NLP for Low resource languages

Raviraj B Joshi

L3Cube Pune

Agenda

Introduction to low resource NLP

Multi-lingual BERT models

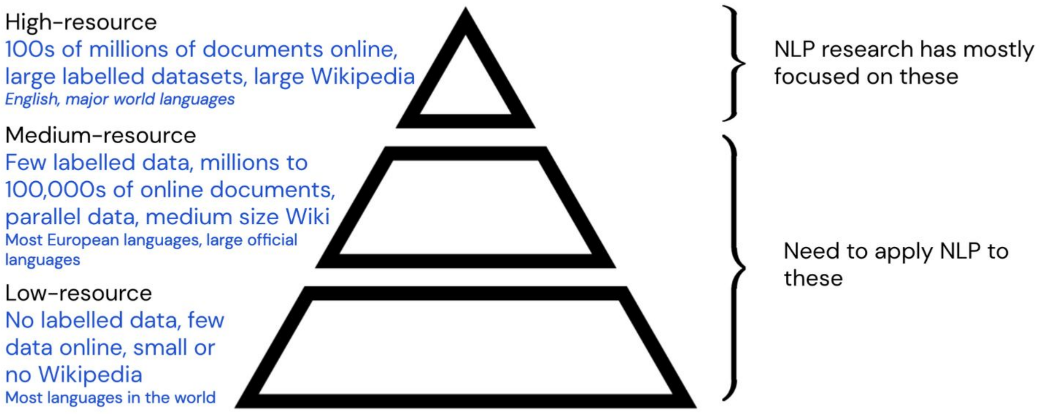
Monolingual BERT models

Cross lingual transfer learning

Cross lingual sentence representations

Case Study in Marathi

NLP Landscape



https://www.ruder.io/unsupervised-cross-lingual-learning/

Why not machine translation ?

- **Limited Availability and Cost:** MT systems for specific language pairs might not be readily accessible and can be expensive to develop, posing a barrier to their widespread use.
- **Underperformance Compared to Multilingual Models:** Models trained on machine-translated data often lag behind advanced deep multilingual models, indicating limitations in the efficacy of MT in certain scenarios.
- **Challenges with Distant Language Pairs and Domain Mismatches:** MT encounters difficulties when dealing with distant language pairs and domain mismatches, impacting its accuracy and effectiveness (Guzmán et al., 2019).
- **Task-Specific Limitations:** Translation-based models, especially in tasks like question answering, heavily rely on the quality of translated named entities, which can significantly influence their performance (Liu et al., 2019).
- **Complexity in Sequence Labelling:** MT faces challenges in projecting annotations across languages, particularly in sequence labelling tasks, presenting a complex problem that is hard to overcome (Akbik & Vollgraf, 2018).

BERT Training Data

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8

Indic language Data

Language	Sentences	Tokens
pa	6.5M	179.4M
hi	62.9M	1199.8M
bn	7.2M	100.1M
or	3.5M	51.5M
gu	7.8M	129.7M
mr	9.9M	142.4M
kn	14.7M	174.9M
te	15.1M	190.2M
ml	11.6M	167.4M
ta	20.9M	362.8M

https://indicnlp.ai4bharat.org/pages/indicnlp-corpus/

Importance of transfer learning for Low Resource

- Transfer learning is a powerful technique that can help improve the performance of natural language processing models for low-resource languages
- It can help models **learn from related languages or tasks**, which can be used to improve their performance on the target language or task
- It can be done using a variety of techniques, including cross-lingual word embeddings, multi-task learning, and **pre-training on related languages**
- It can help overcome the **data scarcity challenge** faced by low-resource languages by leveraging data from other languages or tasks
- It can also help **reduce the amount of labeled data** required to train models for low-resource languages

Pre-training BERT

- Monolingual pre-training
- Multilingual pre-training
  - Less data for individual languages
  - Low-resource languages combined assist each other in pre-training

Multilingual models

mBERT

- mBERT is a pre-trained language model developed by Google that can understand and generate text in 104 languages.
- It was trained on a multilingual corpus of text data from Wikipedia in 100 languages, with most of the data being in English .
- mBERT is an extension of the original BERT model, which was trained only on English text data.
- mBERT has been shown to be effective in various natural language processing tasks, including cross-lingual transfer learning and multilingual question answering .

