Multilingual NLP

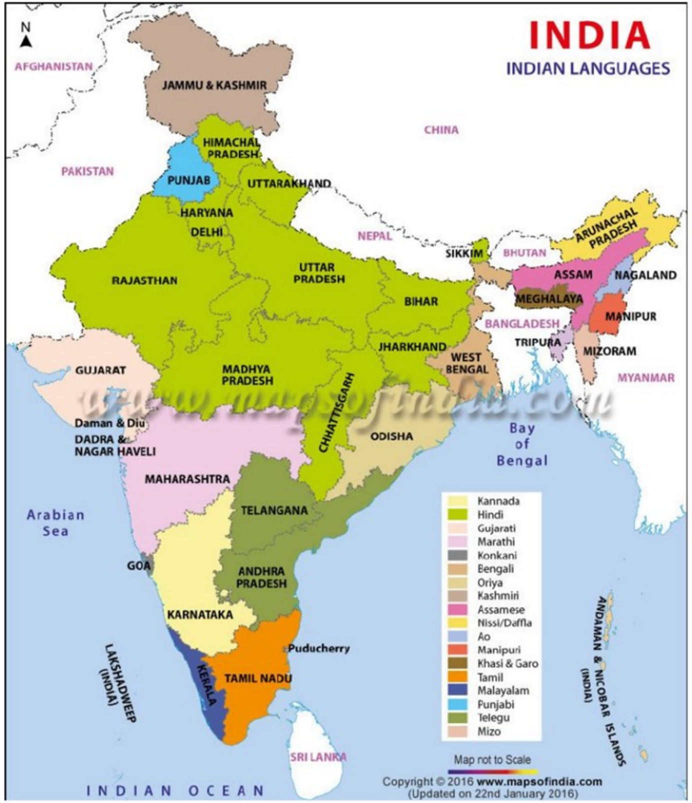
Inspired from Graham Neubig's
CMU CS 11737, Fall 2020

Linguistic diversity: ~7000 languages

Low resource languages

There are about 460 languages in India.

