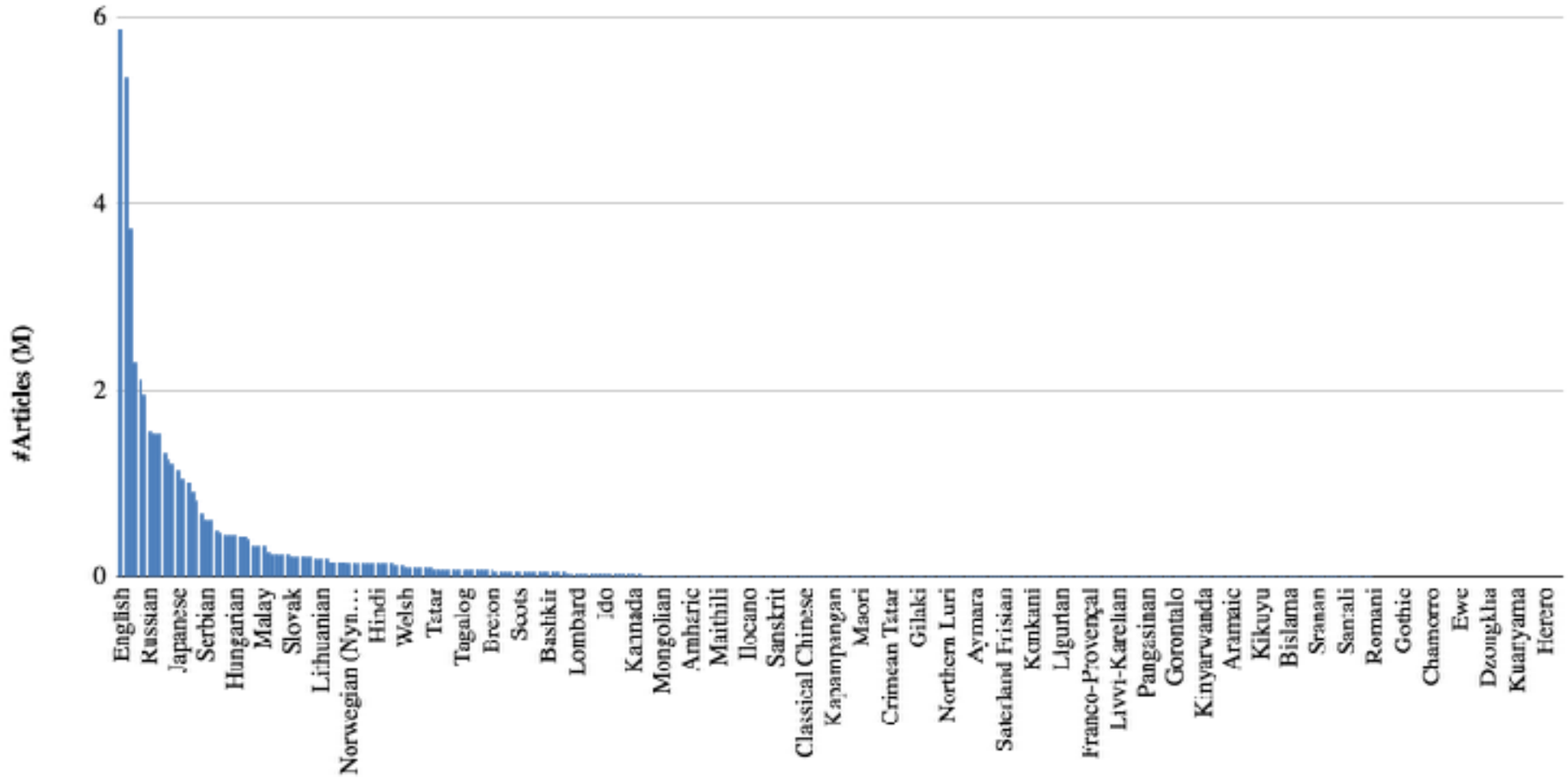
CMU CS11-737: Multilingual NLP

Multilingual Training and Cross-lingual Transfer

Xinyi Wang

Many languages are left behind



- There is not enough monolingual data for many languages
- Even less annotated data for NMT, sequence label, dialogue…

Data Source: Wikipedia articles from different languages

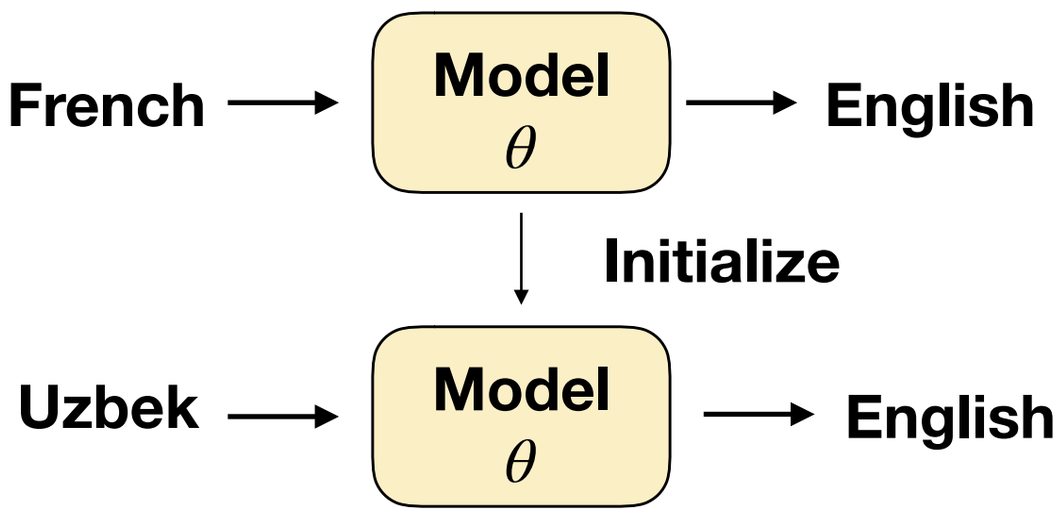
Roadmap

- Two methods: cross-lingual transfer and multilingual training
- Zero-shot transfer
- Open problems with multilingual training

Roadmap

- Two methods: cross-lingual transfer and multilingual training
- Zero-shot transfer
- Open problems with multilingual training

