

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



Tutorial Agenda

Introduction



Dataset Creation



NLG



Emerging Connections



NLU



Conclusion

Tutorial Recap

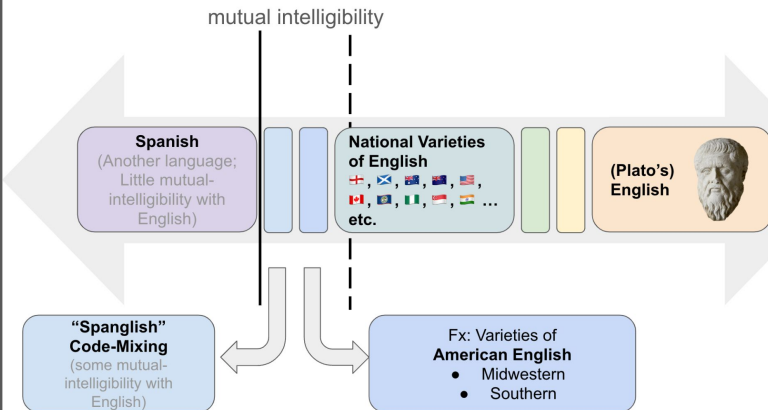
Introduction



Emerging Connections

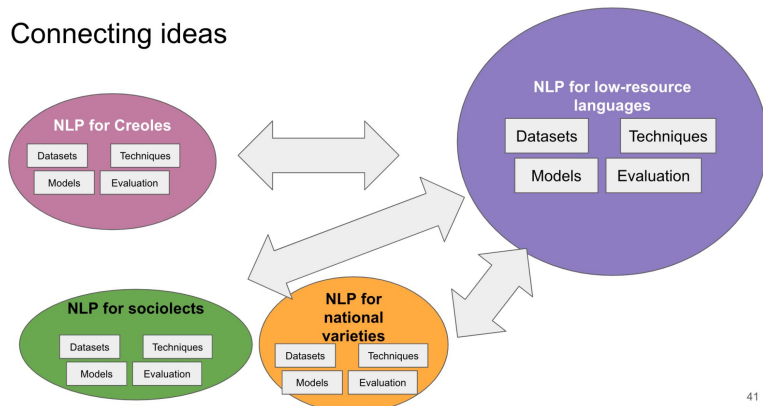


Language as a Continuum



11

Connecting ideas



41

Tutorial Recap

Introduction



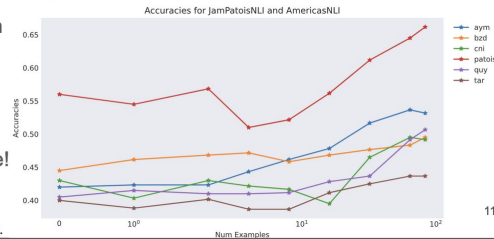
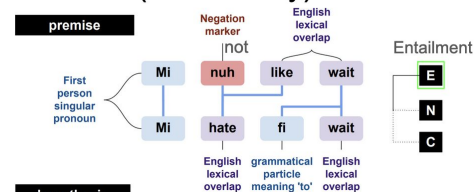
Emerging Connections



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/expressions
- 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.

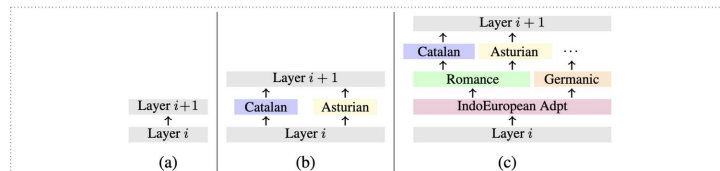


Figure 1: Incorporating phylogeny into neural models with adapters: starting with an unadapted model (a), current practice uses language-specific adapters between layers (b). We instead impose a phylogeny-informed tree hierarchy over adapters as in (c).

[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.

Tutorial Recap

Introduction



Dataset Creation



Emerging Connections



NLU



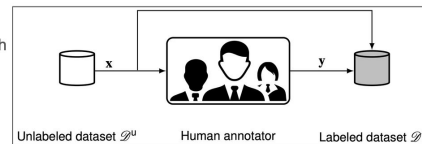
Summary

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

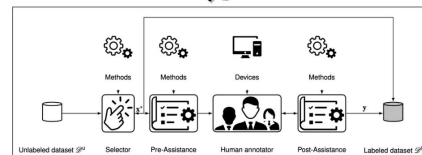
16

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.



