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

Module 2: Emerging Connections

– **Intro**: Sociolinguistic Considerations (Variability and informality of dialects)

– **Identifying the Baseline**: Zero-shot performance

– Getting more from **transfer learning** (Phylogenetic relationships of low-resource languages, language selection, etc.)

– Emerging **common themes** in the tutorial

– **Hands-on Session** (10 mins): Evaluate zero-shot on a dialect dataset

for sarcasm detection highlighting results, challenges, and pitfalls.

– **Q&A & Discussion** (10 mins)

Speaker Dynamics: Singlish Example

"John sibei hum sup one!"

Translation: "*John is so lecherous*"

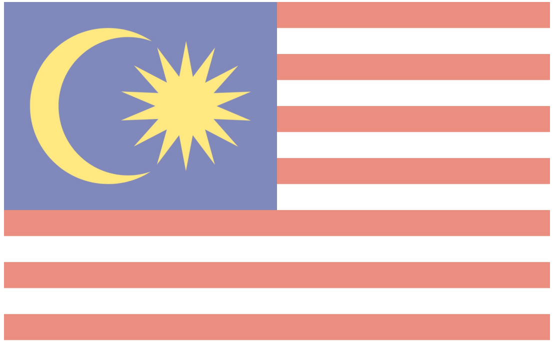
"John very buaya sia!"

Bajpai, R., Poria, S., Ho, D., & Cambria, E. (2016). [Developing a concept-level knowledge base for sentiment analysis in Singlish](). *Conference on Intelligent Text Processing and Computational Linguistics*.

Speaker Dynamics: Singlish Example

"John sibei hum sup one!"

"John very buaya sia!"



Bajpai, R., Poria, S., Ho, D., & Cambria, E. (2016). Developing a concept-level knowledge base for sentiment analysis in Singlish. *Conference on Intelligent Text Processing and Computational Linguistics*.

Speaker Dynamics: Singlish Example

*“**Singlish is avoided in formal contexts**, especially at job interviews, meetings with clients, presentations or meetings, where Standard English is preferred”*

Basilect

“Wah lau! This guy Singlish si beh hiong sia”

Mesolect

“This guy Singlish damn good leh.”

Acrolect (Standard)

“This person’s Singlish is very good.”

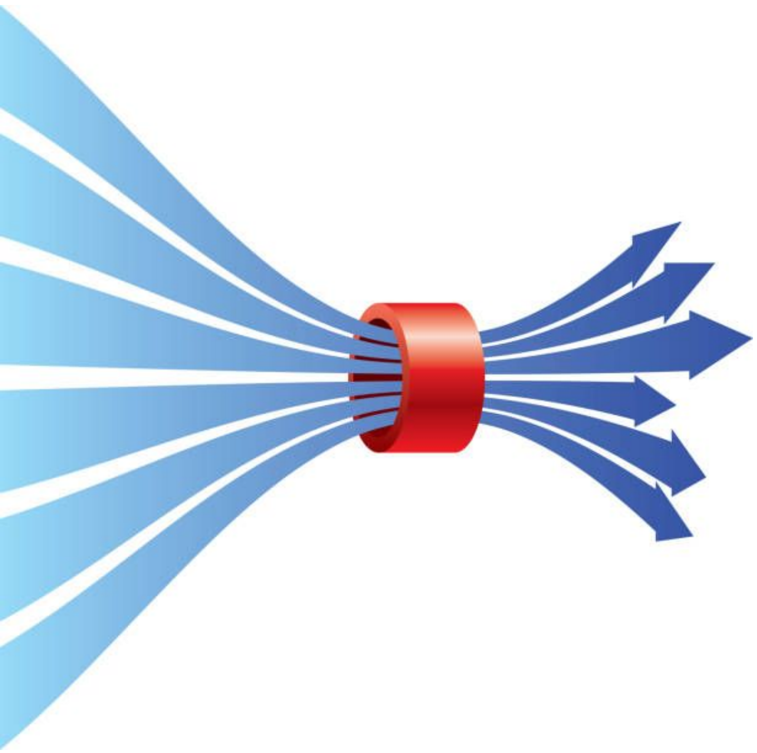
(National Variety!)

*“Although Singlish is officially discouraged in Singaporean schools, in practice, there is often some level of **code-switching present in the classroom**”*

*“In **informal settings**, such as during conversation with friends, or transactions in kopitiams and shopping malls, **Singlish is used without restriction**”*

https://en.wikipedia.org/wiki/Singlish

Connecting Ideas: A Formality Bottleneck



Some dialects, code-mixing, and Creoles tend to be used predominantly in informal contexts: speech, SMS, social media, etc.

They are less likely to be written in formal documents like parliamentary transcripts or encyclopedias.

Gathering data can be a challenge, and the **domain** can differ from LLM pre-training data.

Still, users prefer language technology that mimics their own language usage Bawa et al. (2020).

How effective are out-of-the-box models at processing such cases?

