CS11-737 Multilingual NLP

Multilingual Training and Cross-lingual Transfer

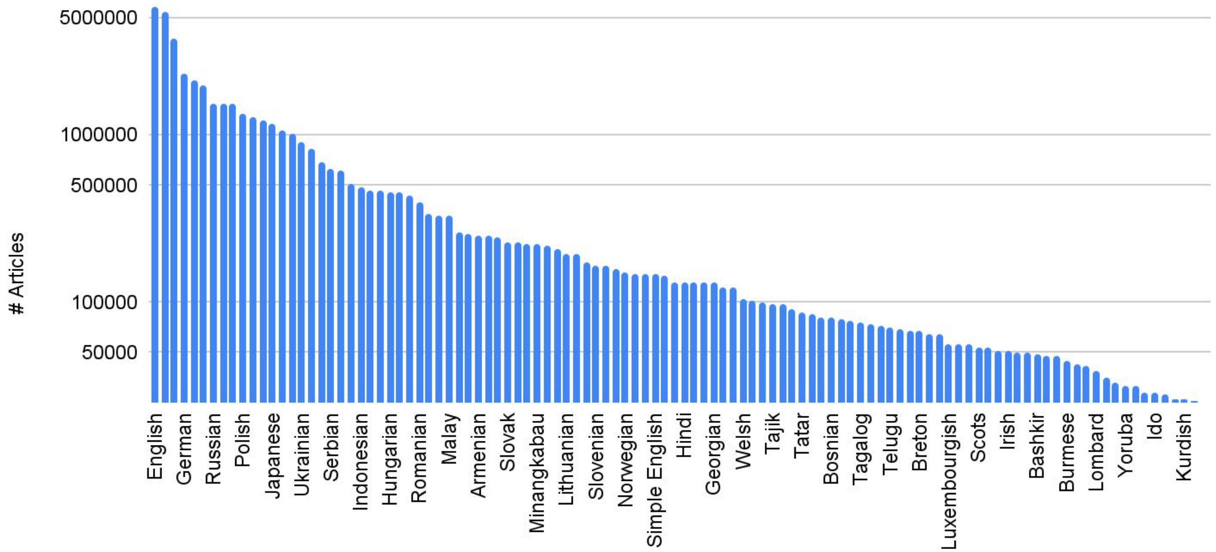
Patrick Fernandes

http://phontron.com/class/multiling2022/

Many slides adapted from Graham Neubig & Xinyi Wang

Carnegie Mellon University
Language Technologies Institute

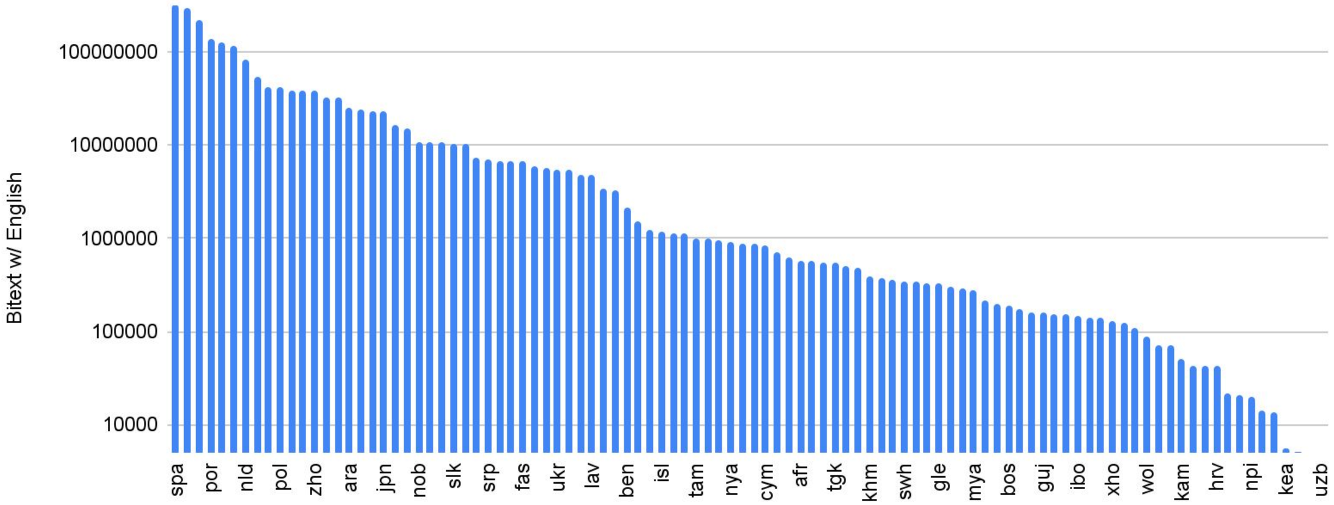
Many languages are left behind



➔ There is not enough *monolingual* data for many languages

Data Source: Wikipedia articles from different languages

Many languages are left behind



➔ The problem is even worst for *annotated* data

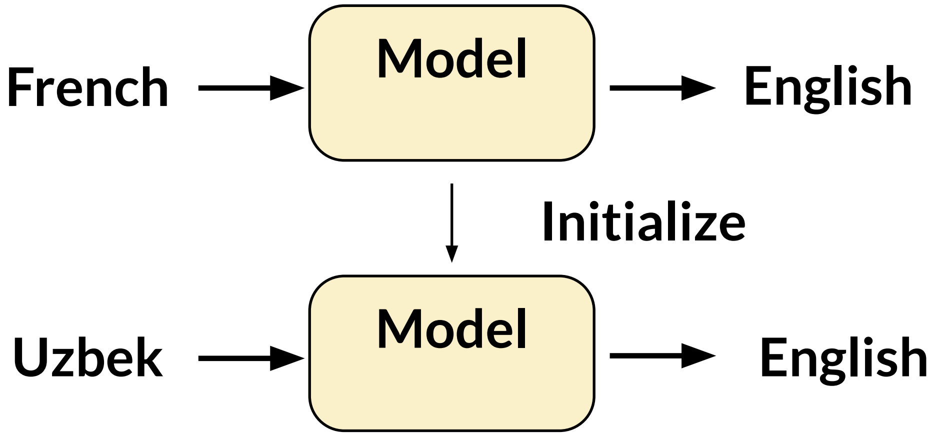
Data Source: FLORES-101 paper

Tackling *data-scarcity* problem in multilingual NLP

➔ How to effectively train (data-hungry) NLP models for low-resource languages?

- Cross-lingual transfer
- Multilingual Training

Cross-lingual Transfer



- ➔ Train a model on high-resource language
- ➔ Finetune on small low-resource language