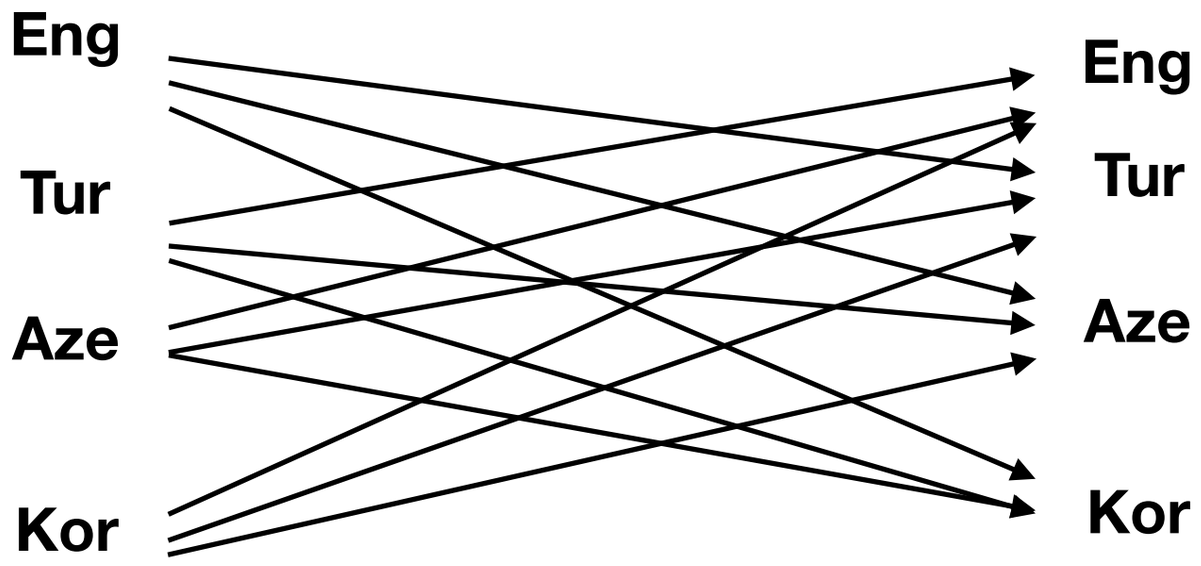
Cross-lingual transfer



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

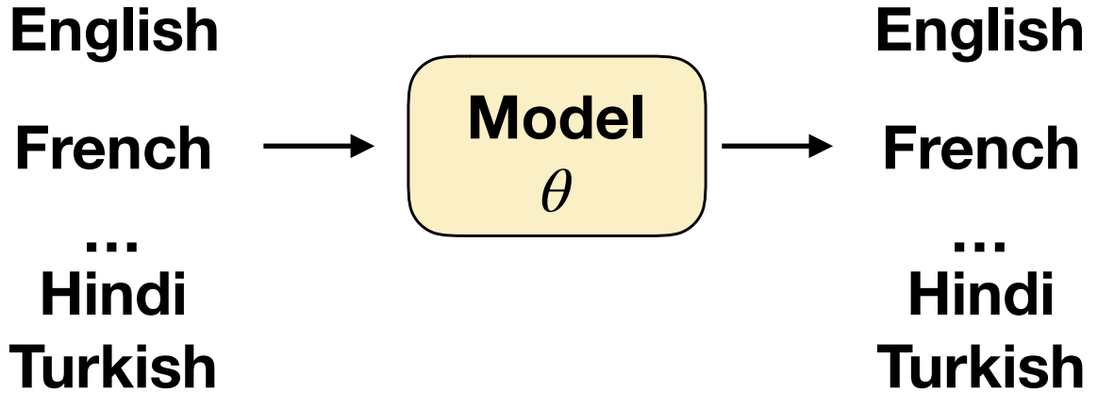
Transfer learning for low-resource neural machine translation. Zoph et. al. 2016

Supporting multiple languages could be tedious



- Supporting just translating from 4 to 4 languages require 4*3=12 NMT models

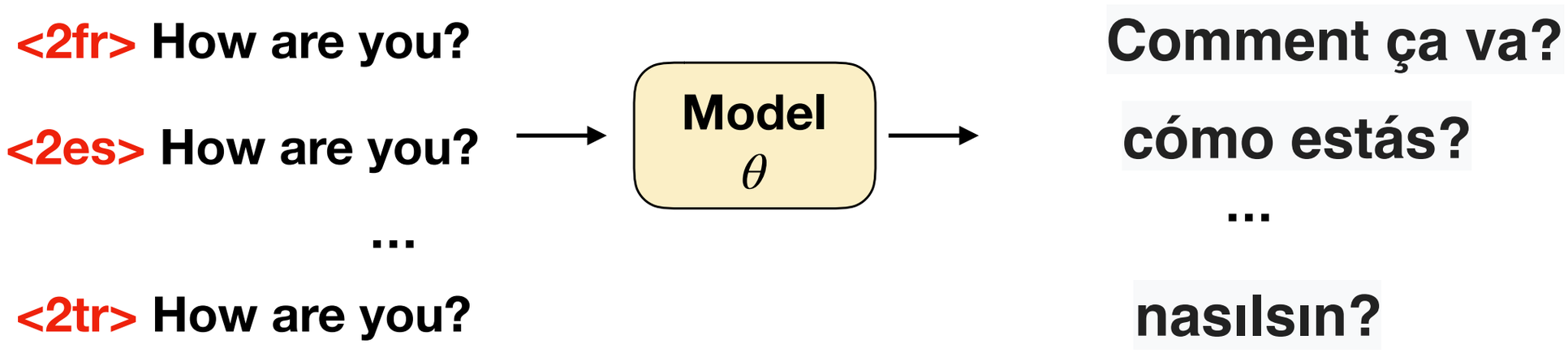
Multilingual training



- Training a single model on a mixed dataset from multiple languages (eg. ~5 languages in the paper)

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

Multilingual training



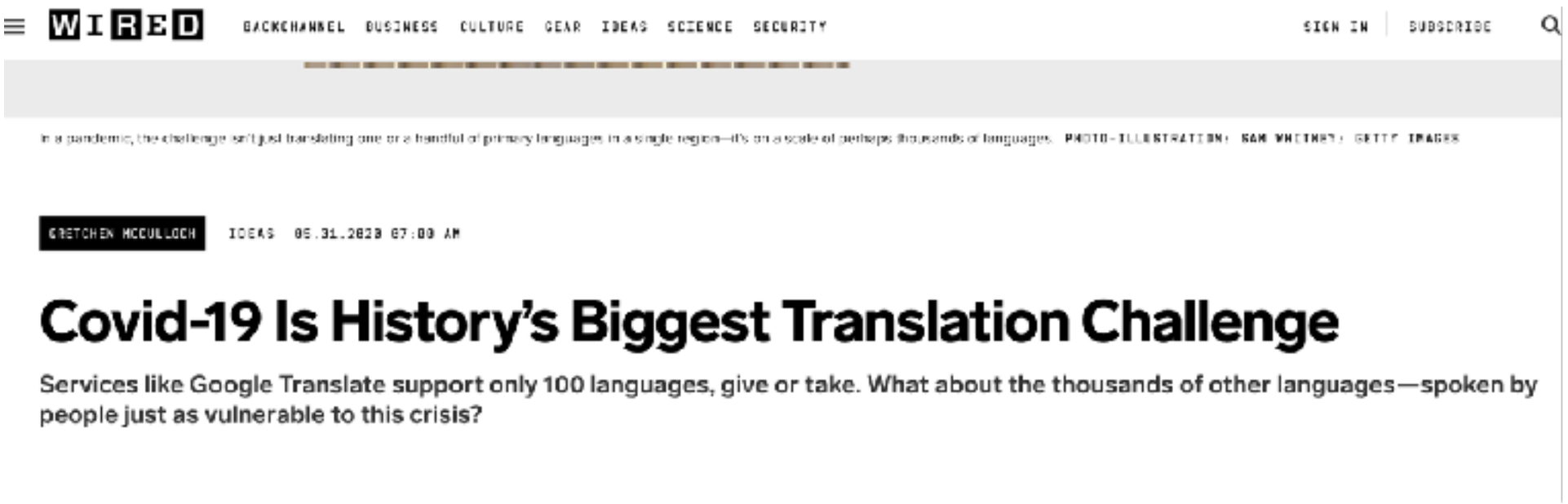
- NMT needs to generate into many languages, simply add target language label

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

Combining the two 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?

Use case: covid-19 response



WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN | SUBSCRIBE

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 06.31.2020 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

https://www.wired.com/story/covid-language-translation-problem/

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.

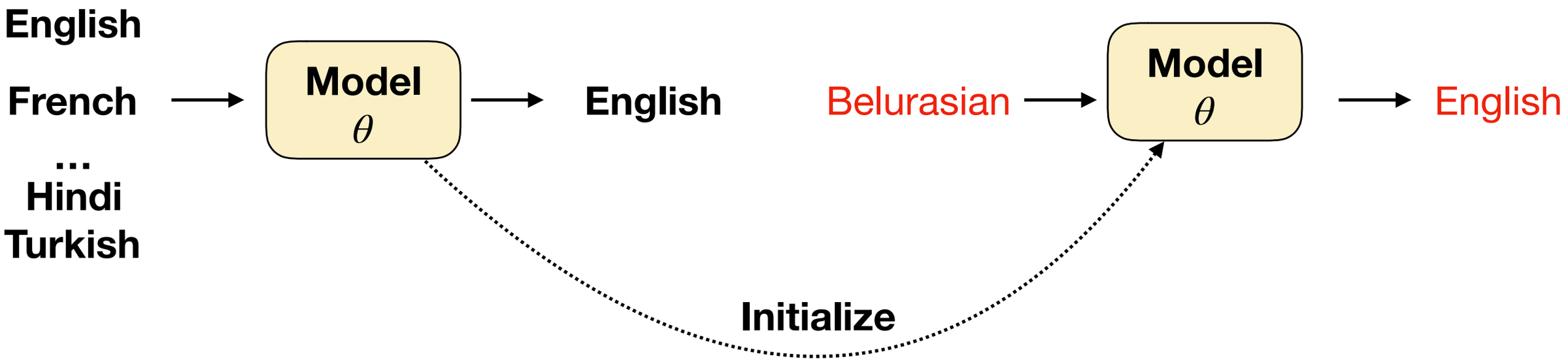
- Additional details and data download here.