Zero-shot Baselines: Dialects

- Task: Vietnamese dialect→English translation.
- Problem: when the input is in a minor Vietnamese dialect, the performance will be unsatisfactory.
  - Reason: vocabulary difference from the standard dialect.
- Possible solutions: style transfer/few-shot prompting/fine-tuning

Minor dialect

	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	

🥲 🤗

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

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

Zero-shot Baselines: NLU for Creoles

Performance quickly drops w/o data!!

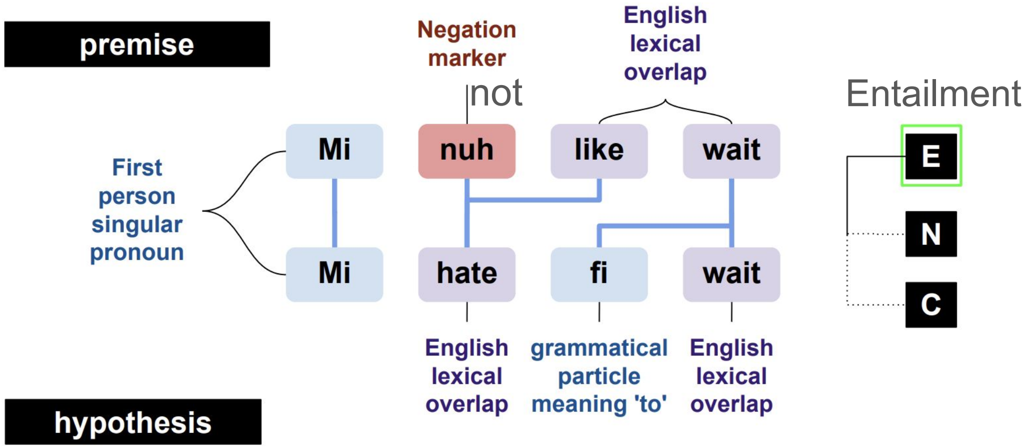
Task	Language	Dataset	Metric	mBERT	XLM-R	mT5
UDPoS (supervised)	pcm	UD_Naija-NSC (Caron et al., 2019)	Acc	0.98	**0.98**	0.98
	singlish	Singlish Treebank (Wang et al., 2017)	Acc	0.91	**0.93**	0.91
NER (supervised)	pcm	MasakhaNER (Adelani et al., 2021)	Span-F1	0.89	0.89	**0.90**
	bis	WikiAnn (Pan et al., 2017)	Span-F1	**0.94**	0.90	0.72
	cbk-zam			**0.96**	0.96	0.94
	hat			0.78	**0.84**	0.48
	pih			**0.90**	0.88	0.61
	sag			0.89	**0.93**	0.79
	tpi			**0.91**	0.89	0.75
	pap			**0.90**	0.89	0.85
SA (supervised)	pcm	AfriSenti (Muhammad et al., 2023b)	Acc	0.66	**0.68**	0.67
	pcm	Naija VADER (Oyewusi et al., 2020)	Acc	0.71	**0.72**	0.72
NLI (few-shot)	jam	JamPatoisNLI (Armstrong et al., 2022)	Acc	0.74	**0.76**	0.66
Sentence Matching (zero-shot)	cbk-eng	Tatoeba (Artetxe and Schwenk, 2019)	Acc	**15.9**	3.9	6.5
	gcf-eng			**12.8**	4.9	6.9
	hat-eng			23.9	18.5	**37.9**
	jam-eng			**19.9**	9.6	10.3
	pap-eng			**22.4**	6.1	15.9
	sag-eng			5.7	2.1	**7.3**
	tpi-eng			7.2	3.3	**7.6**

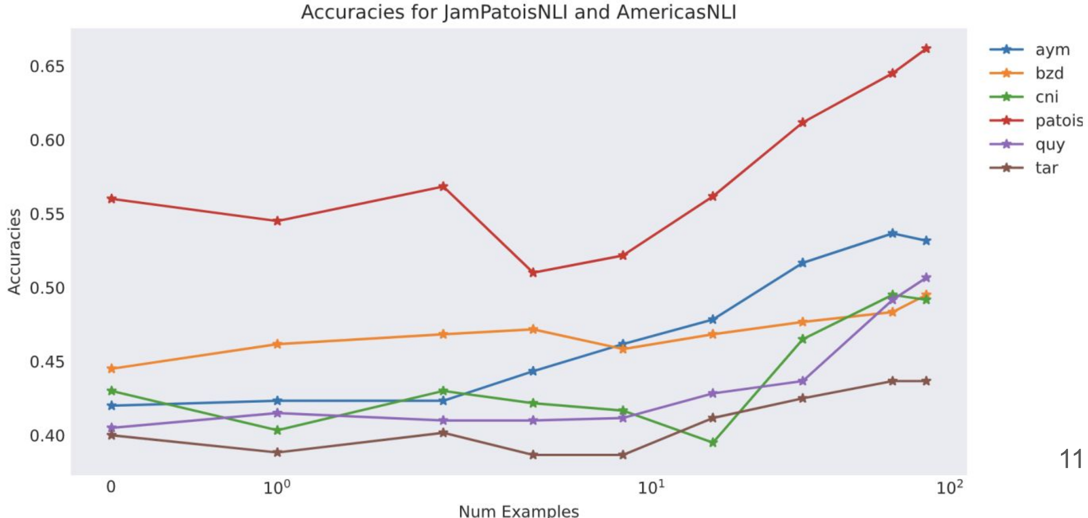
Table 4: Baseline scores for pre-existing NLU tasks for Creoles: dependency parsing (UDPoS), named entity recognition (NER), sentiment analysis (SA), natural language inference (NLI) , and sentence matching. Additional experiments, results, and analysis are included in the CreoleVal repository's documentation.

