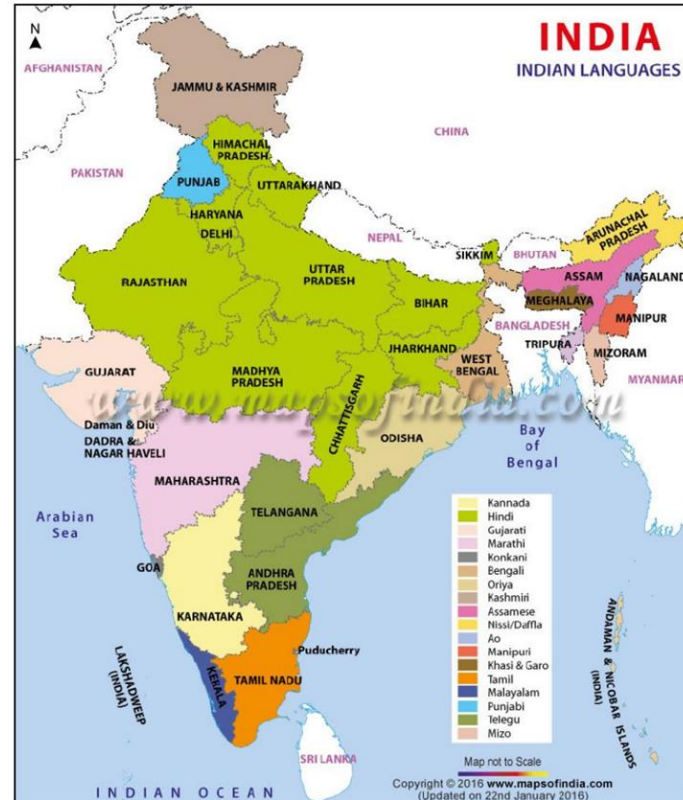

Multilingual NLP

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

Low resource languages

There are about 460 languages in India.