Translated Terminologies

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

https://tico-19.github.io/l

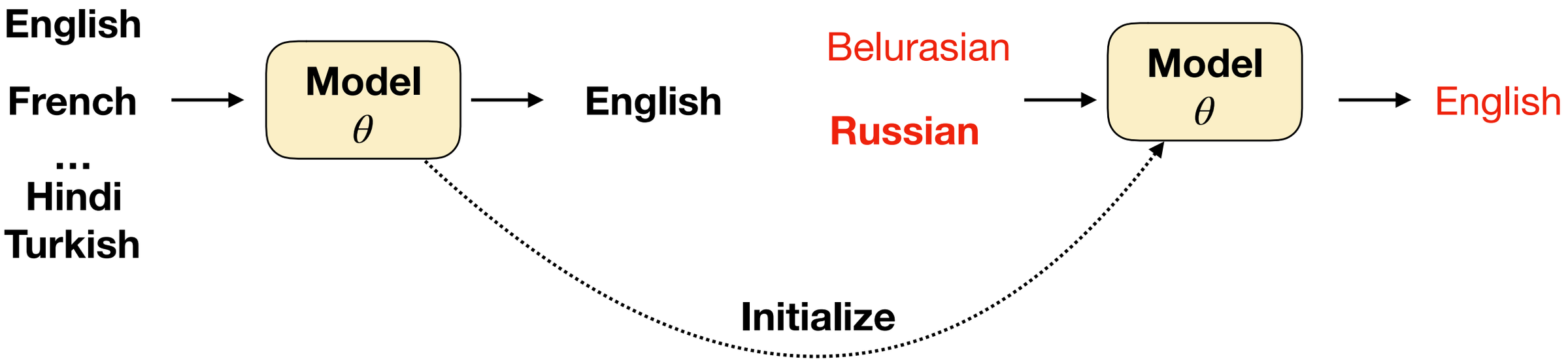
Rapid adaptation of massive multilingual models



- First, do multilingual training on many languages (eg. 58 languages in the paper)
- Next fine-tune the model on a new low-resource language

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

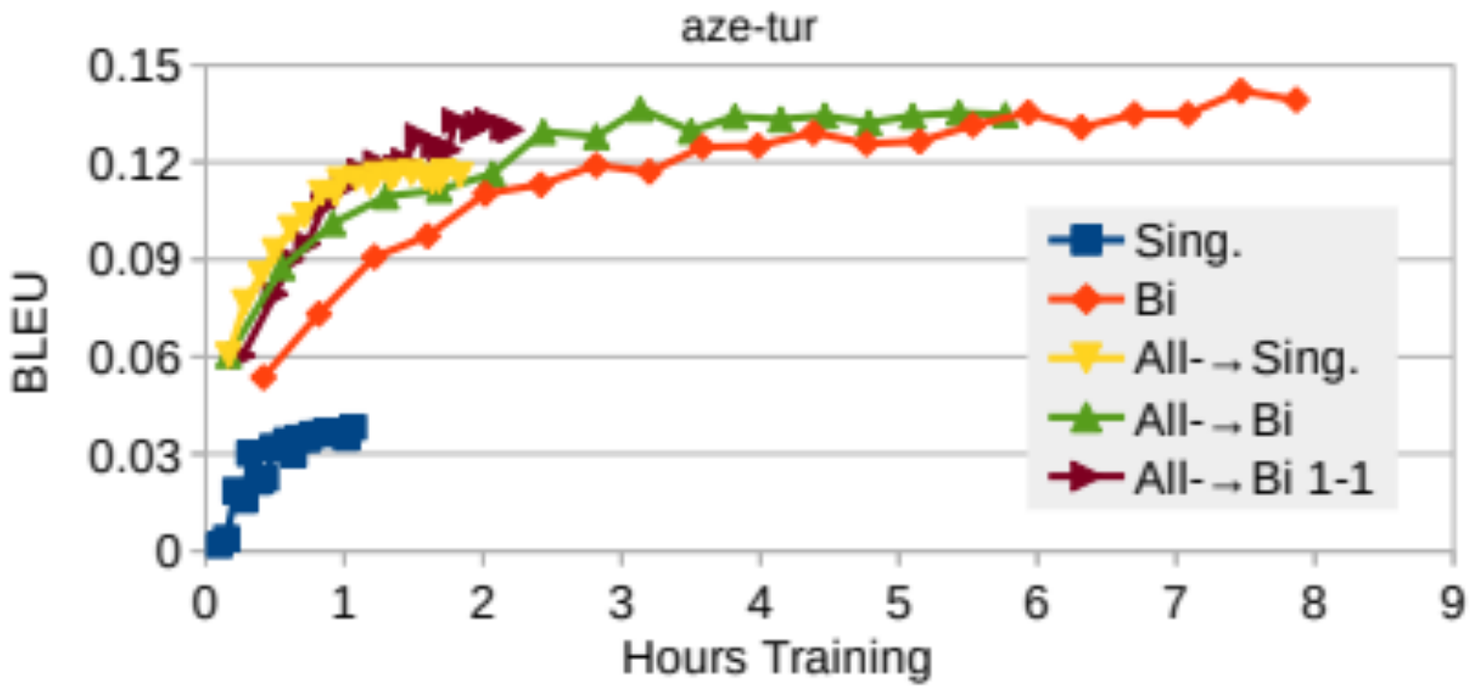
Rapid adaptation of massive multilingual models



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

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

Rapid adaptation of massive multilingual models

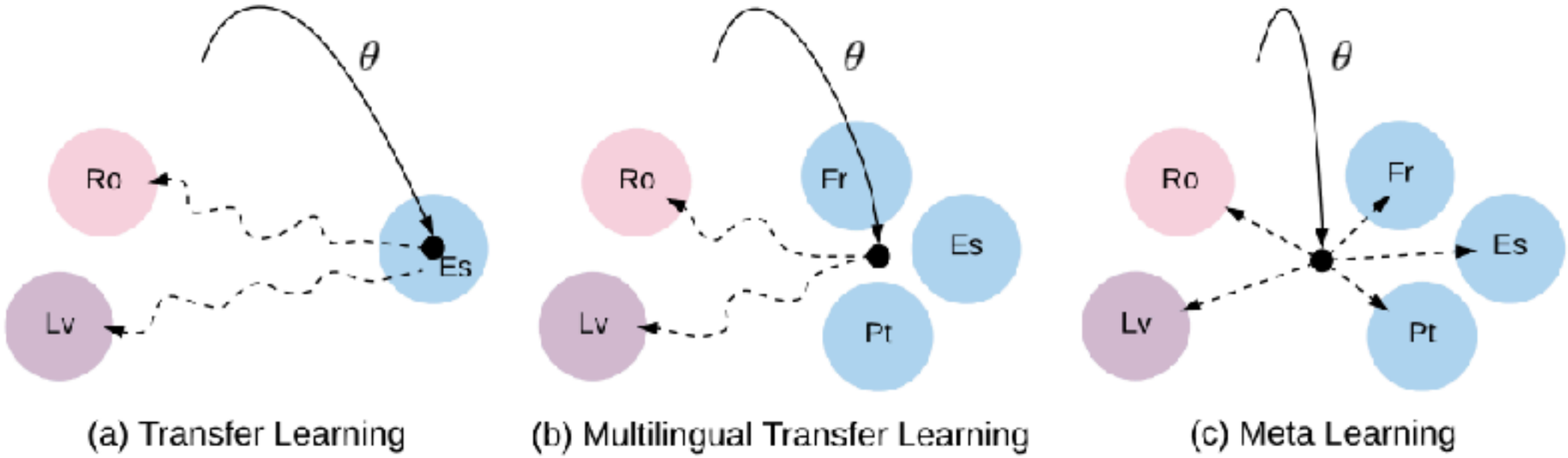


- All- -> xx models: adapting from a multilingual makes convergence faster

- Regularized fine-tuning leads to better final performance

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

Meta-learning for multilingual training



(a) Transfer Learning (b) Multilingual Transfer Learning (c) Meta Learning

- Learning a good initialization of model for fast adaptation to all languages
- Meta-learning: learn how to learn
  - Inner loop: optimize/learn for each language
  - Outter loop (meta objective): learn how to quickly optimize for each language