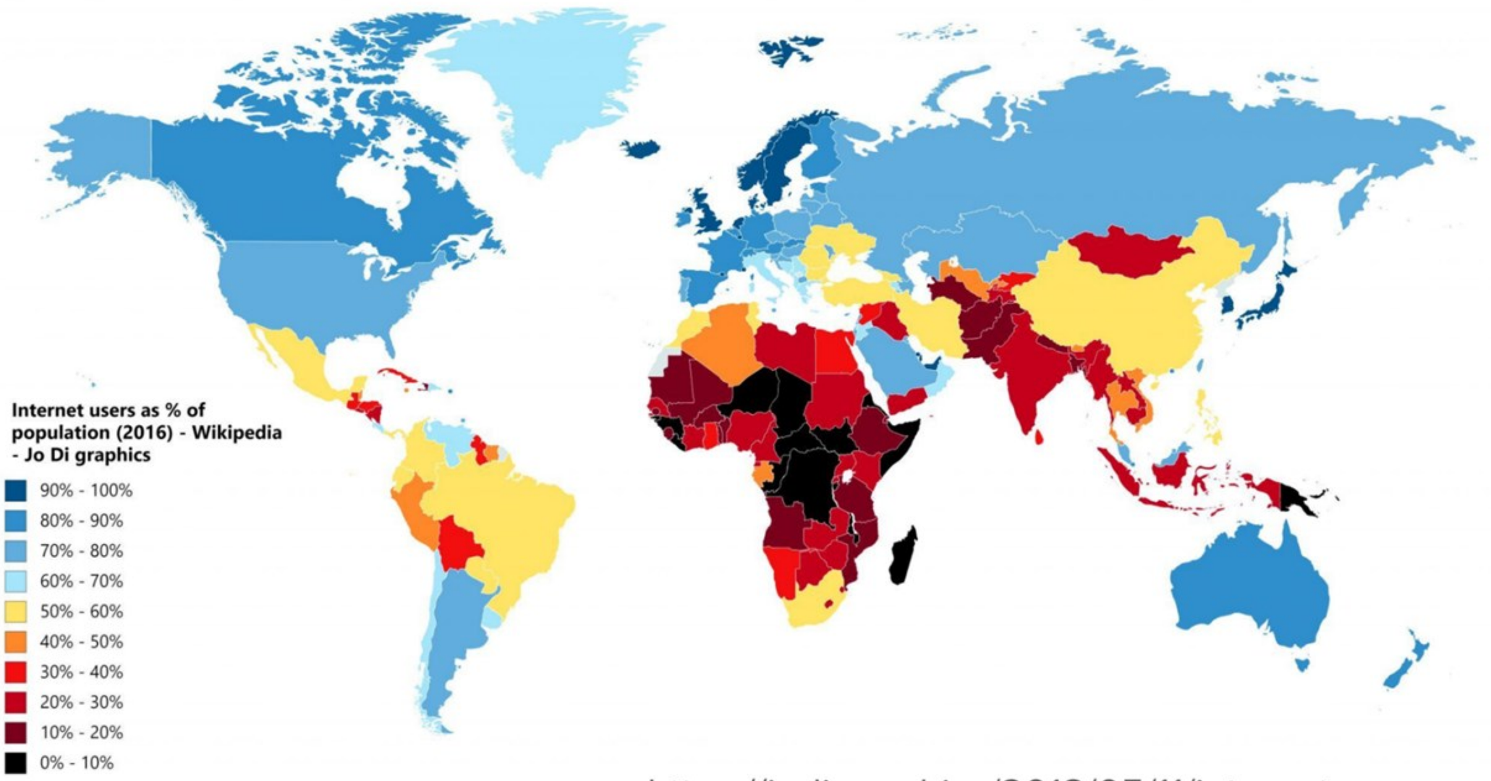
1.38 billion people



Low resource languages

Africa is a continent with a very high linguistic diversity:
there are an estimated 1.5-2K African languages from 6 language families.
1.33 billion people

Low-resource/multilingual NLP



https://jodi.graphics/2018/05/11/internet-users-as-of-population/

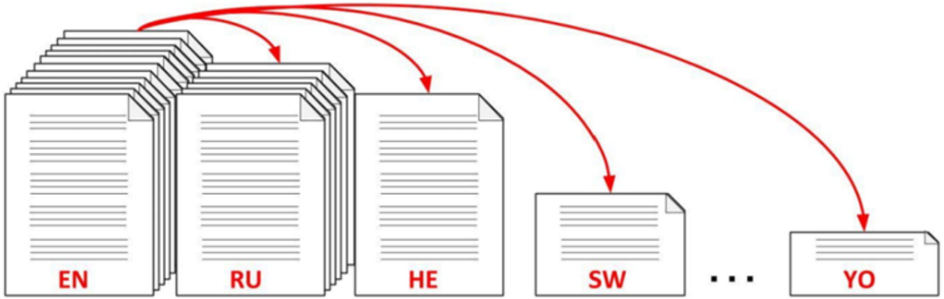
40% of world's population: South Asia - 1.75 billion, Africa - 1.3 billion, etc.

Approaches to low-resource/multilingual NLP

- Manual curation and annotation of large-scale resources for thousands of languages in infeasible or prohibitively expensive

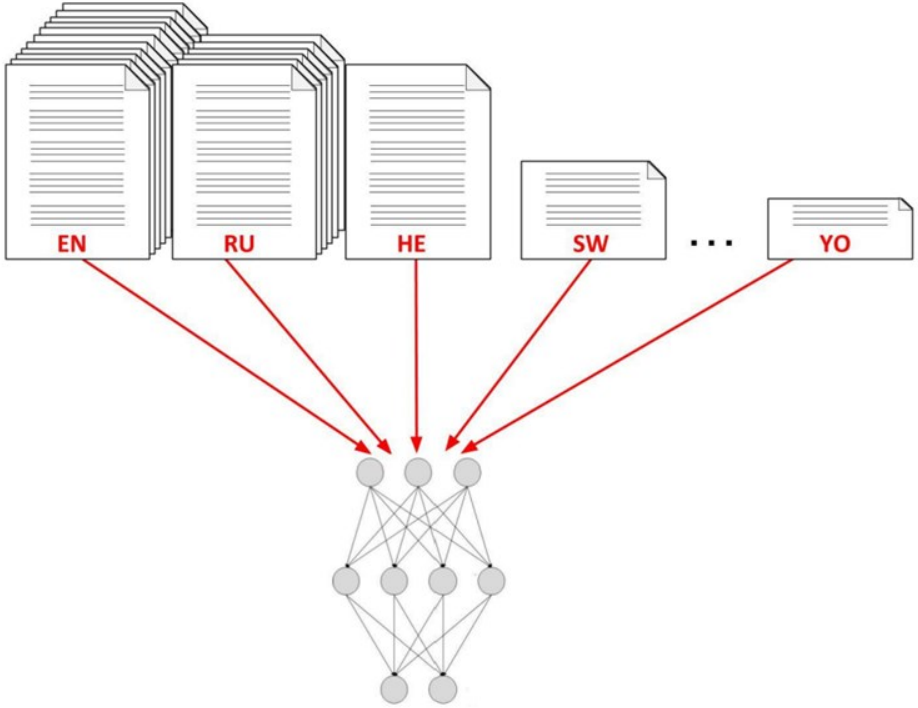
- Unsupervised learning (Snyder and Barzilay 2008; Cohen and Smith, 2009; Snyder, 2010; Vulić, De Smet, and Moens 2011; Spitkovsky et al., 2011; Goldwasser et al., 2011; Titov and Klementiev 2012; Baker et al., 2014, and many others)