https://arxiv.org/abs/1810.04805

Multilingual models

IndicBERT

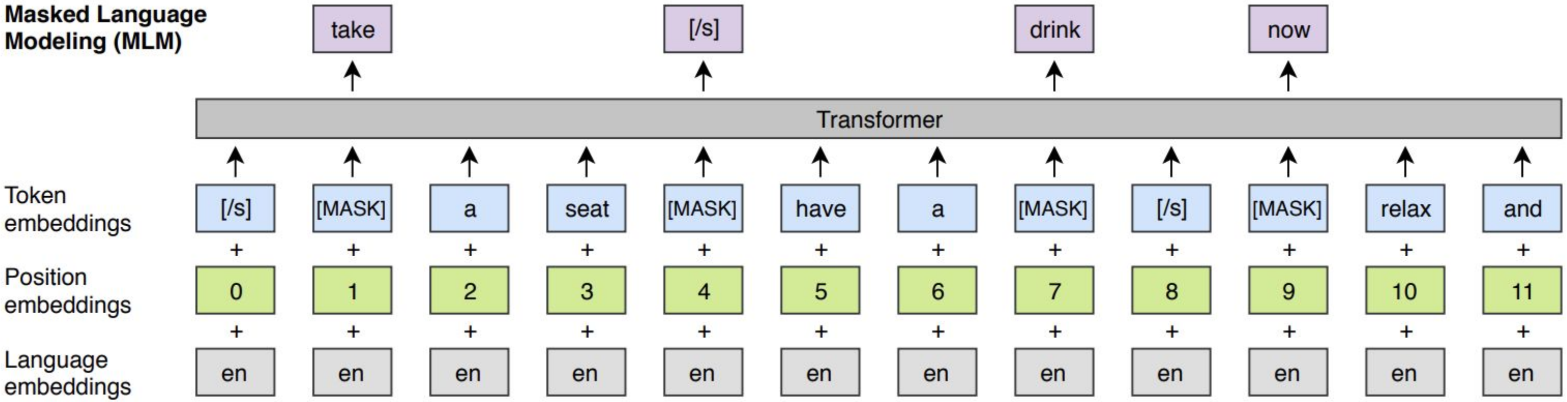
- IndicBERT is a pre-trained language model developed by AI4Bharat that can understand and generate text in **12 major Indian languages**
- It is based on the ALBERT model, which is a derivative of BERT
- IndicBERT was trained on a novel corpus of around 9 billion tokens from IndicNLP in 12 Indian languages, including English
- The model has around 10x fewer parameters than other popular publicly available multilingual models while achieving performance on-par or better than these models
- It has been evaluated on a set of diverse tasks and is available for pre-training, fine-tuning, and evaluation

Multilingual models

MuRIL

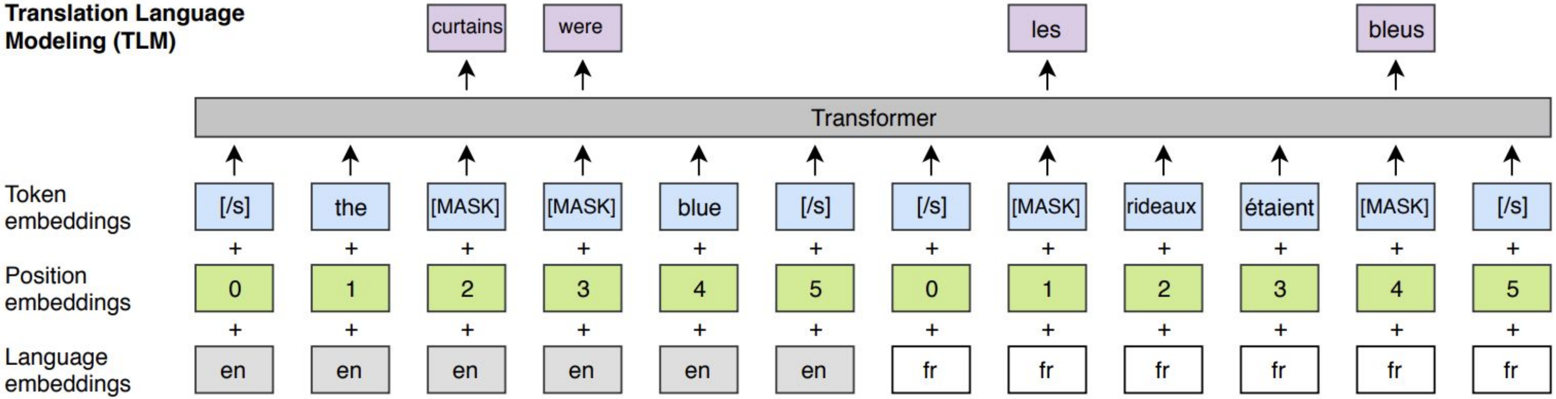
- MuRIL is a BERT model pre-trained on 17 Indian languages and their transliterated counterparts
- It was trained on publicly available corpora from Wikipedia, Common Crawl, PMINDIA, and Dakshina
- The model uses a BERT base architecture and is trained on monolingual segments as well as parallel segments
- MuRIL is intended to be used for a variety of downstream NLP tasks for Indian languages
- The model was released by Google in 2021 under the Apache-2.0 license