Meta-learning for low-resource neural machine translation. Gu et. al. 2018

Roadmap

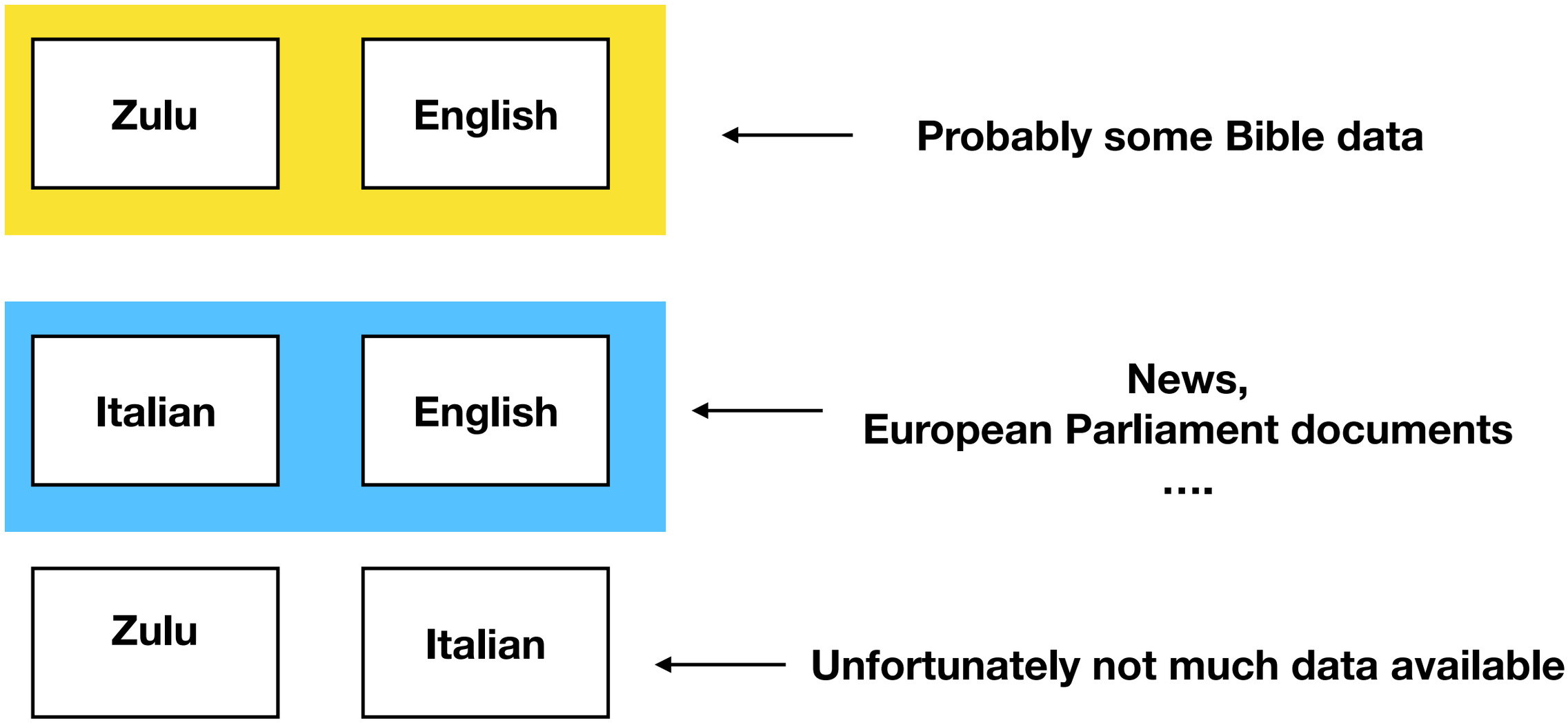
- Two methods: cross-lingual transfer and multilingual training
- Zero-shot transfer
- Open problems with multilingual training

Zero-shot transfer

- Train models that work for a language without **annotated** data in that language

- Allowed to train using **monolingual data** for the test language or **annotated data for other languages**

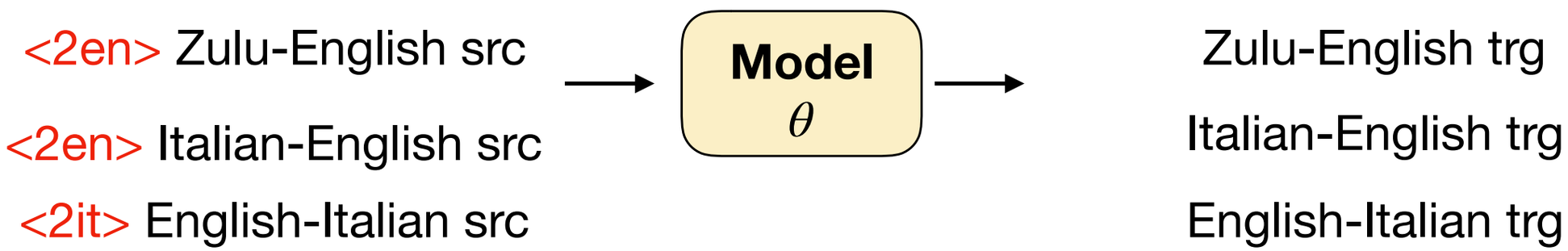
Multilingual NMT



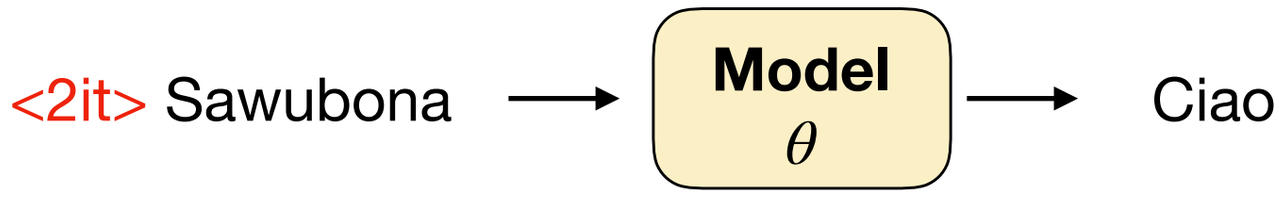
- Parallel data are English centric

Multilingual NMT

Training:



Testing:

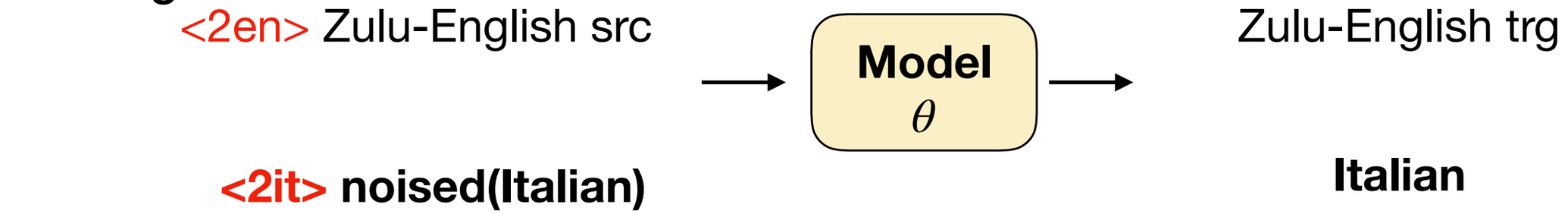


- Multilingual Training allows zero-shot transfer
- Train on {zulu-english, english-zulu, english-italian, italian-english}
- Zero-shot: the model can translate **Zulu to Italian** with out any Zulu-Italian parallel data

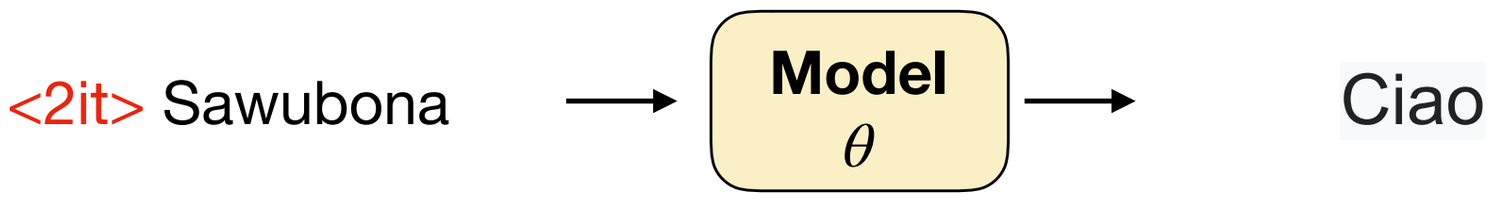
Google's multilingual neural machine translation system . Johnson et. al. 2016

