The Low Resource NLP Toolbox, 2020 Version

Graham Neubig

@ AfricaNLP 4/26/2020

(collaborators highlighted throughout)

Carnegie Mellon University
Language Technologies Institute

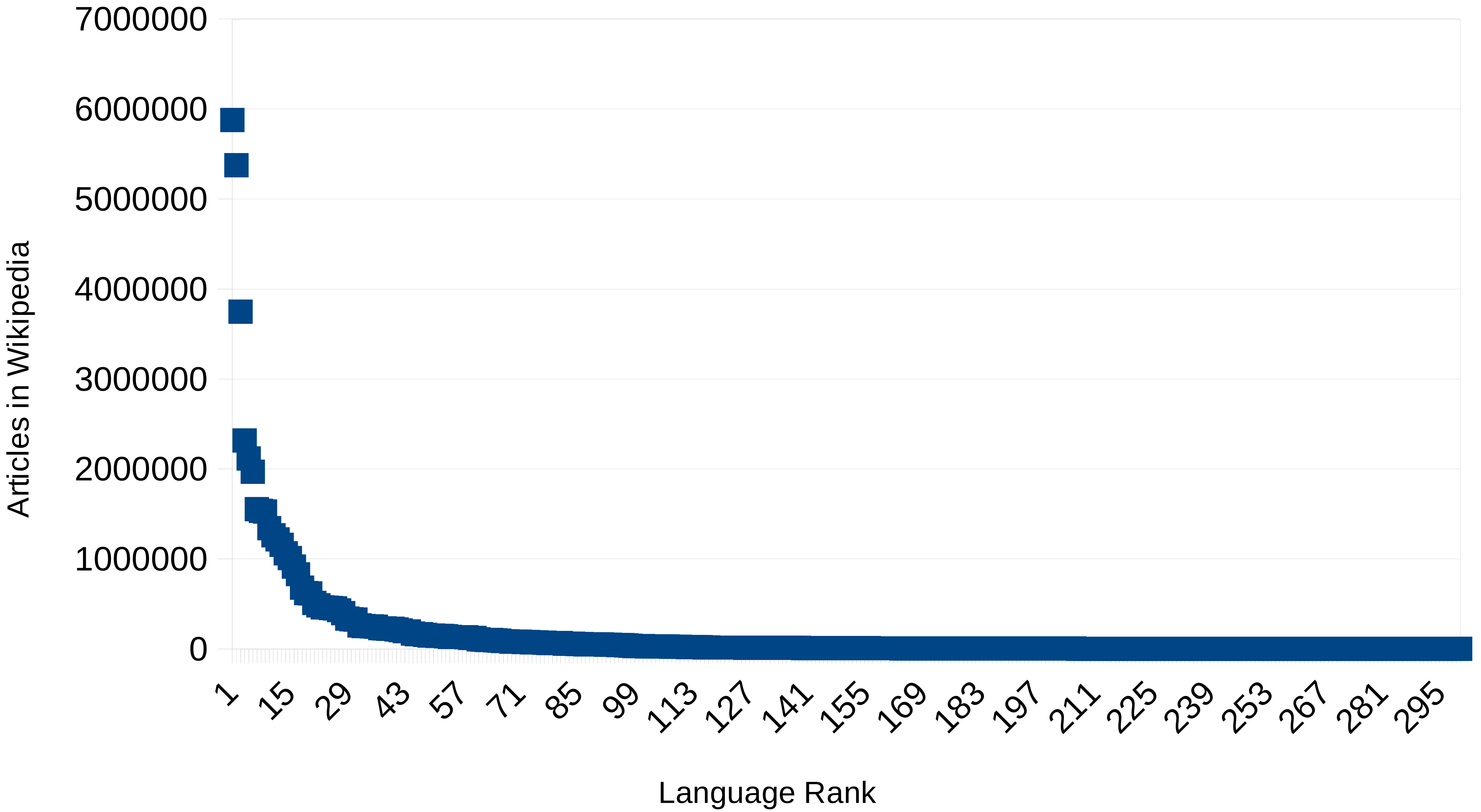
NEULAB

http://endangeredlanguages.com/

How do We Build NLP Systems?

- **Rule-based systems:** Work OK, but require lots of human effort for each language for where they're developed
- **Machine learning based systems:** Work really well when lots of data available, not at all in low-data scenarios

The Long Tail of Data



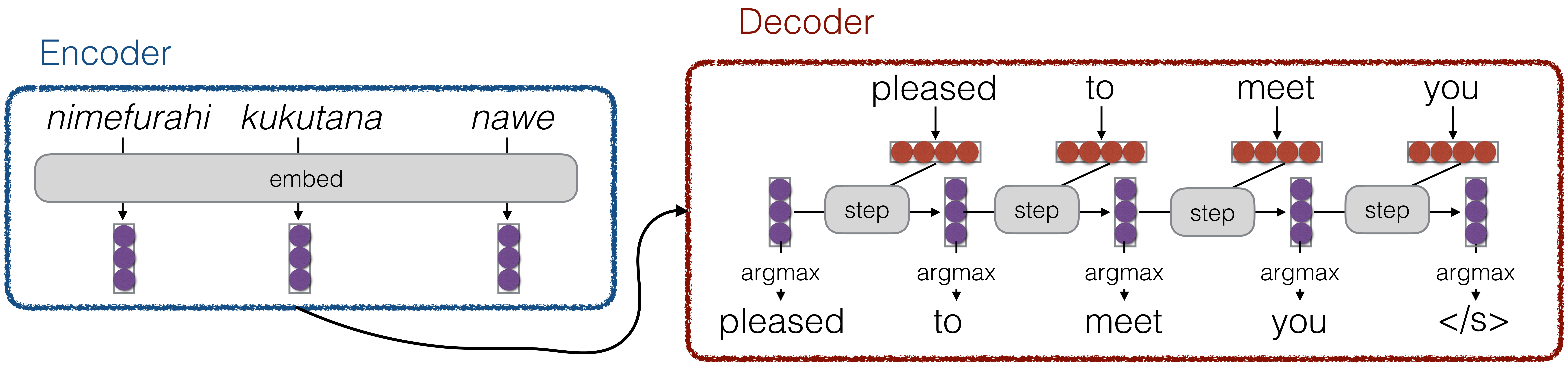
Machine Learning Models

- Formally, map an input X into an output Y. Examples:

Input X	Output Y	Task
Text	Text in Other Language	Translation
Text	Response	Dialog
Speech	Transcript	Speech Recognition
Text	Linguistic Structure	Language Analysis

- To learn, we can use
  - Paired data <X, Y>, source data X, target data Y
  - Paired/source/target data in *similar* languages

Method of Choice for Modeling: Sequence-to-sequence with Attention



- **Various tasks:** Translation, speech recognition, dialog, summarization, language analysis
- **Various models:** LSTM, transformer
- Generally trained using **supervised learning**: maximize likelihood of <X, Y>

