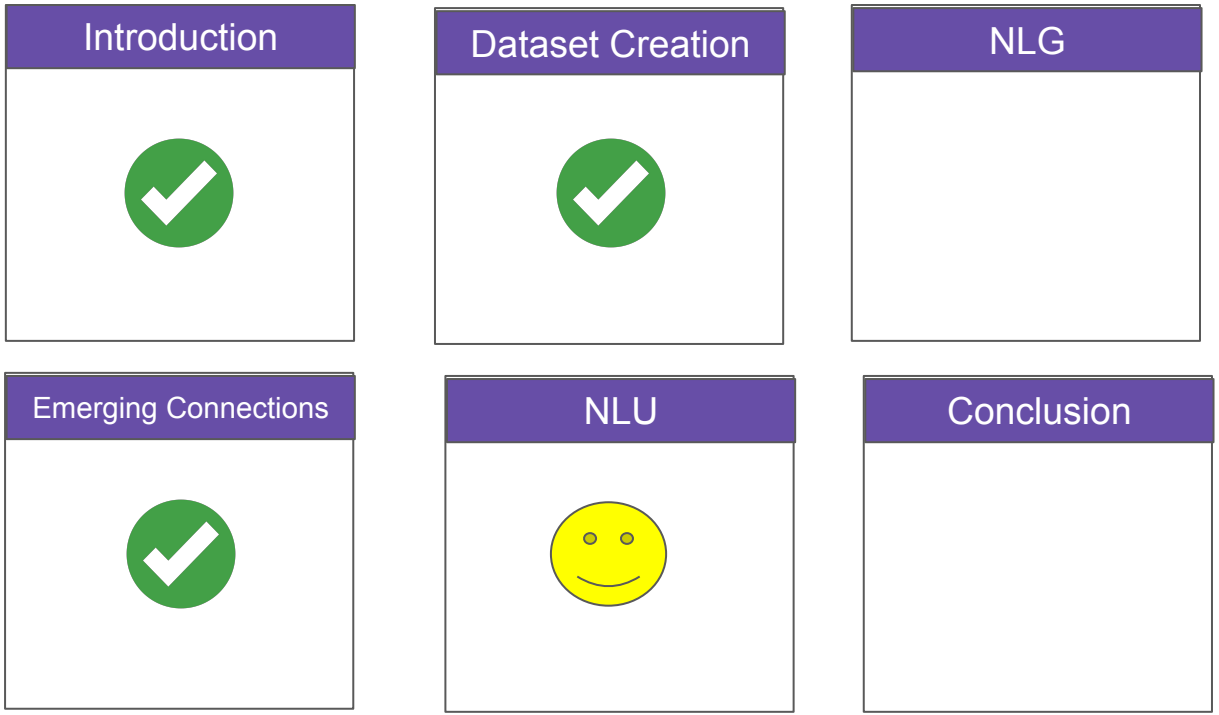
Connecting Ideas in ‘Lower-Resource’ Scenarios: NLP for National Varieties, Creoles and Other Low-resource Scenarios

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

Tutorial Agenda



Natural Language Understanding (NLU)

NLU typically involves extraction of implied or structured information from text, aiming to interpret their meaning in context and understanding intent. *e.g.,*

- ➔ Labels (Sentiment or Emotion Classification)
- ➔ Trees (Dependency Parsing, Constituency Parsing)
- ➔ Token spans (Named Entity Recognition, Abbreviation and Long-form detection)

Sequence Classification	Token Classification
Provide class label(s) to a sequence of tokens or words, typically a sentence, but can be a conversation, paragraph, or document.	Provide token-level or phrase-level labels to a sequence of words.

Natural Language Understanding (NLU) Tasks

Sequence Classification

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

Token Classification

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

Emotion Identification

"I am excited about this tutorial" (Label?)

"Data is the new oil" (Label?)

➔ Considerations for multi-**label** vs. multi-**class**

Abbreviation and Long-form Detection

"ECG reports show reduced pressure"

"Neural Networks are good at generalization but NN explainability is much needed" (Label for each token?)

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

NLU Tasks *vs.* NLP Layers

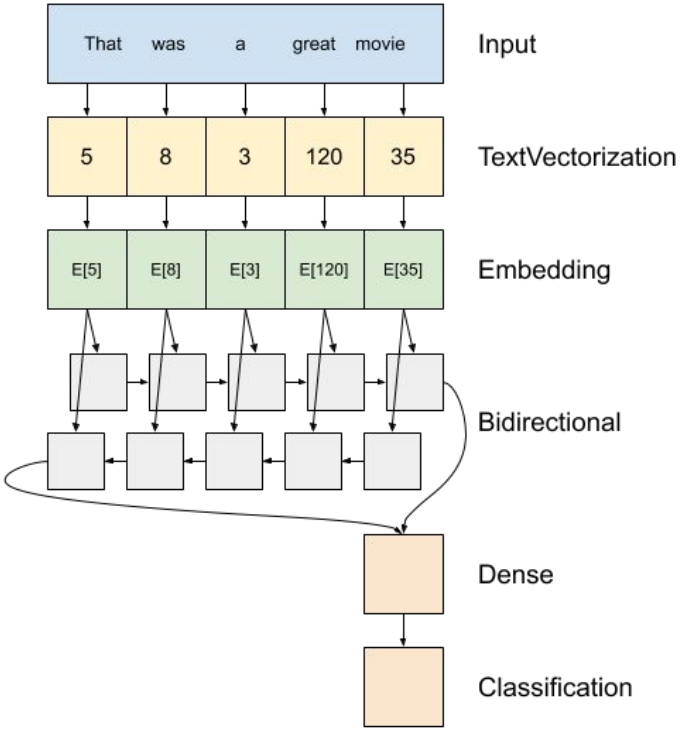
Syntax / Morphology / Semantics / Pragmatics / Discourse

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

Sequence Classification: Low-resource Setting

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

- ➔ Standard sequence classification methods heavily **rely on abundant labeled data, which is scarce** in low-resource scenarios.
  - ◆ Neural classifiers are data hungry but generalize better.
- ➔ Creating labeled datasets is **expensive** and **time-consuming**
  - ◆ For specialized domains, limited training examples may hinder model generalization
  - ◆ Availability of digitized data for less-studied languages.
- ➔ Variations **within** and **across** languages (dialects, code-mixing) pose challenges for generalized pre-trained models.
- ➔ Sequence or text classification requires **input text to be vectorized** or embedded as the **model learns to map these embeddings to a class** label.



Text classification with an RNN | TensorFlow

From Word Embeddings to Sentence Embeddings

For languages, varieties or Creoles which are not a part of any Transformer model or LLM

➔ Use **word2vec and attention within a sequence model** to obtain sentence embeddings

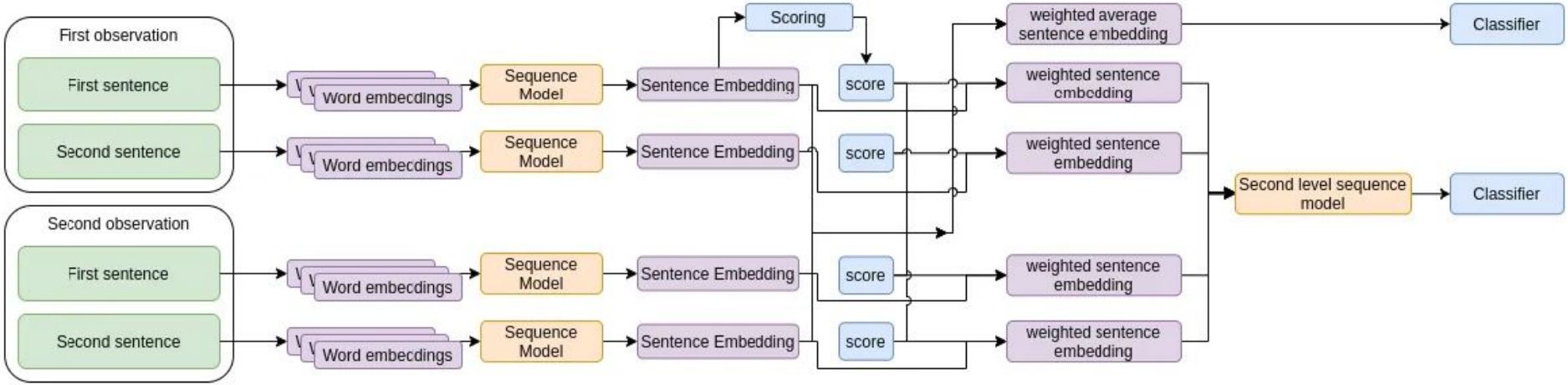


Image Source: Identifying relevant text sections for long sequence classification. | by John Crousse | Medium

Deep Learning for Sequence Classification

- ➔ Traditional word-level embedding models lose out on **word order** and **context**.
- ➔ **Deep learning helps automatically learn complex features** from raw text.
- ➔ **Attention (Hadamard product) based models with sequence order work well** for relatively non-complex classification.

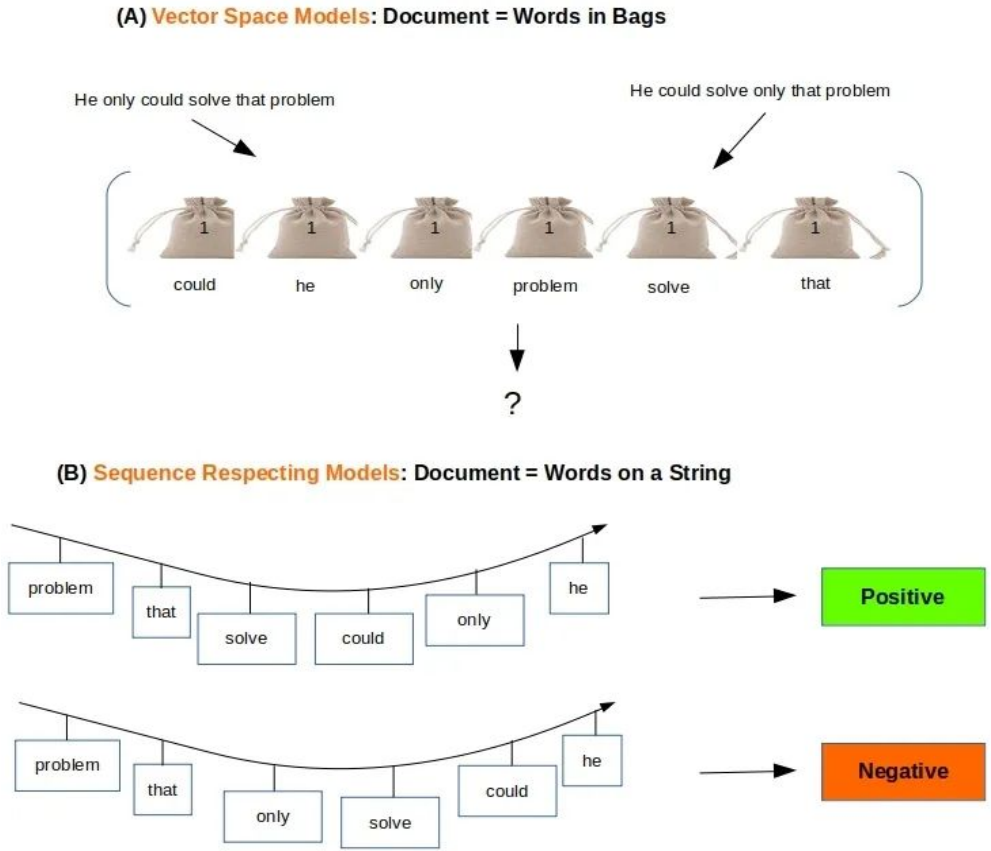
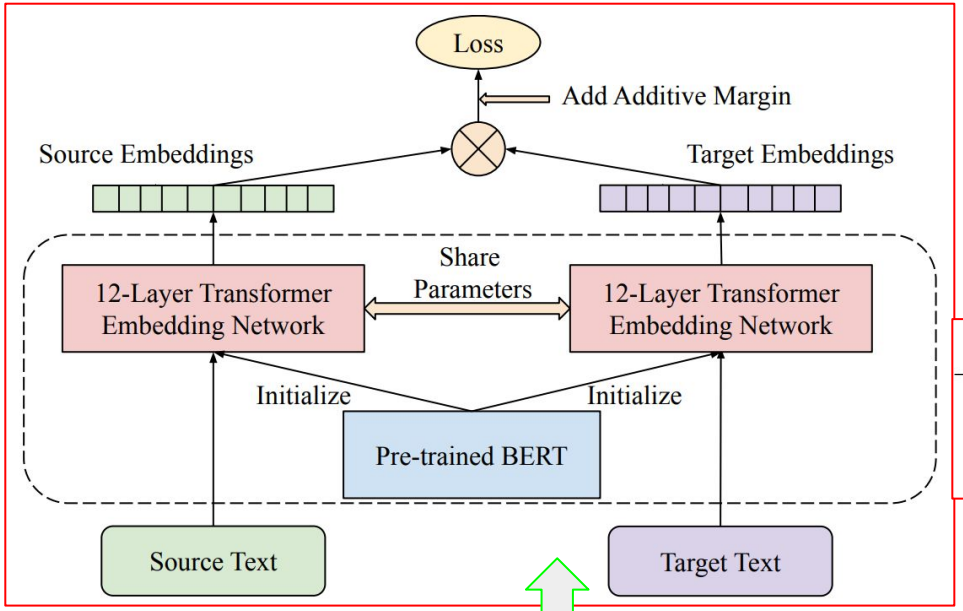