Improve zero-shot NMT: Use monolingual data

Training:



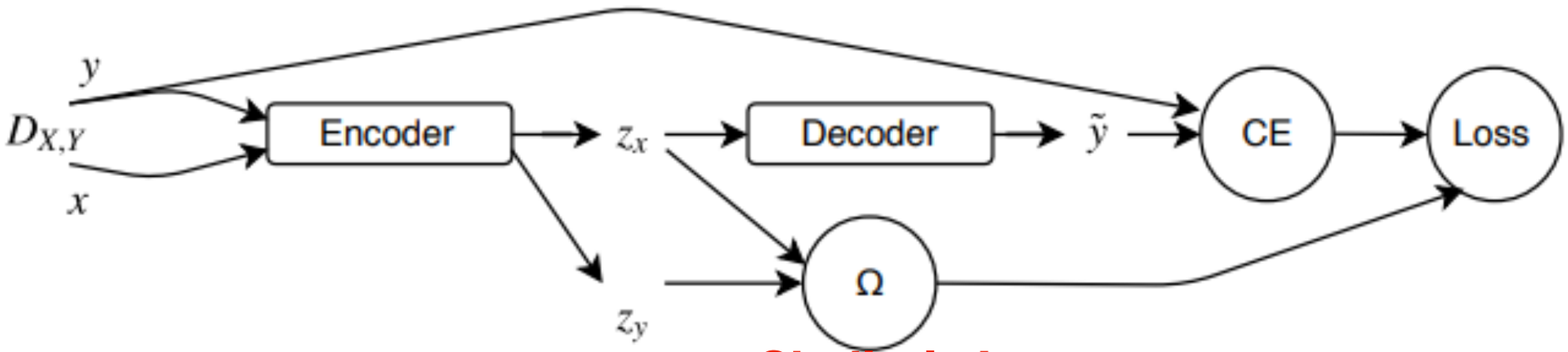
Testing:



- Add monolingual data by asking the model to reconstruct the noisy version of the monolingual data
- Use masked language model objective

Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation . Siddhant et. al. 2019

Improve zero-shot NMT: Align multilingual representation



Similarly Loss
Between 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 for pretrained representations

- Pretrain: large language model using **monolingual data** from many different langauges

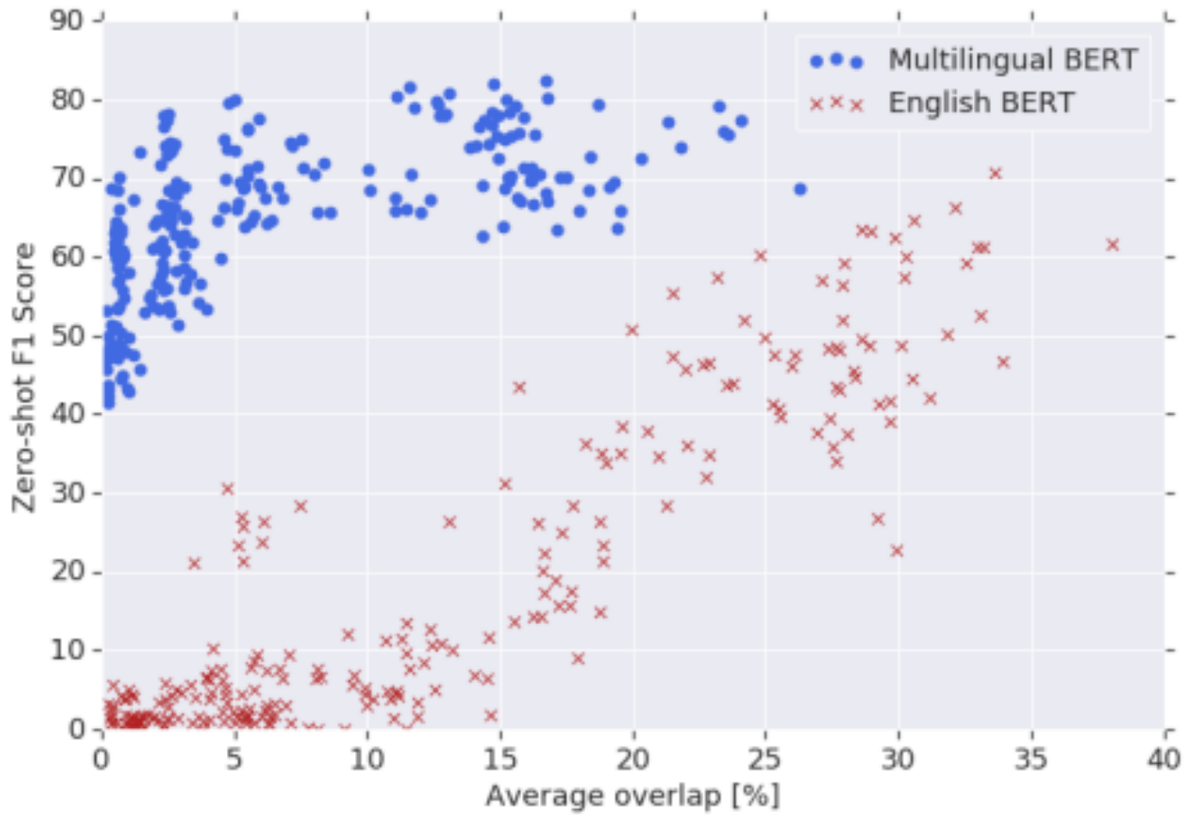
- Fine-tune: using **annotated data** in a given language (eg. English)

- Test: test the fine-tuned model on a **different** language from the fine-tuned language (eg. French)

- **Multilingual pretraining** learns a language-universal representation!

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

Zero-shot transfer for pretrained representations



Vocabulary overlap

- Generalize to language with different scripts: transfer well to languages with little vocabulary overlap
- Does not work well for typologically different languages: fine-tune on English, test on Japanese

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

Roadmap

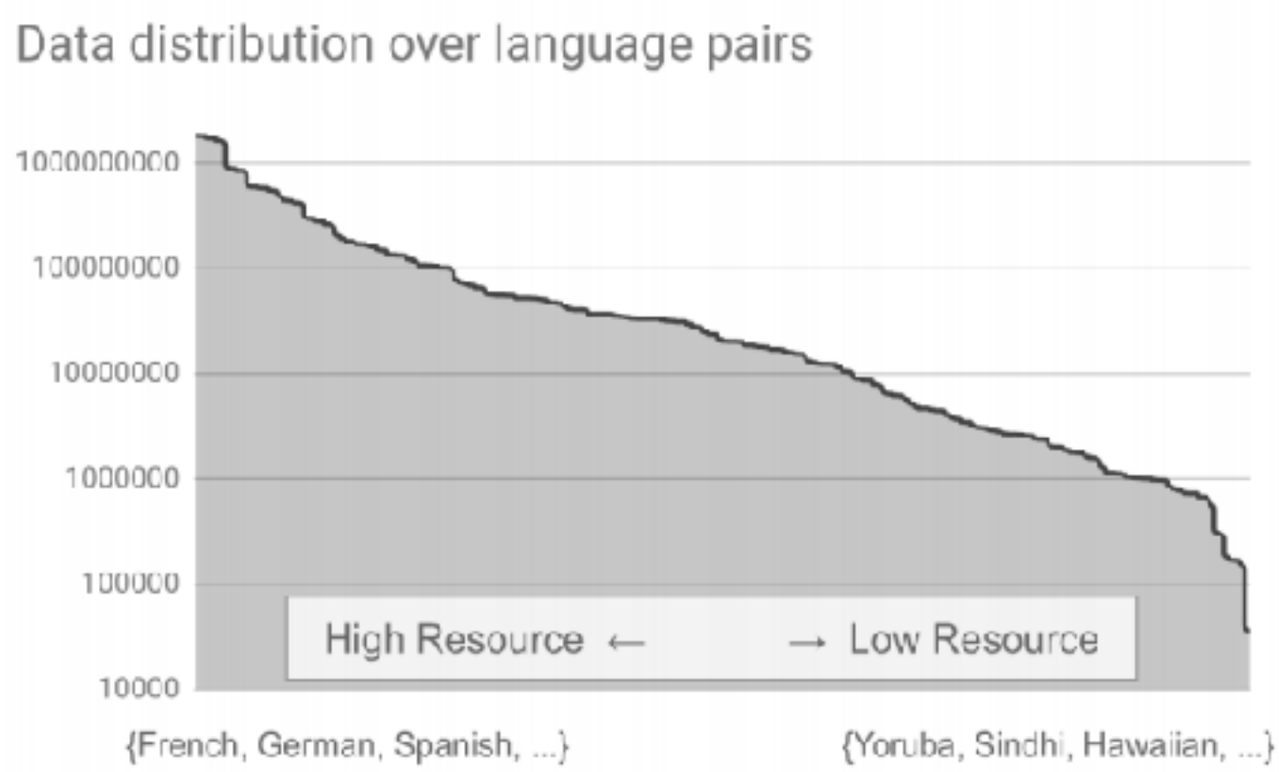
- Two methods: cross-lingual transfer and multilingual training
- Zero-shot transfer
- Open problems with multilingual training