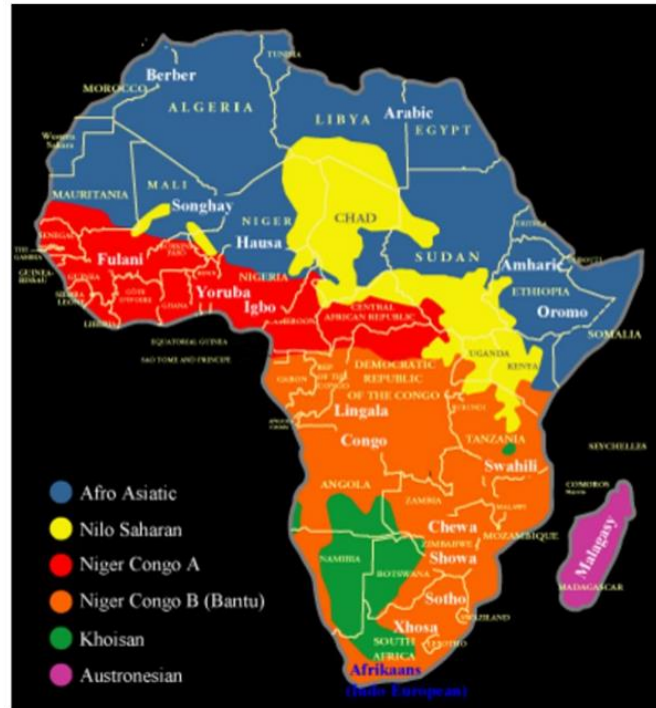
1.38 billion people



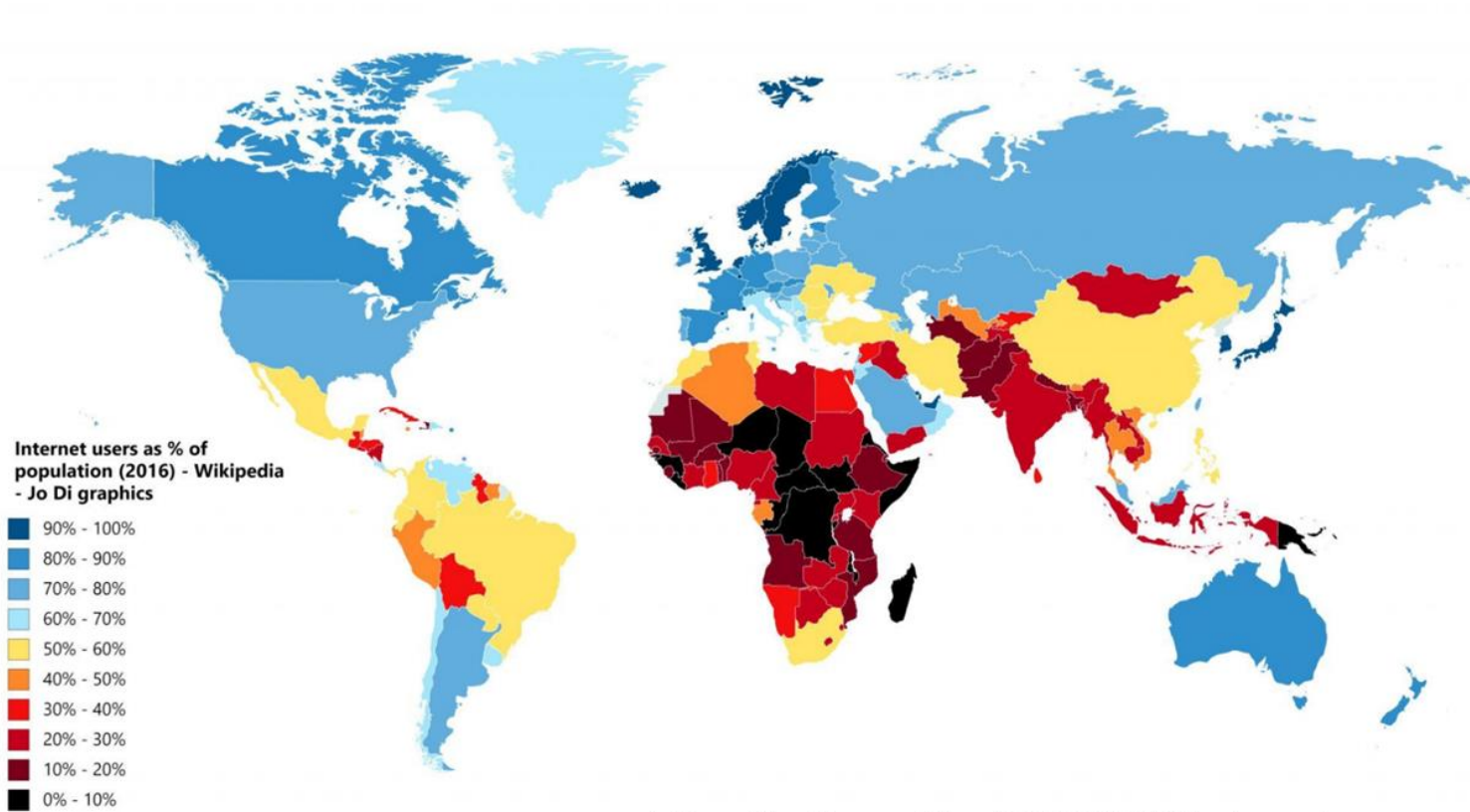
Low resource languages

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



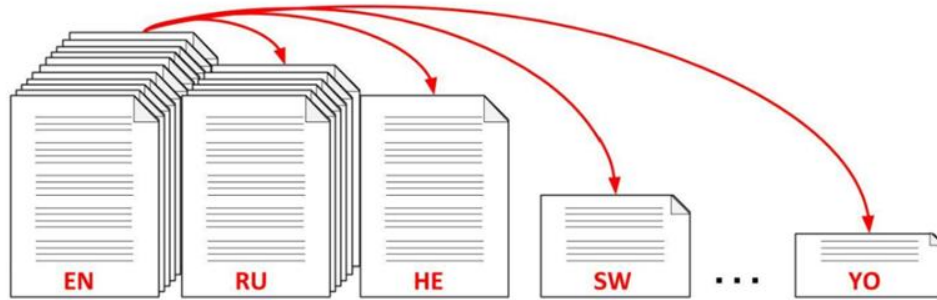
<https://jodi.graphics/2018/05/11/internet-users-as-of-population/>