VS



18

NLU Tasks vs. NLP Layers

Syntax / Morphology / Semantics / Pragmatics / Discourse

Sequence Classification

Provide class label(s) to a sequence of words, typically a sentence; can be a conversation, paragraph, or document.

Emotion Identification

"I am excited about this tutorial" (Happy)

"Data is the new oil" (No evident emotion)

Considerations for multi-label vs. multi-class

Token Classification

Provide token-level or phrase-level labels to a sequence of words.

Abbreviation and Long-form Detection

"ECG_B-AC reports show reduced pressure" [Rest have O labels]

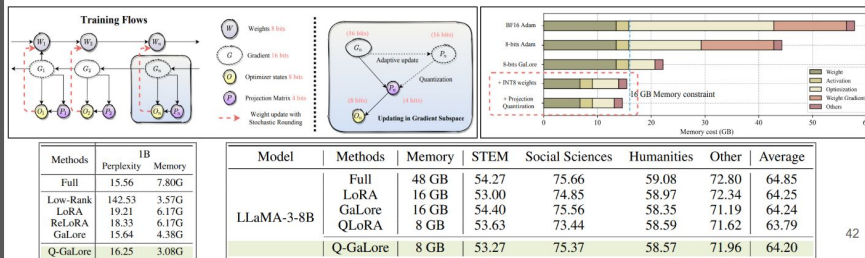
"Neural_B-LF Networks_I-LF are good at generalization but NN_B-AC explainability is the need of the hour" [Rest are O]

Considerations for token/label ratio; hard with real-world data

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.



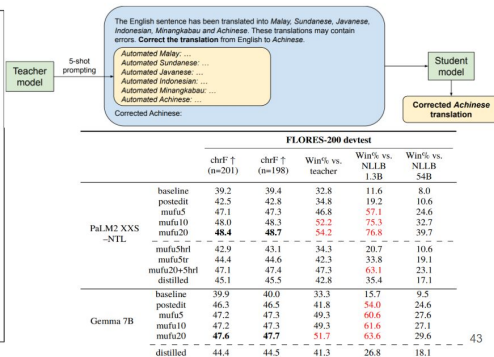
Multilingual Fused Learning for Low-resource Translation

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.



NLU



Conclusion

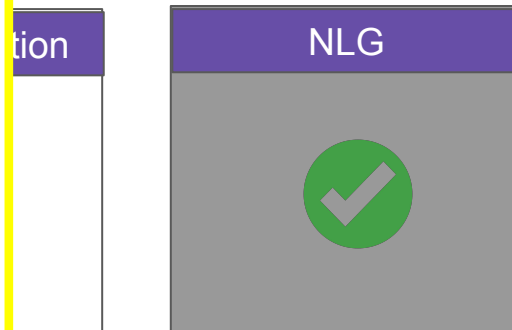
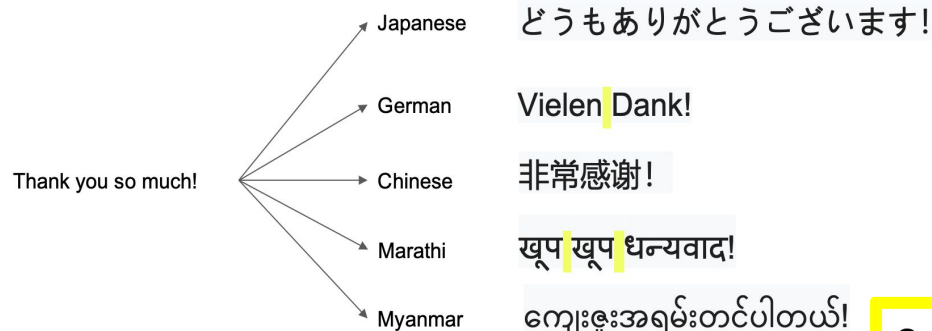
Natural Language Generation: Machine Translation for various (low-resource) languages