Approaches to low-resource/multilingual NLP



- **Cross-lingual transfer learning** – transfer of resources and models from resource-rich source to resource-poor target languages
    - Transfer of annotations (e.g., POS tags, syntactic or semantic features) via cross-lingual bridges (e.g., word or phrase alignments)
    - Transfer of models – train a model in a resource-rich language and adapt (e.g. fine-tune) it in a resource-poor language

- Zero-shot learning – train a model in one domains and assume it generalizes more or less out-of-the-box in a low-resource domain
- Few shot learning – train a model in one domain and use only few examples from a low-resource domain to adapt it



- Joint multilingual learning – train a single model on a mix of datasets in all languages, to enable data and parameter sharing where possible

Multilingual Pre-training

- Extend pre-training to multiple languages
- Pro: Can transfer information across languages

Multilingual Pre-training

- Extend pre-training to multiple languages
- Pro: Can transfer information across languages

- Con: Limited model capacity
  - *Curse of Multilinguality*
- Increases low-resource performance
- Reduces high-resource performance

mBERT on 100 languages

- 110K WordPiece vocabulary
- Rules for handling specific languages like Chinese
  - Chinese, Japanese and Korean don't use whitespaces

- Exponential Weighted Sampling
  - English vs. Icelandic - 1000x → 100x
  - Exponentiate probability by 0.7 and re-normalize

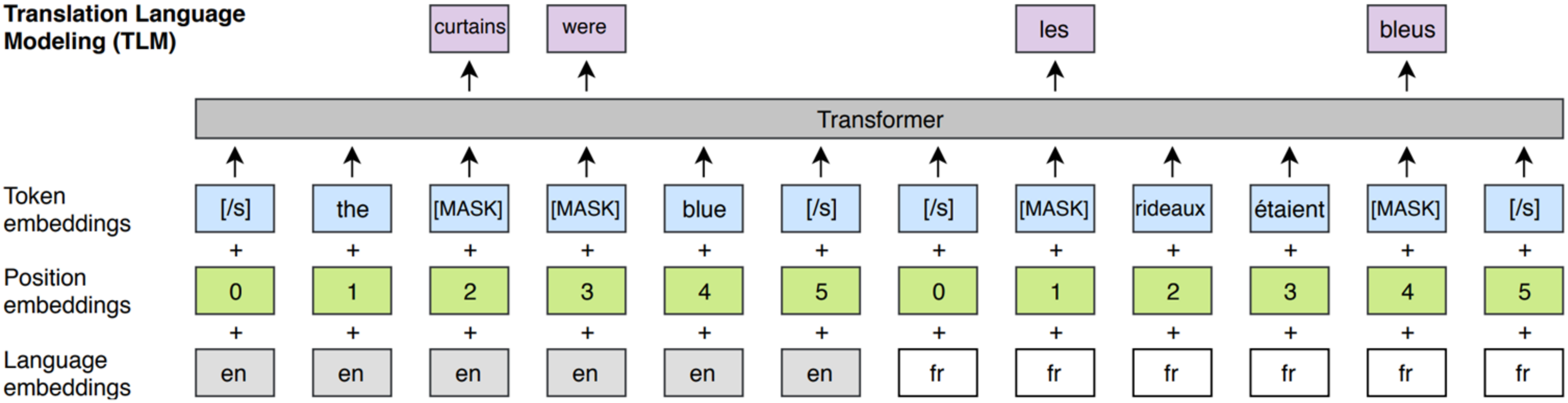
mBERT monolingual performance

- mBERT vs BERT:
  - MNLI: 81.4 vs. 84.2

- mBERT vs BERT-Chinese:
  - XNLI: 74.2 vs. 77.2

XLM, XLM-Roberta from Facebook

- XLM uses Translation Language Modeling (TLM)

