, Sundanese, Javanese, Indonesian, Minangkabau and Achinese*. These translations may contain errors. **Correct the translation** from English to *Achinese*.

Automated Malay: ...
Automated Sundanese: ...
Automated Javanese: ...
Automated Indonesian: ...
Automated Minangkabau: ...
Automated Achinese: ...

Corrected Achinese:

		FLORES-200 devtest				
		chrF ↑ (n=201)	chrF ↑ (n=198)	Win% vs. teacher	Win% vs. NLLB 1.3B	Win% vs. NLLB 54B
PaLM2 XXS –NTL	baseline	39.2	39.4	32.8	11.6	8.0
	postedit	42.5	42.8	34.8	19.2	10.6
	mufu5	47.1	47.3	46.8	57.1	24.6
	mufu10	48.0	48.3	52.2	75.3	32.7
	mufu20	**48.4**	**48.7**	54.2	76.8	39.7
	mufu5hrl	42.9	43.1	34.3	20.7	10.6
	mufu5tr	44.4	44.6	42.3	33.8	19.1
	mufu20+5hrl	47.1	47.4	47.3	63.1	23.1
	distilled	45.1	45.5	42.8	35.4	17.1
Gemma 7B	baseline	39.9	40.0	33.3	15.7	9.5
	postedit	46.3	46.5	41.8	54.0	24.6
	mufu5	47.2	47.3	49.3	60.6	27.6
	mufu10	47.2	47.3	49.3	61.6	27.1
	mufu20	**47.6**	**47.7**	51.7	63.6	29.6
	distilled	44.4	44.5	41.3	26.8	18.1

43

Mufu: Multilingual Fused Learning for Low Resource Translation with LLM

NLU

✓

Conclusion

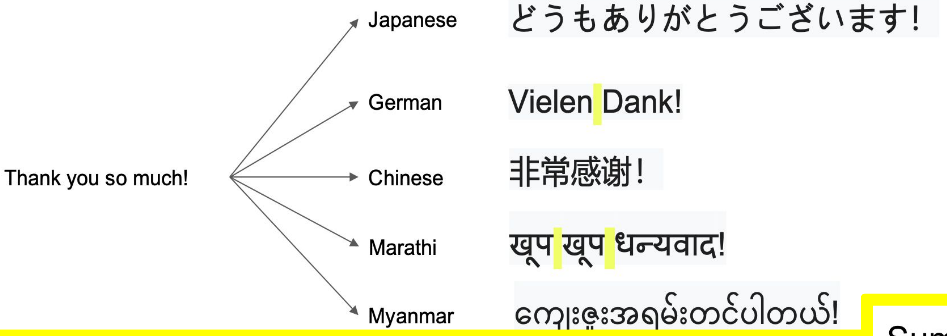
Natural Language Generation:
Machine Translation for various (low-resource) languages

Challenges & solutions

Data scarcity -> Data augmentation

Diverse scripts -> Script normalization

Evaluation -> Language agnostic evaluation



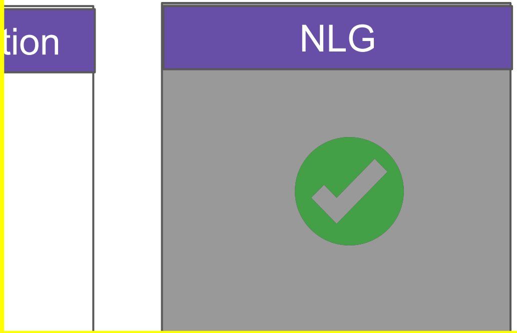
どうもありがとうございます!

Vielen Dank!

非常感谢!

खूप खूप धन्यवाद!

ကျေးဇူးအရမ်းတင်ပါတယ်!



Unified Evaluation and Correction

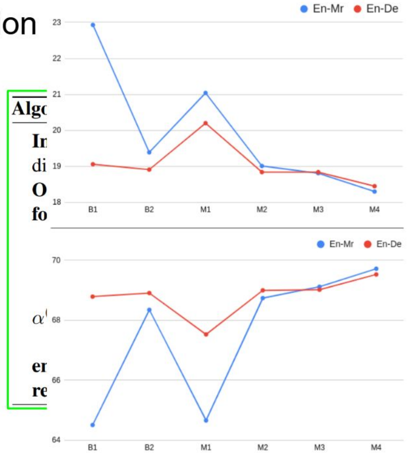
Merging QE and APE - Sentence-level + Word-level + APE, for context-aware unified evaluation and correction.