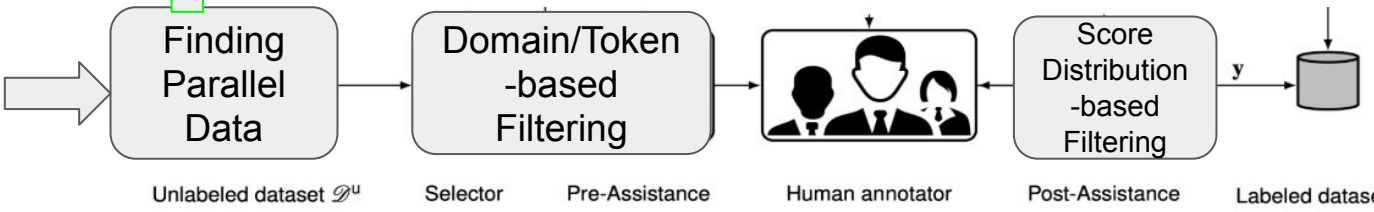
Image Source:W ord Bags vs Word Sequences for Text Classification | by Ashok Chilakapati | Towards Data Science

Language-agnostic BERT Sentence Embedding (LaBSE)



Model	14 Langs	36 Langs	82 Langs	All Langs
m-USE$_{\text{Trans.}}$	93.9	–	–	–
LASER	**95.3**	84.4	75.9	65.5
LaBSE	**95.3**	**95.0**	**87.3**	**83.7**

QE for MT:
Data Annotation



Transformers-based Encoders | Recent updates

Model	Number of parameters	Dataset name	Dataset size	Number of training tokens
~~XLM-R_{Large}~~	~~550M~~	~~CC100~~	~~167B~~	6T
XLM-R_{XL}	3.5B	CC100	167B	0.5T
XLM-R_{XXL}	10.7B	CC100	167B	0.5T
mt5-XL	3.7B	mC4	6.4T	1T
mt5-XXL	13B	mC4	6.4T	1T

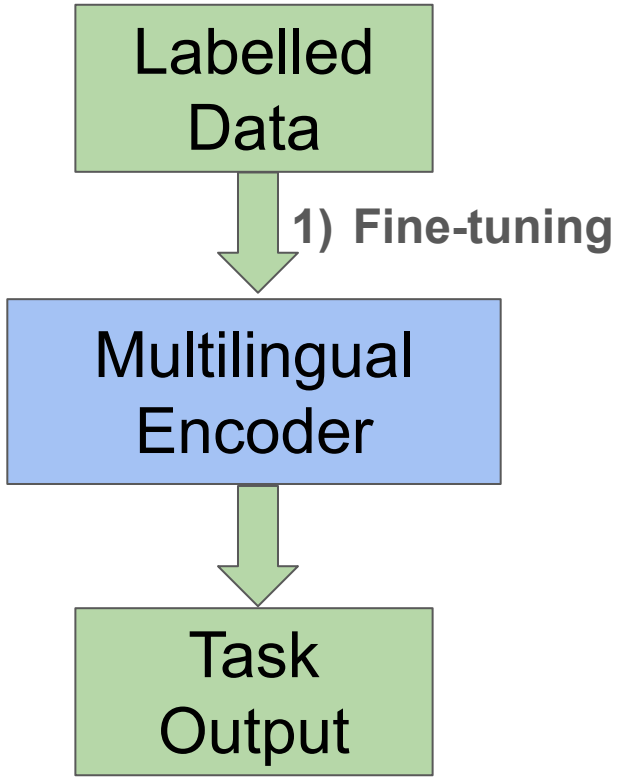
	nDCG@10	R@100
BM25	39.3	78.7
mDPR	41.5	78.8
$mE5_{small}$	60.8	92.4
$mE5_{base}$	62.3	93.1
$mE5_{large}$	**66.5**	94.3
$mE5_{large\text{-}instruct}$	65.7	**94.6**

	# Sampled
Wikipedia	150M
mC4	160M
Multilingual CC News	160M
NLLB	160M
Reddit	160M
S2ORC	50M
Stackexchange	50M
xP3	80M
Misc. SBERT Data	10M
Total	~1B

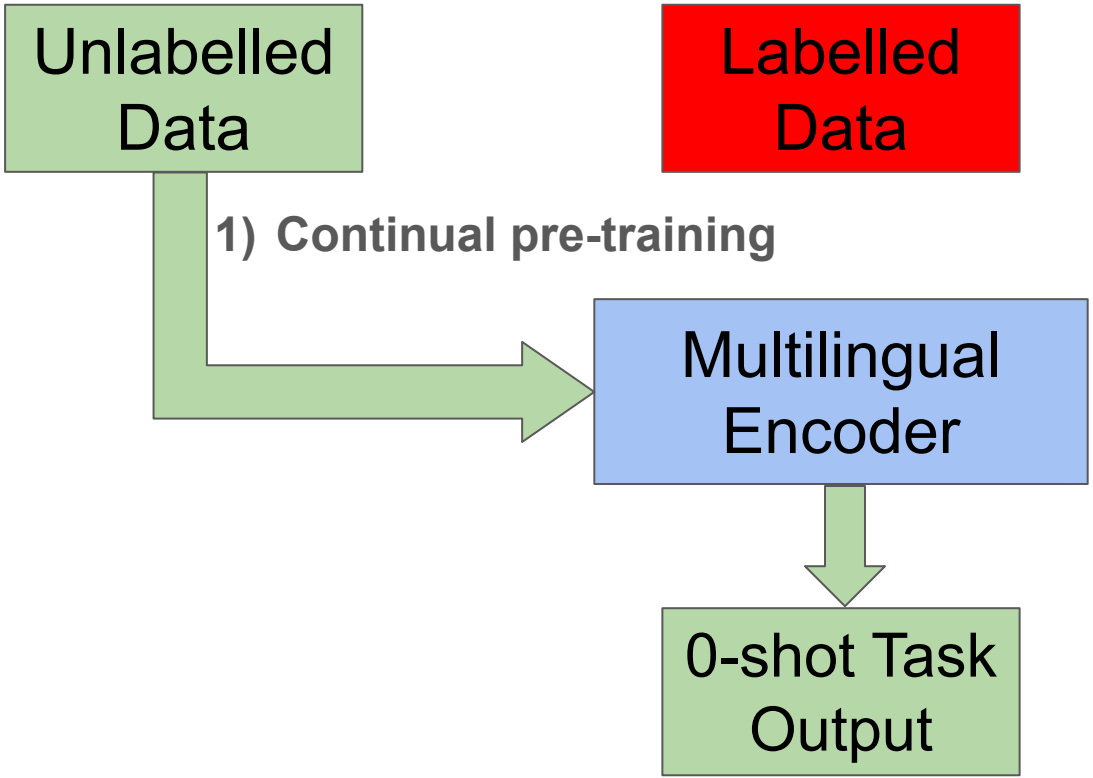
	# Sampled
MS-MARCO Passage	500k
MS-MARCO Document	70k
NQ, TriviaQA, SQuAD	220k
NLI	275k
ELI5	100k
NLLB	100k
DuReader Retrieval	86k
Fever	70k
HotpotQA	70k
Quora Duplicate Questions	15k
Mr. TyDi	50k
MIRACL	40k
Total	~1.6M

	MTEB (56 datasets)
LaBSE	45.2
$Cohere_{multilingual\text{-}v3}$	64.0
$BGE_{large\text{-}en\text{-}v1.5}$	64.2
$mE5_{small}$	57.9
$mE5_{base}$	59.5
$mE5_{large}$	61.5
$mE5_{large\text{-}instruct}$	**64.4**

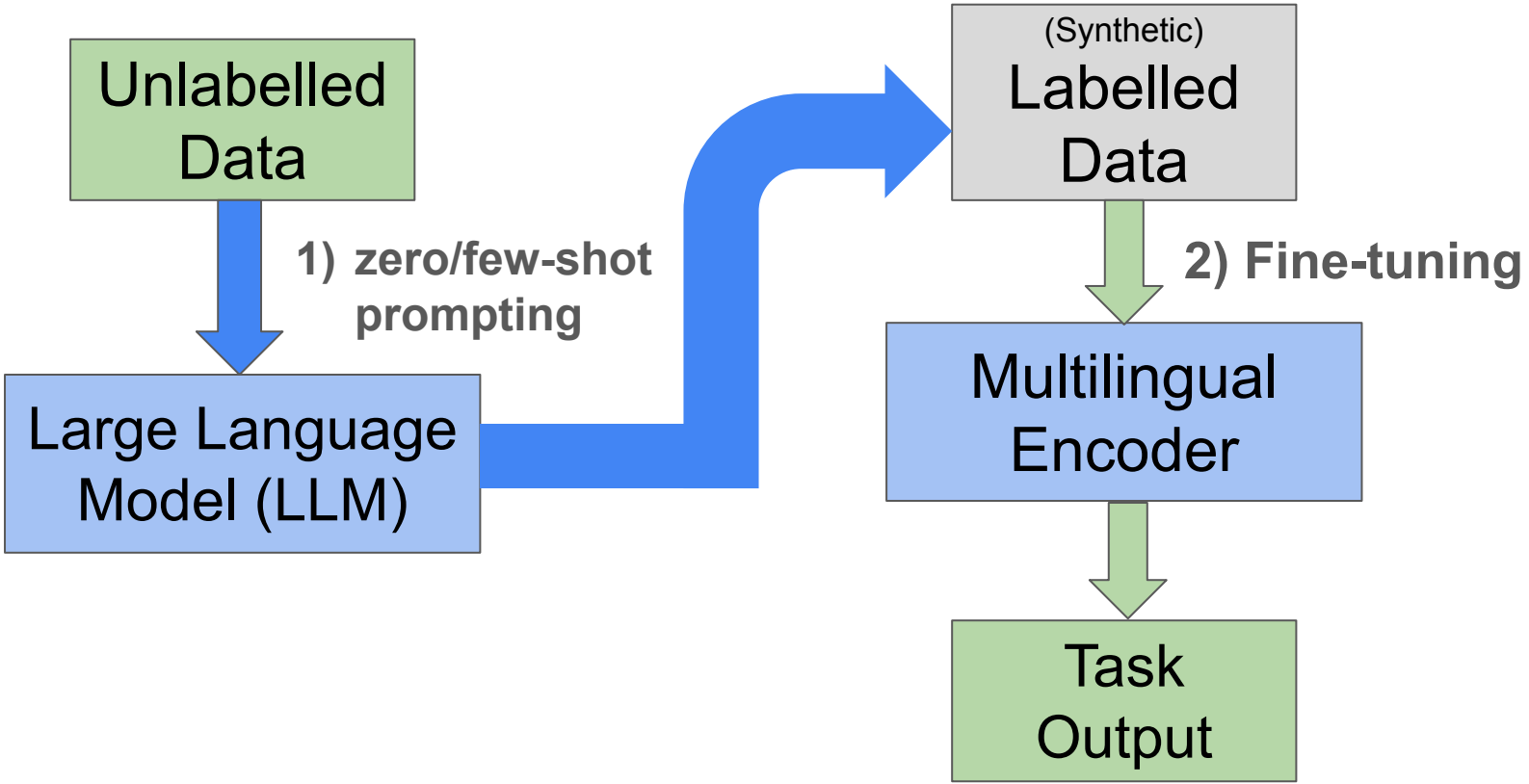
Transfer Learning | Low-resource Scenarios | SFT