Lent, Heather, et al. "CreoleVal: Multilingual Multitask Benchmarks for Creoles." *Transactions of the Association for Computational Linguistics* 12 (2024): 950-978.

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!



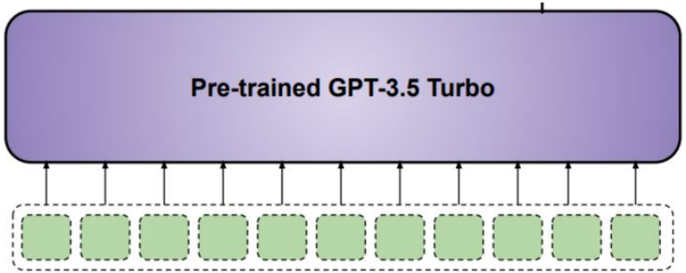


More examples,
Better performance!

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

Zero-shot Baselines: Code-Mixing

- Task: Sarcasm Detection
  - Binary classification: Sarcastic or Non-Sarcastic
  - **Tamil-English** and **Malayalam-English**
- Problem: "*... a pre-trained multilingual model does not necessarily guarantee high quality representations on code-switching, …*" [Winata et al. 2021]
- Approach: GPT-3.5 Turbo in zero-shot mode via prompting
- **Macro F1 not better than random chance!**



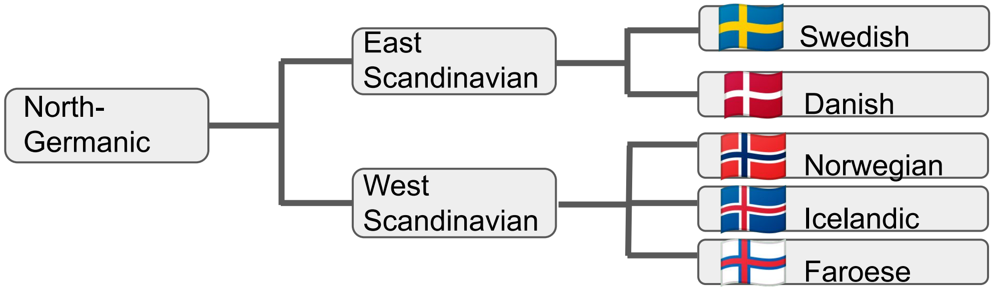
	Precision	Recall	F1-Score	Support
Non-sarcastic	0.79	0.79	0.79	4621
Sarcastic	0.43	0.43	0.43	1717
Micro avg	0.69	0.69	0.69	6338
Macro avg	0.61	0.61	0.61	6338
Weighted avg	0.69	0.69	0.69	6338

Non-sarcastic	0.82	0.73	0.77	2314
Sarcastic	0.18	0.27	0.22	512
Micro avg	0.65	0.65	0.65	2826
Macro avg	0.50	0.50	0.50	2826
Weighted avg	0.70	0.65	0.67	2826

Genta Indra Winata, et al., 2021. [Are Multilingual Models Effective in Code-Switching?]
Deroy and Maity, 2024. [YouTube Comments Decoded: Leveraging LLMs for Low Resource Language Classification]

Approaches to improving the zero-shot baseline via Transfer Learning, without additional LR data?