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

Approaches to low-resource/multilingual NLP

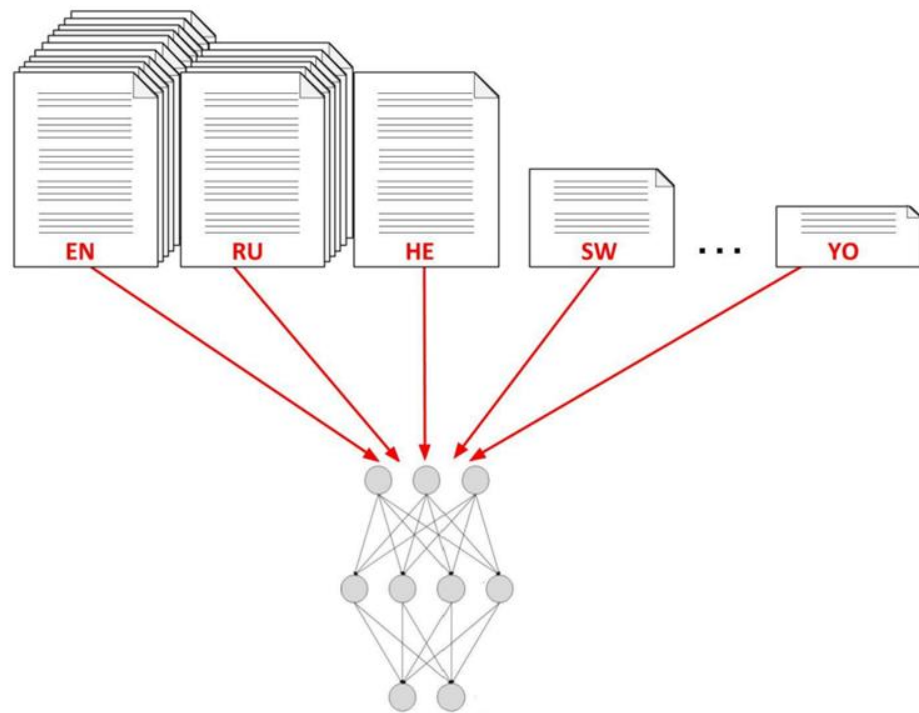
- Manual curation and annotation of large-scale resources for thousands of languages is infeasible or prohibitively expensive
- Unsupervised learning (Snyder and Barzilay 2008; Cohen and Smith, 2009; Snyder, 2010; Vulić, De Smet, and Moens 2011; Spitkovsky et al., 2011; Goldwasser et al., 2011; Titov 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 vs. 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