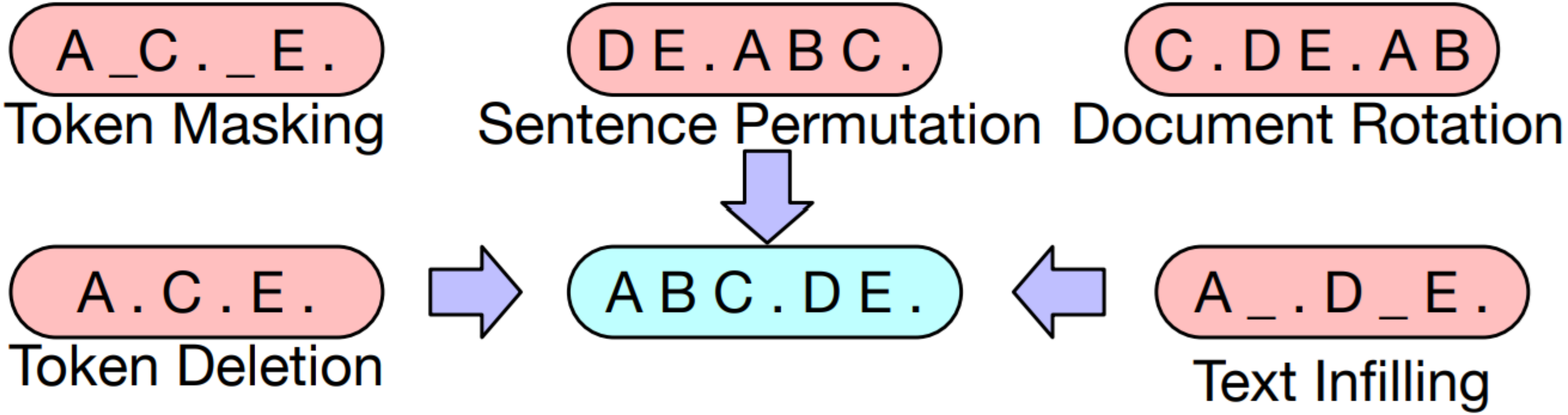


XLM, XLM-Roberta from Facebook

- XLM uses
  - Wikipedia text for MLM
  - Supervised translation data for TLM

- XLM-R scales it up to use CommonCrawl text
- Does not use TLM or language-ids

BART


A _C . _ E .
Token Masking
D E . A B C .
Sentence Permutation
C . D E . A B
Document Rotation
A . C . E .
Token Deletion
A B C . D E .
A _ . D _ E .
Text Infilling

mBART: Multi-Lingual BART

- Training on CC25 corpus
- Corpus of 25 languages
- A subset of *Common Crawl*
- A crawl of the internet

Code	Language	Tokens/M	Size/GB
En	English	55608	300.8
Ru	Russian	23408	278.0
Vi	Vietnamese	24757	137.3
Ja	Japanese	530 (*)	69.3
De	German	10297	66.6
Ro	Romanian	10354	61.4
Fr	French	9780	56.8
Fi	Finnish	6730	54.3
Ko	Korean	5644	54.2
Es	Spanish	9374	53.3
Zh	Chinese (Sim)	259 (*)	46.9
It	Italian	4983	30.2
Nl	Dutch	5025	29.3
Ar	Arabic	2869	28.0
Tr	Turkish	2736	20.9
Hi	Hindi	1715	20.2
Cs	Czech	2498	16.3
Lt	Lithuanian	1835	13.7
Lv	Latvian	1198	8.8
Kk	Kazakh	476	6.4
Et	Estonian	843	6.1
Ne	Nepali	237	3.8
Si	Sinhala	243	3.6
Gu	Gujarati	140	1.9
My	Burmese	56	1.6

Table 1: **Languages and Statistics of the CC25 Corpus.** A list of 25 languages ranked with monolingual corpus size. Throughout this paper, we replace the language names with their ISO codes for simplicity. (*) Chinese and Japanese corpus are not segmented, so the tokens counts here are sentences counts

mT5

- Collect the mC4 corpus over 100 languages
- Train using Span-denoising objective
- Don’t use language-ids
  - Results in “accidental translations”