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

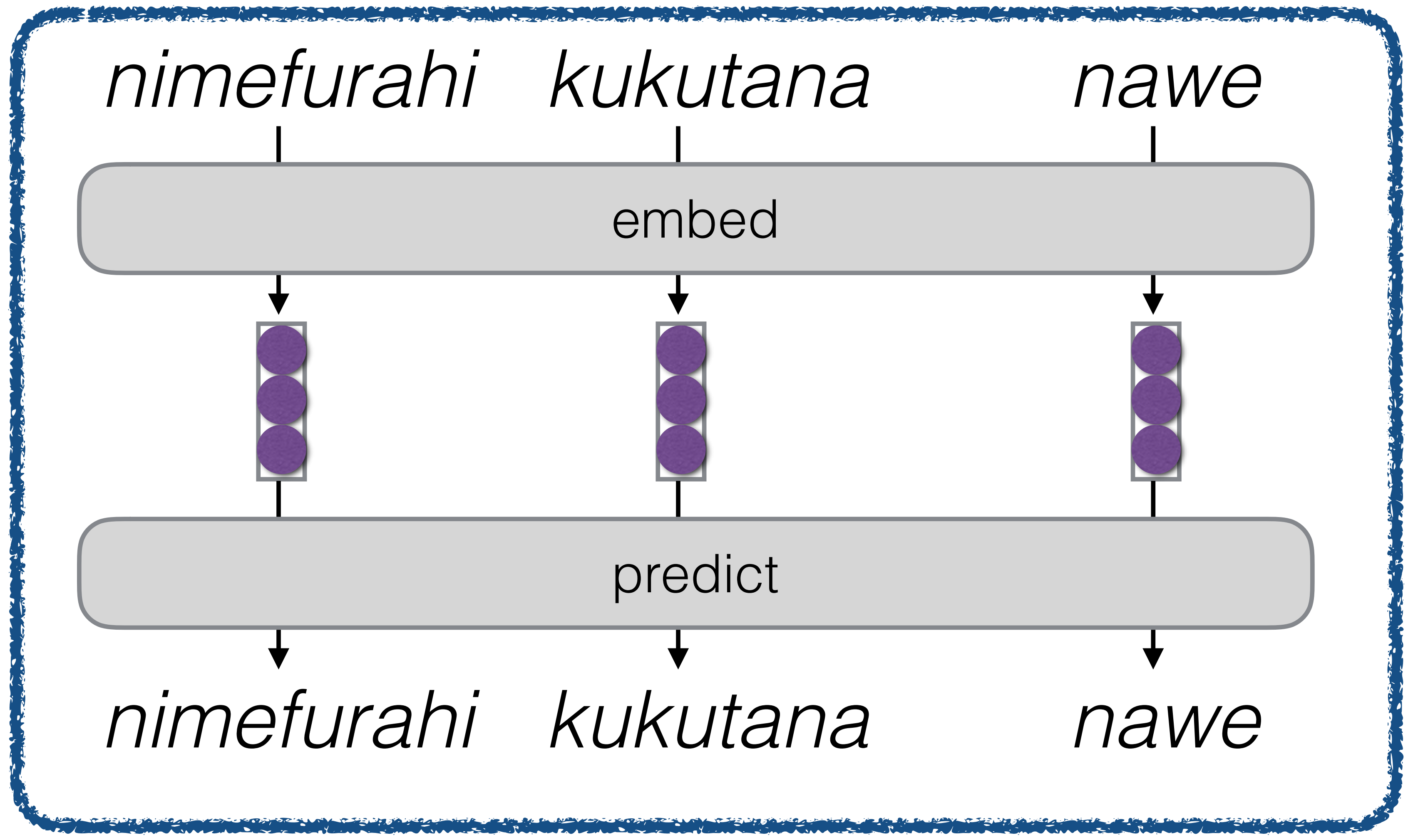
The Low-resource NLP Toolbox

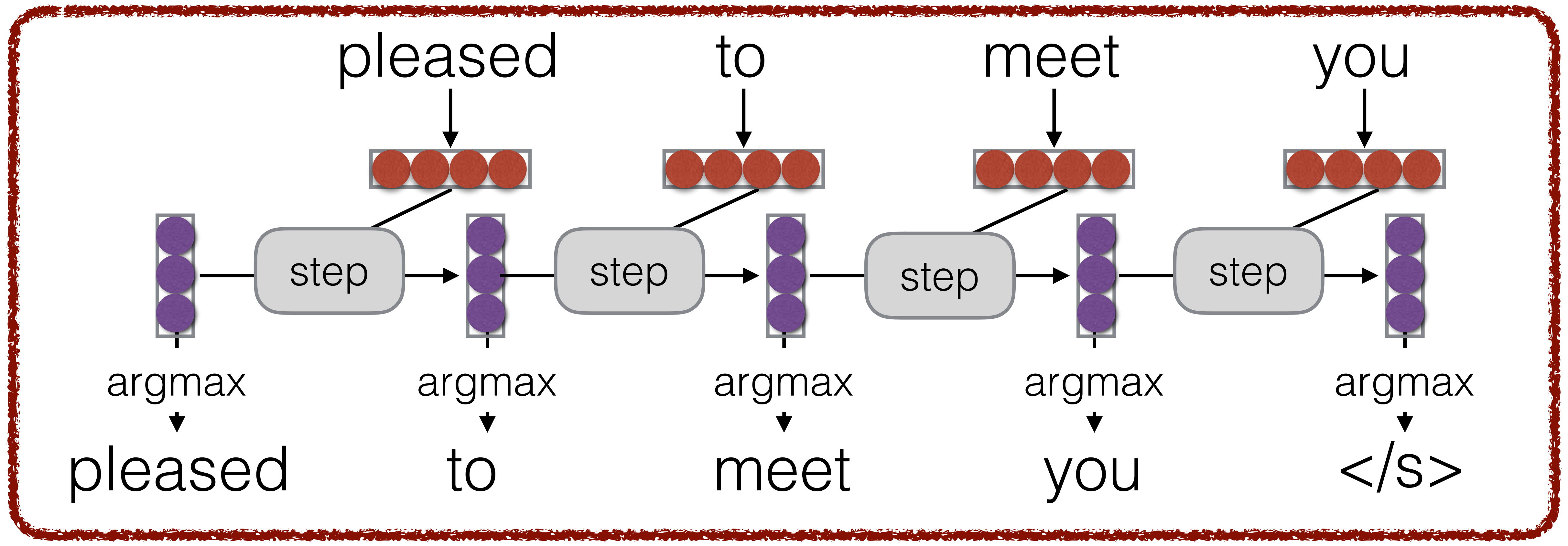
- In cases when we have lots of paired data <*X*, *Y*>
  -> **supervised learning**
- But what if we don't?!
  - Lots of source or target data *X* or *Y*
    -> **monolingual pre-training, back-translation**
  - Paired data in another, similar language <*X'*, *Y*> or <*X*, *Y'*>
    -> **multilingual training, transfer**
  - Can ask speakers to do a little work to generate data
    -> **active learning**

Learning from Monolingual Data

Language-model Pre-training

- Given source or target data *X* or *Y*, train just the encoder or decoder as a language model first



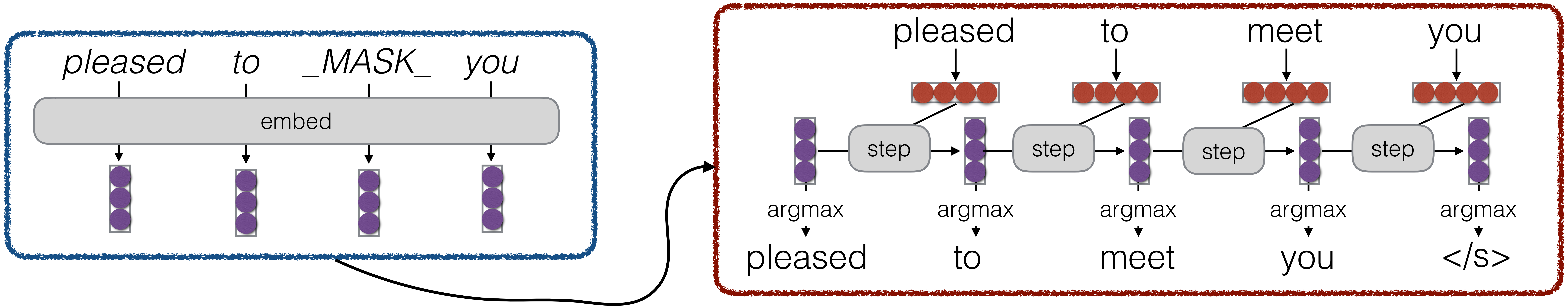


- Many different methods: simple language model, BERT, etc.