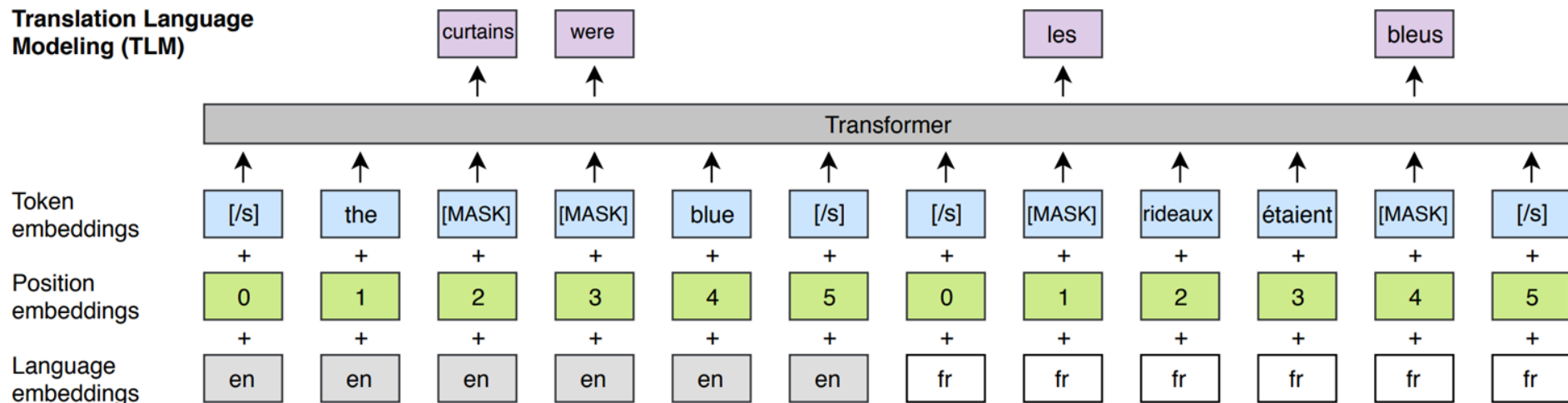
SentencePiece

- WordPiece issues:
 - Individual words to split (*'hello', 'world', '.'*)
 - Does not work for all languages like Chinese
 - Not open-source
- SentencePiece:
 - Directly works on sentences
 - Treats spaces as another character to tokenize

XLM, XLM-Roberta from Facebook

- XLM uses Translation Language Modeling (TLM)

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

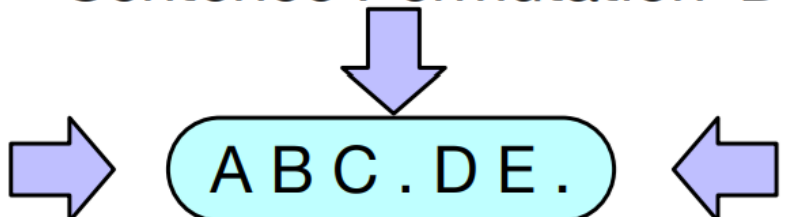
A.C.E.
Token Deletion

DE.ABC.
Sentence Permutation

ABC.DE.

C.DE.AB
Document Rotation

A_.D_E.
Text Infilling



mBART: Multi-Lingual BART

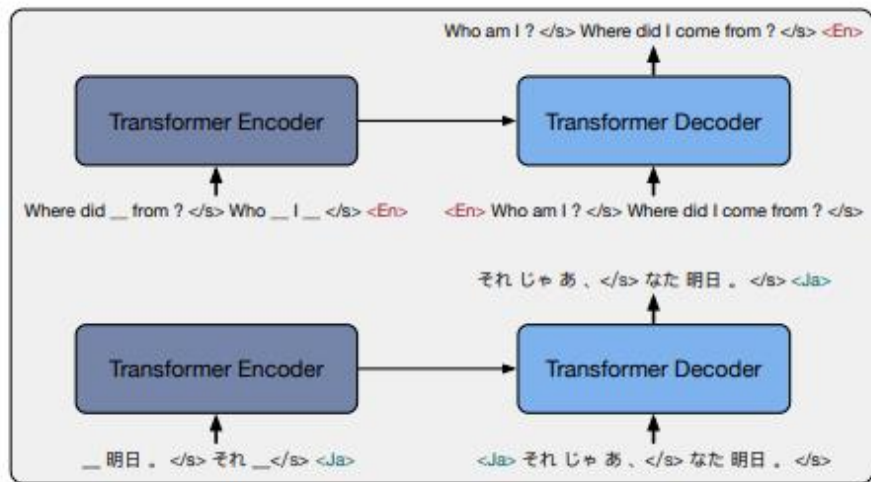
- Training on CC25 corpus
- Corpus of 25 languages
- A subset of *Common Crawl*
- A crawl of the internet