Transfer learning for low-resource neural machine translation. Zoph et. al. 2016

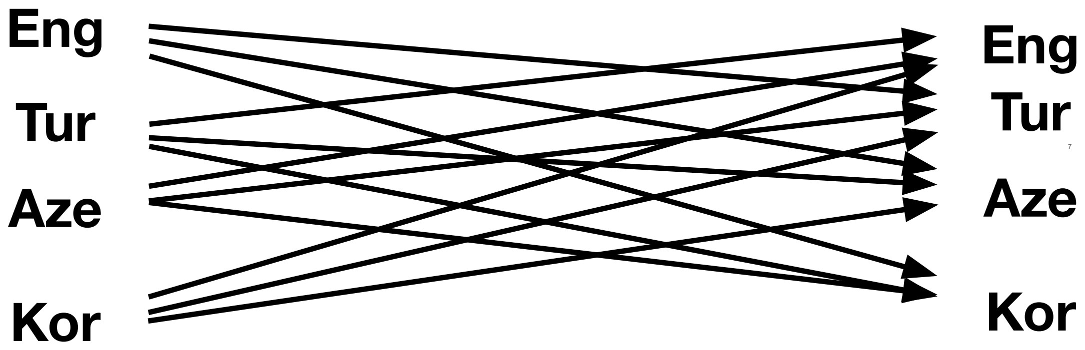
Supporting multiple languages can be difficult

➔ Cross-lingual transfer can be effective for small number of languages

➔ However, supporting a moderately large number of languages is problematic

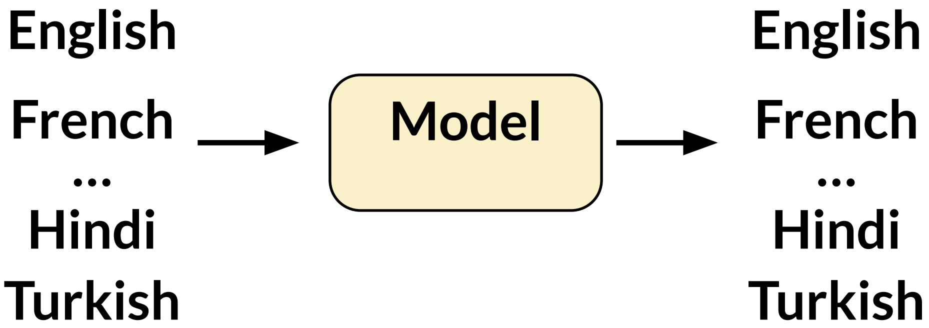
Transfer learning for low-resource neural machine translation. Zoph et. al. 2016

Supporting multiple languages can be difficult



➔ Supporting for 4 languages in all directions requires 12 NMT models!

Multilingual Training



➔ Training a single model on a mixed dataset from multiple languages

Google's multilingual neural machine translation system . Johnson et. al. 2016

Multilingual Training

<2fr> How are you? → **Model** → **Comment ça va?**

<2es> How are you? → **cómo estás?**

… → **…**

<2tr> How are you? → **nasılsın?**

➔ To specify target language, simply add a language tag!