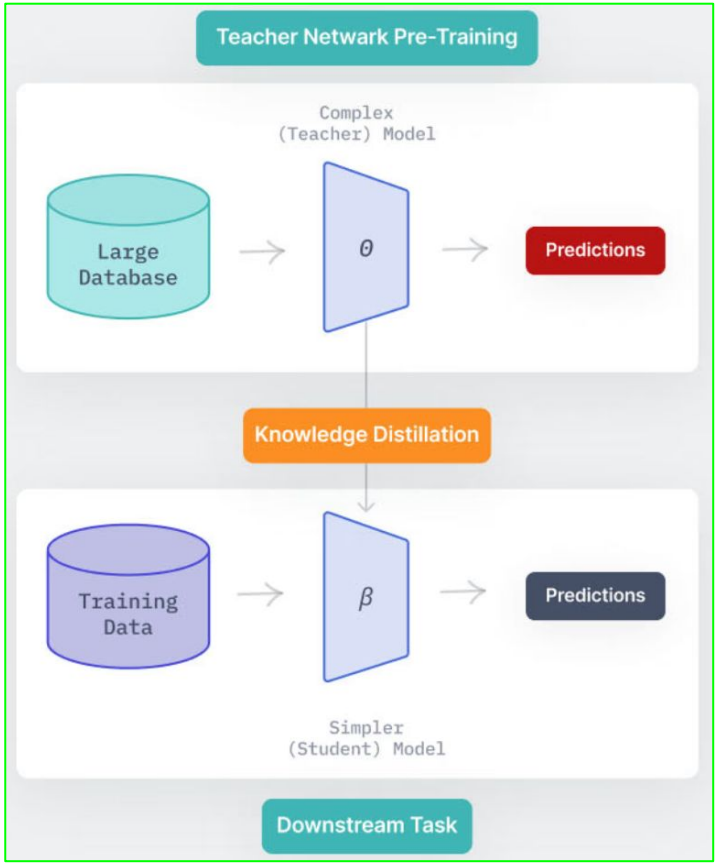
Transfer Learning | Low-resource Scenarios | SFT

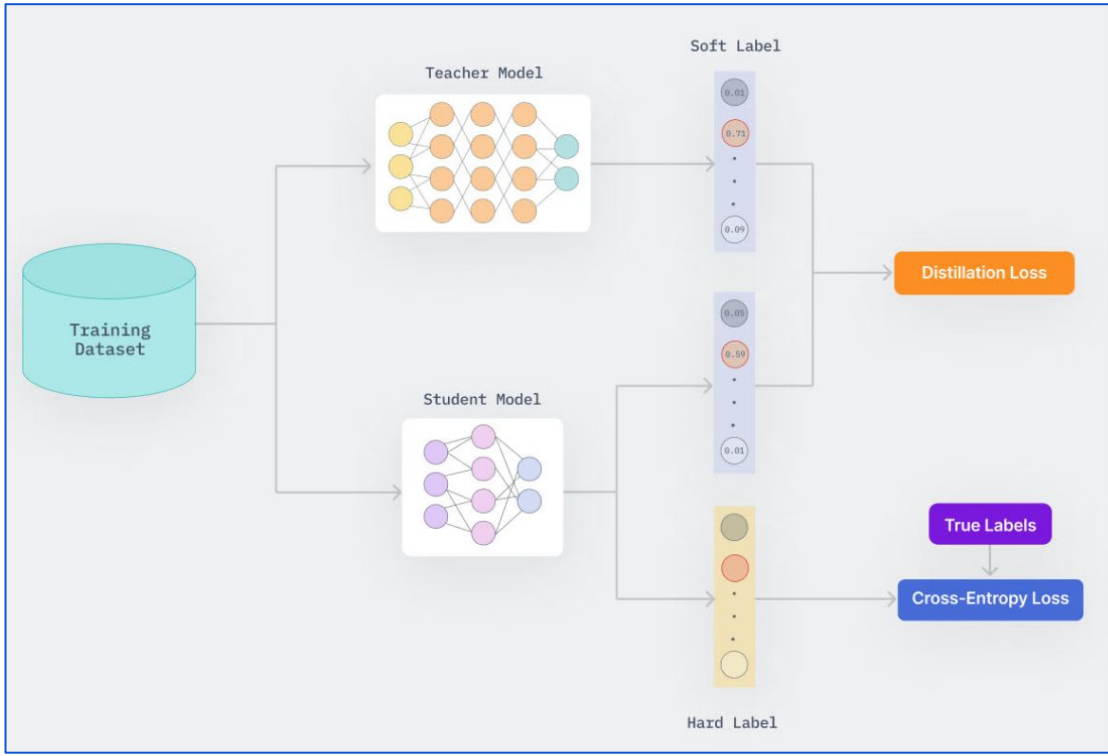


Transfer Learning | Low-resource Scenarios | SFT



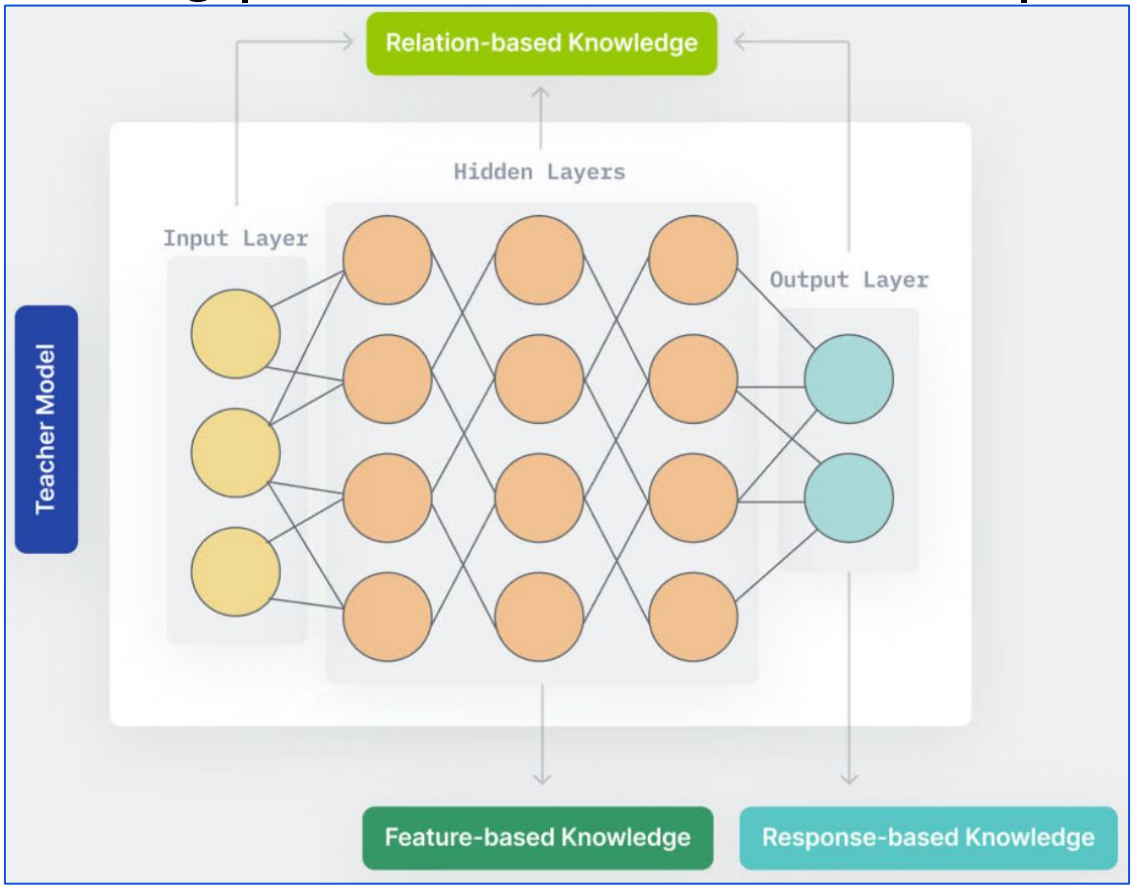
Transfer Learning | Low-resource Scenarios | SFT





V7Labs: Knowledge Distillation: Principles & Algorithms [+Applications]

Transfer Learning | Low-resource Scenarios | SFT



V7Labs: Knowledge Distillation: Principles & Algorithms [+Applications]

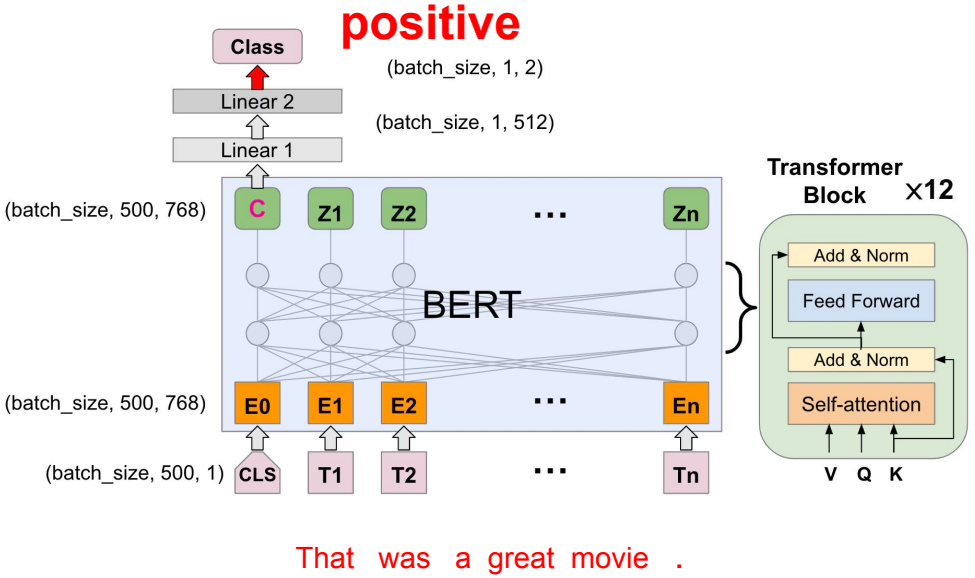
Autoencoders *vs.* LLMs | Zero-shot Classification

Prompt:

```
Classify the text into neutral, negative or positive.
Text: I think the food was okay.
Sentiment:
```