Translation language modeling (TLM)



https://arxiv.org/pdf/1901.07291.pdf

Translation language modeling (TLM)



https://arxiv.org/pdf/1901.07291.pdf

Bi-lingual models

DevBERT, DevRoBERTa, DevAlBERT

- DevBERT, DevRoBERTa, and DevAlBERT are bilingual BERT models developed by L3Cube that have been fine-tuned on publicly available Hindi and Marathi monolingual datasets
- The model has been trained on Devanagari-based Hindi and Marathi languages
- DevBERT has shown significant improvements over multi-lingual MuRIL, IndicBERT, and XLM-R models in downstream tasks

https://arxiv.org/abs/2211.11418

Monolingual models

Hindi BERT, Hindi RoBERTa, Hindi AlBERT

- These are **monolingual BERT** models that are fine-tuned on publicly available Hindi monolingual datasets

MahaBERT, MahaRoBERTa, MahaAlBERT

- These are **monolingual BERT** models that are fine-tuned on publicly available Marathi monolingual datasets
- These monolingual models are developed by L3Cube

https://arxiv.org/abs/2211.11418
https://arxiv.org/abs/2202.01159

Other monolingual models

- Kannada BERT
- Telugu BERT
- Malayalam BERT
- Tamil BERT
- Gujarati BERT
- Oriya BERT
- Bengali BERT
- Punjabi BERT
- Assamese BERT

https://arxiv.org/abs/2211.11418

Comparison on MahaHate Corpus

Model	Variant	2-Class Accuracy	4-Class Accuracy
CNN	Random	**0.880**	0.703
	Trainable	0.866	0.710
	Non-Trainable	0.870	**0.751**
LSTM	Random	0.857	0.681
	Trainable	0.860	0.691
	Non-Trainable	**0.869**	**0.751**
BiLSTM	Random	0.858	0.699
	Trainable	0.860	0.664
	Non-Trainable	**0.870**	**0.761**
BERT	IndicBERT	0.865	0.711
	mBERT	0.903	0.783
	xlm-RoBERTa	0.894	0.787
	MahaALBERT	0.883	0.764
	MahaBERT	**0.909**	**0.803**
	MahaRoBERTa	0.902	0.803

https://arxiv.org/pdf/2203.13778.pdf

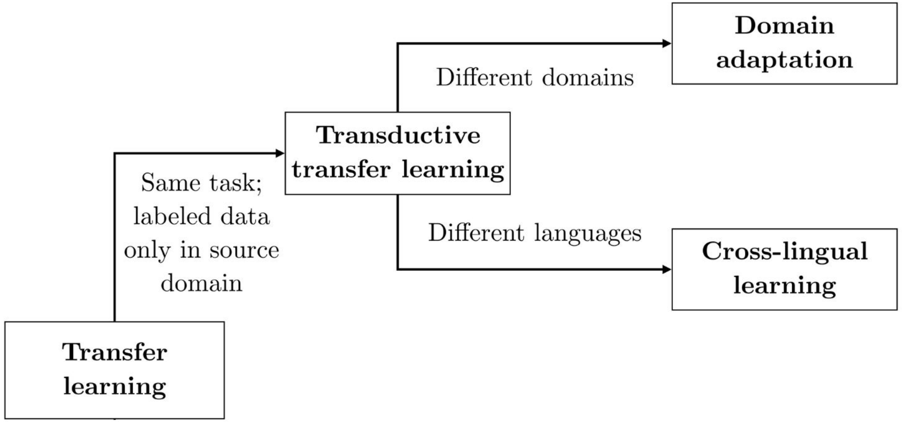
Cross-lingual representations

- **Mapping Language to a Common Space**: Cross-lingual representation learning involves mapping words, phrases, or entire documents from different languages into a shared vector space. This mapping ensures that semantically similar content from diverse languages is represented closely in the vector space, enabling meaningful cross-lingual comparisons.
- **Preserving Semantic Relationships**: Cross-lingual representation learning techniques aim to preserve semantic relationships across languages. Words or phrases with similar meanings in different languages should have similar vector representations, capturing the essence of the words' meanings regardless of the language.
- **Utilizing Parallel Data and Multilingual Contexts**: These techniques often leverage parallel corpora, where the same content is available in multiple languages, to align the representations. Additionally, they can utilize multilingual contexts, exploiting the relationships between different languages to enhance the quality of cross-lingual representations.