Google's multilingual neural machine translation system . Johnson et. al. 2016

Combining both methods

- ➔ We just covered the two main paradigms for multilingual methods
  - ◆ Cross-lingual transfer
  - ◆ Multilingual training
- ➔ What’s the best way to use the two to train a good model for a new language?

Example use case: COVID-19 response

In a pandemic, the challenge isn't just translating one or a handful of primary languages in a single region—it's on a scale of perhaps thousands of languages. PHOTO-ILLUSTRATION: SAM WHITNEY; GETTY IMAGES

GRETCHEN MCCULLOCH IDEAS 05.31.2020 07:00 AM

Covid-19 Is History's Biggest Translation Challenge

Services like Google Translate support only 100 languages, give or take. What about the thousands of other languages—spoken by people just as vulnerable to this crisis?

➔ Quickly translate covid-19 related info for speakers of various languages

Example use case: COVID-19 response

Translation Initiative for COVID-19

Providing machine-readable translation data related to the COVID-19 pandemic

In response to the on-going crisis, several academic (Carnegie Mellon University, Johns Hopkins University) and industry (Amazon, Appen, Facebook, Google, Microsoft, Translated) partners have partnered with the Translators without Borders to prepare COVID-19 materials for a variety of the world's languages to be used by professional translators and for training state-of-the-art Machine Translation (MT) models. The focus is on making emergency and crisis-related content available in as many languages as possible. The collected, curated and translated content across nearly 90 languages will be available to the professional translation as well the MT research community.

To this end, we have so far created:

Translation Memories for the Translation Community

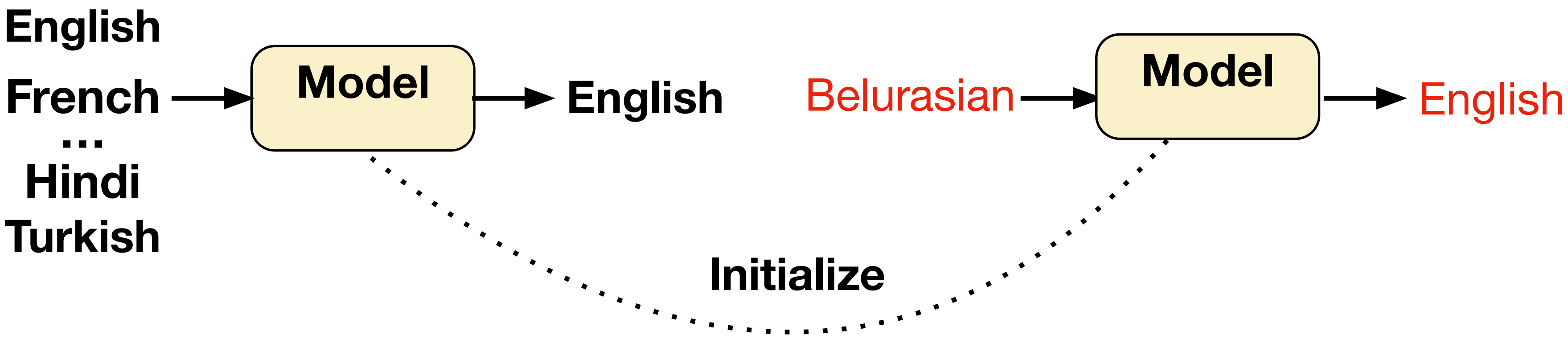
We have combined the terminologies and other translation data to create translation memories in .tmx format for the majority of the language pairs.

- Additional details and data download here.

Translated Terminologies

➔ Quickly translate covid-19 related info for speakers of various languages

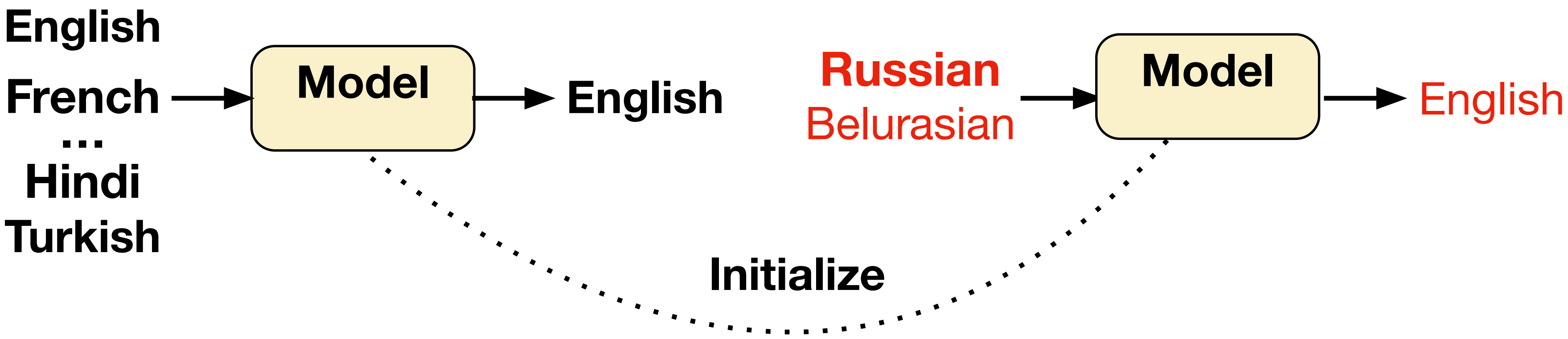
Rapid adaptation of massive multilingual models