Massively multilingual training

- How about we scale up to over **100** languages?
  - Many-to-one: translate from many languages to one target
  - One-to-many: translate from one source language to many languages
  - Many-to-many: translate from many source to many target languages

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

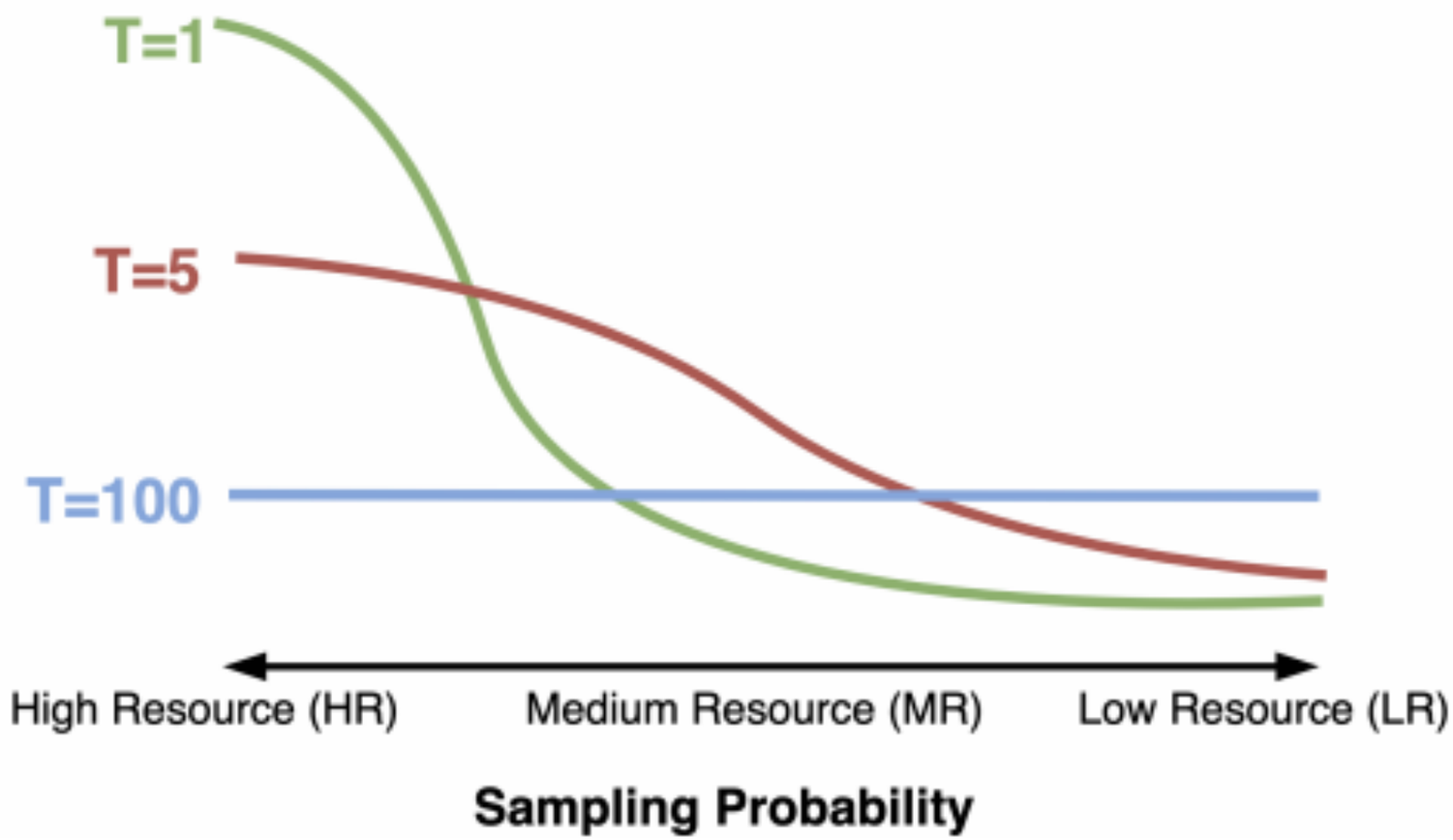
Training data highly imbalanced



- Again, data distribution is highly imbalanced
- Important to upsample low-resource data in this setting!