Ramachandran, Prajit, Peter J. Liu, and Quoc V. Le. "Unsupervised pretraining for sequence to sequence learning." *arXiv preprint arXiv:1611.02683* (2016).
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Sequence-to-sequence Pre-training

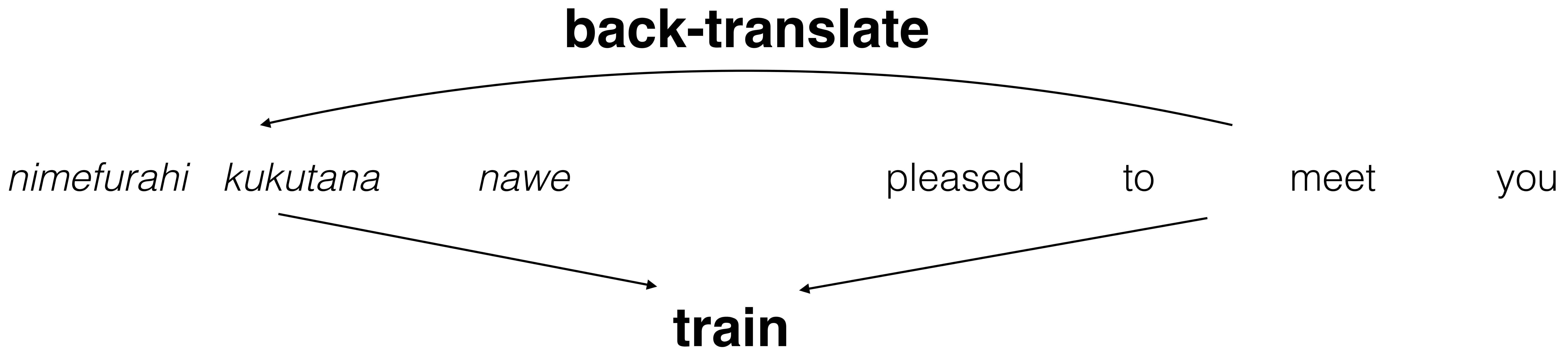
- Given just source, or just target data *X* or *Y*, train the encoder and decoder together



Song, Kaitao, et al. "Mass: Masked sequence to sequence pre-training for language generation." *arXiv preprint arXiv:1905.02450* (2019).
Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).

Back Translation

- Translate target data *Y* into *X* using a target-to-source translation system, then use translated data to train source-to-target system



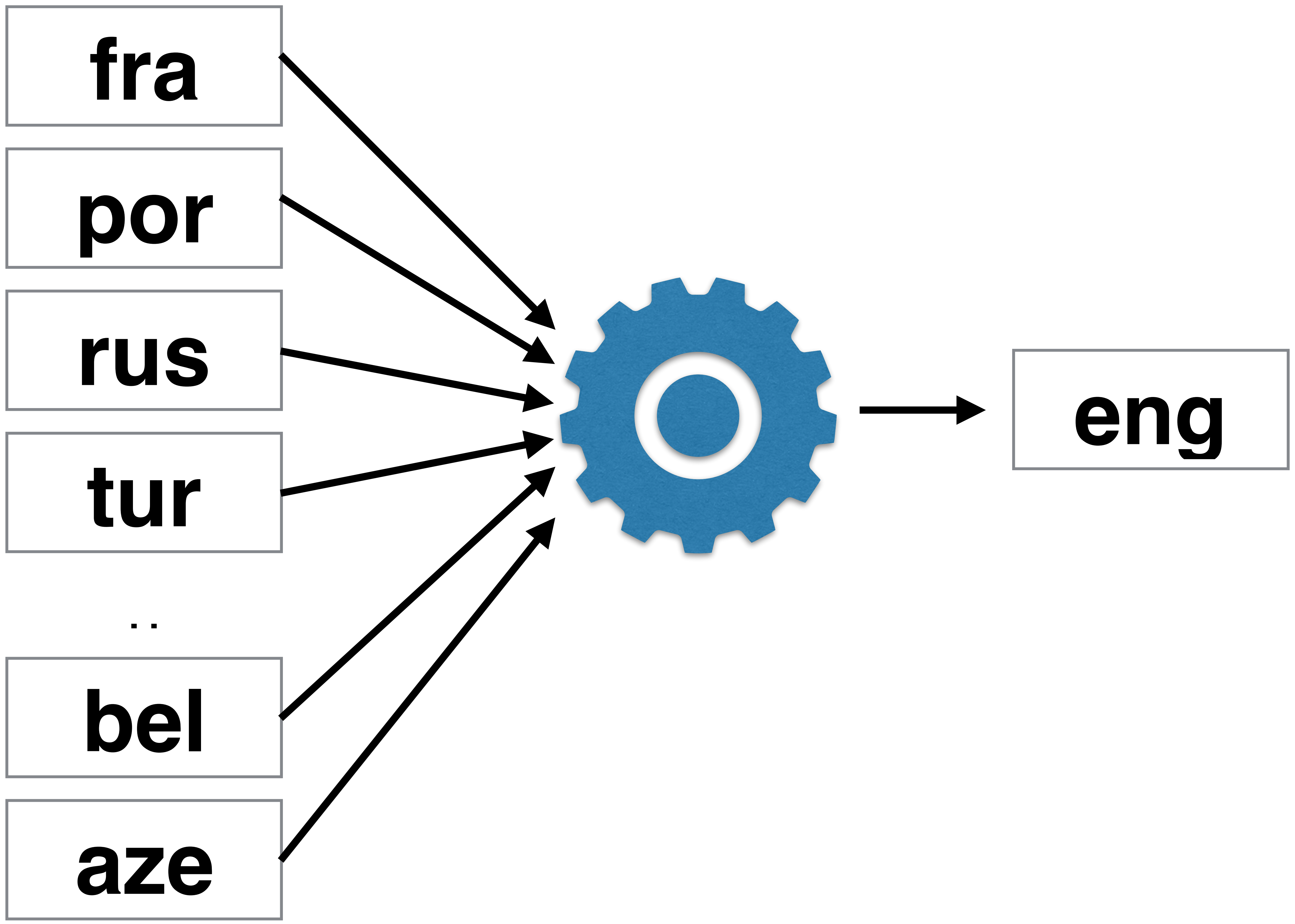
- **Iterative back-translation:** train src-to-trg, trg-to-src, src-to-trg etc
- **Semi-supervised translation:** many iterations of iterative translation, weighting confident instances

Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Improving neural machine translation models with monolingual data." *arXiv preprint arXiv:1511.06709* (2015).
Hoang, Vu Cong Duy, et al. "Iterative back-translation for neural machine translation." WNGT. 2018.
Cheng, Yong. "Semi-supervised learning for neural machine translation." ACL 2016. 25-40.

Multilingual Learning, Cross-lingual Transfer

Multilingual Training [Johnson+17, Ha+17]