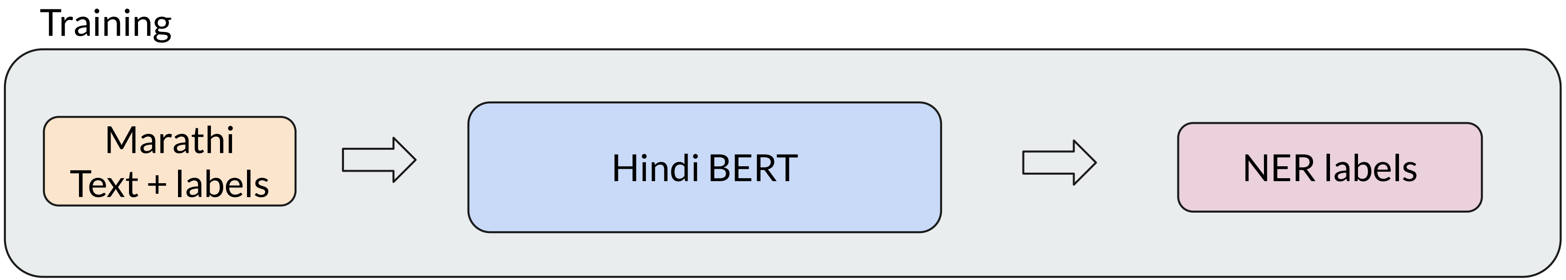
Cross-lingual representations

- **Enabling Transfer Learning**: Cross-lingual representation learning facilitates transfer learning across languages. Models pretrained on cross-lingual representations can be fine-tuned for specific tasks in different languages, leveraging the shared linguistic knowledge encoded in the representations.
- **Enhancing Multilingual NLP Applications**: By providing a common representation space for multiple languages, cross-lingual representation learning empowers various multilingual NLP applications, such as machine translation, sentiment analysis, and named entity recognition. These applications benefit from the ability to operate seamlessly across different languages, improving their accuracy and versatility.