Transfer Learning via Phylogeny

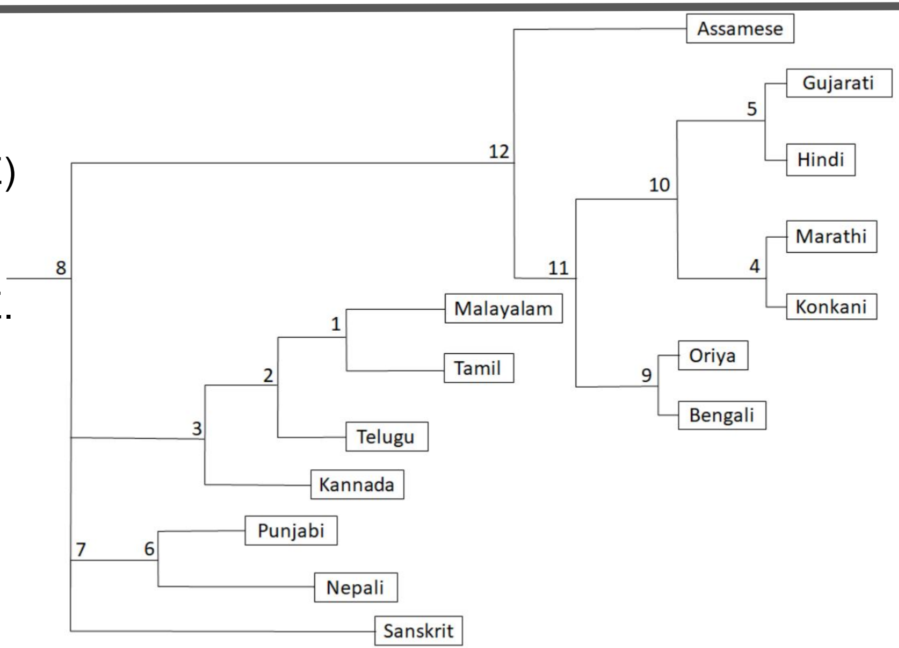


- Target: Faroese (~70k speakers)
- Best language for transfer?
  - Icelandic (300k speakers)
  - Norwegian (4.3m speakers)

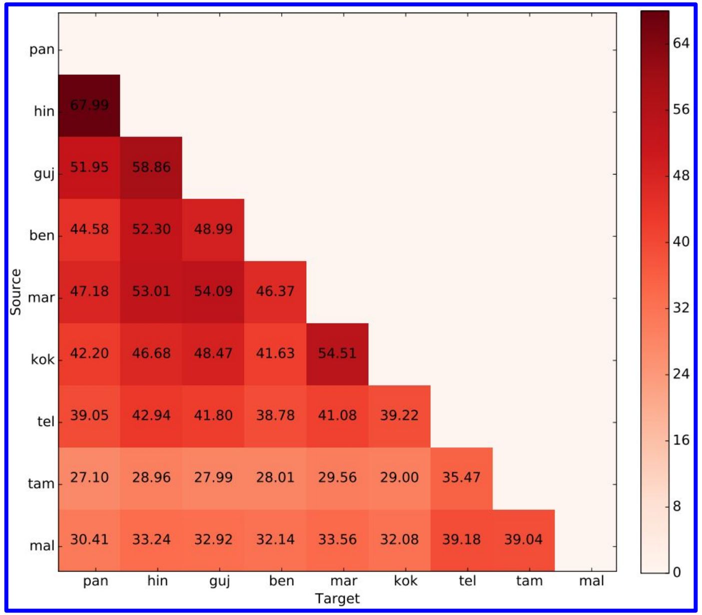
[Harnessing Deep Cross-lingual Word Embeddings to Infer Accurate Phylogenetic Trees [CODS-COMAD 2020]]

➔ Proposes using cross-lingual word embeddings (CWE) for phylogenetic reconstruction.
➔ **Dataset** - by Unicode offsetting IndoWordnet data.
➔ Sub-word embeddings using fastText; MUSE for CWE.
➔ Fares **better than simple lexical similarity** based approach - how? Cognates and borrowed vocabulary

Language distance - **Average word-pair cosine distances** for '**synset distance**', and average parallel synset distances for ***interlanguage distance***.