- ➔ First, do multilingual training on many languages
- ➔ Next fine-tune the model on a new low-resource language

Rapid adaptation of Neural Machine Translation to New Languages. Neubig et. al. 2018

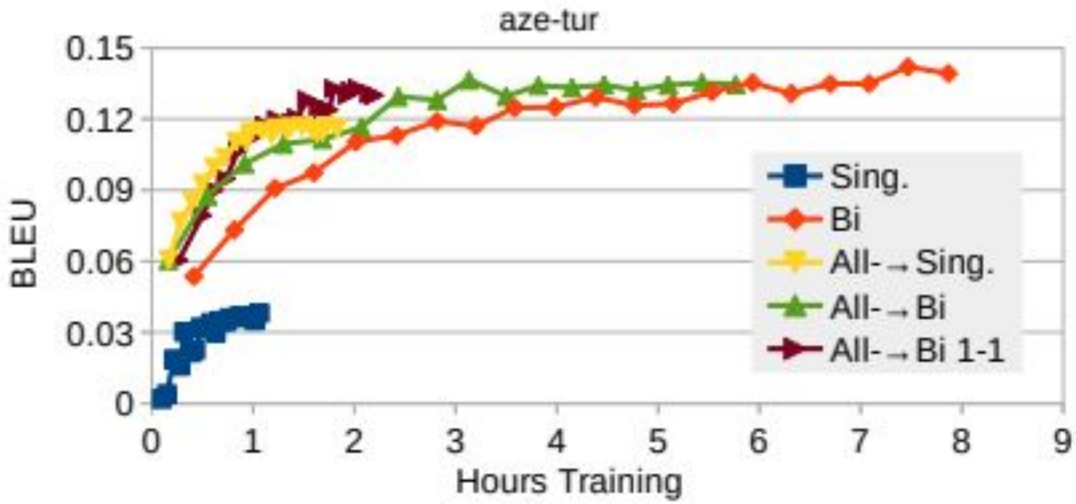
Rapid adaptation of massive multilingual models



- Regularized fine-tuning:
  - Fine-tune on low-resource language and a related high-resource one to avoid overfitting

Rapid adaptation of Neural Machine Translation to New Languages. Neubig et. al. 2018

Rapid adaptation of massive multilingual models



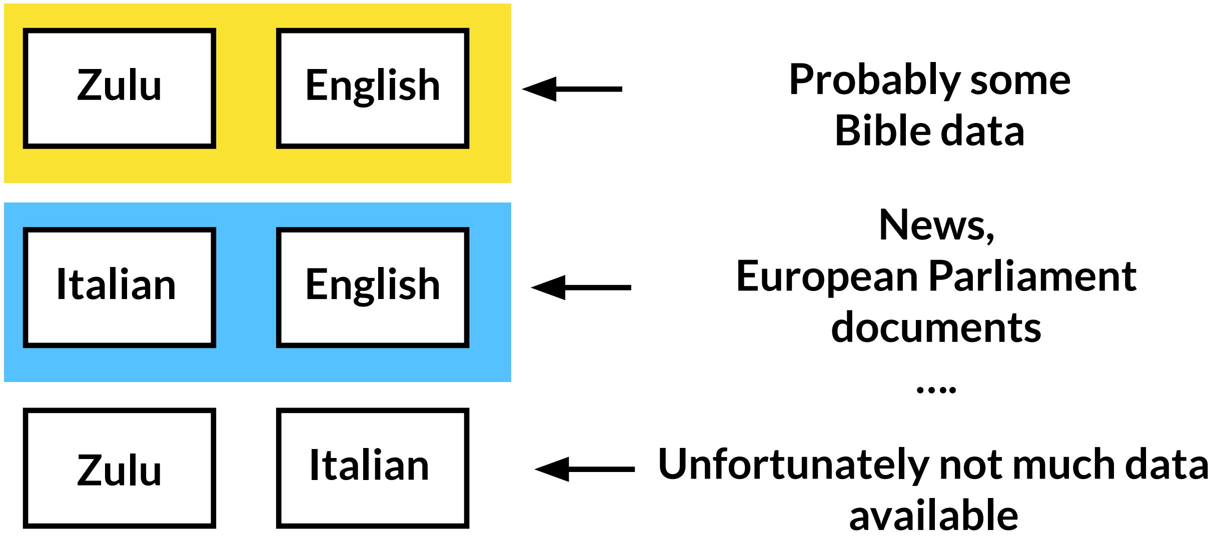
- ➔ All- -> xx models: adapting from a multilingual makes convergence faster
- ➔ Regularized fine-tuning leads to better final performance