Output:

```
Neutral
```



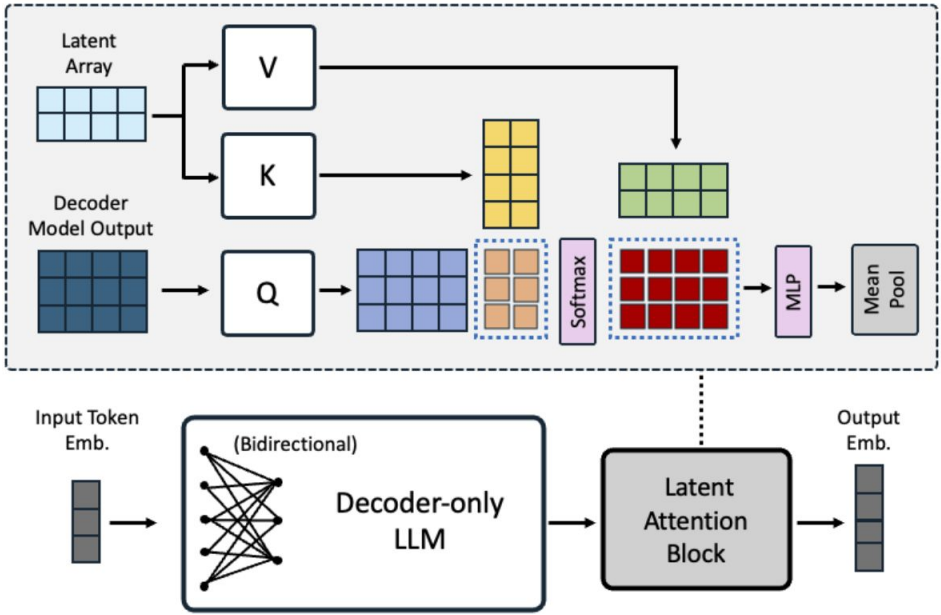
Autoencoders to Generalist Embedding Models

NV-Embed-v2 is a generalist embedding model performing well on the Massive Text Embedding Benchmark (MTEB benchmark).

➔ Each query needs to be accompanied by an corresponding instruction describing the task.

[MTEB Leaderboard - a Hugging Face Space by mteb]

[Multilingual models for inference]



Few-shot or In-context Learning w/ LLMs

Traditional Supervised Learning vs. Few-Shot/In-Context Learning

- **Few-Shot/In-Context Learning** learns from a few examples (demonstrations) provided in the input prompt *without* updating the model's weights.
  - Enables rapid adaptation to new tasks.

Task: Classify the sentiment of these movie reviews as positive or negative.
Review: "This movie was incredibly moving. The acting was superb, and the story was captivating."
Sentiment: Positive

Review: "I found this film to be quite boring and predictable. The plot was weak."
Sentiment: Negative

Review: "The special effects were amazing, and the cinematography was stunning. A visual masterpiece!"
Sentiment: Positive

Review: "The dialogue was terrible, and the characters were one-dimensional."
Sentiment: **<LLM PREDICTS SENTIMENT HERE>**

QE for MT with LLMs :: Zero-shot *vs.* ICL

LP	Template	Gemma-7B	Llama-2-7B	Llama-2-13B	OC-3.5-7B
En-Gu	0-shot-GEMBA	0.113	0.006	0.019	0.249*
	0-shot-TE	-0.102†	-0.008	-0.052	0.117†
	0-shot-AG	-0.079	-0.007	0.008	0.164†
	3-ICL-AG	-0.005	0.036	-0.036	0.223
	5-ICL-AG	0.023	-0.008	0.095	0.151
	7-ICL-AG	0.071	-0.053	-0.108	**0.260**
En-Te	0-shot-GEMBA	0.081	-0.016	0.121†	0.145*
	0-shot-TE	0.018	0.013	0.010	0.072
	0-shot-AG	0.065	0.083	0.045	0.121†
	3-ICL-AG	0.092	0.027	0.015	0.152
	5-ICL-AG	0.021	0.051	0.073	0.126
	7-ICL-AG	-0.033	0.021	-0.028	**0.196**
Et-En	0-shot-GEMBA	0.289	0.168	0.185	0.571
	0-shot-TE	0.086	0.100	0.146	0.455
	0-shot-AG	0.098	0.064	0.319	0.619*
	3-ICL-AG	0.226	0.268	-0.058	0.613
	5-ICL-AG	0.327	0.269	0.438	**0.636**
	7-ICL-AG	0.306	0.033	0.169	0.616

When LLMs Struggle: Reference-less Translation Evaluation for Low-resource Languages (Sindhujan et al., 2025)

Fine-tuning LLMs with Instructions

Sindhujan et al. (2025) **instruction fine-tuned LLMs** for QE regression task.

Instruction fine-tuning contained a **prompt based on annotation guidelines, which helped improve existing LLM performance compared to 0-shot and ICL**.

We need to Evaluate the machine translated sentences of **<Source language>**(Source) to **<Target language>** (Translation), with quality scores ranging from 0 to 100.

Source: <Source Sentence>

Translation: <Translated Sentence>

Scores of 0-30 indicate that the translation is mostly unintelligible, either completely inaccurate or containing only some keywords. Scores of 31-50 suggest partial intelligibility, with some keywords present but numerous grammatical errors. A score between 51-70 means the translation is generally clear, with most keywords included and only minor grammatical errors. Scores of 71-90 indicate the translation is clear and intelligible, with all keywords present and only minor non-grammatical issues. Finally, scores of 91-100 reflect a perfect or near-perfect translation, accurately conveying the source meaning without errors.

The evaluation criteria focus on two main aspects: Adequacy (how much information is conveyed) and Fluency (how grammatically correct the translation is). Predict the quality score in the range of 0 to 100 considering the above instructions. Predict only the score, no need for explanation.

Lang-pair	Gemma-7B	Llama-2-7B	Llama-2-13B	OC-3.5-7B	TransQuest
En-Gu	0.440	0.214	0.421	0.520	**0.653**
En-Hi	0.375	0.282	0.336	**0.474**	0.119
En-Mr	0.557	0.509†	0.501	0.554†	**0.629**
En-Ta	0.475	0.375	0.441	**0.509**	0.303
En-Te	0.217	0.263	0.261	**0.271**	0.087
Et-En	0.648	0.589	0.598	0.652	**0.806**
Ne-En	0.612	0.497	0.543†	0.614	**0.746**
Si-En	0.387	0.332	0.346	0.441	**0.581**

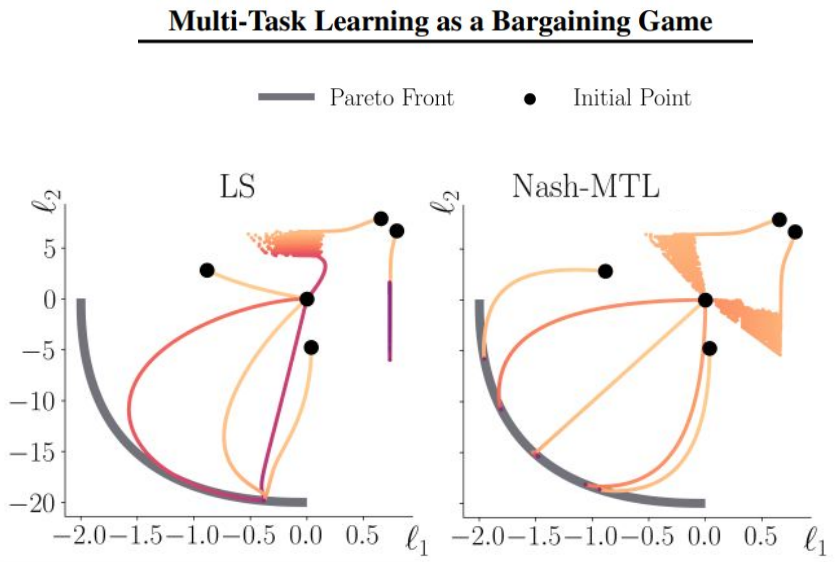
The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities

Multitask Learning

➔ Consider using **different tasks from the same lower-resource scenario**, but for the same language/language pair.

https://arxiv.org/pdf/2202.01017, and https://aclanthology.org/2023.findings-acl.585.pdf

Suggested Read: Multi-task Active Learning for Pre-trained Transformer-based Models | TACL



LP	Word-Level					Sentence-Level				
	STL	LS-MTL	+/- %	Nash-MTL	+/- %	STL	LS-MTL	+/- %	Nash-MTL	+/- %
En-Mr	0.3930	0.4194	2.64%	**0.4662**	7.32%	0.5215	0.5563	3.48%	**0.5608**	3.93%
Ne-En	0.4852	0.5383	5.31%	**0.5435**	5.83%	0.7702	0.7921	2.19%	**0.8005**	3.03%
Si-En	0.6216	0.6556	3.40%	**0.6946**	7.30%	0.6402	0.6533	1.31%	**0.6791**	3.89%
Et-En	0.4254	0.4971	7.17%	**0.5100**	8.46%	0.7646	0.7905	2.59%	**0.7943**	2.97%
Ro-En	0.4446	0.4910	4.64%	**0.5273**	8.27%	0.8952	**0.8985***	0.33%	0.8960*	0.08%
Ru-En	0.3928	0.4208	2.80%	**0.4394**	4.66%	0.7864	0.7994	1.30%	**0.8000**	1.36%
En-De	0.3996	0.4245	2.49%	**0.4467**	4.71%	0.4005	0.4310	3.05%	**0.4433**	4.28%

Evaluation Metrics *vs.* Human Evaluation

Sequence Classification

- **Accuracy, Precision, Recall, and F1-score, AUC-ROC**
- **Macro-F1 is** averaged across all classes; Important for imbalanced datasets.

Token Classification

- **Per-token Accuracy and Precision, Recall, F1-score:** Calculated for each entity type (e.g., Person, Organization, Location in NER, abbreviation and long-form in Abbreviation Detection) and then averaged

Direct Assessment: Rating the quality of outputs on a Likert scale.

Relative Ranking: Comparing outputs from different models and ranking them.

Pairwise Testing: Presenting users with two different outputs and asking them to choose the better one.

Correlation with Automatic Metrics: Analyze the correlation between human judgments and automatic evaluation scores. This can help determine the validity of automatic metrics.

Regression Task (*e.g.*, QE for MT, sentiment valence/intensity prediction, . . .)

Spearman's Rank Correlation measures the *monotonic* relationship (whether the rankings are consistent) between predicted and human scores. *Useful when the relationship might not be linear.*

Kendall's Tau, a correlation coefficient, measures *ordinal* association between two measured quantities; more robust to outliers.

Pearson Correlation measures the *linear correlation* between predicted and human quality judgments.

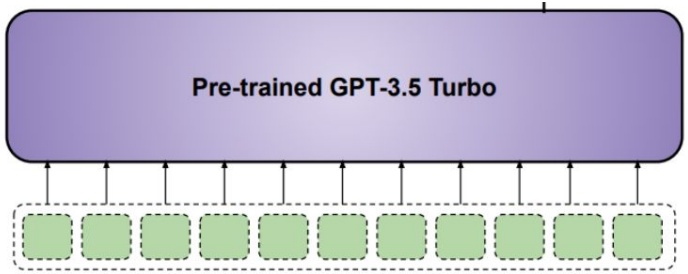
F1-score (for error span identification at word level) can measures the accuracy of identifying error spans.

NLU Case Study: Low-resource + Code-mixed

Sequence Classification :: Sarcasm Detection

- ➔ Provide sarcastic or non-sarcastic class label.
- ➔ **Tamil-English** and **Malayalam-English** code-mixed.

- ➔ Zero-shot with GPT-3.5T macro avgs.!
- ➔ Fine-tuned models perform better but hit a skyline for F1.