Progressively Integrating QE with APE

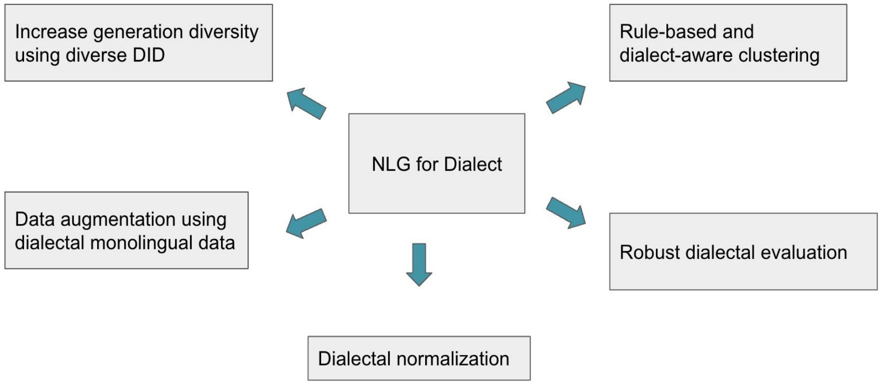
- ➔ QE as APE Activator
- ➔ QE as MT/APE Selector
- ➔ QE as APE Guide
- ➔ **Joint Training over QE and APE**
  - ◆ **Linear Scalarization (LS-MTL) vs. Nash-MTL**

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

Quality Estimation-Assisted Automatic Post-Editing [Findings of EMNLP 2023]



Summary



NLG for Dialect

Increase generation diversity using diverse DID

Rule-based and dialect-aware clustering

Data augmentation using dialectal monolingual data

Robust dialectal evaluation

Dialectal normalization

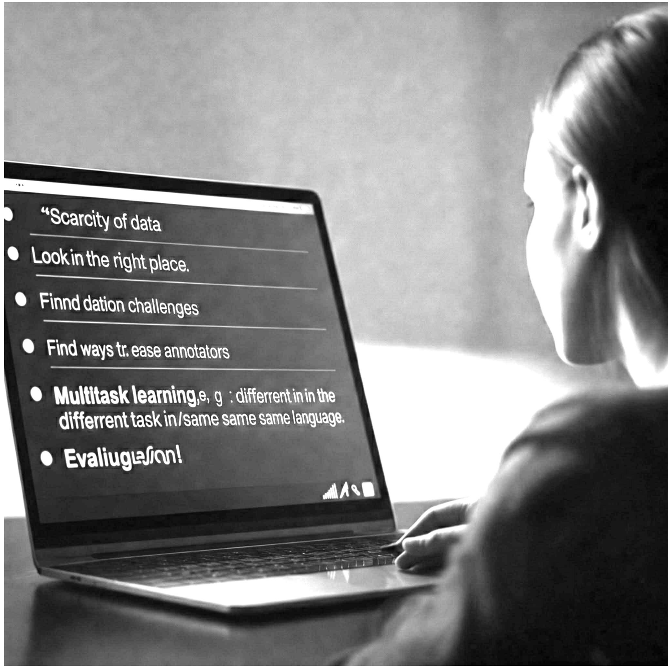
A question for all the presenters…

A question for all the presenters…

What new idea from another lower-resource scenario did you learn as a part of making this tutorial?

Common Ideas

- Motivations
    - Out-of-the-box NLP tools don’t work for low-resource scenarios.
        - Sometimes they’re not designed too (*e.g.*, LID)
- Data
    - Scarcity of data → Look in the right place!
    - Annotation challenges → Find ways to ease the burden for annotators.
    - Multilingual data for same task [related languages]
- Method(s)
    - Adaptation
    - Multitask learning [same language (pair), different tasks]
- Evaluation
    - Challenging test sets


“Scarcity of data
Look in the right place.
Finnd dation challenges
Find ways tr. ease annotators
Multitask learning,e, g : differrent in in the differrent task in/same same same language.
Evaliug⊿ʃion!

Future “common” directions

- Dataset creation
  - Let’s create datasets of our languages/dialects!
  - Translate (or style transfer) existing datasets into low-resource languages/dialects
  - Generating synthetic data by leveraging LLMs
- Domain/language adaptation:
  - Few-shot prompting
  - Instruction fine-tuning
- Incorporating linguistic information and intuitions
- Evaluation
  - Low-resource language/dialect aware NLU and NLG evaluation metrics.
  - Evaluating on different a low-resource scenario to understand a method’s generalizability.

Future “Common” Directions: Dataset Creation

Let’s create datasets of our languages/dialects!

Use assisted annotation workflow & ensure regular validation of data curation

Generating synthetic data by leveraging LLMs

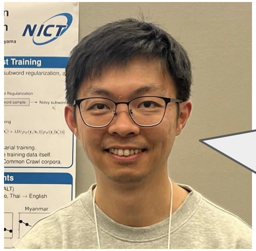
Translate (or style transfer) existing datasets into low-resource languages/dialects

Future “Common” Directions: Methodology

Domain/language adaptation:
– Few-shot prompting
– Instruction fine-tuning

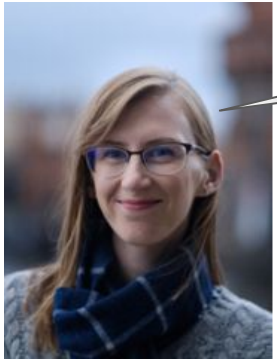
Incorporating linguistic information and intuitions (when possible)

Leverage Multilinguality, Cognates, and Multi-task Learning



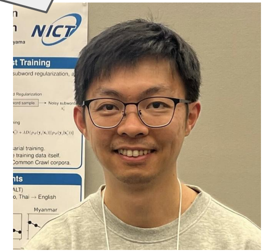
Universal LID to identify new language without training.