Rapid adaptation of Neural Machine Translation to New Languages. Neubig et. al. 2018

Data-scarcity taken to the extreme: zero-shot transfer

- ➔ Suppose that we **no** data for a language (pair) of interest
- ➔ Can we leverage multilingual training to do **zero-shot** transfer to this language (pair)?

Zero-shot transfer in NMT



Zero-shot transfer in NMT

Training

<2en> Zulu-English src
<2en> Italian-English src → **Model** →
<2it> English-Italian src

Zulu-English tgt
Italian-English tgt
English-Italian tgt

Testing

<2it> Sawubona → **Model** → **Ciao**

➔ Multilingual training allows zero shot transfer!

- Train on {Zulu-English, English-Zulu, English-Italian, Italian-English}
- The model is able to translate **Zulu-Italian** *without* any parallel data

Google's multilingual neural machine translation system . Johnson et. al. 2016

Data-scarcity taken to the extreme: zero-shot transfer

- ➔ Suppose that we **no** data for a language (pair) of interest
- ➔ Can we leverage multilingual training to do **zero-shot** transfer to this language (pair)?
- ➔ Can we improve **zero-shot** transfer with monolingual/unlabeled data ?

Zero-shot transfer in NMT with monolingual data

Training

<2en> Zulu-English src → **Model** → **Zulu-English tgt**

<2it> add_noise(Italian tgt) → **Model** → **Italian tgt**

Testing

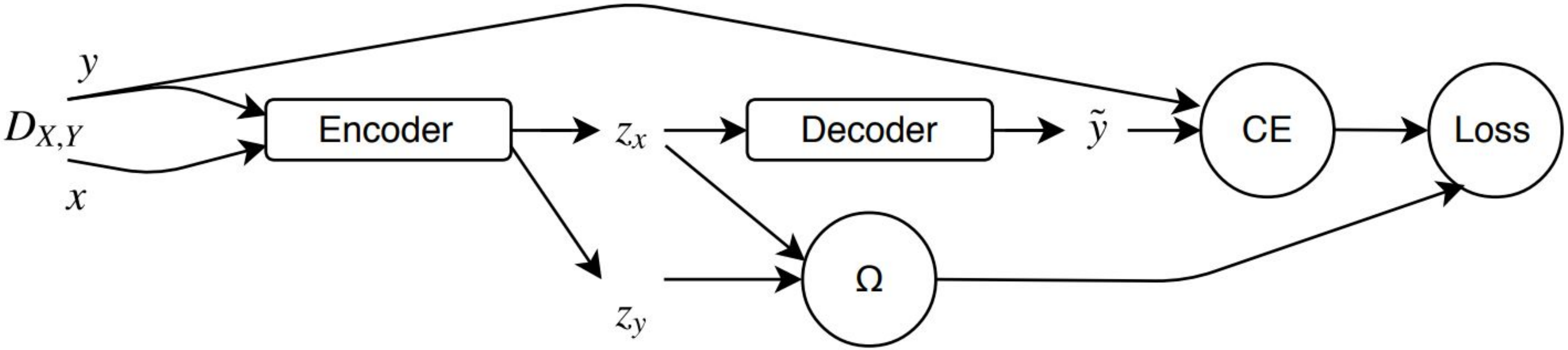
<2it> Sawubona → **Model** → **Ciao**

➔ Add a *denoising* objective for the monolingual data

➔ Masked Language Modeling, Denoising Autoencoder

Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation . Siddhant et. al. 2019
Multilingual Translation from Denoising Pre-Training . Tang et. al. 2020

Zero-shot transfer in NMT with aligned representations



➔ Translation objective alone might not encourage language-invariant representation