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

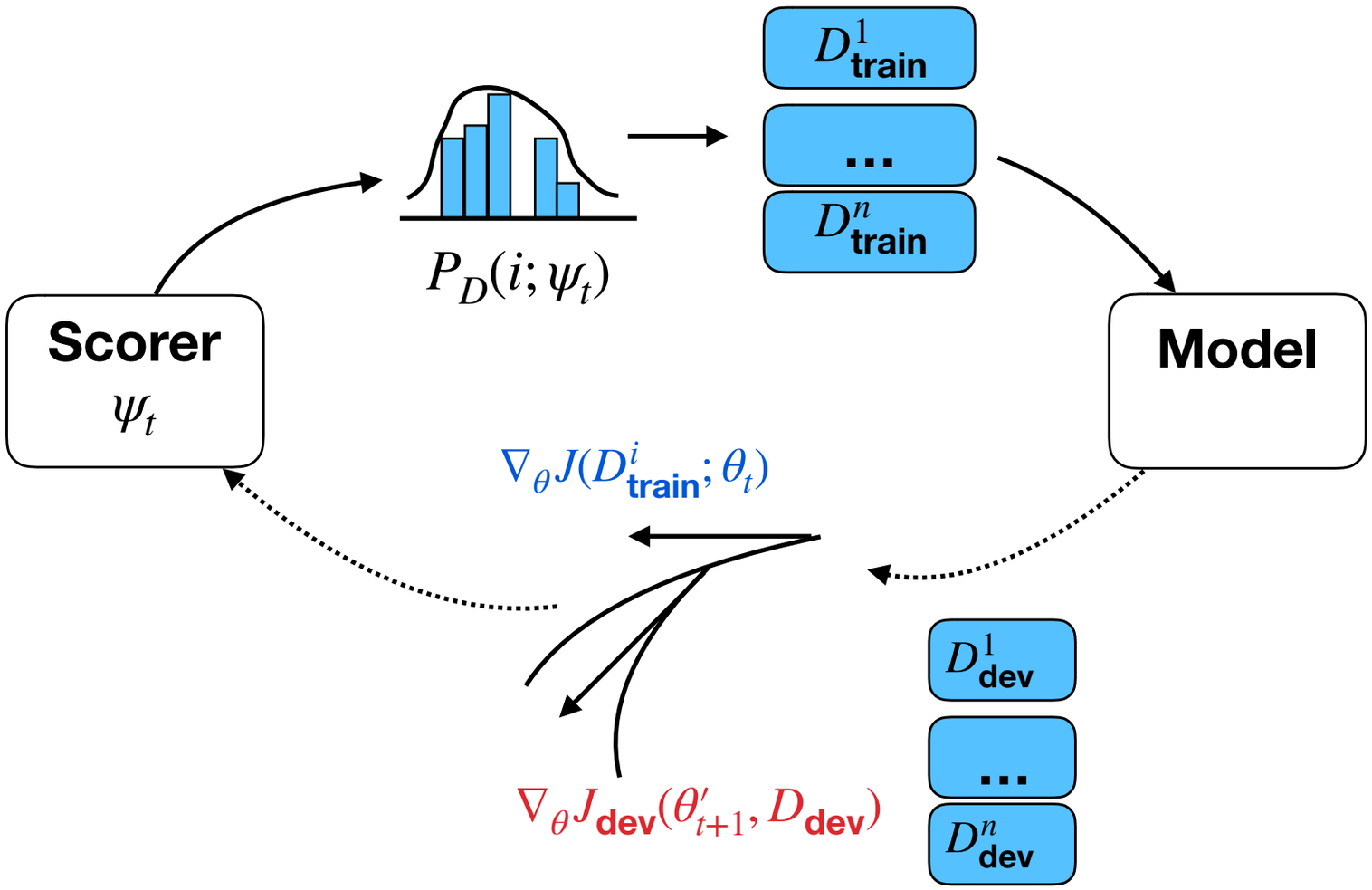
Heuristic Sampling of Data



- 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

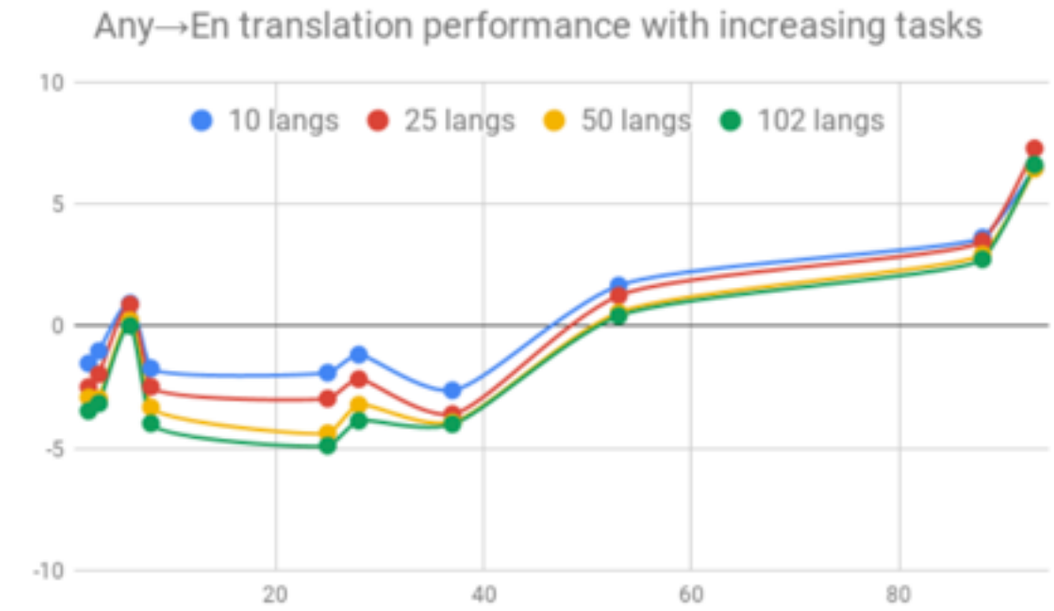
Learning to balance data



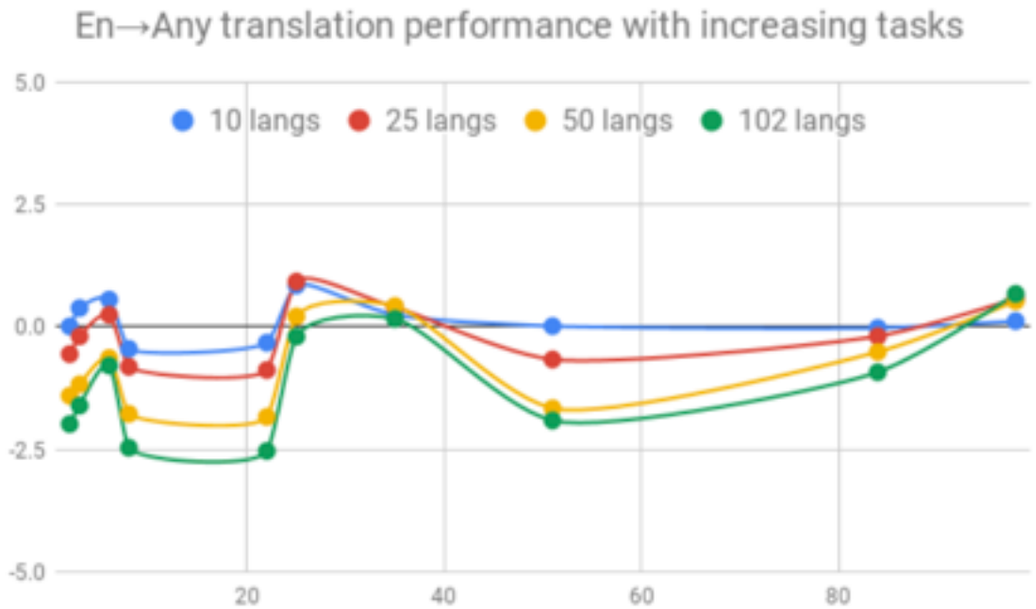
- Optimize the data sampling distribution during training
- Upweight languages that have similar gradient with the multilingual dev set

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

Problem: sometimes underperforms bilingual model



High-resource **Low-resource**



High-resource **Low-resource**

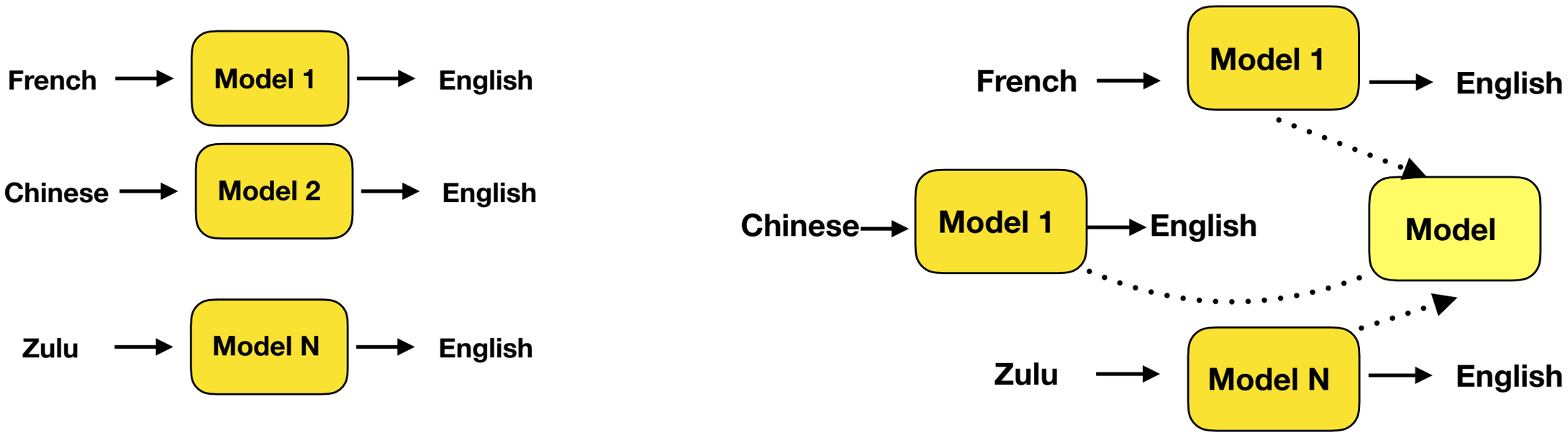
- Multilingual training degrades high-resource language

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

Problem: sometimes underperforms bilingual model

- Possible solutions:
  - Instead of training a single multilingual model, train one model for each language cluster
  - Make models bigger and deeper?
  - Use extra monolingual data
  - …..