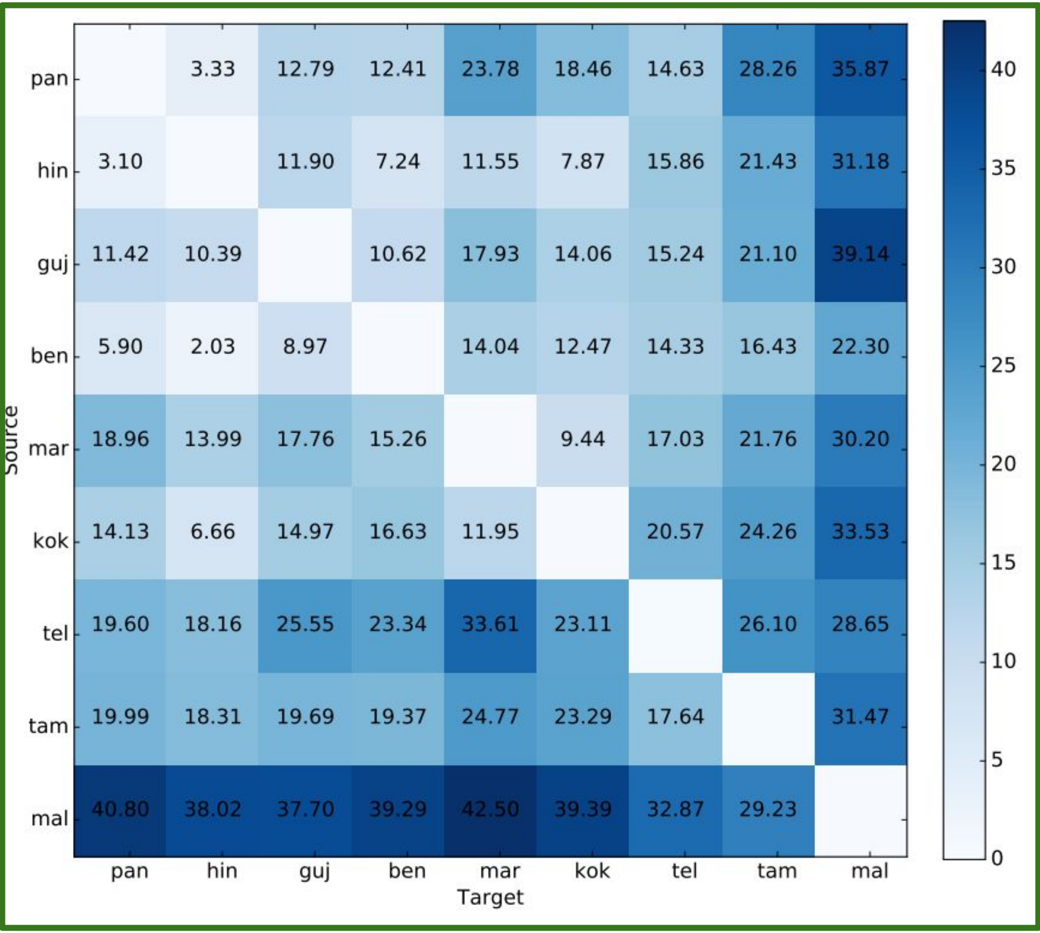


Transfer Learning via Phylogeny - Language Relatedness



रोटी (
मछली
भाषा (
दस (
Sans
चक्रम्
मत्स्यः
अश्वः (
जलम्

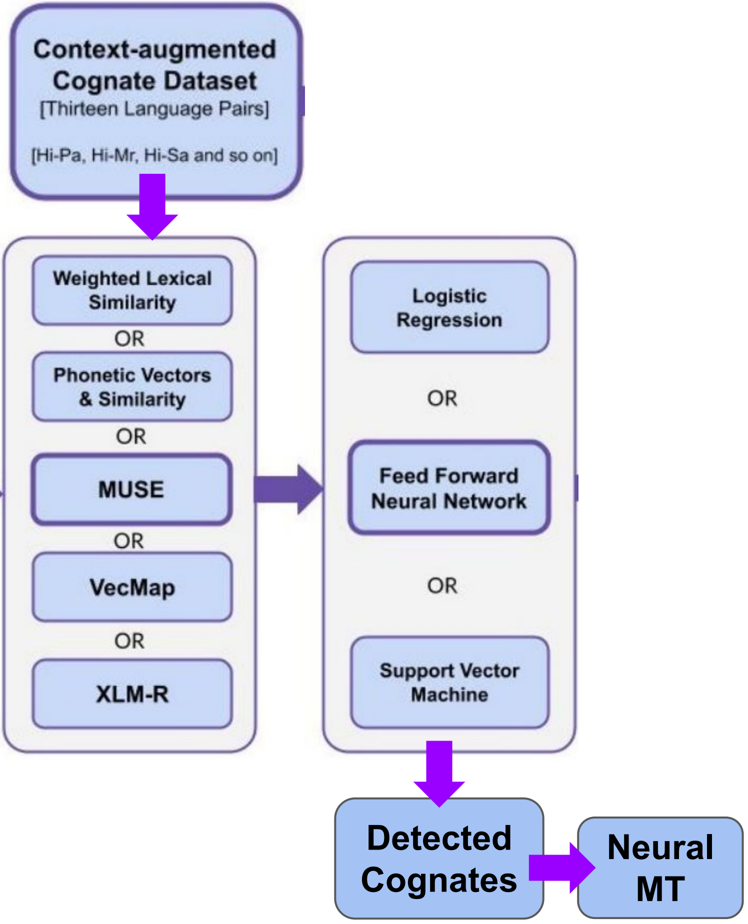
→ T
→ H
b
w



- ➔ **Average increase of 20.2% in BLEU** compared to word-level models.
- ➔ **Larger increase for** translation involving **Dravidian languages** (27.7% improvement)
- ➔ BPE-level models show 6.1% improvement over morpheme-level models.

[2003.08925] Utilizing Language Relatedness to improve Machine Translation: A Case Study on Languages of the Indian Subcontinent

Transfer Learning via Phylogeny [Cognates help MT]



[Harnessing Cross-lingual Features to Improve Cognate Detection for Low-resource Languages [COLING 2022]]

For Hi-Pa, improvement of **2.76 BLEU**; where **15001** cognates were detected. Consistent improvement for all the language pairs, even when **930** cognates (Hi-Te) are added, an improvement of **0.4 BLEU**.