| 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 |

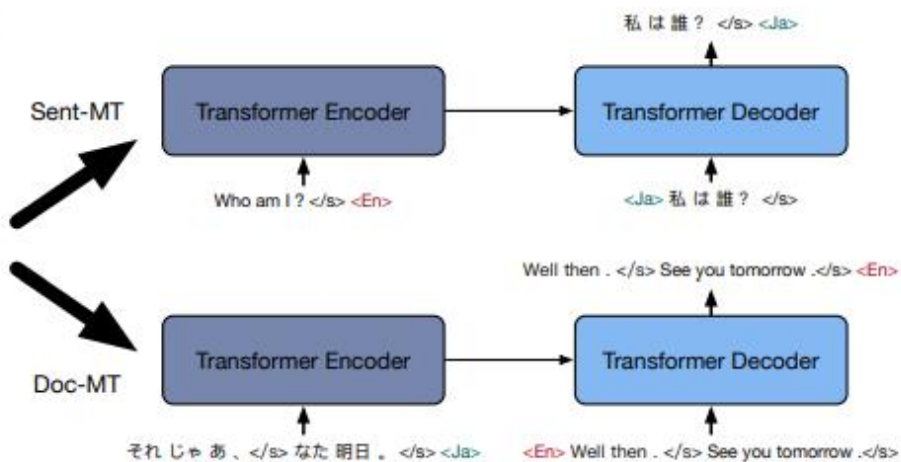
Table 1: **Languages and Statistics of the CC25 Corpus.** A list of 25 languages ranked with monolingual corpus size. Throughout this paper, we replace the language names with their ISO codes for simplicity. (*) Chinese and Japanese corpus are not segmented, so the tokens counts here are sentences counts

mT5

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



Multilingual Denoising **Pre-Training** (mBART)



Fine-tuning on Machine Translation

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

| | |
|-----------------|---------|
| ম##িল##ন | মিলন |
| স##ং##ব##ং##দি | সংবন্দি |
| অ##ন##্বে##ষ##ণ | অন্বেষণ |
| ন##িল##ত | নিলত |
| tu##mh##ara | tumhara |
| و##ال##و | والون |

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

A4 Primer - Hindi and Tamil

Context:

मानव कंकाल शरीर की आन्तरिक संरचना होती है। यह जन्म के समय 300 हड्डियों से बना ... म मजबूत होते हैं। अन्य प्राणियों से भिन्न, मानव पुरुष का लिंग स्तंभास्थि रहित होता है। [2]

Question:

जन्म के समय शिशु के शरीर में कितनी हड्डियाँ होती है?

Answer:

300

A4 Possible Ideas

- Start from multilingual models
- Utilise English QA datasets like SQuAD
 - Train on English + Hindi/Tamil
 - Translate from English to Hindi/Tamil
 - Context/Question

Machine Translation

Be the change you want to see in the world

वह परिवर्तन बनो जो संसार में देखना चाहते हो



Government: administrative requirements, education, security.

Enterprise: product manuals, customer support

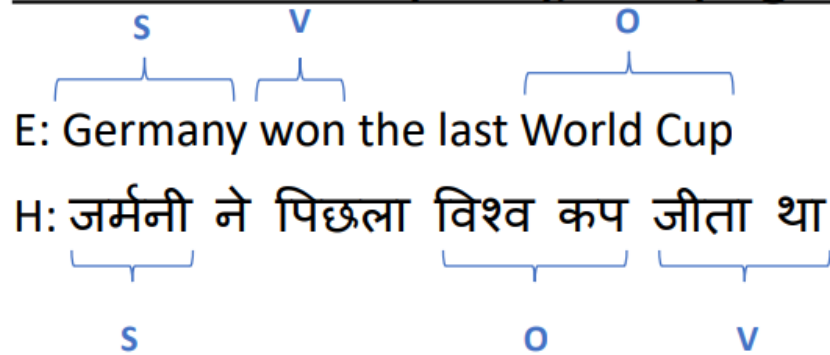
Social: travel (signboards, food), entertainment (books, movies, videos)

Translation under the hood

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

What is Machine Translation?