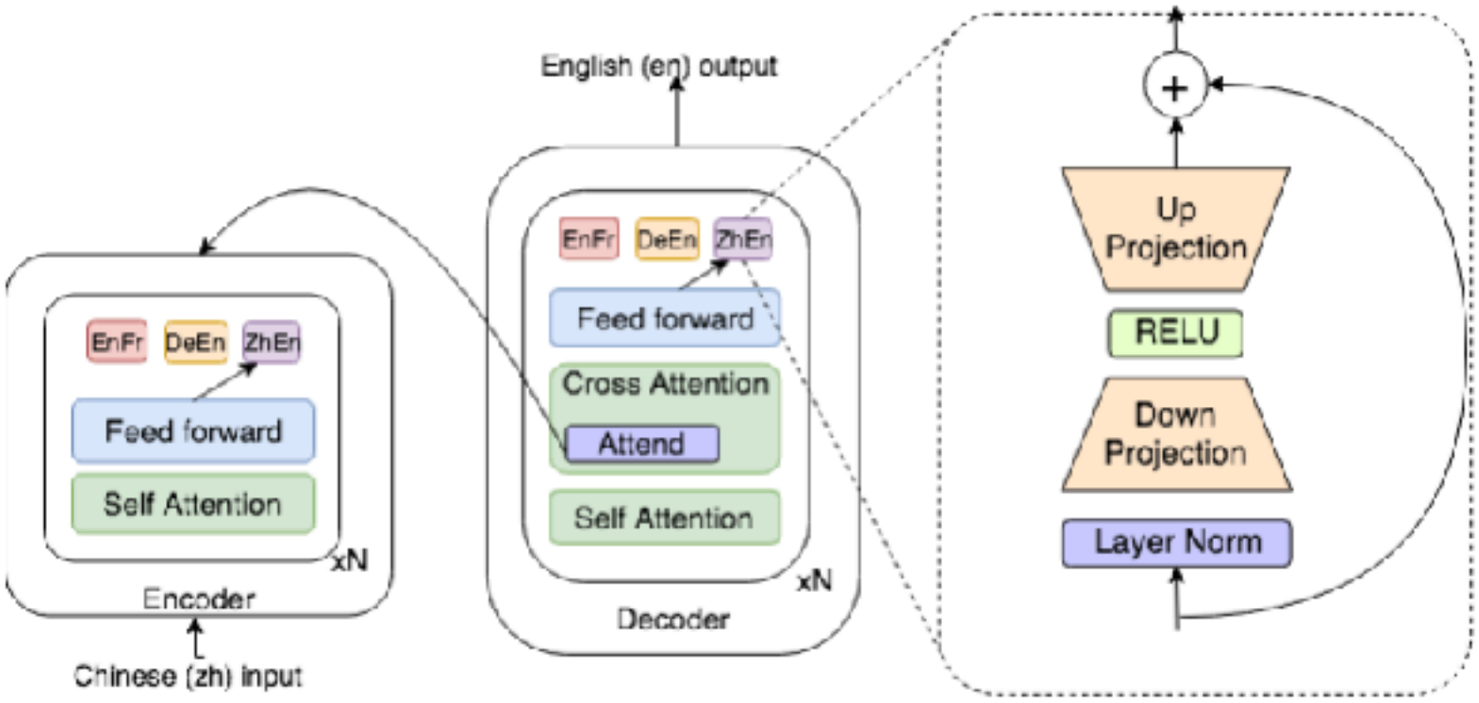
Multilingual Knowledge Distillation



- First train individual model on each language pair
- Then “distill” the individual models for a single multilingual model
- However, takes much efforts to train many different models

Multilingual Neural Machine Translation with Knowledge Distillation. Tan et. al. 2019

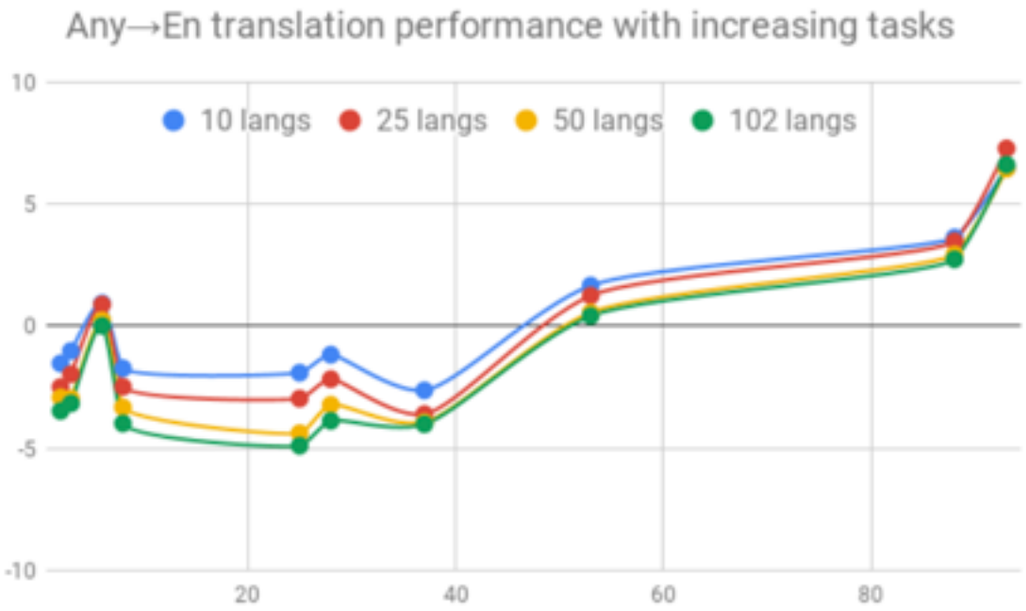
Adding Language-specific layers



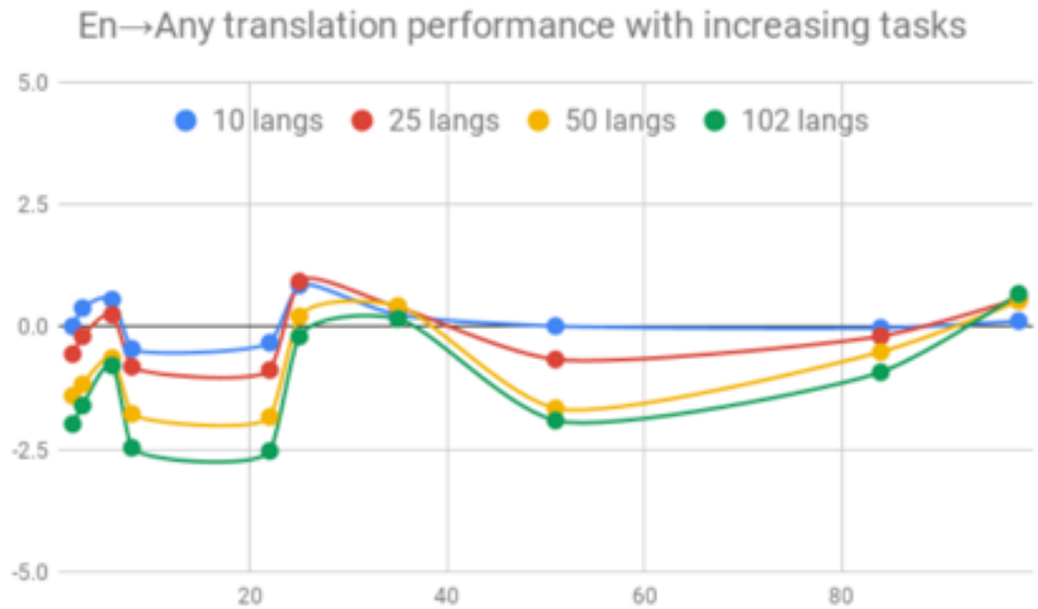
- Add a small module for each language pair
- Much better at matching bilingual baseline for high-resource languages

Simple, Scalable adaptation for neural machine translation. Bapna et. al. 2019

Problem: one-to-many transfer



High-resource **Low-resource**



High-resource **Low-resource**

- Transfer is much **harder for one-to-many** than many-to-one

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

Problem: one-to-many transfer

- Transfer is much **harder for one-to-many** than many-to-one
    - One-to-many is closer to a multitask problem, while the decoder of many-to-one benefits more from the same target language
    - Language specific module?
    - How to decide what parameter to share and what to separate?

Problem: multilingual vocabulary construction

- Vocabulary construction for massively multilingual data is non-trivial
  - Standard approach: upsample low-resource languages and do joint BPE on all the data
  - Problem: over-segment low-resource or morphologically rich languages

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-zh: 20)
  - Is BLEU score between two languages comparable? Does +5 BLEU on en-zh has the same “benefit” as +5 BLEU on en-fr?

Discussion question

- Read “Massively Multilingual Neural Machine Translation in the Wild (https://arxiv.org/pdf/1907.05019.pdf)”

- Question: what is one interesting problem with multilingual NMT, and what is the experiment or analysis from the paper that explains this problem? Can you think about any potential solutions?