Approaches / LP	Hi-Pa	Hi-Bn	Hi-Gu	Hi-Mr	Hi-Ta	Hi-Te	Hi-Ml
NMT-BPE Baseline	62.79	28.75	52.17	31.66	13.78	19.18	10.4
Cognate-aware NMT-BPE	**65.55**	**29.43**	**52.39**	**32.41**	**13.85**	**19.58**	**11.18**

	Baseline Approaches									Cross-lingual Embeddings based Approaches									Best Combination		
LP	WLS w/ FFNN			PVS w/ Siamese CNN (Rama, 2016)			WLS w/ RNN (Kanojia et al., 2019)			XLM-R w/ FFNN			MUSE w/ FFNN			VecMap w/ FFNN			MUSE + WLS w/ FFNN		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Hi-Bn	0.51	0.28	0.36	0.68	0.62	0.65	0.67	0.69	0.68	0.81	0.76	**0.78**	0.77	0.75	0.76	0.72	0.74	0.73	0.80	0.75	0.77
Hi-As	0.48	0.26	0.34	0.72	0.71	0.71	0.72	0.70	0.71	0.70	0.72	0.71	0.80	0.75	**0.77**	0.74	0.73	0.73	0.84	0.75	**0.79**
Hi-Or	0.51	0.30	0.38	0.65	0.58	0.61	0.66	0.58	0.62	0.65	0.61	0.63	0.72	0.68	**0.70**	0.67	0.70	0.68	0.81	0.69	**0.75**
Hi-Gu	0.43	0.16	0.23	0.70	0.65	0.67	0.81	0.71	0.76	0.80	0.73	0.76	0.80	0.84	**0.82**	0.77	0.74	0.75	0.83	0.85	**0.84**
Hi-Ne	0.50	0.16	0.24	0.72	0.84	0.78	0.78	0.73	0.75	0.75	0.75	0.75	0.86	0.83	**0.84**	0.78	0.73	0.75	0.86	0.83	**0.84**
Hi-Mr	0.51	0.20	0.29	0.70	0.68	0.69	0.74	0.70	0.72	0.76	0.71	**0.73**	0.70	0.73	0.71	0.71	0.71	0.71	0.72	0.73	0.72
Hi-Ko	0.47	0.24	0.32	0.63	0.63	0.63	0.63	0.59	0.61	0.66	0.58	0.62	0.69	0.73	**0.71**	0.61	0.60	0.60	0.70	0.75	**0.72**
Hi-Pa	0.28	0.17	0.21	0.51	0.44	0.47	0.76	0.72	0.74	0.75	0.71	0.73	0.83	0.78	**0.80**	0.71	0.74	0.72	0.83	0.78	**0.80**
Hi-Sa	0.34	0.19	0.24	0.55	0.51	0.53	0.73	0.71	0.72	0.75	0.70	0.72	0.77	0.76	**0.76**	0.73	0.71	0.72	0.80	0.77	**0.78**
Hi-Ml	0.49	0.20	0.28	0.59	0.66	0.62	0.66	0.66	0.66	0.72	0.63	0.67	0.76	0.71	**0.73**	0.69	0.71	0.70	0.77	0.71	**0.74**
Hi-Ta	0.22	0.19	0.20	0.49	0.58	0.53	0.49	0.58	0.53	0.63	0.51	0.56	0.72	0.68	**0.70**	0.66	0.72	0.69	0.72	0.70	**0.71**
Hi-Te	0.18	0.15	0.16	0.60	0.71	0.65	0.62	0.71	0.66	0.65	0.70	0.67	0.70	0.72	**0.71**	0.67	0.67	0.67	0.73	0.72	**0.72**
Hi-Kn	0.19	0.18	0.18	0.54	0.60	0.57	0.58	0.60	0.59	0.60	0.58	0.59	0.69	0.73	**0.71**	0.65	0.64	0.64	0.70	0.73	**0.71**

Transfer Learning via Phylogeny

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

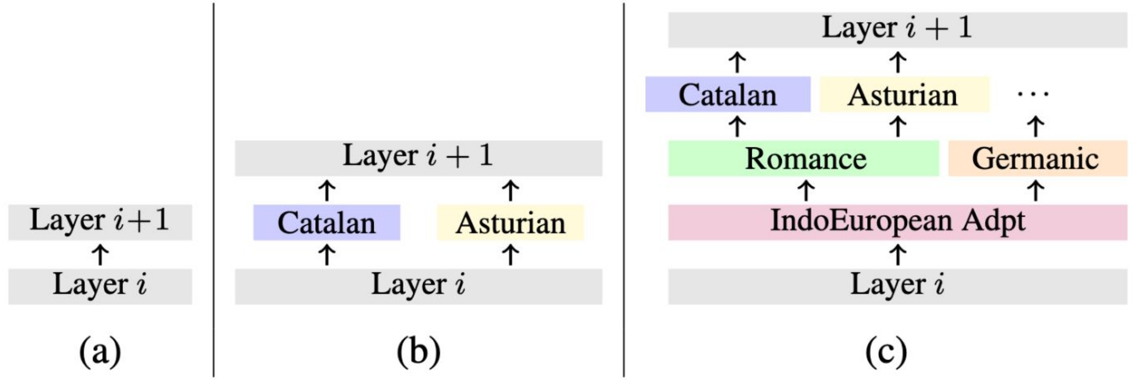


Figure 1: Incorporating phylogeny into neural models with adapters: starting with an unadapted model (a), current practice uses language-specific adapters between layers (b). We instead impose a phylogeny-informed tree hierarchy over adapters as in (c).