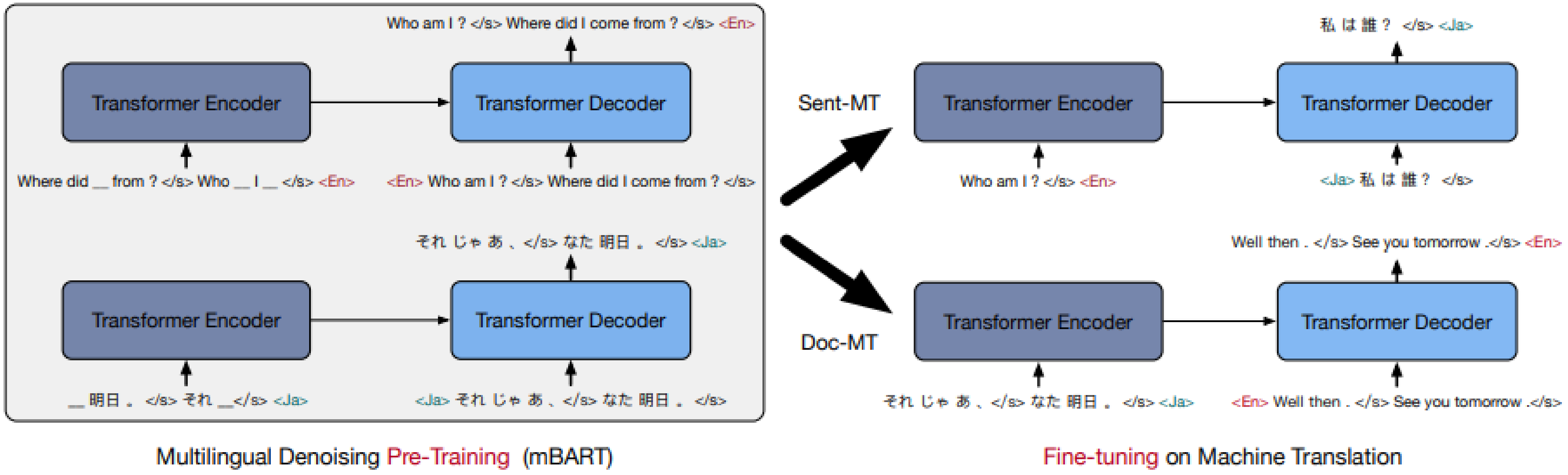


Figure 1: Framework for our Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right), where we use (1) sentence permutation (2) word-span masking as the injected noise. A special language id token is added at both the encoder and decoder. One multilingual pre-trained model is used for all tasks.

Indian Languages

- MuRIL from Google: 16 IN and EN
  - MLM and TLM
  - Vocabulary of 197K
- Better than mBERT

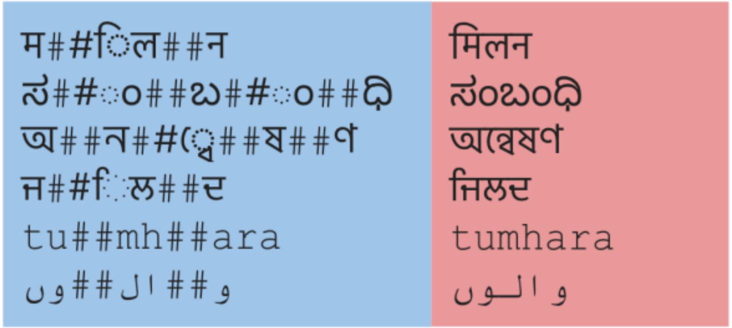


Figure 3: *IN language words tokenized using mBERT (blue) and MuRIL (Red).*

- Available in TFHub and Huggingface

Indian Languages

- IndicBERT from IITM: 12 IN and EN
  - Only MLM
  - Vocabulary of 200K
- Better than mBERT, XLM-R for IndicGLUE

Machine Translation

Be the change you want to see in the world

वह परिवर्तन बनो जो संसार में देखना चाहते हो

Microsoft

Google Translate

Government: administrative requirements, education, security.

Enterprise: product manuals, customer support

Social: travel (signboards, food), entertainment (books, movies, videos)

Translation under the hood

- Cross-lingual Search
- Cross-lingual Summarization
- Building multilingual dictionaries

What is Machine Translation?

Word order: SOV (Hindi), SVO (English)