	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

Model	Acc	M-F1	F1(S)	P(S)	R(S)	ROC-AUC
MURIL	**0.781**	**0.743**	**0.644**	**0.571**	**0.738**	**0.767**
m-BERT	0.776	0.737	0.637	0.563	0.733	0.762

Data	Class		
	Non sarcastic	Sarcastic	Total
Train	19,866	7,170	27,036
Validation	4,939	1,820	6,759
Test	6,186	2,263	8,449
Total	30,991	11,253	*42,244*

Class		
Non sarcastic	Sarcastic	Total
9,798	2,259	12,057
2,427	588	3,015
3,083	685	3,768
15,308	3,532	*18,840*

Model	Acc	M-F1	F1(S)	P(S)	R(S)	ROC-AUC
MURIL	**0.850**	**0.731**	**0.553**	**0.604**	0.510	0.718
m-BERT	0.813	0.709	0.536	0.489	**0.594**	**0.728**

Overview of Sarcasm Identification of Dravidian Languages in DravidianCodeMix@FIRE-2023
Sarcasm Detection in Dravidian Code-Mixed Text Using Transformer-Based Models
[2411.05039] YouTube Comments Decoded: Leveraging LLMs for Low Resource Language Classification
Findings of Shared task on Sarcasm Identification in Code-Mixed Dravidian Languages

NLU Case Study: **Native Samples** + **Mid-resource** + Code-mixed

Sequence Classification :: Sarcasm Detection [2412.12761]

Adding samples from the native language to code-mixed data

➔ Best F1 achieved by adding English or English + Hindi data samples to code-mixed data.

➔ Significant improvement observed on adding samples anyways

Sarcasm

NLD →		NHD			iSarc			SC-V2		
Model ↓	CM	CM+Hi+En	CM+En	CM+Hi	CM+Hi+En	CM+En	CM+Hi	CM+Hi+En	CM+En	CM+Hi
NB	(0.74)	0.35	0.37	0.41	0.38	0.41	0.42$^{\#}$	0.32	0.34	0.37
RF	(0.69)	0.43	0.51	0.60$^{\#}$	0.51	0.57	0.57	0.49	0.60$^{\#}$	0.58
SVM	(0.74)	0.59	0.59	0.73$^{\#}$	0.64	0.64	0.71	0.68	0.68	0.73$^{\#}$
mBERT	0.80	0.83$^{*\#}$	0.79	0.82*	0.79	0.81	0.78	0.78	0.82	(0.84*)
XLM-R	0.81$^{\#}$	(0.83*)	0.79	(0.83*)	0.81$^{\#}$	0.81$^{\#}$	0.81$^{\#}$	0.80	(0.83)	0.81$^{\#}$
MuRIL	**0.83**	**0.86***	(**0.89***)	0.82	**0.84**	0.85*	**0.84**	**0.87**$^{*\#}$	**0.84**	**0.87**$^{*\#}$
IndicBERT	0.81	**0.86**$^{*\#}$	0.86$^{*\#}$	**0.85***	0.81	(**0.88***)	0.83	0.83	0.83	0.82

NLU Case Study: **Multi-task** + **Mid-resource** + Code-mixed

Sequence Classification :: Sarcasm Detection

[2412.12761]

Multitask learning on with Sarcasm as primary task, where humor and hate classification are auxiliary tasks

- ➔ Best F1 achieved when 3 tasks are combined with the multi-task gating mechanism
- ➔ Most multi-tasking results show an improvement.

Sarcasm

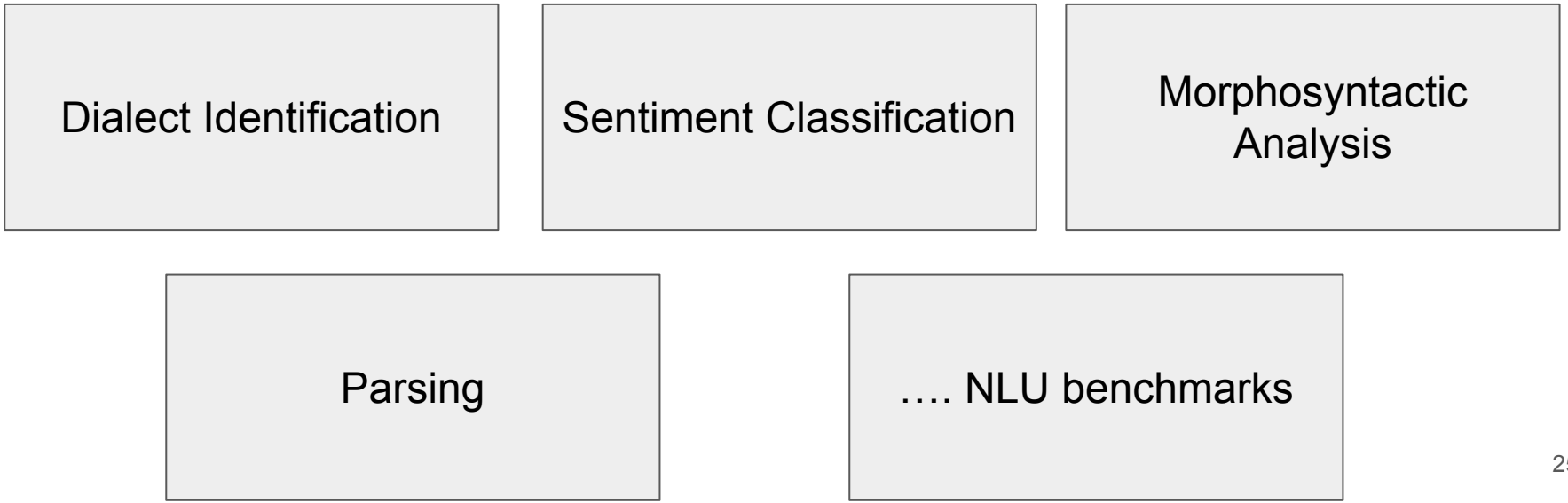
NLD : *NHD*		mBERT$_{MTL}$		XLM-R$_{MTL}$		MuRIL$_{MTL}$	
Hate	Humor	Gate	w/o Gate	Gate	w/o Gate	Gate	w/o Gate
☐	☑$_{Col}$	0.82	0.83	**0.85***	0.82	0.83	0.78
☐	☑$_{POTD}$	0.84	0.83	**0.85***	0.82	0.82	0.76
☐	☑$_{HaHa}$	0.84	0.83	**0.85***	0.82	0.84	0.78
☐	☑$_{16000}$	0.81	0.83	**0.90**$^{*\#}$	0.82	0.86*	0.79
☑	☐	0.81	0.85$^{*\#}$	**0.86***	0.82	**0.86**	0.78
☑	☑$_{Col}$	0.83*	(0.86*)	0.84*	**0.88**$^{*\#}$	0.86*	(0.84)
☑	☑$_{POTD}$	0.84*	0.81	**0.88***	0.87*	0.84	0.82
☑	☑$_{HaHa}$	0.84*	0.83*	**0.88***	0.87*	0.81	0.81
☑	☑$_{16000}$	0.85$^{*\#}$	0.85$^{*\#}$	**0.88***	0.82	0.87$^{*\#}$	0.81

NLD : *SC-V2*							
☐	☑$_{Col}$	**0.84***	**0.84***	0.83	0.83	0.81	0.83$^{\#}$
☐	☑$_{POTD}$	0.82	0.82	0.84*	0.83	**0.85**	0.83$^{\#}$
☐	☑$_{HaHa}$	**0.85**$^{*\#}$	0.83*	0.82	0.83	**0.85**	0.83$^{\#}$
☐	☑$_{16000}$	0.81	0.84*	**0.85***	0.83	0.84	0.83$^{\#}$
☑	☐	0.83*	0.84*	**0.88***	0.83	0.85	0.79
☑	☑$_{Col}$	0.80	0.82	**0.89***	0.87*	0.84	0.81
☑	☑$_{POTD}$	0.83*	0.83*	**0.89***	0.85*	0.86*	(0.84)
☑	☑$_{HaHa}$	(0.86*)	0.84*	0.85*	(**0.89***)	0.87$^{*\#}$	0.79
☑	☑$_{16000}$	0.83*	0.81	(**0.91***)	0.86*	0.85	0.81

Typical tasks within Dialectal NLP

Existing Dialectal NLP is quite focused Dialect Identification and is well explored in terms of NLU compared to NLG.

Suggested Read: [2401.05632] Natural Language Processing for Dialects of a Language: A Survey (Accepted at ACM CSUR)



Dialect Identification

Prediction of the dialect of input text

Long history (Chitturi et al. 2008)

Several shared tasks (VarDial; ArabicNLP have been leading the tasks)

Shared Task	Dialects/Languages
(Zampieri et al., 2014)	Brazilian Portuguese and European Portuguese; American and British English; and Argentinian Spanish and Castilian Spanish
(Zampieri et al., 2015)	American and British English; and Argentinian Spanish and Castilian Spanish
(Malmasi et al., 2016)	Dialects of English, Spanish, French and Arabic
(Zampieri et al., 2019)	Dialects of German, Chinese, Romanian
(Gaman et al., 2020)	Dialects of Romanian, Geolocation-based Varieties
(Aepli et al., 2022)	Dialects of French and Italian
(Aepli et al., 2023)	Dialects of Indo-European and Ural languages (and other tasks)
(Abdul-Mageed et al., 2023)	Dialects of Arabic

Table 3. Shared tasks related to dialect identification.

Rahul Chitturi and John Hansen. 2008.Dialect Classification for Online Podcasts Fusing Acoustic and Language Based Structural and Semantic Information. In ACL

Examples of dialectal datasets

Italian Alan Ramponi and Camilla Casula. 2023a.DIATOPIT: A Corpus of Social Media Posts for the Study of Diatopic Language Variation in Italy. In VarDial

Norwegian Jeremy Barnes, Petter Mæhlum, and Samia Touileb. 2021.NorDial: A Preliminary Corpus of Written Norwegian Dialect Use. In 23rd Nordic Conference on Computational Linguistics (NoDaLiDa).

Indonesia Alham Fikri Aji,. 2022.One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. In ACL. Dublin, Ireland

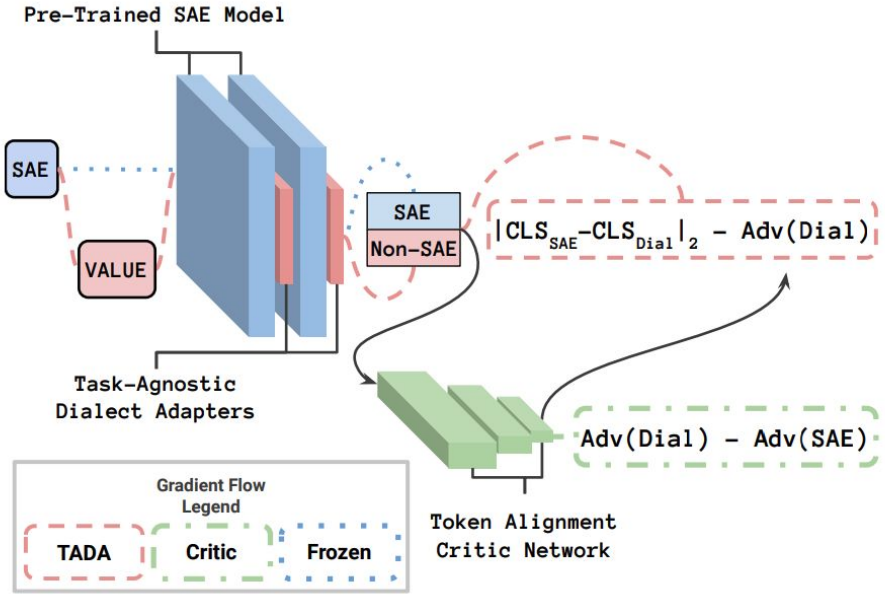
Arabic Bashar Talafha et al, 2024.Casablanca: Data and Models for Multidialectal Arabic Speech Recognition. In EMNLP