- Train a large multi-lingual NLP system

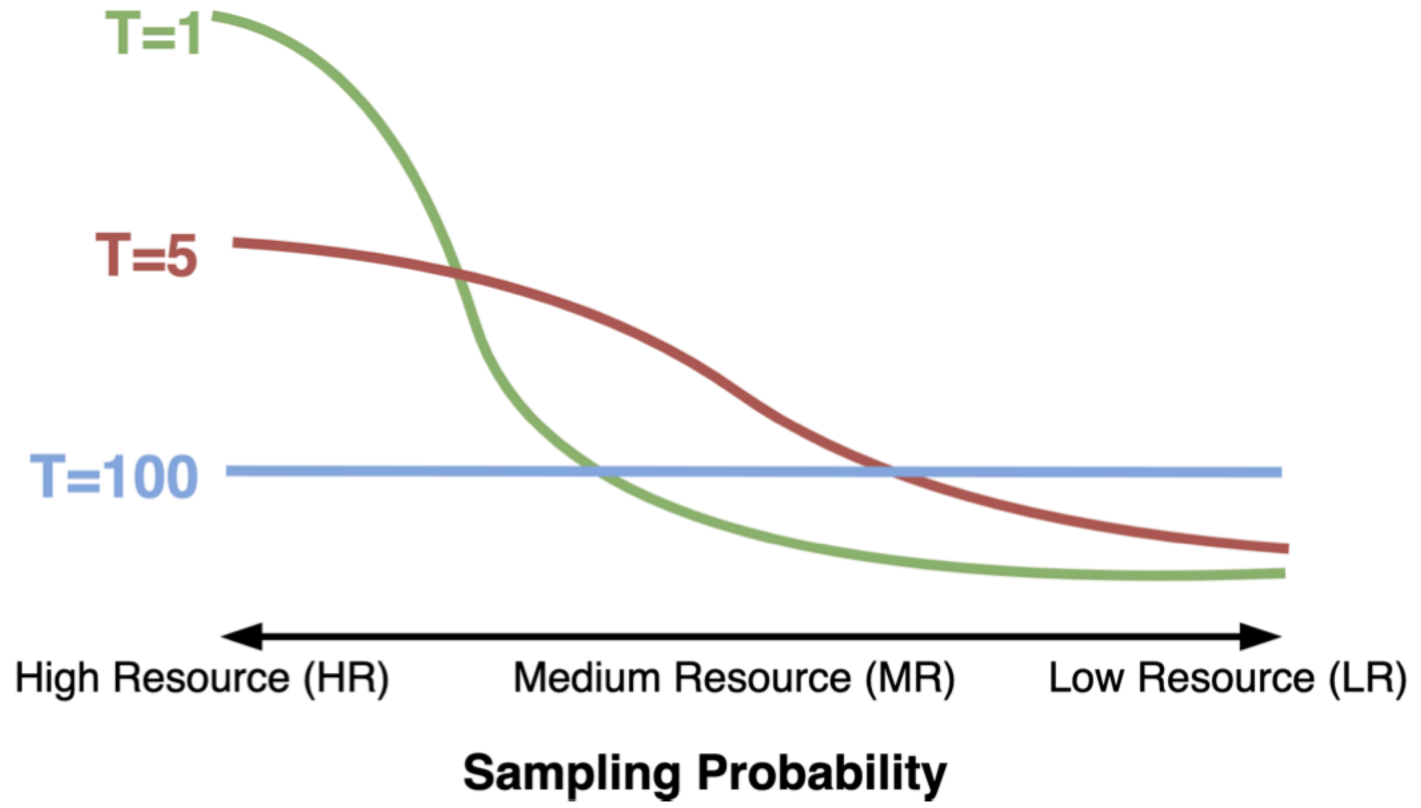


Johnson, Melvin, et al. "Google's multilingual neural machine translation system: Enabling zero-shot translation." *Transactions of the Association for Computational Linguistics* 5 (2017): 339-351.
Ha, Thanh-Le, Jan Niehues, and Alexander Waibel. "Toward multilingual neural machine translation with universal encoder and decoder." *arXiv preprint arXiv:1611.04798* (2016).

Massively Multilingual Systems

- Can train on 100, or even 1000 languages (e.g. Multilingual BERT, XLM-R)

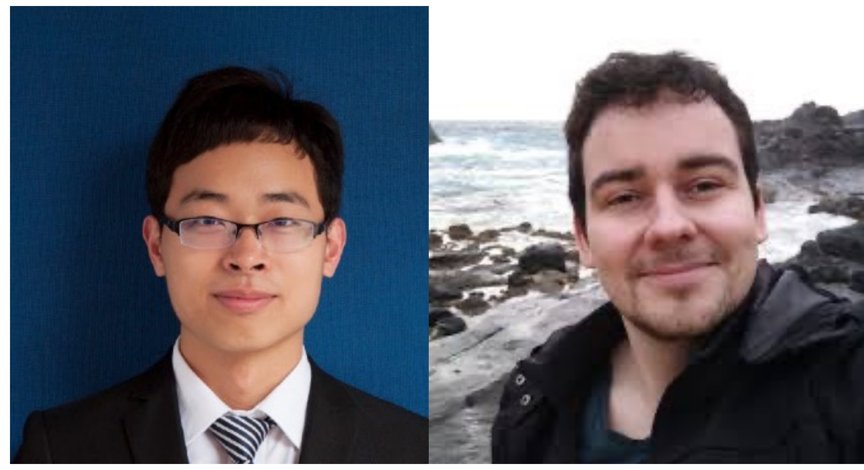
- Hard to balance multilingual performance, careful data sampling necessary



- **Multi-DDS:** Data sampling can be *learned automatically* to maximize accuracy on all languages

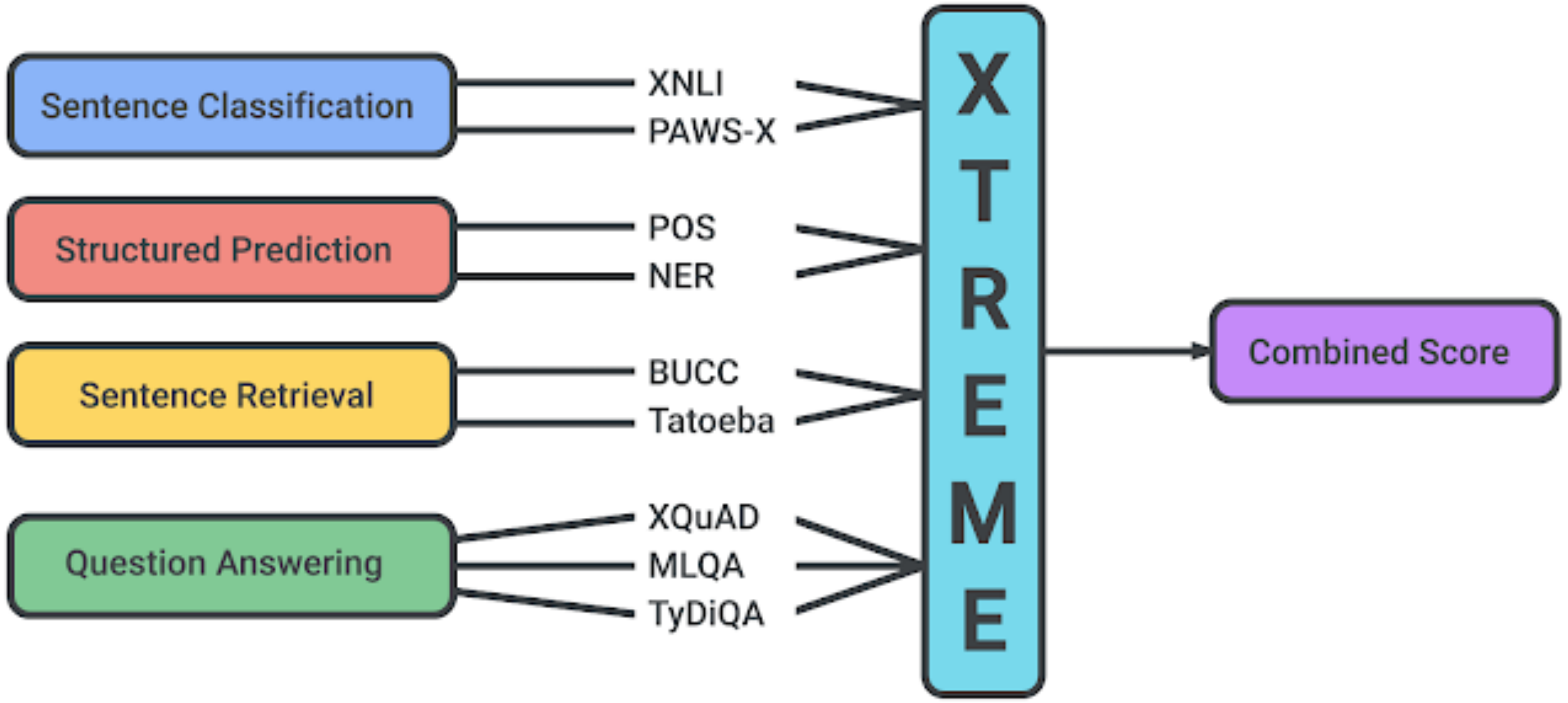
Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges." *arXiv preprint arXiv:1907.05019* (2019).
Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." arXiv preprint arXiv:1911.02116 (2019).
Wang, Xinyi, Yulia Tsvetkov, and Graham Neubig. "Balancing Training for Multilingual Neural Machine Translation." arXiv preprint arXiv:2004.06748 (2020).