S V O

E: Germany won the last World Cup

H: जर्मनी ने पिछला विश्व कप जीता था

S O V

Free (Hindi) vs rigid (English) word order

पिछला विश्व कप जर्मनी ने जीता था *(correct)*

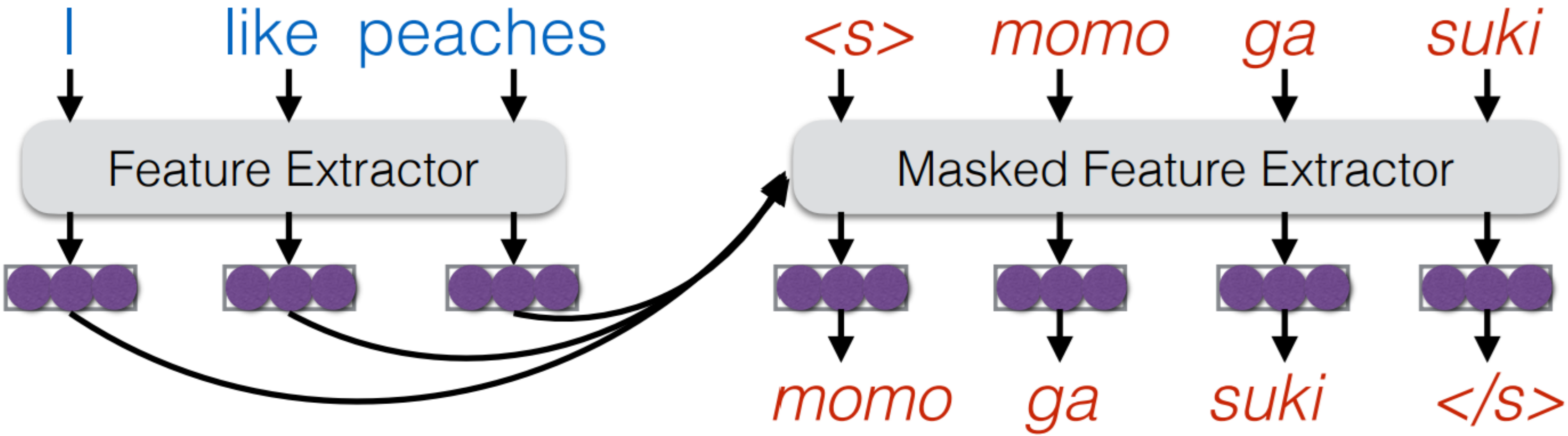
The last World Cup Germany won *(grammatically incorrect)*

The last World Cup won Germany *(meaning changes)*

Language Divergence ➔ *the great diversity among languages of the world*

The central problem of MT is to bridge this language divergence

- Sequence-to-sequence modeling



Parallel corpora

... the open parallel corpus

OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving ...
Contributions are very welcome! Please contact <jorg.tiedemann@helsinki.fi >

Search & download resources: -- select -- ▾ -- select -- ▾ all ▾

Latest News

- 2018-02-15: New corpora: ParaCrawl, XhosaNavy
- 2017-11-06: New version: OpenSubtitles2018
- 2017-11-01: New server location: http://opus.nlpl.eu
- 2016-01-08: New version: OpenSubtitles2016
- 2015-10-15: New versions of TED2013, NCv9
- 2014-10-24: New: JRC-Acquis
- 2014-10-20: NCv9, TED talks, DGT, WMT
- 2014-08-21: New: Ubuntu, GNOME
- 2014-07-30: New: Translated Books
- 2014-07-27: New: DOGC, Tanzil
- 2014-05-07: Parallel coref corpus ParCor

Search & Browse

- OPUS multilingual search interface
- Europarl v7 search interface
- Europarl v3 search interface
- OpenSubtitles 2016 search interface
- EUconst search interface
- Word Alignment Database (old DB)

Tools & Info

- OPUS Wiki
- OPUS API by Yonathan Koren
- Uplug at bitbucket

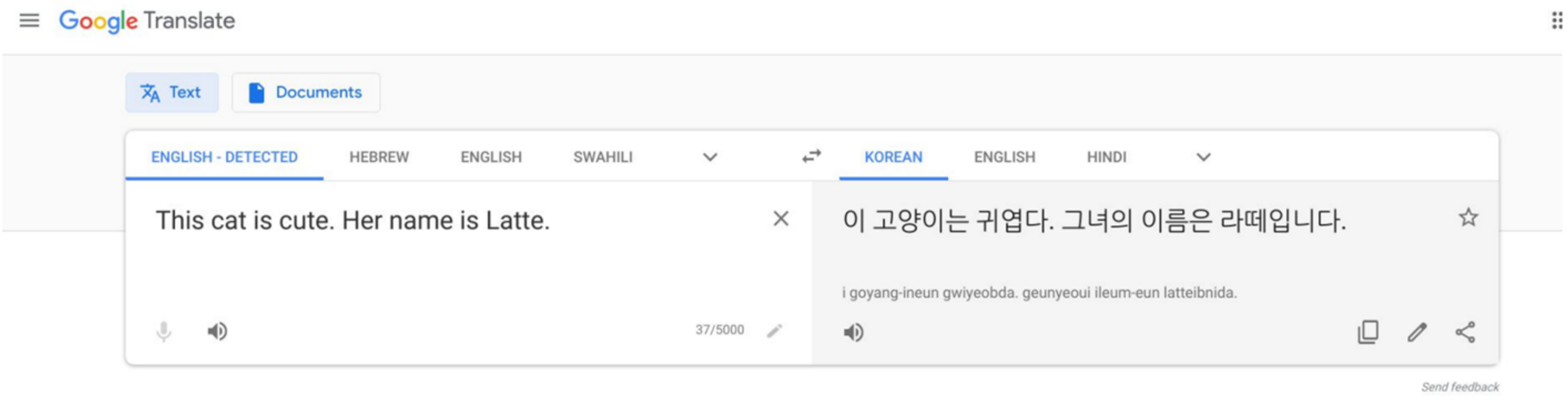
Some Projects using OPUS

- Let'sMT! - On-line SMT toolkit

Sub-corpora (downloads & infos):