Typical Approaches

Statistical: Phonological and linguistic features (e.g., verb constructions)

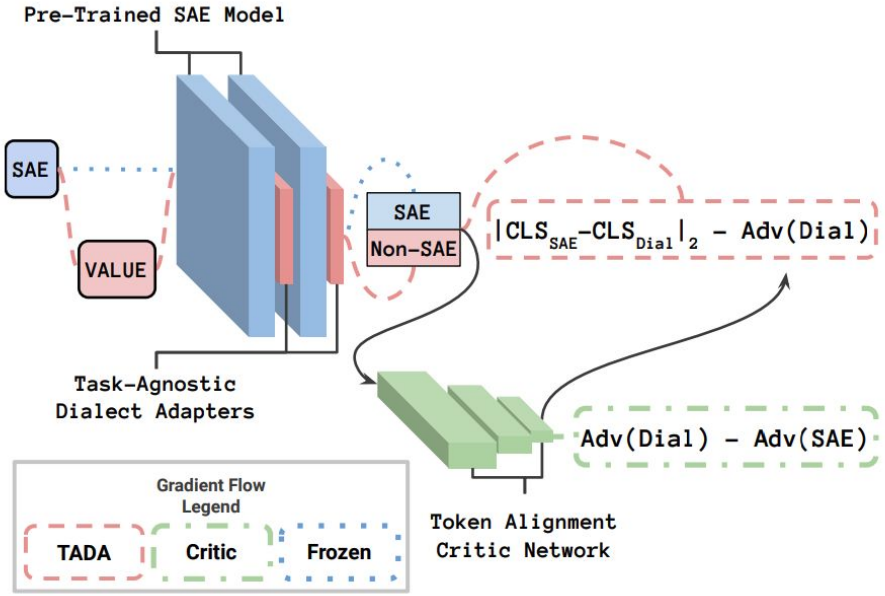
Fine-tuning BERT, etc. (Ramponi et al; 2023b)

LoRA-based dialect adapter



William Held, Caleb Ziems, and Diyi Yang. 2023.TADA : Task Agnostic Dialect Adapters for English. In Findings of ACL

LoRA-based dialect adapter



Dialect Adaptation Details				AAE Glue Performance							
Approach	Method	Task-Agnostic	Dialect Params.	COLA	MNLI	QNLI	RTE	QQP	SST2	STS-B	Mean
N/A	Finetuning	✓	0	13.5	82.0	89.3	71.8	87.1	92.0	89.9	75.1
N/A	Adapters	✓	0	14.1	83.7	90.3	67.1	86.8	92.1	88.7	74.7
VALUE	Finetuning	✗	$T \times 110M$	19.8	84.9	90.8	74.4	89.6	92.4	90.9	77.5
VALUE	Adapters	✗	$T \times 895K$	40.2	85.8	92.2	73.6	89.7	93.6	90.3	80.8
TADA	Adapters	✓	$895K$	29.5+	84.8+	91.7+	67.2+	88.1+	91.9	89.6+	77.5+

Table 1: **Dialect Adaptation GLUE** results of RoBERTa Base (Liu et al., 2019) for the 7 GLUE Tasks (Matthew's Corr. for CoLA; Pearson-Spearman Corr. for STS-B; Accuracy for all others). T is the number of target tasks for dialect adaptation. Tasks where TADA improves the performance of task-specific SAE adapters, are marked with +.

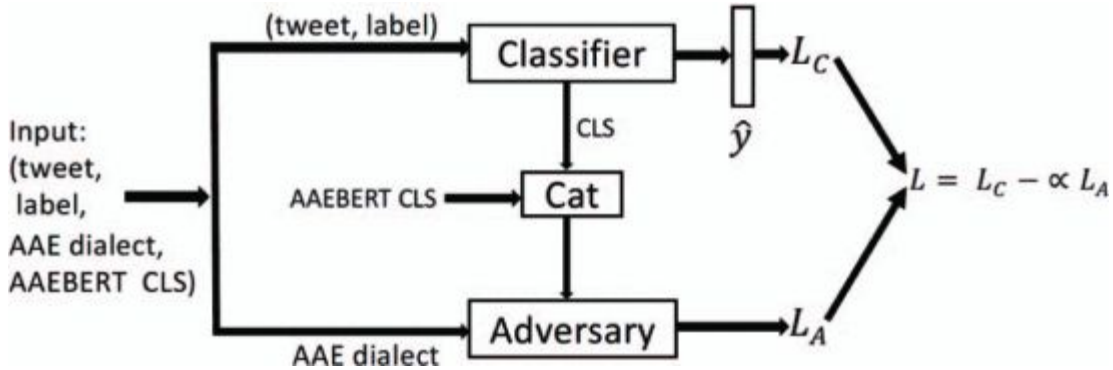
William Held, Caleb Ziems, and Diyi Yang. 2023.TADA : Task Agnostic Dialect Adapters for English. In Findings of ACL

Sentiment classification

Farha et al (2022): Dialect familiarity results in better annotation for sarcasm classification

Kaseb et al (2022): SAIDS: Dialect and sarcasm-informed sentiment analysis

Okpala et al (2022): Dialect prediction as an adversary task for sentiment classification



Ibrahim Abu Farha and Walid Magdy. 2022.The Effect of Arabic Dialect Familiarity on Data Annotation. In Arabic Natural Language Processing Workshop. 399–408.
Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, and Feng Luo. 2022.AAEBERT: Debiasing BERT-based Hate Speech Detection Models via Adversarial Learning. In ICMLA
Abdelrahman Kaseb and Mona Farouk. 2022.SAIDS: A Novel Approach for Sentiment Analysis Informed of Dialect and Sarcasm.WANLP 2022

Token Classification

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

Goal- To identify and classify specific units of information within a text.

Part-of-Speech (POS) Tagging identifies the grammatical role of each word (e.g., noun, verb, adjective).

➔ **Example:** "The/DET cat/NOUN sat/VERB on/PREP the/DET mat/NOUN ./PUNCT"

Named Entity Recognition (NER) identifies and helps extract named entities (*e.g.,* person, organization, location, date).

➔ **Example:** "[Barack Obama]/PERSON, the former president of the [United States]/GPE, visited [London]/GPE on [May 5th, 2023]/DATE."

Chunking (Shallow Parsing) is grouping tokens into phrases (e.g., noun phrases, verb phrases).

➔ **Example:** "[The big red ball]/NP [was bouncing]/VP [in the park]/PP."

Token Classification: Transformer + Conditional Random Field

➔ Fine-tune an encoder, optionally, with Conditional Random Field (CRF) for non-local information gathering.

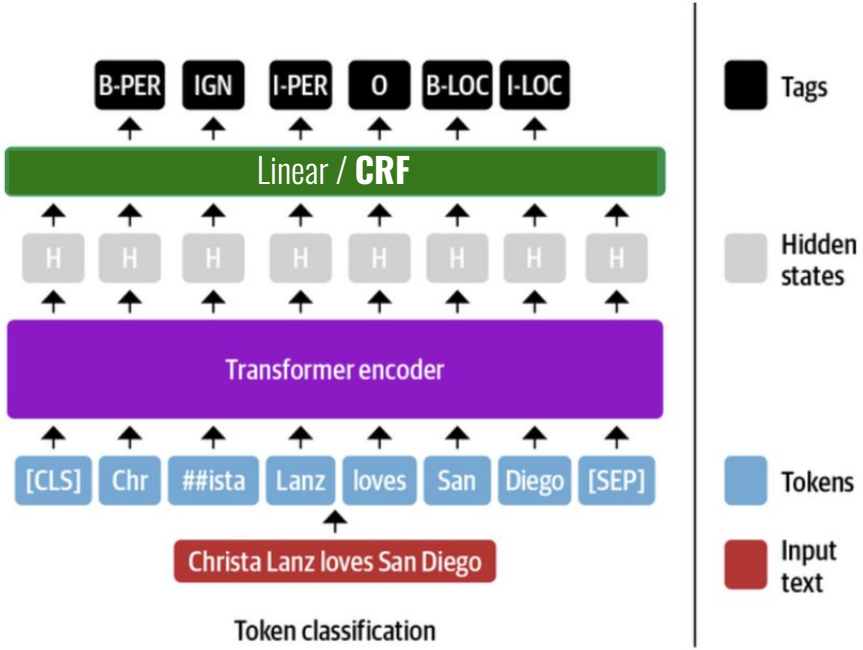


Image Source: LLM-RM at SemEval-2023 Task 2: Multilingual Complex NER using XLM-RoBERTa

Token Classification: Transformer + Conditional Random Field

➔ Fine-tune an encoder, optionally, with Conditional Random Field (CRF) for non-local information gathering.

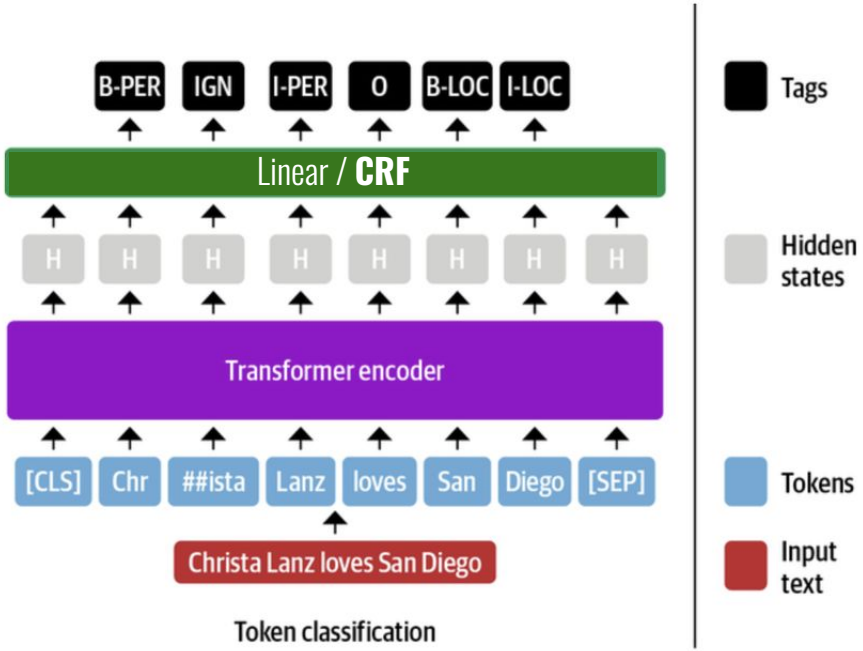


Image Source: LLM-RM at SemEval-2023 Task 2: Multilingual Complex NER using XLM-RoBERTa

CRF helps build a probabilistic model

$$p(y_0 \ldots y_{m-1} \mid \overrightarrow{x_0} \ldots \overrightarrow{x_{m-1}}) = p(\vec{y} \mid \vec{x})$$

where y_i is token label and x_i is token embedding obtained using BERT.