➔ Add an extra supervision to align source and target encoder representation

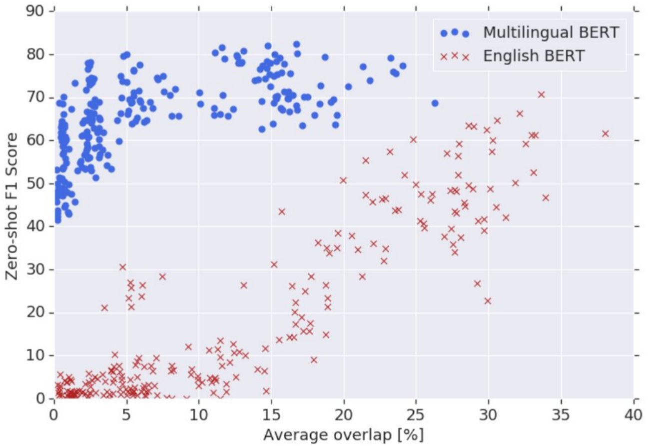
The missing ingredient in zero-shot Neural Machine Translation . Arivazhagan et. al. 2019

Zero-shot transfer in pretrained language models

- ➔ Zero-shot transfer also works for multilingual (masked) language models
  - ◆ Pretrain model on monolingual data from languages
  - ◆ Finetune model on annotated data in one language for downstream task
  - ◆ Test the finetune model in the same task on a **different** language
- ➔ Multilingual models learn language-invariant representations

How multilingual is multilingual BERT? Pires et. al. 2019

Zero-shot transfer in pretrained language models



➔ Generalize to language with different scripts:
  - transfer well to languages with little vocabulary overlap

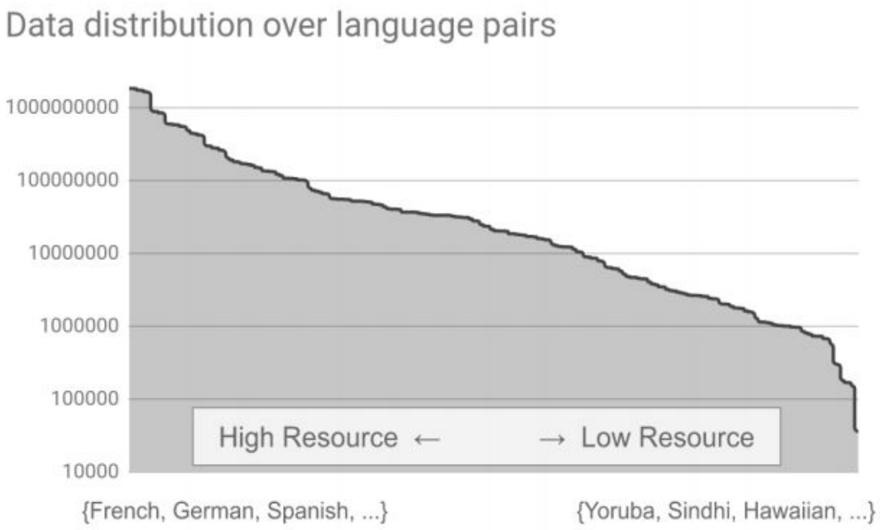
➔ Does not work well for typologically different languages:
  - Eg: fine-tune in English, test on Japanese

How multilingual is multilingual BERT? Pires et. al. 2019

Open problems in multilingual training

- ➔ Despite their benefits in the low-resource setting, multilingual training still has problems
- ➔ Consider the problem of scaling multilingual MT to >100 languages

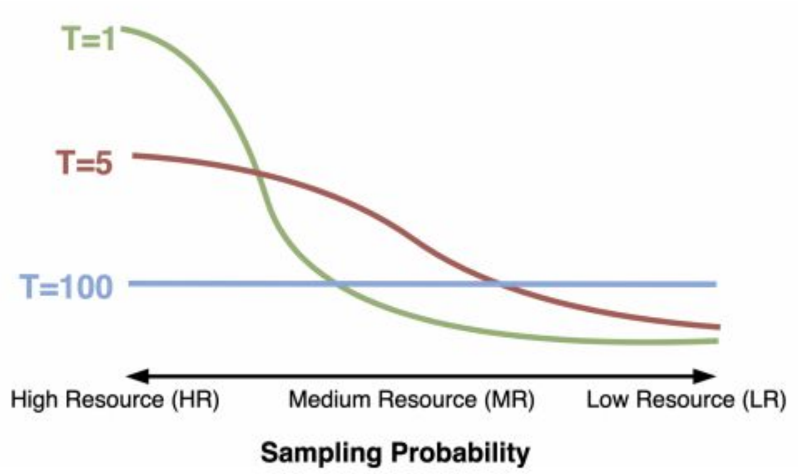
Problem: training data highly imbalanced



- ➔ High resource languages have much more data than low-resource ones
- ➔ Important to upsample low-resource data in this setting!