- Books - A collection of translated literature (Books.tar.gz - 535 MB)
- DGT - A collection of EU Translation Memories provided by the JRC
- DOGC - Documents from the Catalan Goverment (DOGC.tar.gz - 2.8 GB)
- ECB - European Central Bank corpus (ECB.tar.gz - 3.0 GB)
- EMEA - European Medicines Agency documents (EMEA.tar.gz - 13.0 GB)
- The EU bookshop corpus (EUbookshop.tar.gz - 42 GB)
- EUconst - The European constitution (EUconst.tar.gz - 82` MB)
- EUROPARL v7 - European Parliament Proceedings (Europarl.tar.gz - 2(GB)
- GNOME - GNOME localization files (GNOME.tar.gz - 9 GB)
- Global Voices - News stories in various languages (GlobalVoices.tar.gz - 1.2 GB)
- The Croatian - English WaC corpus (hrenWaC.tar.gz - 59 MB)
- JRC-Acquis- legislative EU texts (JRC-Acquis.tar.gz - 11 GB)

Is it a good translation?



MT evaluation is hard

- MT Evaluation is a research topic on its own
- Language variability: there is no single correct translation
    - Is system A better than system B?
- Human evaluation is subjective

Automatic evaluation

- The BLEU score proposed by IBM (Papineni et al., 2002)
  - Count n-grams overlap between machine translation output and reference reference translations
  - Compute precision for ngrams of size 1 to 4
  - No recall (because difficult with multiple references)
  - To compensate for recall: “brevity penalty”. Translations that are too short are penalized
  - Final score is the geometric average of the n-gram precisions, times the brevity penalty

$$\text{BLEU} = min(1, \frac{output\ length}{reference\ length})(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}}$$

  - Calculate the aggregate score over a large test set

Automatic evaluation

- Embedding based
  - BertScore, **chrF, YISI-1, ESIM**, ...

Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics

Nitika Mathur **Timothy Baldwin** **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia

nmathur@student.unimelb.edu.au {tbaldwin,tcohn}@unimelb.edu.au

Massively Multilingual Machine Translation

- One model for Translating between 100 languages!
- Goal: Maintain same performance in high-resource and improve low-resource langs
- M2M-100 from FAIR and MMNMT from Google

Indian Languages

- IIT-B Hi Corpus is one benchmark

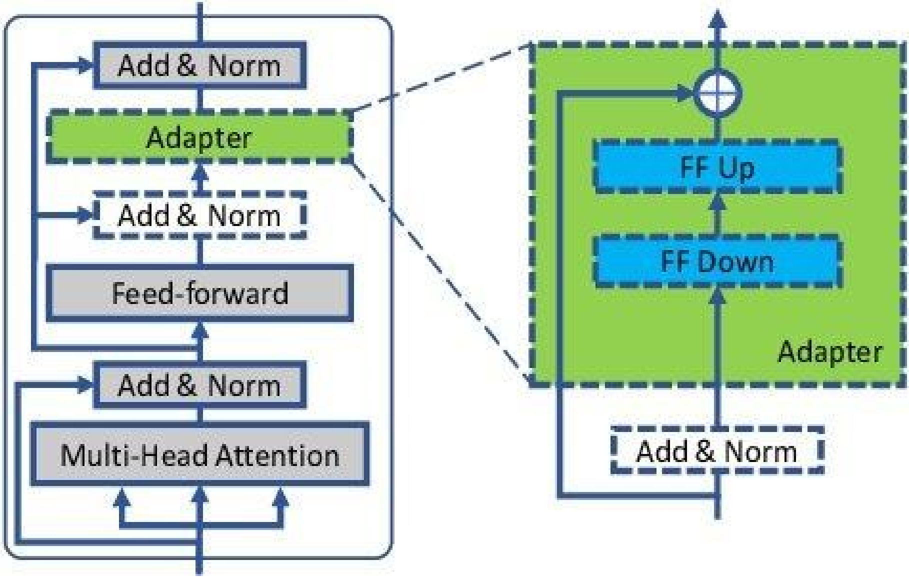
- Recently released: Samanantar (from IIT-M)
  - Largest corpus for 11 Indian languages
  - Automatically mined from web
  - Trained mT5 outperforms Google Translate

Language Adapter

- Trained using Masked Language Modeling (MLM) on the unlabeled Corpus of a language (E.g. Wikipedia)

- Serves as language encoder for a specific language while all other parameters of transformer frozen