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



Free (Hindi) vs rigid (English) word order

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

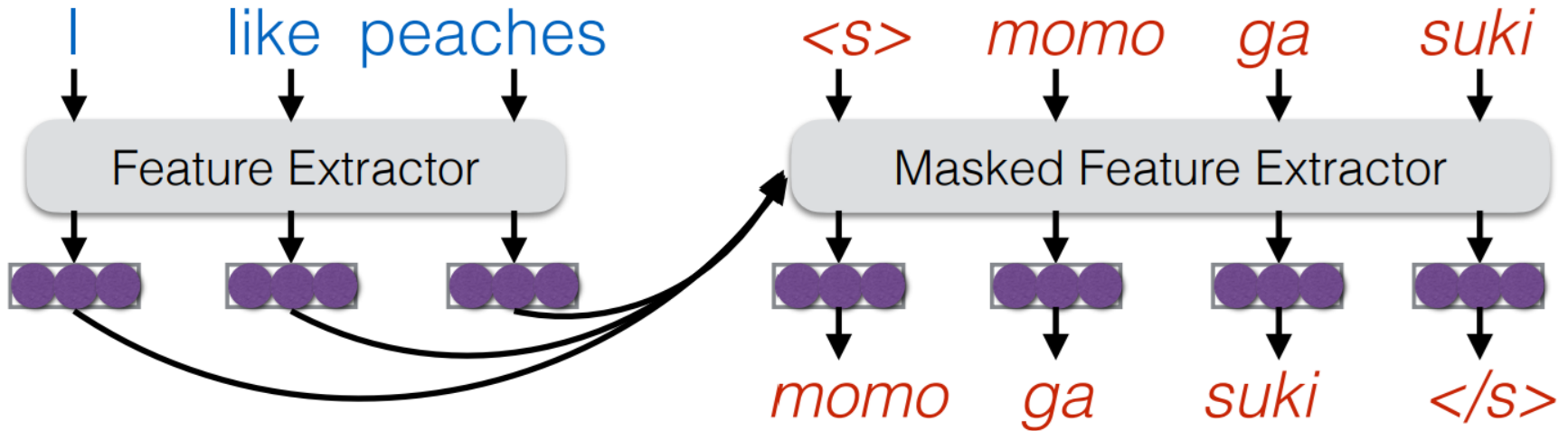
The last World Cup Germany won *(grammatically incorrect)*

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



www.un.org
http://www.un.org/english/

We the peoples

Daily Briefing | Radio, TV, Photo | Documents, Maps | Publications, Stamps, Databases | UN Works | Search
Peace & Security | Economic & Social Development | Human Rights | Humanitarian Affairs | International Law

Welcome to the United Nations

UN Millennium Development Goals
United Nations News Centre
About the United Nations
Main Bodies
Conferences & Events
Member States
General Assembly President

Secretary-General
Situation in Iraq
Mideast Roadmap
Renewing the UN
UN Action against Terrorism
Issues on the UN Agenda
Civil Society / Business
UN Webcast
CyberSchoolBus

8 September 2005 >>

[Home](#) | [Recent Additions](#) | [Employment](#) | [UN Procurement](#) | [Comments](#) | [D & A](#) | [UN System Sites](#) | [Index](#)
عربي | 中文 | **English** | Français | Русский | Español

Copyright, United Nations, 2000-2005 | Use of UN60 Logo | Terms of Use | Privacy Notice | Help
[Text version]

Live and On-Demand Webcasts, 24 Hours a Day. Click on UN Webcast



联合国主页
http://www.un.org/chinese/

我们人民

每日简报 | 多媒体 | 文件与地图 | 出版物 | 邮票 | 数据库 | 服务全球 | 网址搜索
和平与安全 | 经济与社会发展 | 人权 | 人道主义事务 | 国际法

欢迎来到联合国

联合国千年发展目标
联合国新闻
联合国概况
联合国主要机关
会议与活动
联合国会员国
联合国大会主席

联合国秘书长
伊拉克局势
中东路线图
更新联合国
反恐主义
联合国日常议题
民间团体/商业
联合国网络直播
空中校车

联大第60届会议一般性辩论

新增内容 | 工作机会 | 联合国采购 | 建议 | 问题与解答 | 其他网址 | 网址索引
عربي | **中文** | English | Français | Русский | Español

联合国2000-2005年版权 | 联合国80周年徽标使用准则 | 使用条件 | 隐私通告 | 帮助
[纯文字版]

联合国实况直播



... 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:

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

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: New: [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](#)

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 Government ([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](#) - 21 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?

Google Translate

Text Documents

ENGLISH - DETECTED HEBREW ENGLISH SWAHILI

KOREAN ENGLISH HINDI

This cat is cute. Her name is Latte.

이 고양이는 귀엽다. 그녀의 이름은 라떼입니다.

i goyang-ineun gwiyeobda. geunyeoui ileum-eun latteibnida.

37/5000

Send feedback

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

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

- 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