Define feature vector: $\vec{\Phi}(\vec{x}, \vec{y}) \in R^d$

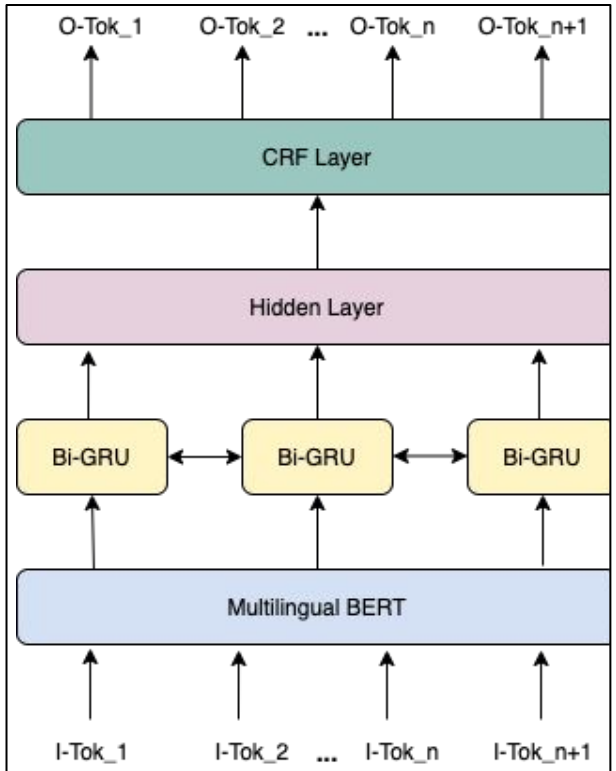
Build a giant log-linear model:

$$p(\vec{y} \mid \vec{x}) = \frac{\exp(\vec{\Phi}(\vec{x}, \vec{y}))}{\sum_{\vec{y'} \in Y^m} \exp(\vec{\Phi}(\vec{x}, \vec{y'}))}$$

Decoding in CRF is challenging, to effectively calculate the sum over all possible sequences y′ in the denominator, we use dynamic programming.

Token Classification: Transformer + BiGRU + CRF

➔ Train a custom architecture by adding BiLSTM or Bi-GRU layer.



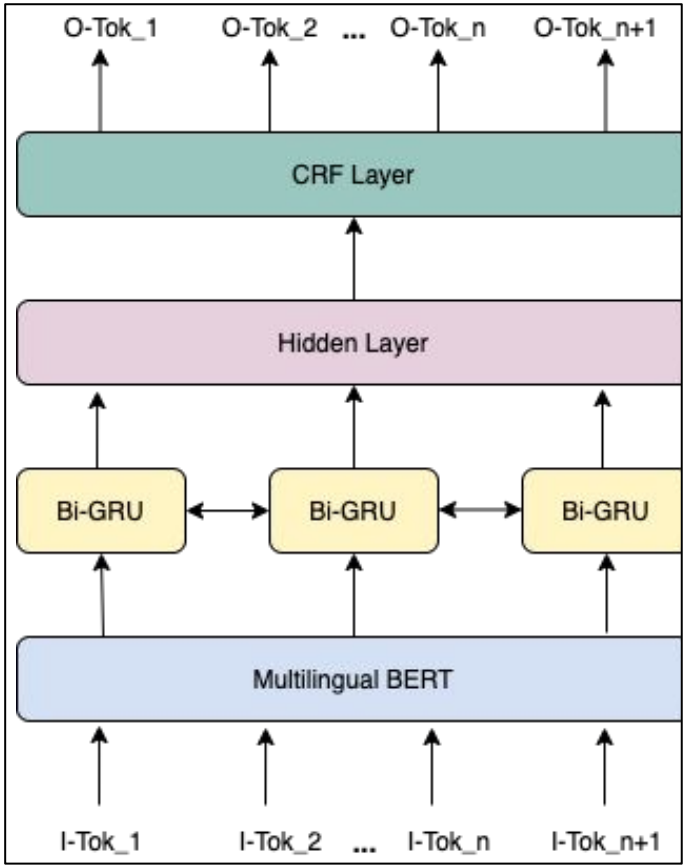
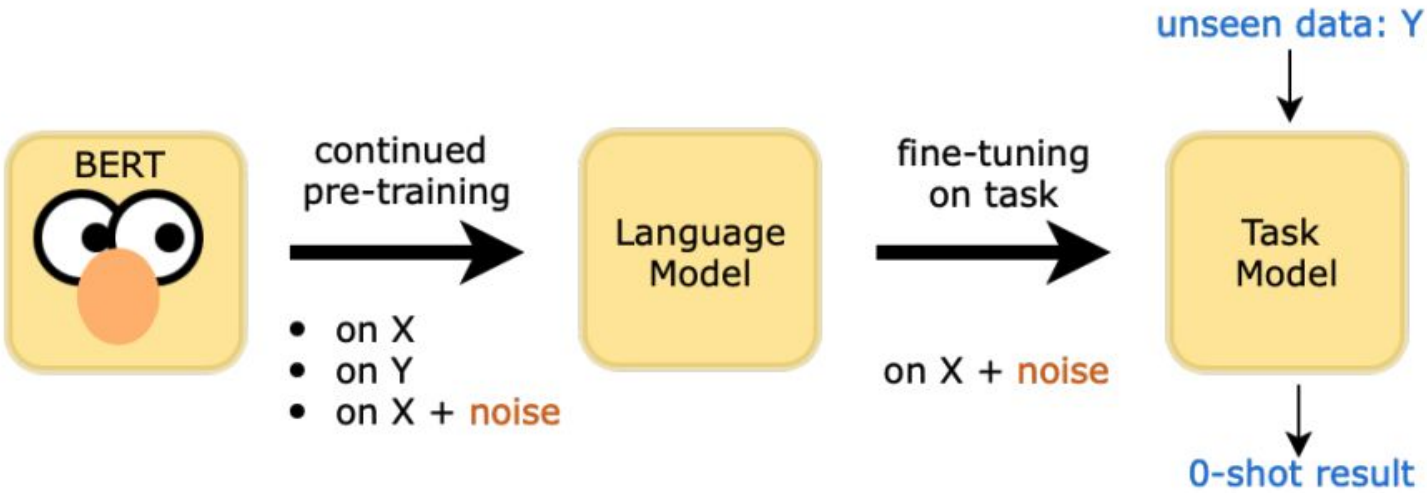


Image Source: Detecting Potential Topics In News Using BERT, CRF and Wikipedia

Morphosyntactic Analysis: POS Taggers

Long history; for example for Arabic: Habash et al (2006)!

Recent approaches include Aepli et al (2022): Character-level noise injection for improved POS tagging



Nizar Habash and Owen Rambow. 2006.MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In ACL
Noëmi Aepli and Rico Sennrich. 2022.Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise. In Findings of ACL

Morphosyntactic analysis: Parsers

Development of treebanks

Example: Norwegian (Kåsen et al (2022); Manx Gaelic (Scannell; 2020).

Adaptation of an existing parser (Wang et al 2017)

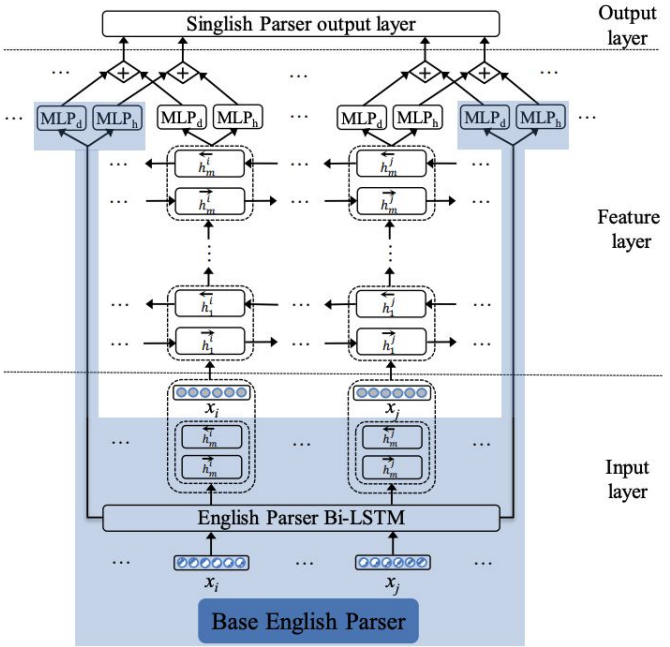


Figure 6: Parser with neural stacking

Andre Kåsen, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg, and Dag Trygve Truslew Haug. 2022.The Norwegian Dialect Corpus Treebank. In LREC.
Kevin Scannell. 2020.Universal Dependencies for Manx Gaelic. In Workshop on Universal Dependencies (UDW 2020)
Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017.Universal Dependencies Parsing for Colloquial Singaporean English. In ACL.

Pre-training / Continual Pre-training: Practical Considerations

Pre-training LMs, even Transformers-based encoder models, is a costly affair

BERT-like (Transformer encoder) models with a 32GB GPU can take weeks of training.

Pre-training a **contextual character-level embeddings** model is cheaper.

In terms of both training time and compute requirements.

Training these models on **monolingual Indic language text took 6-7 days**, per language pair, bringing the **perplexity scores down from ~14-18 to ~2-3**.

Let's delve deeper into contextual character-level embeddings.

Contextual Character-level Embeddings

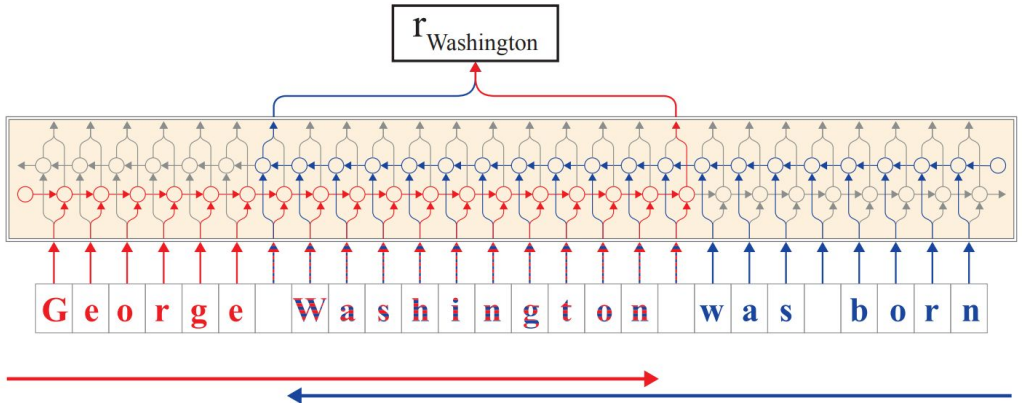
Contextual character embeddings vary based on the surrounding characters and words.

Akbik et al. (2018) propose a character-level BiLSTM models. The approach **passes sentences as sequences of characters** into a character-level language model to form word-level embeddings.

Propose using Pooled (Akbik et al., 2019) and Stacked embeddings with **good performance for token classification.**

Model learns an estimate of the predictive distribution over the next character, given past characters.

$$P(\mathbf{x}_{0:T}) = \prod_{t=0}^{T} P(\mathbf{x}_t | \mathbf{x}_{0:t-1})$$



Character BiLSTM network for embedding extraction (Akbik et al., 2018)

Pre-training with Limited Resources

FlairNLP supports pre-training contextual character level embedding models.

Documentation: https://flairnlp.github.io/flair/

- ➔ Model training is cheaper, faster.
- ➔ Readily available code for adoption of new languages/varieties in the framework.
- ➔ Library supports Transformers encoder from Huggingface too.
- ➔ Flair Stackoverflow: https://stackoverflow.com/questions/tagged/flair

Home — Flair Documentation

```
from flair.embeddings import TransformerWordEmbeddings

# init embedding
embedding = TransformerWordEmbeddings('roberta-base')

# create a sentence
sentence = Sentence('The grass is green .')

# embed words in sentence
embedding.embed(sentence)
```

Use the Huggingface Model hub to find any open source text embedding model to use.

flair/resources/docs/TUTORIAL_9_TRAINING_LM_EMBEDDINGS.md at master · flairNLP/flair

NLU Case Study: (Continual) Pre-training

PLOD Dataset (Zilio et al., 2022)

Lightweight **Character-level Language Models** for the **Biomedical Domain**.

Continually pre-trained PubMed model on the PLOS Journal dataset - PLOD

SoTA performance at detecting abbreviations and long forms.

Models publicly available on SurreyNLP Huggingface and Github.