[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*.

Transfer Language Selection

- Data-dependent features
  - Data size, Type-Token Ratio, Word Overlap, and Subword Overlap
  - **Important for the MT task**
- Linguistic properties
  - Genetic, inventory, syntactic, and phonological distance
  - **Important for the linguistic tasks**
- Proposed
  - A language ranking model (LangRank)

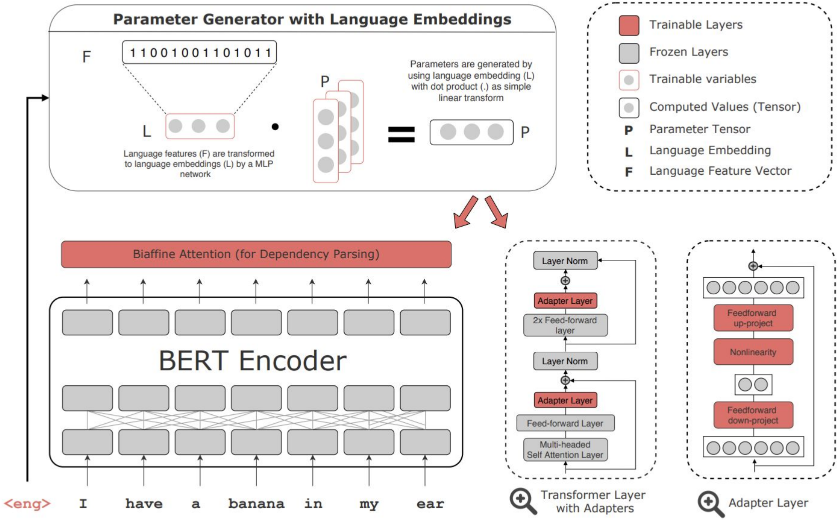
	Method	MT	EL	POS	DEP
dataset	word overlap o_w	28.6	30.7	13.4	52.3
	subword overlap o_{sw}	29.2	–	–	–
	size ratio s_{tf}/s_{tk}	3.7	0.3	9.5	24.8
	type-token ratio d_{ttr}	2.5	–	7.4	6.4
ling. distance	genetic d_{gen}	24.2	50.9	14.8	32.0
	syntactic d_{syn}	14.8	46.4	4.1	22.9
	featural d_{fea}	10.1	47.5	5.7	13.9
	phonological d_{pho}	3.0	4.0	9.8	43.4
	inventory d_{inv}	8.5	41.3	2.4	23.5
	geographic d_{geo}	15.1	49.5	15.7	46.4
	LangRank (all)	51.1	**63.0**	**28.9**	**65.0**
	LangRank (dataset)	**53.7**	17.0	26.5	**65.0**
	LangRank (URIEL)	32.6	58.1	16.6	59.6

Table 1: Our LangRank model leads to higher average NDCG@3 over the baselines on all four tasks: machine translation (MT), entity linking (EL), part-of-speech tagging (POS) and dependency parsing (DEP).

Yu-Hsiang Lin, et al., (2019). [Choosing Transfer Languages for Cross-Lingual Learning.]

Parameters Sharing via Linguistic Properties

- Linguistic properties embedding
- Parameter generator
    - Generate biaffine attention and adapter layers
    - based on linguistic properties of a language
- “benefits low resource languages without hurting high resource ones”



