MUNI FI

Low-Resource Machine Translation

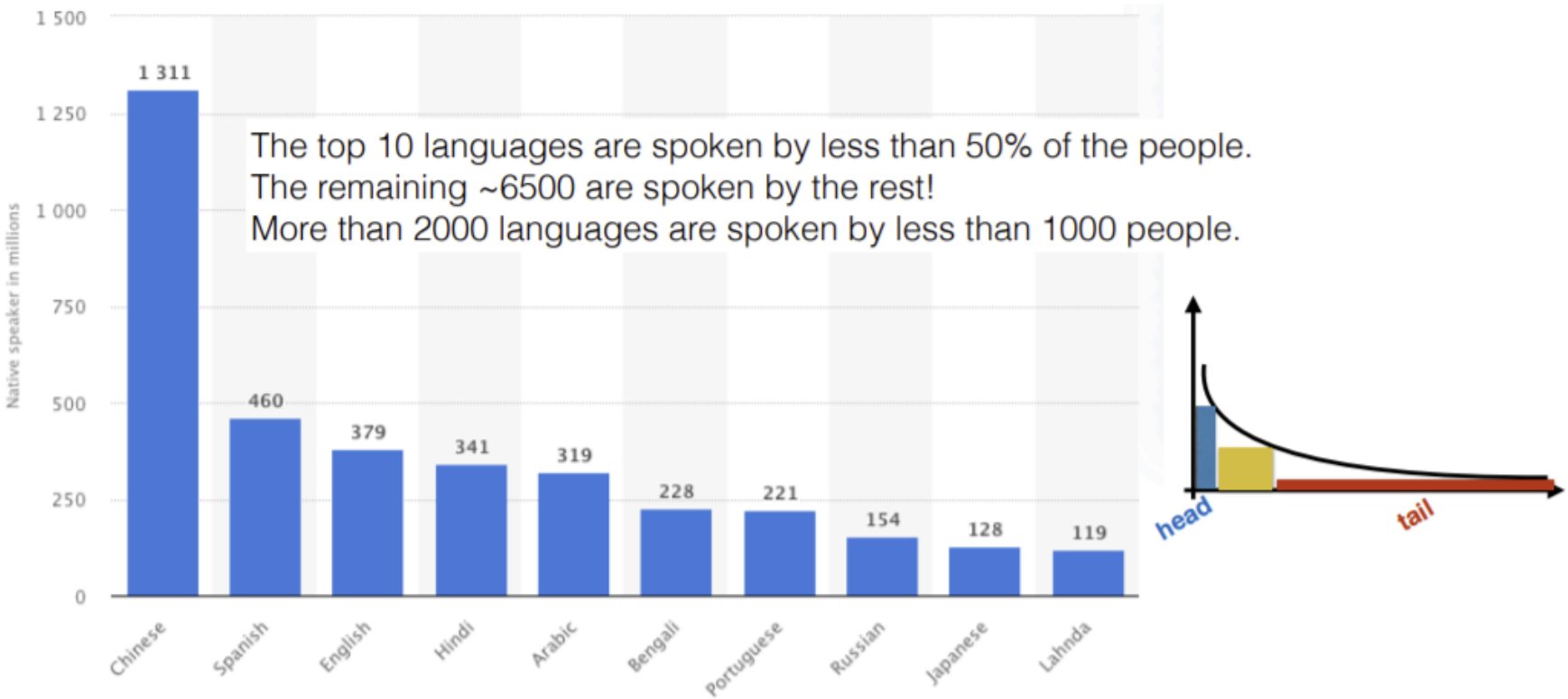
PV061

Pavel Rychlý

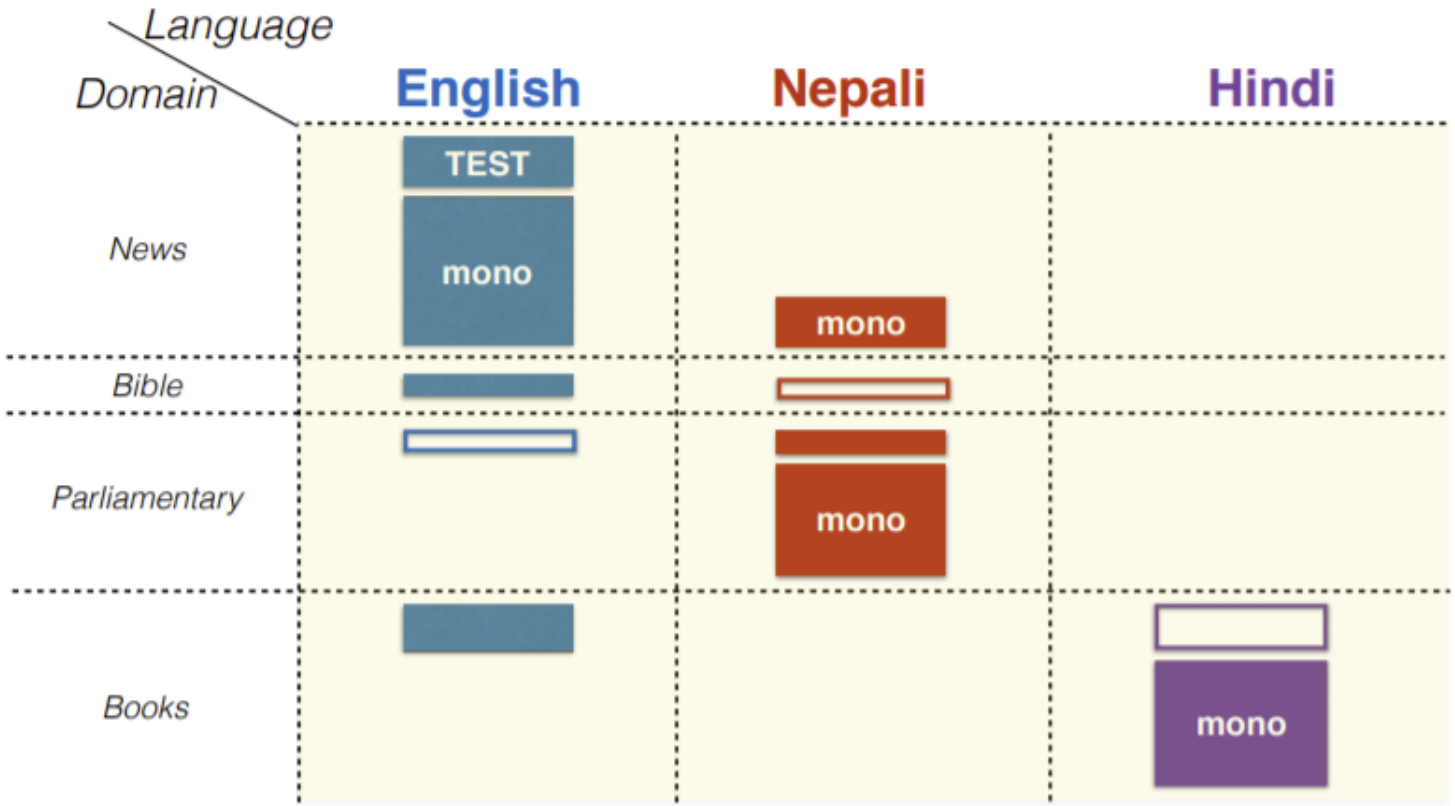
NLP Centre, FI MU

20 Sep 2023

Low-Resource Languages



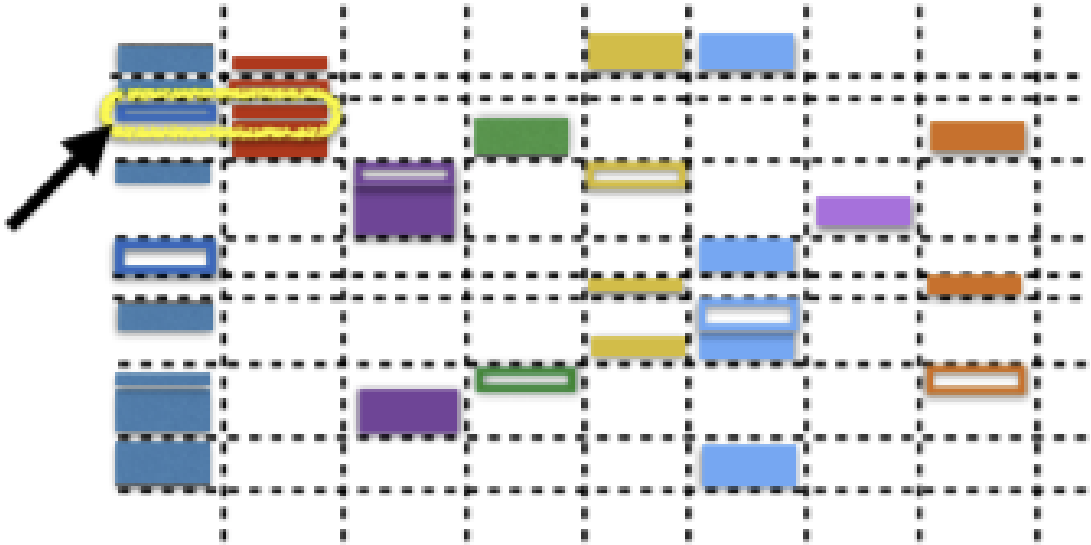
Parallel Data in Practice



- There might be parallel and monolingual data with a high resource language close to the low resource language of interest. This data may belong to a different domain.

Finding data

Low-resource MT is about **large**-scale learning!



General ML Tip: whenever you lack supervised data (typical case), come up with auxiliary tasks or even fantasize it.

Data Augmentation

- word or phrase replacement
- back-translation (iterative back-translation)
- reducing noise
  - monolingual data selection
  - synthetic parallel data filtering
  - distinguishing between original and back-translated data

Unsupervised MT

- alignment of monolingual embeddings
  - generating word translations
- multilingual embeddings
- fine-tuning on monolingual data

Multilingual MT

- single encoder-decoder for all the languages
  - single target
  - multiple target langs with annotation
    - target lang ID in source
    - target lang ID as first token in target
- per-language encoder-decoder
- single encoder with per-language decoder
- per-language encoder with single decoder

Zero-shot NMT

- pivoting
  - pivot language (English)
- many-to-many multilingual MT

Evaluation

- Pretrained models (COMET) not supported
- ChrF++
- datasets: FLoRes Evaluation Benchmark
  - FLORES-101 (3000 English sentences)
  - FLOREW-200
  - Open Language Data Initiative (OLDI)

WMT Shared Tasks

Shared Tasks are the main part of the Conference on Machine Translation (WMT).

2023: http://www2.statmt.org/wmt23/index.html

- general translation task (former News task),
- terminology translation task,
- literary translation task,
- word-level autocompletion task,
- sign language translation task,
- biomedical translation task,
- indic translation task,
- african translation task,
- metrics evaluation task.

Summary 1

- statistical MT
  - IBM Model 1
  - language modelling
  - phrase-based models
  - decoding
- neural MT
  - language modelling, RNN
  - RNN MT
  - attention
  - decoding

Summary 2

- transformers
- subwords
- evaluation
- data acquisition

Next semester

- PA107 Corpus Tools Project
  - WMT 2024 Shared Tasks

MUNI
FACULTY
OF INFORMATICS