Cross lingual transfer learning



Cross lingual transfer



Cross lingual transfer

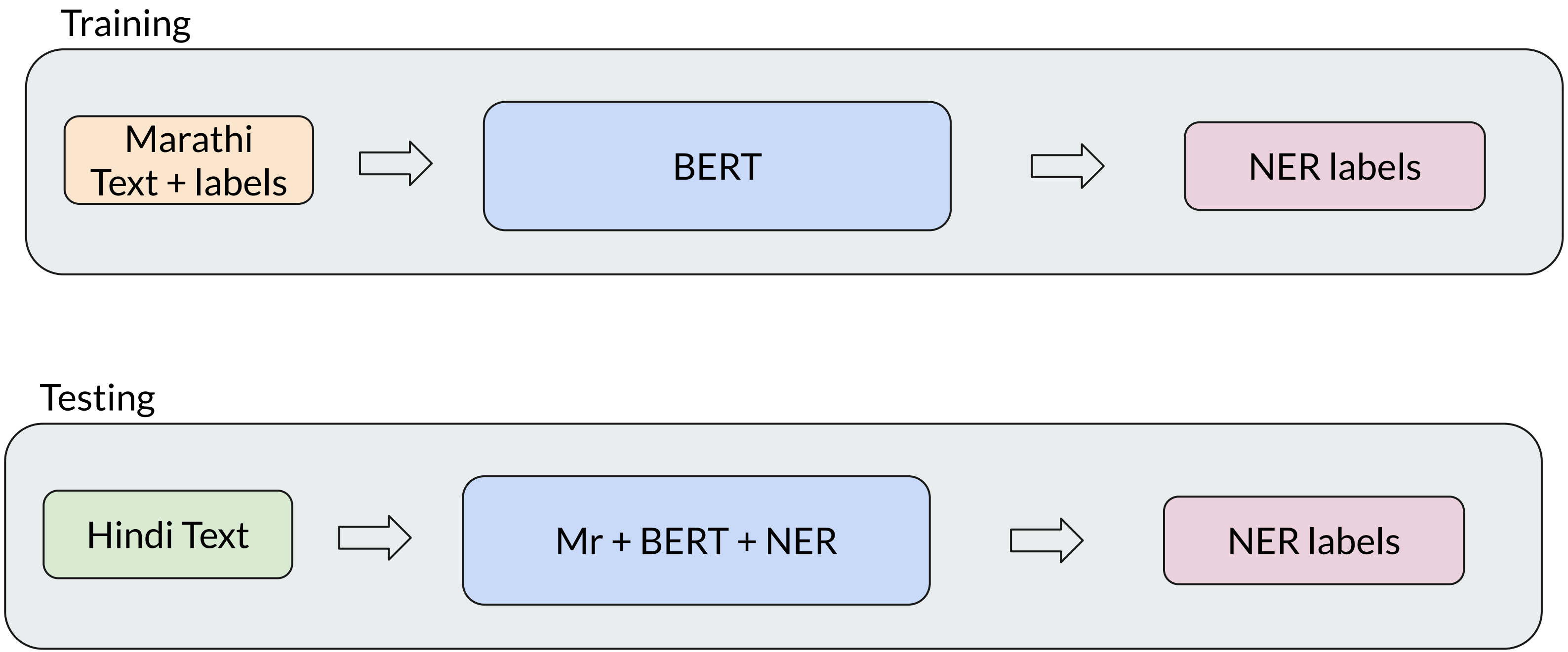
Training

Dataset	IIT Bombay				WikiAnn			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
Multicase BERT	58.35	63.67	54.58	92.42	86.49	86.25	86.73	95.18
Indic BERT	60.79	66.05	53.76	92.57	87.03	87.06	87.00	95.13
Xlm-Roberta	62.32	64.14	60.60	93.00	87.38	86.92	87.85	95.48
Roberta-Marathi	43.81	42.64	45.03	91.34	82.00	80.26	83.82	93.73
MahaBERT	62.57	64.67	60.61	92.97	88.18	88.22	88.14	95.77
MahaRoBERTa	**64.34**	65.64	63.08	92.90	**88.90**	88.59	89.20	96.06
MahaAlBERT	60.00	63.77	56.52	92.52	87.15	87.19	87.11	95.14
RoBERTa Hindi	42.19	41.52	42.88	91.10	82.50	81.69	83.33	95.29
Indic-transformers -hi-roberta	36.80	36.81	36.7	90.49	80.00	78.73	81.32	94.27

Table 6: F1 score(macro), precision and recall of various transformer models using the Marathi datasets.

https://arxiv.org/pdf/2203.12907.pdf

Zero-shot transfer



IndicSBERT

- L3Cube-IndicSBERT is based on multilingual BERT models that map different languages to a common representation space and are useful for cross-language similarity and mining tasks
- It exhibits strong cross-lingual capabilities and performs significantly better than alternatives like LaBSE, LASER, and paraphrase-multilingual-mpnet-base-v2 on Indic cross-lingual and monolingual sentence similarity tasks
- It supports the following 10 Indian regional languages: Hindi, Marathi, Kannada, Telugu, Malayalam, Tamil, Gujarati, Odia, Bengali, and Punjabi

https://arxiv.org/abs/2304.11434

IndicSBERT

Source Sentence

हम दीपावली उत्साह के साथ मनाते हैं

Sentences to compare to

हम दीपावली खुशियों से मनाते हैं

दिवाली रोशनी का त्योहार है

Add Sentence

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.091 s

हम दीपावली खुशियों से मनाते हैं	0.923
दिवाली रोशनी का त्योहार है	0.590

IndicSBERT

Source Sentence

आम्हाला भारतीय असल्याचा अभिमान आहे

Sentences to compare to

हमें भारतीय होने पर गर्व है

భారతీయులమైనందుకు గర్విస్తున్నాం

અમને ભારતીય હોવાનો ગર્વ છે

Add Sentence

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

हमें भारतीय होने पर गर्व है	0.790
భారతీయులమైనందుకు గర్విస్తున్నాం	0.646
અમને ભારતીય હોવાનો ગર્વ છે	0.706

Monolingual SBERT Models

Marathi SBERT

Hindi SBERT

Kannada SBERT

Telugu SBERT

Malayalam SBERT

Tamil SBERT

Gujarati SBERT

Oriya SBERT

Bengali SBERT

Punjabi SBERT

Indic SBERT (multilingual)

https://arxiv.org/abs/2304.11434

Monolingual SBERT Models tuned for Similarity

Marathi Similarity

Hindi Similarity

Kannada Similarity

Telugu Similarity

Malayalam Similarity

Tamil Similarity

Gujarati Similarity

Oriya Similarity

Bengali Similarity

Punjabi Similarity

Indic Similarity (multilingual)

https://arxiv.org/abs/2304.11434

About code-mixed NLP

- L3Cube-HingCorpus and HingBERT models (Hindi-English)
- L3Cube-MeCorpus and MeBERT models (Marathi-English)
- Code-mixing vs code-switching
- Resources: https://github.com/l3cube-pune/code-mixed-nlp

https://arxiv.org/abs/2204.08398

https://arxiv.org/abs/2306.14030

Break