Massively Multilingual Neural Machine Translation in the Wild. Arivazhagan et. al. 2019

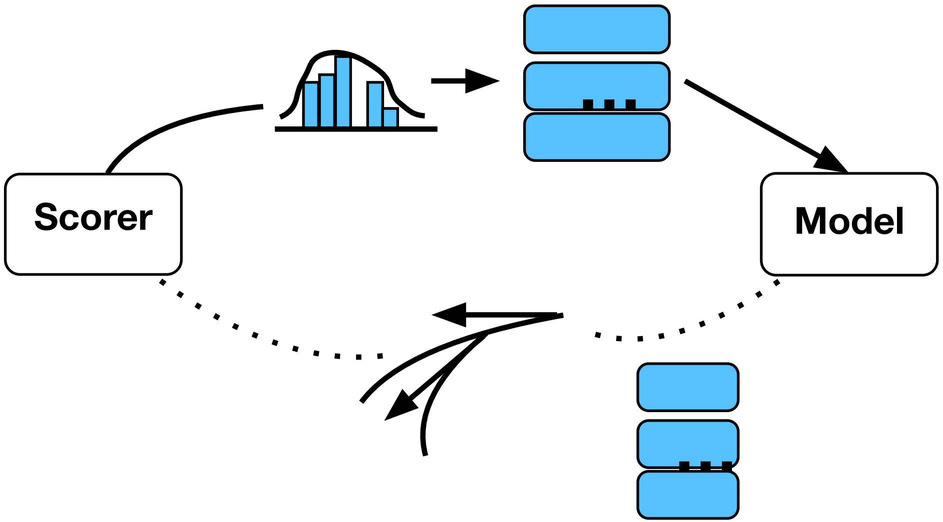
Problem: training data highly imbalanced



- ➔ Sample data based on dataset size scaled by a temperature term
- ➔ Easy control of how much to upsample low-resource data

Massively Multilingual Neural Machine Translation in the Wild. Arivazhagan et. al. 2019

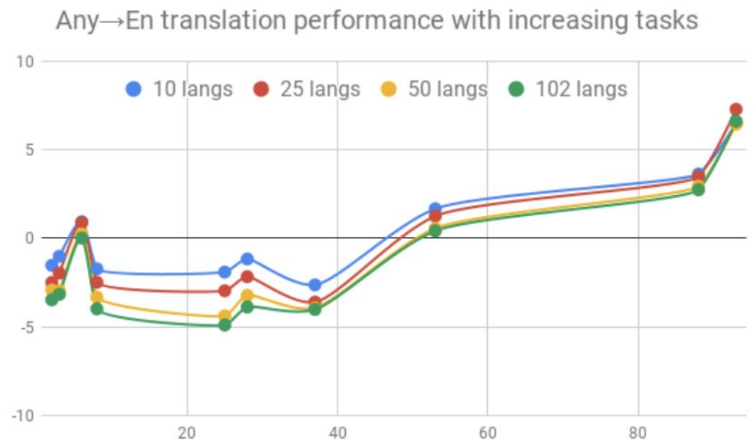
Learning to balance data



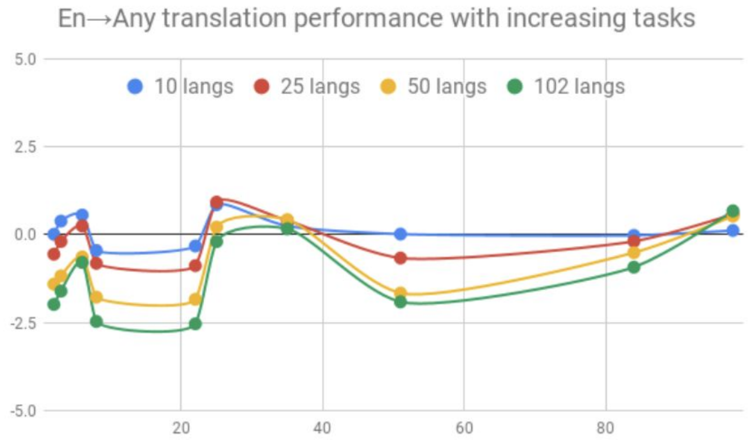
- ➔ Optimize the data sampling distribution during training
- ➔ Upweight languages that have similar gradient with the multilingual dev set

Balancing Training for multilingual neural machine translation. Wang et. al. 2020