- Highy parameter efficient
  - 1 % parameters of the standard mBERT model

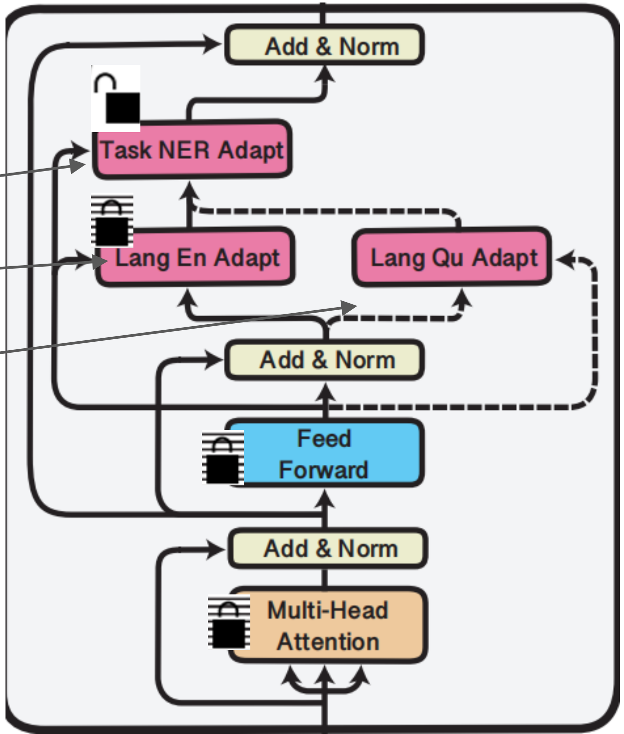


Language Adapters for Cross-Lingual Transfer from English to Target (Pfeiffer et al., 2020)

Task adapter inserted during training (only trainable Module)

Use English adapter (frozen) during training

Replace with target language adapter during inference

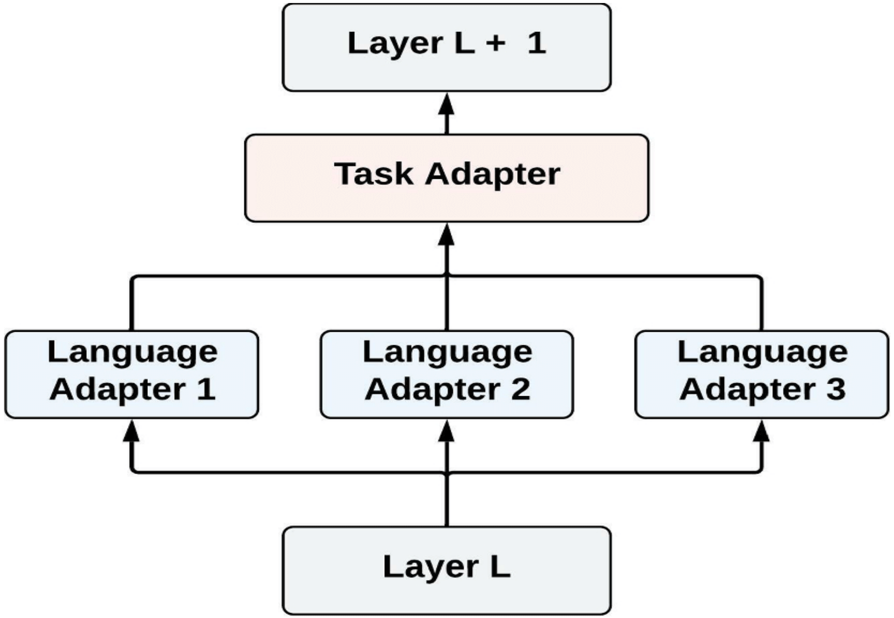


Strong Results for zero-shot transfer (He et al., 2021)

Model	POS All	POS Target	POS Distant	NER All	NER Target	NER Distant	XNLI All	XNLI Target	XNLI Distant
XLMR-ft (Hu et al., 2020)	73.80	73.14	64.34	65.40	64.87	58.21	79.24	78.56	76.73
XLMR-ft (reproduced)	74.29	73.61	64.90	63.85	63.32	56.85	79.28	78.64	77.03
XLMR-adapter$_{256}$	**75.82**	**75.20**	**68.05**	**66.40**	**65.95**	**59.01**	**80.08**	**79.43**	**77.60**

Zero-shot cross-lingual results (reported by He et al., 2021). Target is the average test result of all target languages except English. Distant is the average test result of the languages not in the Indo-European family.

Using Multiple Language Adapters



Placing Language Adapters in Parallel (He et al., 2021)

Best Practices with Adapters!!!

- Keep a higher learning rate than the one used with standard BERT/mBERT models
    - 1e-4 vs 2e-5
- Might have to train for longer than the standard BERT/mBERT fine-tuning