XTREME: Benchmark for Multilingual Learning



[Hu, Ruder+ 2020]

- Difficult to examine performance of systems on many different languages
- XTREME benchmark makes it easy to evaluate on existing datasets over 40 languages
    - Some coverage of African languages -- Afrikaans, Swahili, Yoruba

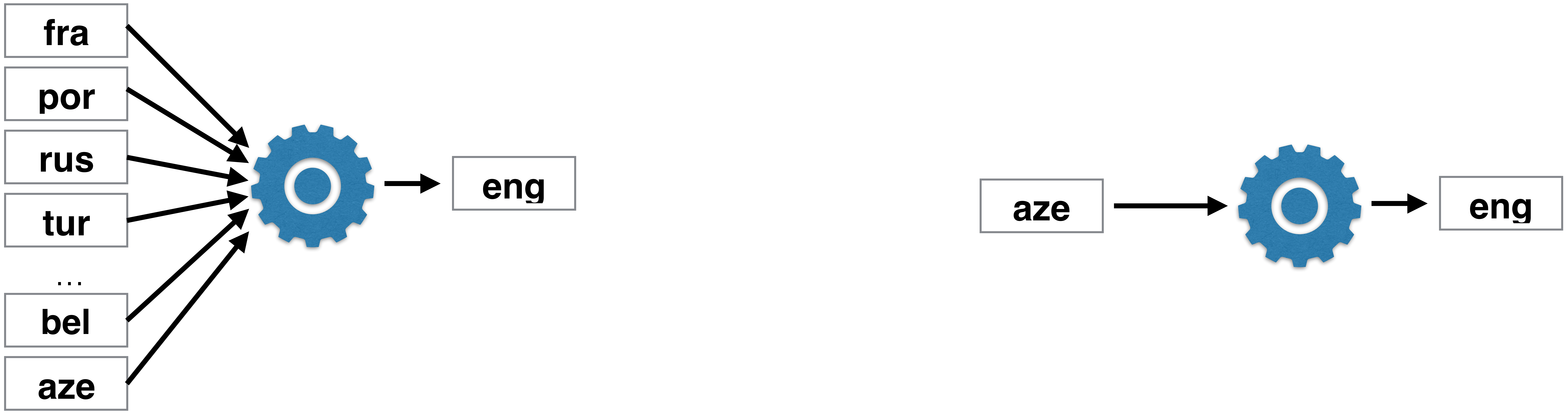


Hu, Junjie, et al. "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization." *arXiv preprint arXiv:2003.11080* (2020)

Cross-lingual Transfer

- Train on one language, transfer to another

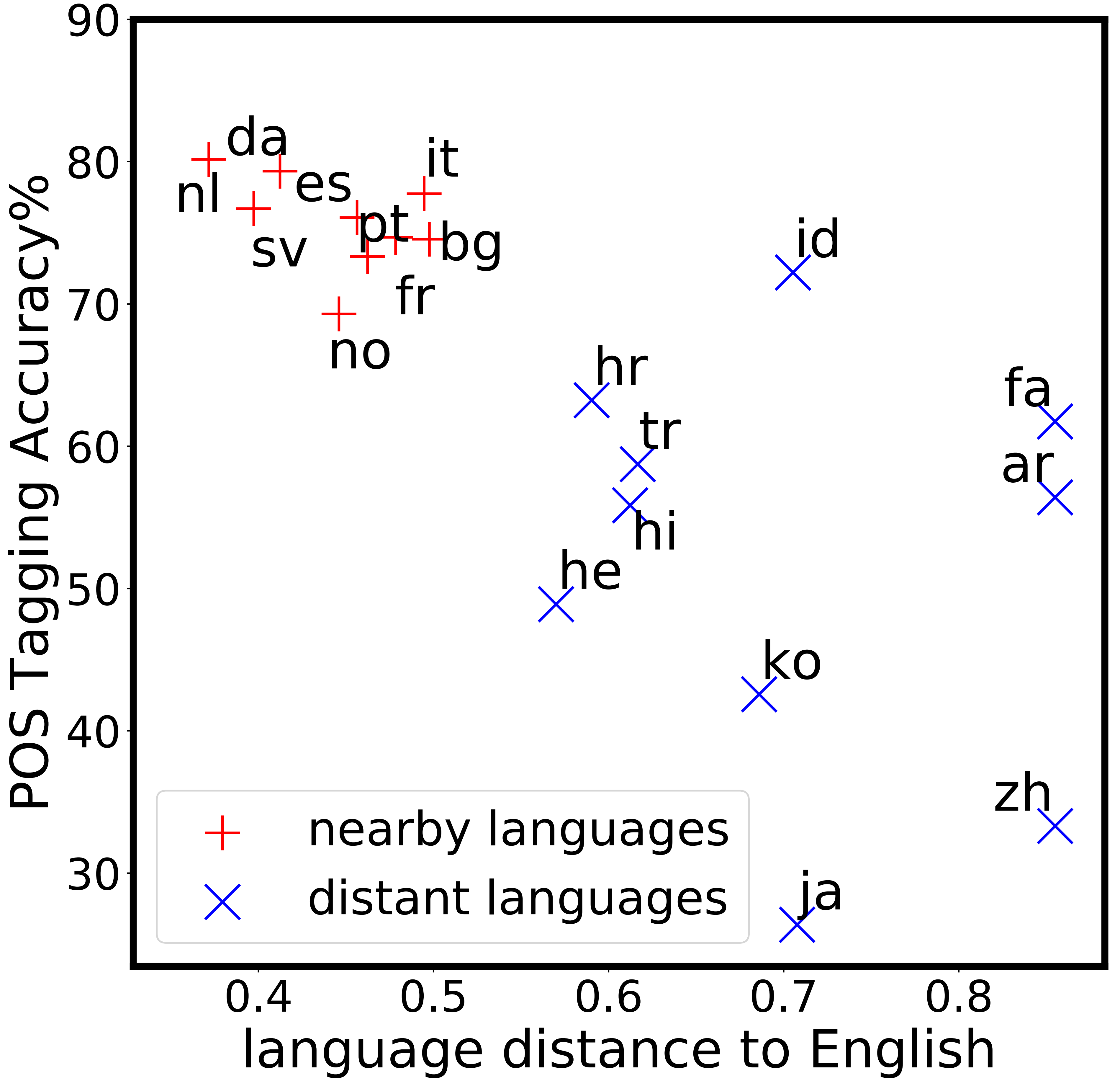
- Train on many languages, transfer to another