		AC	LF
PLOD v2 *filtered*	train	334294	183347
	valid	71838	39247
	test	72143	39127
PLOD v2 *unfiltered*	train	341894	186402
	valid	72888	39877
	test	72174	39187
SDU	train+valid	9167	6411

positron	NOUN	B-LF
FP-CIT	PROPN	B-AC
emission	NOUN	I-LF
tomography	NOUN	I-LF
(	PUNCT	B-O
PET	PROPN	B-AC
)	PUNCT	B-O

Pre-trained Language Model (PTLM)	Abbreviations			Long Forms		
	P	R	F	P	R	F
Baselines						
ALBERT$_{base}$ (baseline on PLOD)	*0.8450*	*0.8980*	*0.8710*	*0.7580*	*0.8120*	*0.7840*
RoBERTa$_{large}$ (SOTA on PLOD)	*0.9110*	*0.9350*	*0.9220*	*0.8760*	*0.9210*	*0.8980*
Stacked Embeddings						
GloVe + CLM-PubMed	0.8729	0.9272	0.8992	0.8368	0.8830	0.8593
GloVe + CLM-PLOS	0.8750	0.9295	0.9015	0.8365	0.8795	0.8575
FastText + CLM-PLOS	0.8782	0.9312	0.9039	0.8396	0.8817	0.8601
CLM-PubMed + CLM-PLOS	0.8811	0.9333	0.9065	0.8398	0.8828	0.8608
BERT + CLM-PLOS	0.8978	0.9338	0.9154	0.8465	0.8991	0.8720
RoBERTa$_{large}$ + CLM-PubMed-PLOS	**0.9164**	**0.9456**	**0.9308**	0.8833	**0.9308**	**0.9065**
FastText + RoBERTa$_{large}$ + CLM-PubMed-PLOS	0.9145	0.9413	0.9277	**0.8841**	0.9263	0.9047
GloVe + RoBERTa$_{large}$ + CLM-PubMed-PLOS	0.8800	0.9317	0.9051	0.8316	0.8959	0.8625

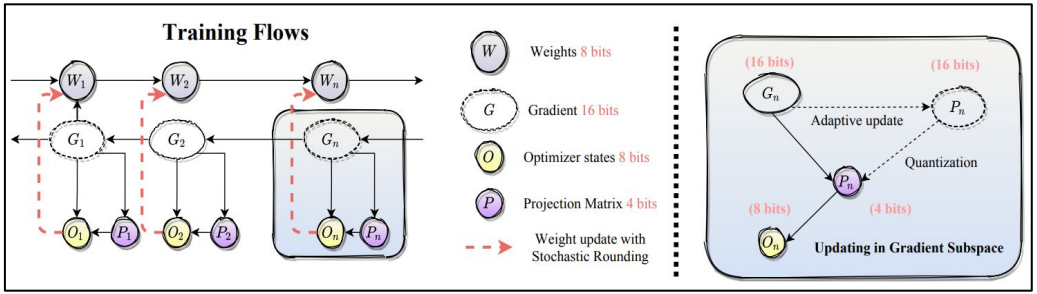
Language Model	PLOD Test *unfiltered*						SDU (Train+Dev Set)					
	Abbreviations			Long Forms			Abbreviations			Long Forms		
	P	R	F	P	R	F	P	R	F	P	R	F
RoBERTa$_{large}$ + CLM-PubMed-PLOS	**0.8977**	0.9351	**0.9160**	0.8726	0.9260	0.8985	0.9012	0.8207	0.8590	0.7896	0.7348	0.7613
Ensemble	0.8893	**0.9362**	0.9121	**0.8758**	**0.9482**	**0.9106**	**0.9148**	0.8574	0.8852	**0.8449**	**0.8524**	**0.8487**

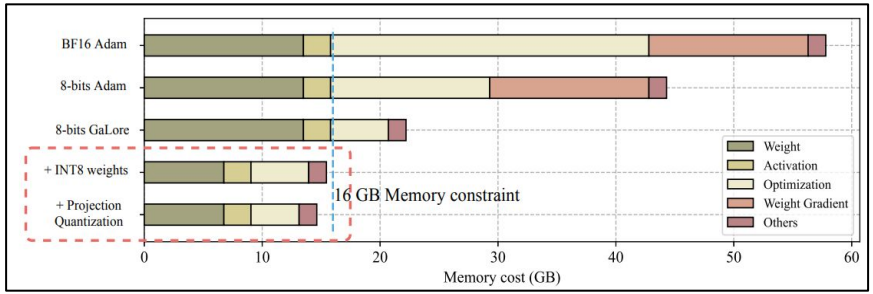
Character-level Language Models for Abbreviation and Long-form Detection - ACL Anthology

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	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

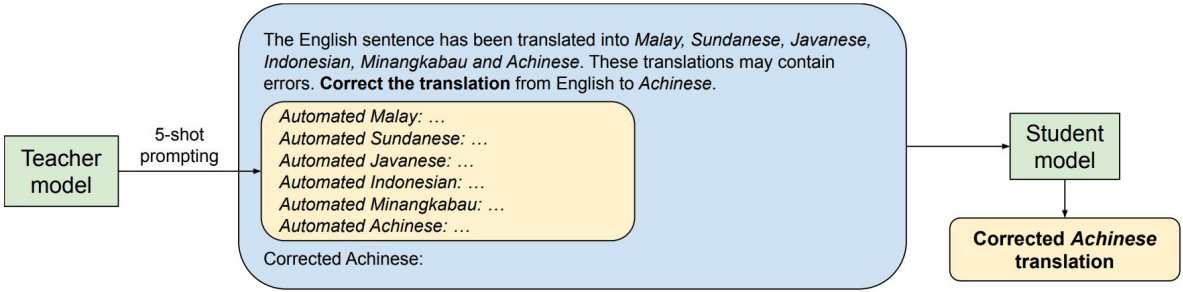
Multilingual Fused Learning for Low-resource Translation

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.



		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

Mufu: Multilingual Fused Learning for Low Resource Translation with LLM

NLU Benchmarks :: Low-resource Scenarios

ScandEval Benchmark - both NLU and NLG tasks in Scandinavian languages.

Hard-Bench - evaluate the learning ability in low-resource settings; includes 11 datasets - 3 from computer vision (CV), and 8 from NLP.

AdvGLUE Benchmark - comprehensive robustness evaluation benchmark that focuses on the adversarial robustness evaluation

XTREME-R - benchmark data on ten natural language understanding tasks, including challenging language-agnostic retrieval tasks, and covers 50 typologically diverse languages.

Quality Estimation for Low-resource Indic Languages - covers five language pairs; En-XX direction.

MLQE-PE - Quality Estimation and Automatic Post-editing datasets for many En-XX and XX-En language pairs.

NLU Benchmarks :: Code-mixed

LinCE - Centralized Benchmark for Linguistic Code-switching Evaluation

DravidianCodeMix - curated from YouTube comments, for analyzing sentiment and identifying offense.

Prabhupadavani - multi-domain, multilingual code-mixed speech translation dataset for 25 languages.

Aggression Detection (TRAC) - Facebook and YouTube comments labelled for detection of Aggression.

Political Aggression Detection - Twitter data in code-mixed English-Hindi within political domain.

Aggression and Offensive Language Identification - Twitter data in code-mixed English-Hindi with both aggression and offense labels.

NLU Benchmarks :: Dialectal NLP

Paper	Dialects	Approach
[Ziems et al. 2022]	African-American English	Perturbation to create variants
[Dacon et al. 2022]	African-American English	Adversarial learning
[Held et al. 2023]	Dialects of English	Contrastive loss, Morphosyntactic loss
[Xiao et al. 2023]	Dialects of English	Hypernetworks as LoRA adapters

Table 6. Dialect-aware approaches evaluated on NLU benchmarks.

Our NLU Benchmark: BESSTIE

Benchmark	Sent.	Sarc.	Eng.	Var.
Cieliebak et al. (2017) [6]	✓	✗	✗	✗
Wang et al. (2018) [37]	✓	✗	✓	✗
Alharbi et al. (2020) [4]	✓	✗	✗	✓
Abu Farha et al. (2021) [3]	✓	✓	✗	✓
Elmadany et al. (2023) [10]	✓	✓	✗	✓
Faisal et al. (2024) [11]	✓	✗	✗	✓
BESSTIE	✓	✓	✓	✓

Table 1: Comparison of BESSTIE with past benchmarks for sentiment or sarcasm classification. 'Sent.' indicates sentiment classification, 'Sarc.' denotes sarcasm classification, 'Eng.' denotes English, and 'Var.' denotes language varieties. A checkmark (✓) denotes the availability of a particular feature, while a cross (✗) indicates its absence.

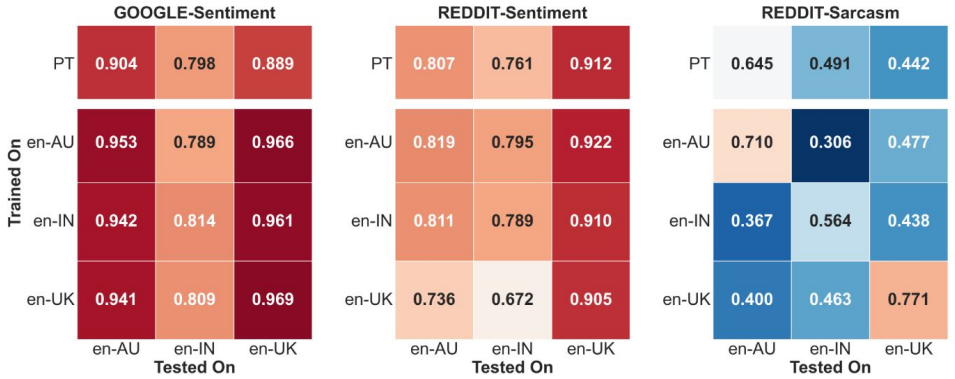


Figure 5: Cross-variety performance analysis of MISTRAL. The figure compares three different scenarios: pre-trained (PT), in-variety fine-tuning, and cross-variety fine-tuning for sentiment and sarcasm classification across all varieties.

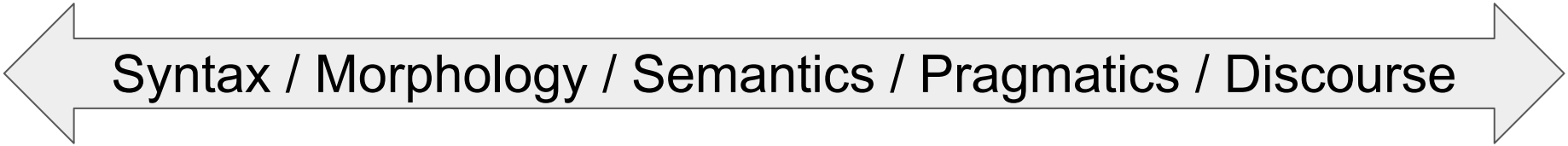
https://arxiv.org/pdf/2412.04726

Some Final Thoughts

NLU involves extraction of implied or structured information from text, aiming to **interpret meaning within context** and **understanding intent**.

Primarily **modelled as sequence/token classification**, or **regression** in ML.

Model/architecture selection of extraction is important - **consider NLP layers**



Consider **multi-task learning** methods and involve related tasks.

Moreover, **consider leveraging multilingual data** for exploiting linguistic **Cognates**.

Contribute to challenging benchmark datasets, ensure fair evaluation of AI models.

Thank you

Questions?

Discussion

How are the challenges in NLP for language varieties different from those in NLP for low-resource languages?

Discussion

How are the challenges in NLP for language varieties different from those in NLP for low-resource languages?

Potentially contentious question: If there is not enough new data for English for newer LLMs, what does the future of lower-resource scenarios look like?

Tutorial Agenda