Problem: underperforms bilingual models



High-resource Low-resource



High-resource Low-resource

➔ Multilingual training degrades high-resource languages' performance

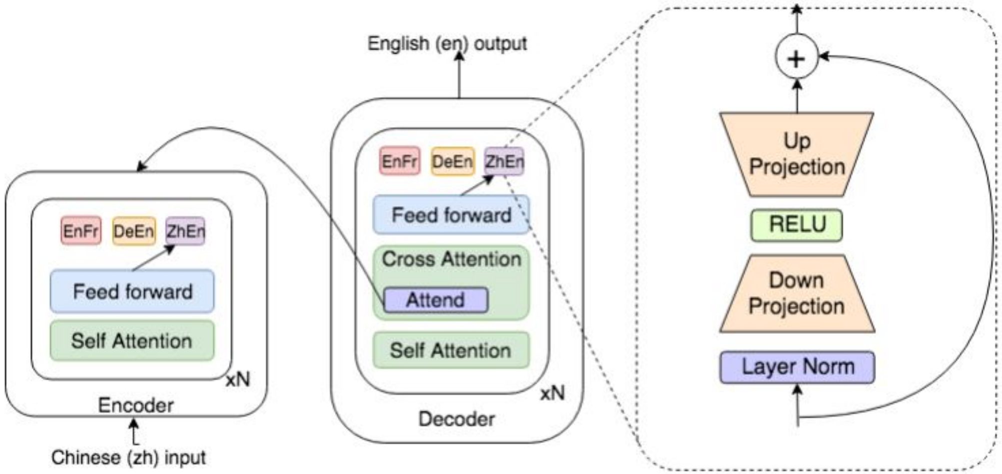
➔ **One-to-many** settings is *much harder*

Massively Multilingual Neural Machine Translation in the Wild. Arivazhagan et. al. 2019

Problem: underperforms bilingual models

- Multiple hypothesis for this phenomena:
  - *Interference* in learning between languages

Adding language-specific layers



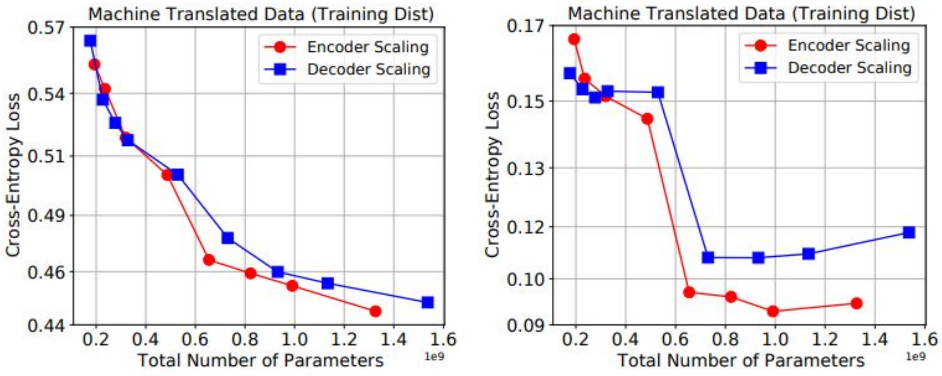
- ➔ Add a small module for each language pair
- ➔ Much better at matching bilingual baseline for high-resource languages

Problem: underperforms bilingual models

- Multiple hypothesis for these phenomena:
  - *Interference* in learning between languages
  - Language generation/decoding is inherently harder than encoding

The difficulty in decoding

- ➔ Self-supervision approaches are more effective with *source* data vs *target* data
- ➔ *Synthetic* data is more effective if the *target* side is natural



***back*-translated** ***forward*-translated**

Scaling Laws for Neural Machine Translation. Ghorbani et. al. 2021

Glimmers of hope for universal multilingual models

- ➔ Back-translation is somewhat effective at improving decoding
- ➔ In WMT21, a massive multilingual model beat (almost) all bilingual models

Comparative Performance vs. Best Entry

MODEL	CZECH	GERMAN	HAUSA	ICELANDIC	JAPANESE	RUSSIAN	CHINESE
FROM ENGLISH							
Our Submission	36.1	31.3	20.1	33.3	46.8	46.0	49.9
WMT2021 Best	33.6	31.3	20.4	30.6	46.9	45.0	49.2
Difference	+2.5	+0.0	-0.3	+2.7	-0.1	+1.0	+0.7
TO ENGLISH							
Our Submission	43.5	53.3	21.0	41.7	27.7	57.1	32.1
WMT2021 Best	43.1	53.0	18.8	40.6	27.8	56.3	33.4
Difference	+0.4	+0.3	+2.1	+1.1	-0.1	+0.8	-1.3

https://ai.facebook.com/blog/the-first-ever-multilingual-model-to-win-wmt-beating-out-bilingual-models/

Problem: multilingual evaluation

- How to evaluate the multilingual model?
  - Average BLEU for all languages? But how to choose between (en-fr: 40, en-zu: 15) vs. (en-fr: 35, en-zu: 20)
  - Is BLEU score between two languages comparable? Does +5 BLEU on en-zu have the same “benefit” as +5 BLEU on en-fr?

Discussion question

- ➔ Read *"Towards the Next 1000 Languages in Multilingual Machine Translation: Exploring the Synergy Between Supervised and Self-Supervised Learning"* (https://arxiv.org/abs/2201.03110)
- ➔ This paper presents a variety of empirical results around the impacts of monolingual and parallel data on multilingual and zero-shot NMT at scale.
  - Choose one (or more) that you found interesting and discuss its implications for building practical systems.
  - Think of an experiment that was not in the paper that you would have liked to see and explain what you hypothesize the result might be.