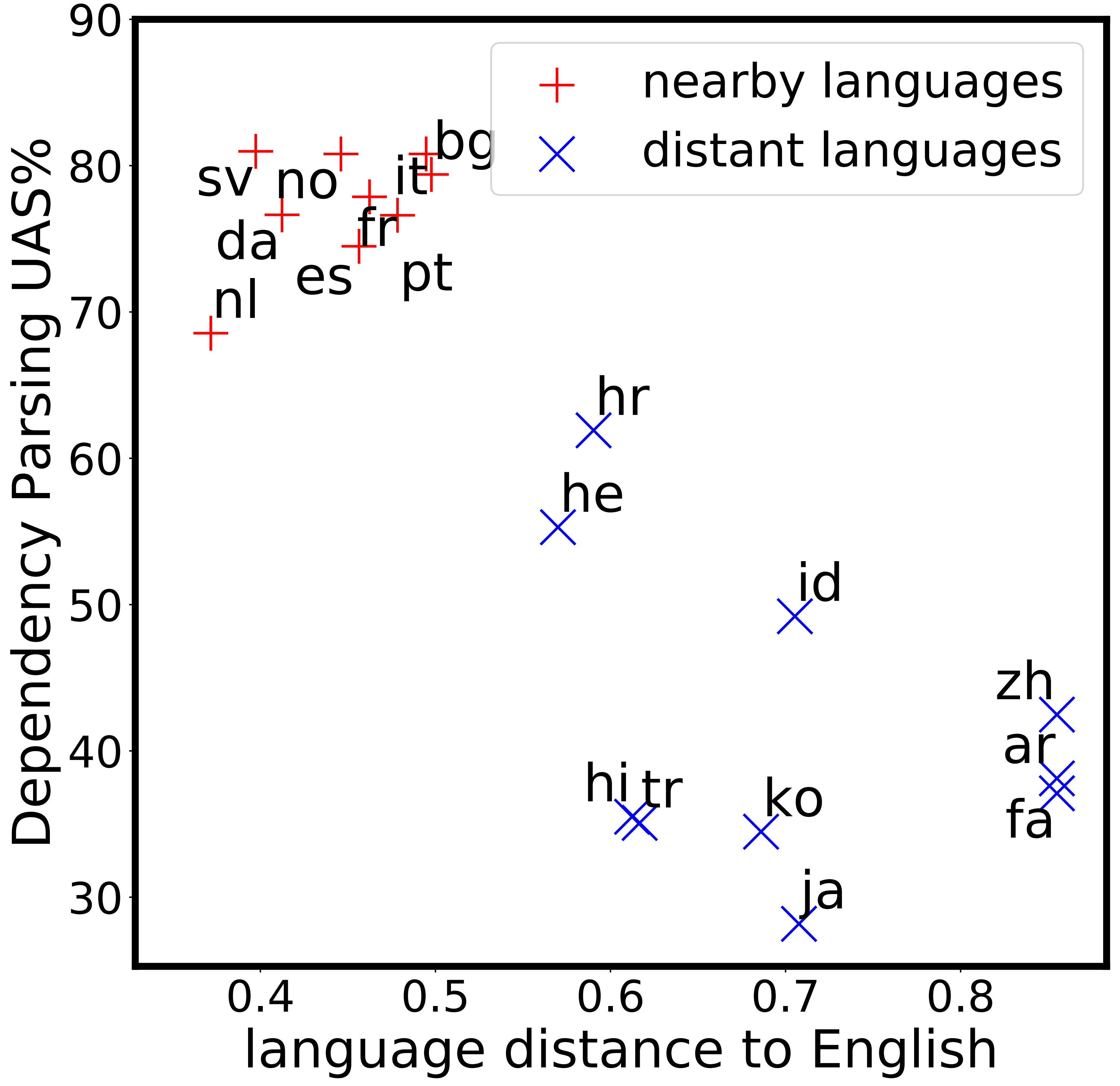
Zoph, Barret, et al. "Transfer learning for low-resource neural machine translation." *arXiv preprint arXiv:1604.02201* (2016).
Neubig, Graham, and Junjie Hu. "Rapid adaptation of neural machine translation to new languages." arXiv preprint arXiv:1808.04189 (2018).

Challenges in Multilingual Transfer

Problem: Transfer Fails for Distant Languages



(a) POS tagging



(a) Dependency parsing

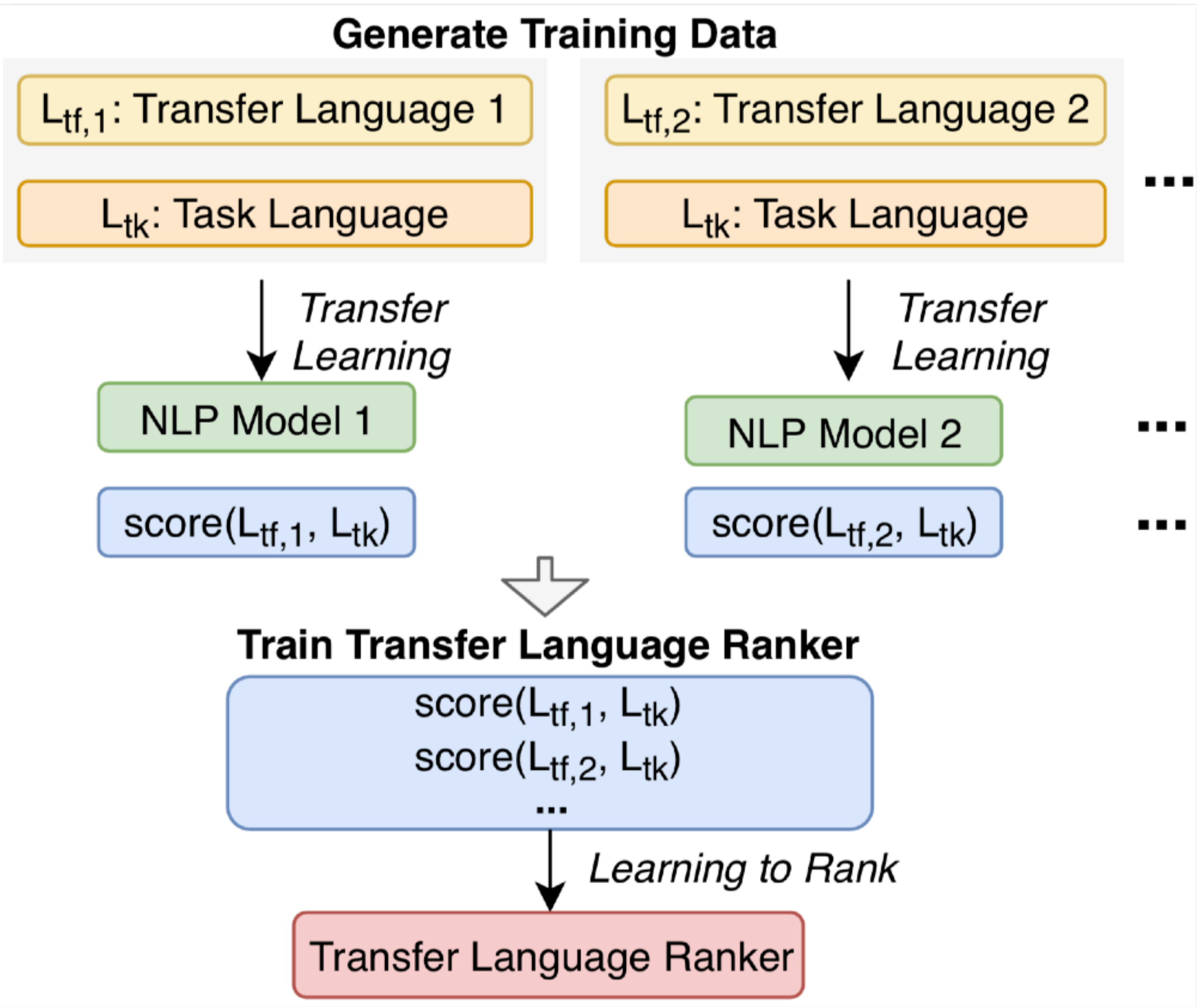
He, Junxian, et al. "Cross-Lingual Syntactic Transfer through Unsupervised Adaptation of Invertible Projections." *arXiv preprint arXiv:1906.02656* (2019).

How can We Transfer Across Languages Effectively?

- Select similar languages, add to training data.
- Model lexical/script differences
- Model syntactic differences

Which Languages to Use for Transfer?

- Similar languages are better for transfer when possible!
- But when want to transfer, what language do we transfer from?
  (various factors: language similarity, available data, etc.)
- **LangRank:** Automatically choose transfer languages data, language similarity features



Task Lang	LangRank	Best Dataset	Best URIEL	True Best
MT aze	tur (1) fas (3) hun (4)	o_w tur (1) hrv (5) ron (31)	d_{fea} ara (32) fas (3) sqi (22)	tur (1) kor (2) fas (3)
MT ben	hun (1) tur (2) fas (4)	o_w vie (3) ita (20) por (18)	d_{geo} mya (30) hin (27) mar (41)	hun (1) tur (2) vie (3)
EL tel	amh (6) orm (40) msa (7)	o_w amh (6) swa (32) jav (9)	d_{inv} pan (2) hin (1) ben (5)	hin (1) pan (2) mar (3)

Lin, Yu-Hsiang, et al. "Choosing transfer languages for cross-lingual learning." *arXiv preprint arXiv:1905.12688* (2019).