Challenges & solutions

Data scarcity -> Data augmentation

Diverse scripts -> Script normalization

Evaluation -> Language agnostic evaluation



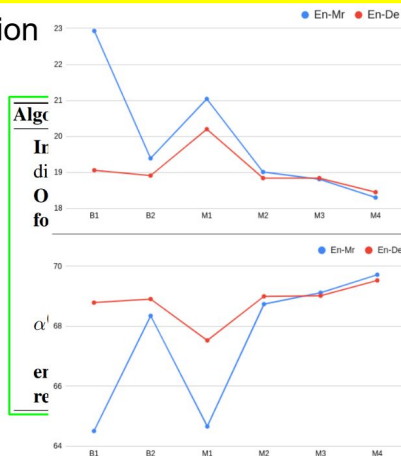
Unified Evaluation and Correction

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

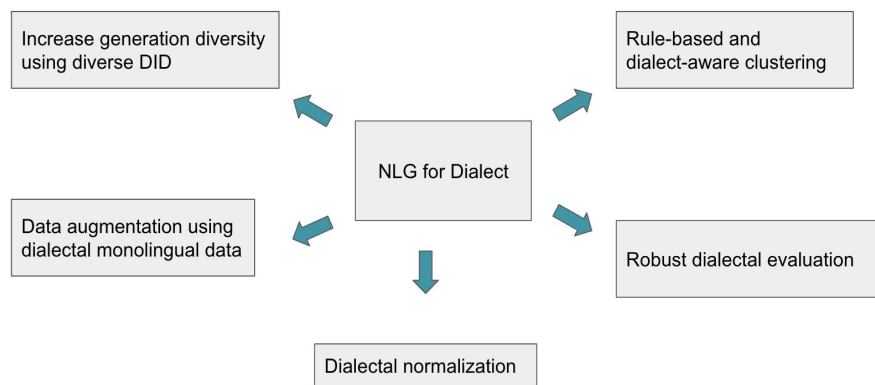
Progressively Integrating QE with APE

- 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

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



Summary



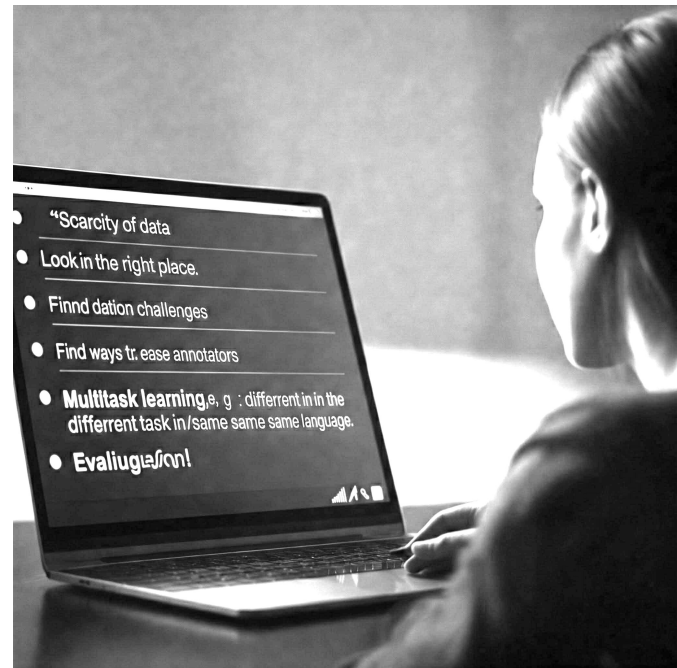
A question for all the presenters...

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




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



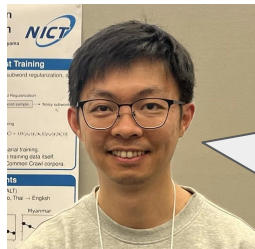
Incorporating linguistic information and intuitions (when possible)



Leverage Multilinguality, Cognates, and Multi-task Learning



Universal LID to identify new language without training.

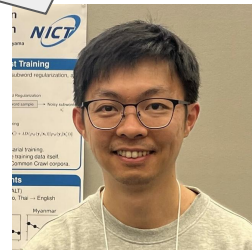
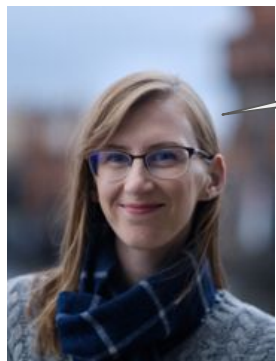


Future “Common” Directions: Evaluation

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

Can MT evaluation leverage a Retrieval Augmented Generation pipeline?

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




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

 **main**  1 Branch  0 Tags


 **shyyhs** Update README.md

 Module_2


Add files via upload

 Module_3


Added hf space data annotation link

 Module_4


updated text classification notebook a

 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 ACL (2020).




[Diptesh Kanodia](#)
University of Surrey, United Kingdom.
Works on quality estimation, social NLP, and low-resource NLP. Previously presented a tutorial 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.