Future “Common” Directions: Evaluation



Evaluating on different a low-resource scenario to understand a method’s generalizability.

Low-resource language/dialect aware NLU and NLG evaluation metrics.



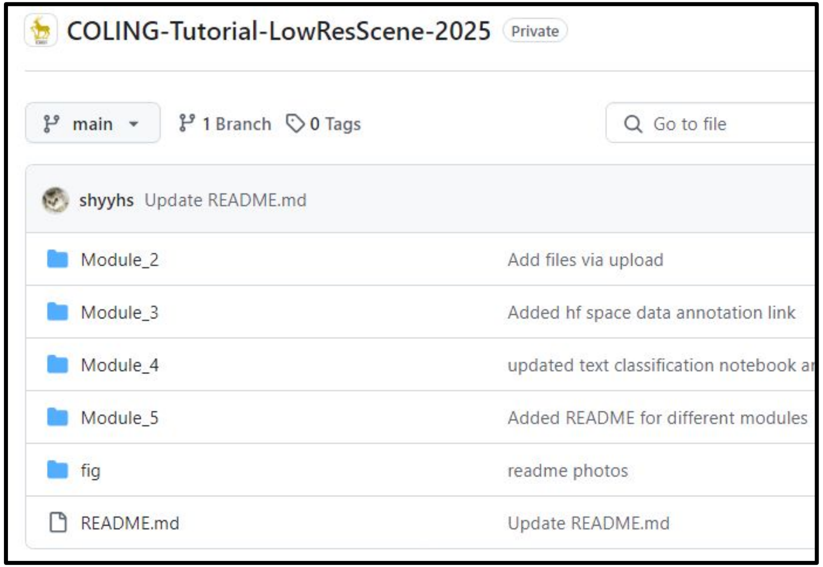
Can MT evaluation leverage a Retrieval Augmented Generation pipeline?

Future “common” directions

- Universal LID: can LID identify new language without training on it before?
  - Predicting language features instead of predicting a fix number of languages (intermediate)
  - Map the predicted features to languages (how?)
- Can MT evaluation leverage a Retrieval Augmented Generation pipeline?
  - Use parallel corpus
  - Does parallel corpus from a related language help?
  - Translation Memories? Phrase tables?

Tutorial material available at:

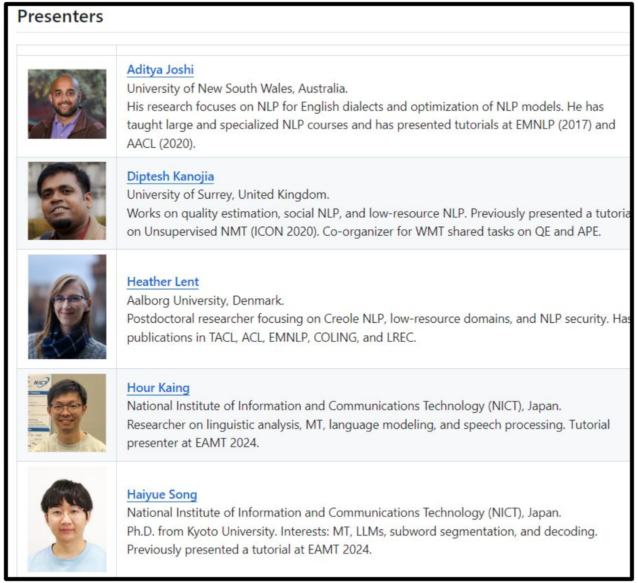
https://github.com/surrey-nlp/COLING-Tutorial-LowResScene-2025



COLING-Tutorial-LowResScene-2025 Private

shyyhs Update README.md

Module_2	Add files via upload
Module_3	Added hf space data annotation link
Module_4	updated text classification notebook ar
Module_5	Added README for different modules
fig	readme photos
README.md	Update README.md



Presenters

Aditya Joshi
University of New South Wales, Australia.
His research focuses on NLP for English dialects and optimization of NLP models. He has taught large and specialized NLP courses and has presented tutorials at EMNLP (2017) and AACL (2020).

Diptesh Kanojia
University of Surrey, United Kingdom.
Works on quality estimation, social NLP, and low-resource NLP. Previously presented a tutoria on Unsupervised NMT (ICON 2020). Co-organizer for WMT shared tasks on QE and APE.

Heather Lent
Aalborg University, Denmark.
Postdoctoral researcher focusing on Creole NLP, low-resource domains, and NLP security. Has publications in TACL, ACL, EMNLP, COLING, and LREC.

Hour Kaing
National Institute of Information and Communications Technology (NICT), Japan.
Researcher on linguistic analysis, MT, language modeling, and speech processing. Tutorial presenter at EAMT 2024.

Haiyue Song
National Institute of Information and Communications Technology (NICT), Japan.
Ph.D. from Kyoto University. Interests: MT, LLMs, subword segmentation, and decoding. Previously presented a tutorial at EAMT 2024.