Problems w/ Word Sharing in Cross-lingual Learning

- Spelling variations (esp. in subword models)

bel		rus		eng
word	subword	word	subword	
фінансавыя	фінансавы я	финансовых	финансовы х	financial
стадыён	стады ён	стадион	стадион	stadium
розных	розны х	разных	разны х	different
паказаць	паказа ць	показать	показать	show

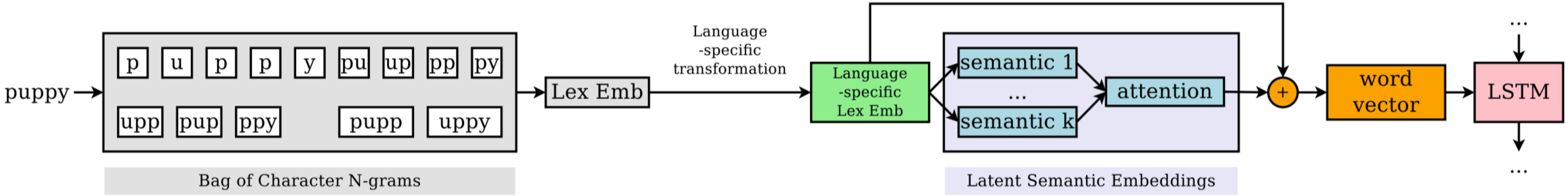
- Script differences / morphology (conjugation) differences

Units	Turkish	Uyghur
Graphemes	<yetmiyor> it is not enough	< قارييالمايدۇ> s/he can’t care for
Phonemes	/qarijalmajdu/	/jetmijoɾ/
Morphemes	/qari-jal-ma-jdu/	/jet-mi-joɾ/
Conjugations	qari + Verb + Pot + Neg + Pres + A3sg	jet + Verb + Neg + Prog1 + A3sg

Better Cross-lingual Models of Words

[Wang+19]

- A method for word encoding particularly suited for cross-lingual transfer



Handles spelling similarity

Handles consistent variations b/t languages

Attempts to capture latent "concepts"

- On MT for four low-resource languages, we find that:
  - SDE is better than other options such as character n-grams
  - SDE improves significantly over subword-based methods (e.g. used in multilingual BERT)

Wang, Xinyi, et al. "Multilingual Neural Machine Translation With Soft Decoupled Encoding." *ICLR 2019* (2019).

Morphological and Phonological Embeddings

[Chaudhary+18]

- A skilled linguist can create a "reasonable" morphological analyzer and transliterator for a new language in short order

- Our method: represent words by bag of
  - phoneme n-grams
  - lemma
  - morphological tags

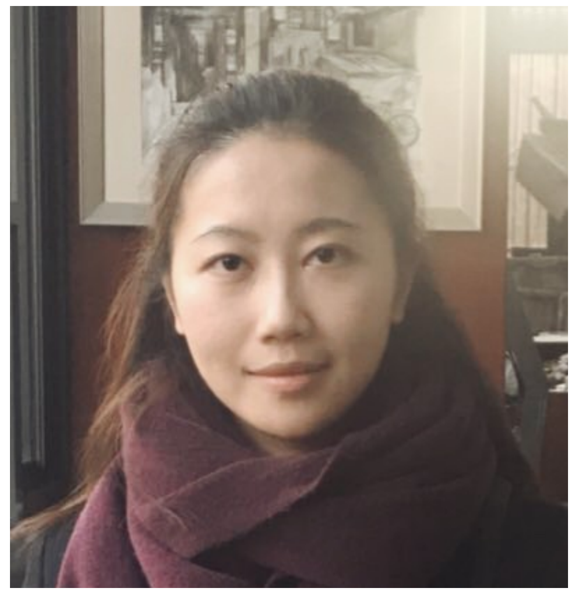
/jetmijor/ jet + Verb + Neg + Prog1 + A3sg

- Good results on NER/MT for Turkish->Uyghur, Hindi->Bengali transfer

Chaudhary, Aditi, et al. "Adapting word embeddings to new languages with morphological and phonological subword representations." *EMNLP 2018* (2018).

Data Augmentation via Reordering

[Zhou+ 2019]



- **Problem:** Source-target word order can differ significantly in methods that use monolingual pre-training

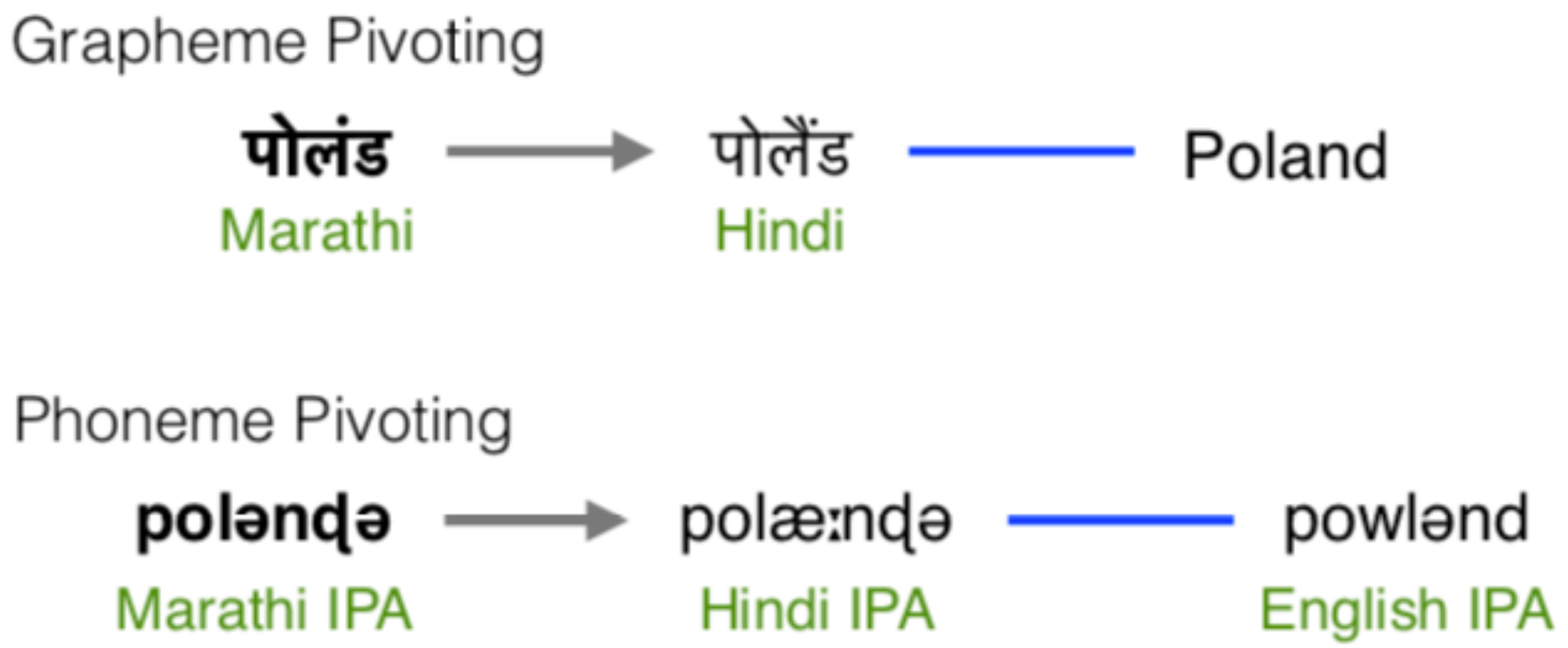
- **Solution:** Do re-ordering according to grammatical rules, followed by word-by-word translation to create pseudo-parallel data

Zhou, Chunting, et al. "Handling Syntactic Divergence in Low-resource Machine Translation." arXiv preprint arXiv:1909.00040 (2019).