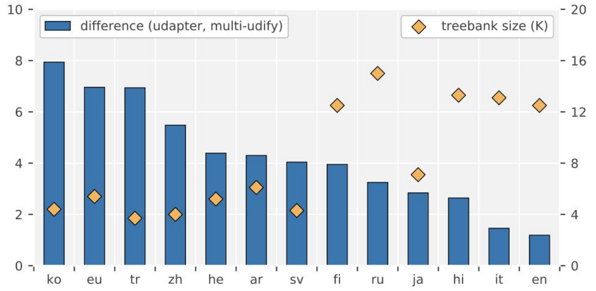
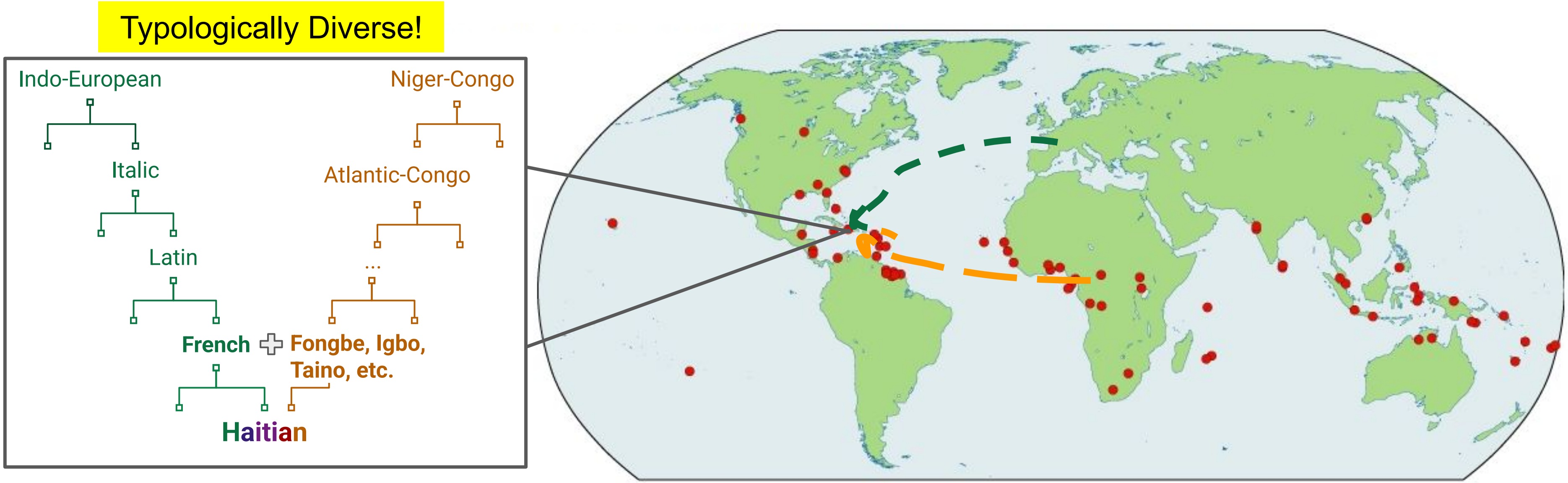


Figure 2: Difference in LAS between UDapter and multi-udify in the high-resource setting. Diamonds indicate the amount of sentences in the corresponding treebank.

Ahmet Üstün, et al., (2020). [UDapter: Language Adaptation for Truly Universal Dependency Parsing.]

Transfer Learning via Phylogeny: **Counter Arguments**

- Creoles are languages, found all around the world
- Arose from linguistic contact between **diverse, unrelated** languages
  - Creoles don't fit into traditional phylogenetic models of language.
  - Data from closely-related languages haven't been helpful for transfer learning for Creoles.



Lent, H., Bugliarello, E., & Søgaard, A. (2022). [Ancestor-to-Creole Transfer is Not a Walk in the Park](). *Proceedings of the Third Workshop on Insights from Negative Results in NLP.*

Transfer Learning via Phylogeny: **Counter Arguments**

- "**Successful transfer** often happens between **unrelated languages**"
- "We show that language written in **non-Latin and non-alphabetic scripts are the best choice** … in a diverse set of 30 low-resource languages"
  - E.g., Japanese is useful for Quechua.
- Proposed explanation: **Subword Evenness**

Target	Transfer	Avg PPL Change	Lang Family (WALS)
Arabic	Hebrew	+0.04	Afro-Asiatic
Burmese	Mandarin	+0.11	Sino-Tibetan
Chamorro	Indonesian	+7.15	Austronesian
	Tagalog	+16.45	
Fijian	Indonesian	+3.14	Austronesian
	Tagalog	+5.72	
Hausa	Hebrew	+1.2	Afro-Asiatic
Khalkha	Turkish	+9.23	Altaic
Malagasy	Indonesian	+0.17	Austronesian
	Tagalog	+0.63	
Oromo	Hebrew	+0.39	Afro-Asiatic

Table 5: Average change in perplexity (across 3 models), when using a genealogically close language instead of a language with low SuE (best PPL option among top 5 is chosen). Higher numbers mean worse performance.

Pelloni, O., Shaitarova, A., & Samardžić, T. (2022). Subword Evenness (SuE) as a Predictor of Cross-lingual Transfer to Low-resource Languages. *EMNLP*.

🔗 Emerging Common Themes

- Data collection can be difficult for low-resource languages/varieties due to unique sociolinguistic settings.

- Zero-shot alone is insufficient across low-resource contexts (hence this tutorial! ☺)
  - As it's unlikely to *solve* the data inequality problem, we need better data and better computational methods!
    - Critical: Getting the *most value* out of the *least data*?

- [something about language relatedness and phylogeny]
  - Are there some non-obvious relations?

- **What do *you* think are some other emerging common themes?**

Hands on Session

Evaluate zero-shot on a dialect dataset for sarcasm detection highlighting results, challenges, and pitfalls.

TODO: add acknowledgement for students :-)

Tutorial Agenda

Introduction	Dataset Creation	NLG
Emerging Connections	NLU	Conclusion