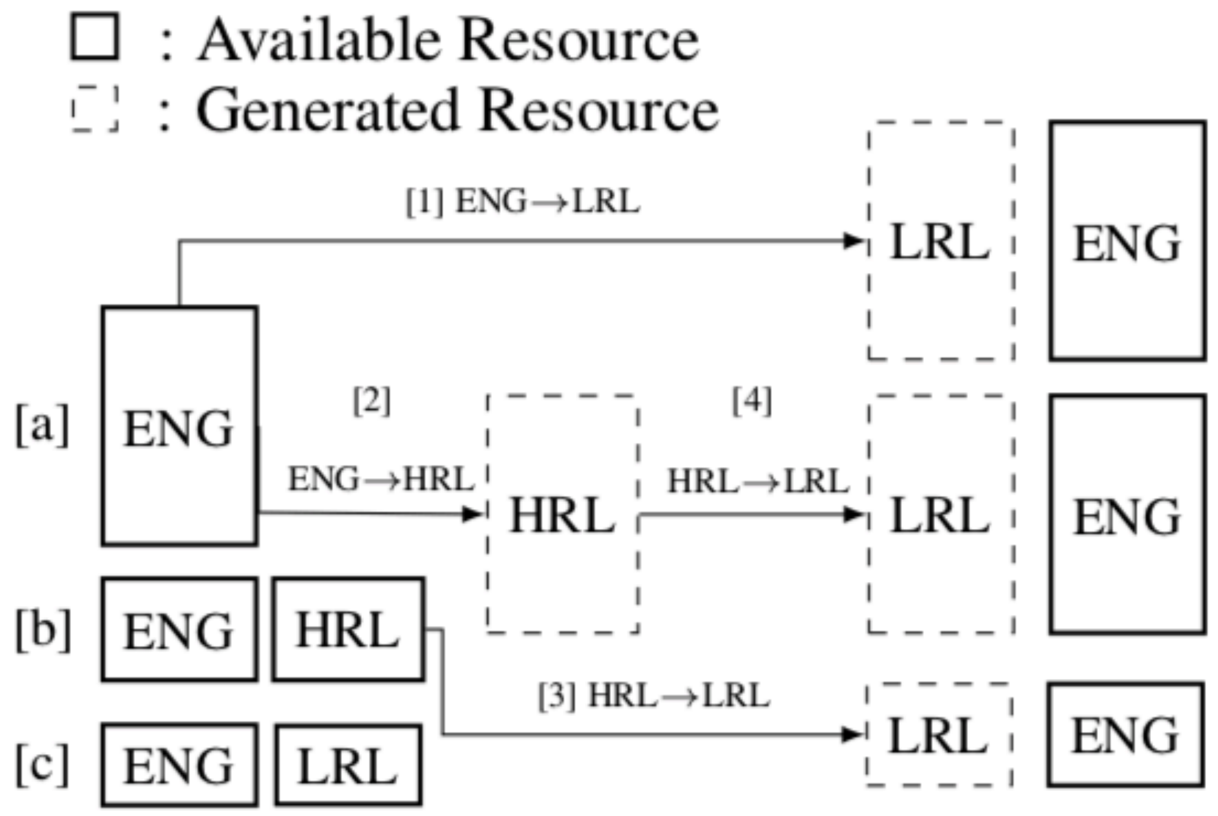
Pivoting Methods

- Tons of data in English, fair amount of data in a relatively high-resourced language (HRL) and want to process a low-resourced language (LRL)
- Pivoting through HRL can take advantage of available resources!

Zero-shot entity linking by pivoting through related language w/ phonetic representations [Rijhwani+19]



Data augmentation for NMT using related language and unsupervised lexicon induction [Xia+19]



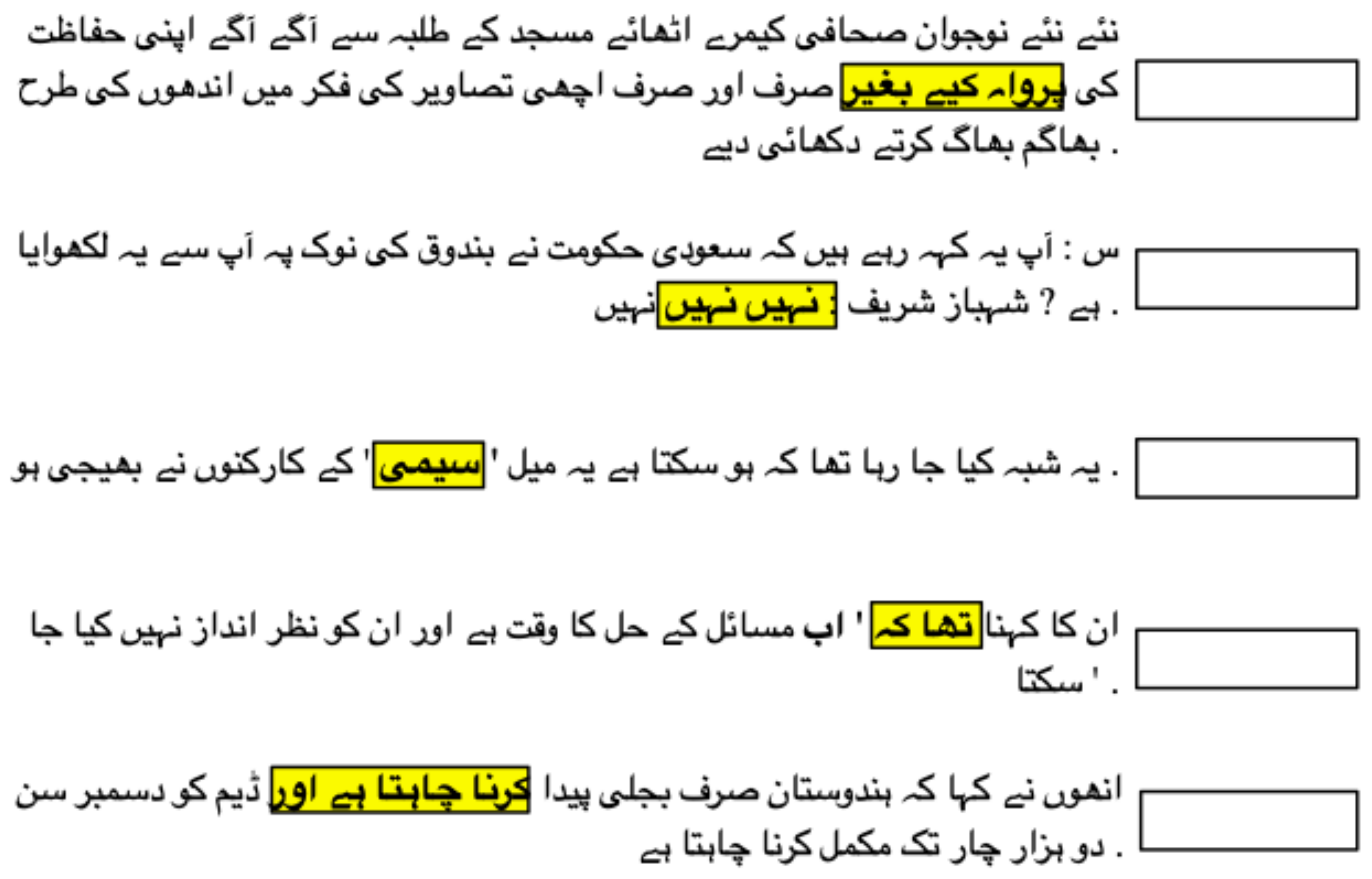
Rijhwani, Shruti, et al. "Zero-shot Neural Transfer for Cross-lingual Entity Linking." *AAAI 2019* (2019).
Xia, Mengzhou, et al. "Generalized Data Augmentation for Low-Resource Translation." *ACL 2019* (2019).

Active Learning

Creating Data

- Cross-lingual transfer is great, but no substitute for actual annotated data!
- **Active learning:** Ask human annotators to create data that maximally improves performance
- **What level of annotation?:**
  - *Sentence level* -- select hard-looking sentences
  - *Phrase-level* -- select hard-looking phrases
- **What criterion for selection?:**
  - *Uncertainty* -- phrases/sentences that look hard for the current model
  - *Representativeness* -- how well does it cover examples in the data?

Simple Example of MT



- Phrase-level annotation

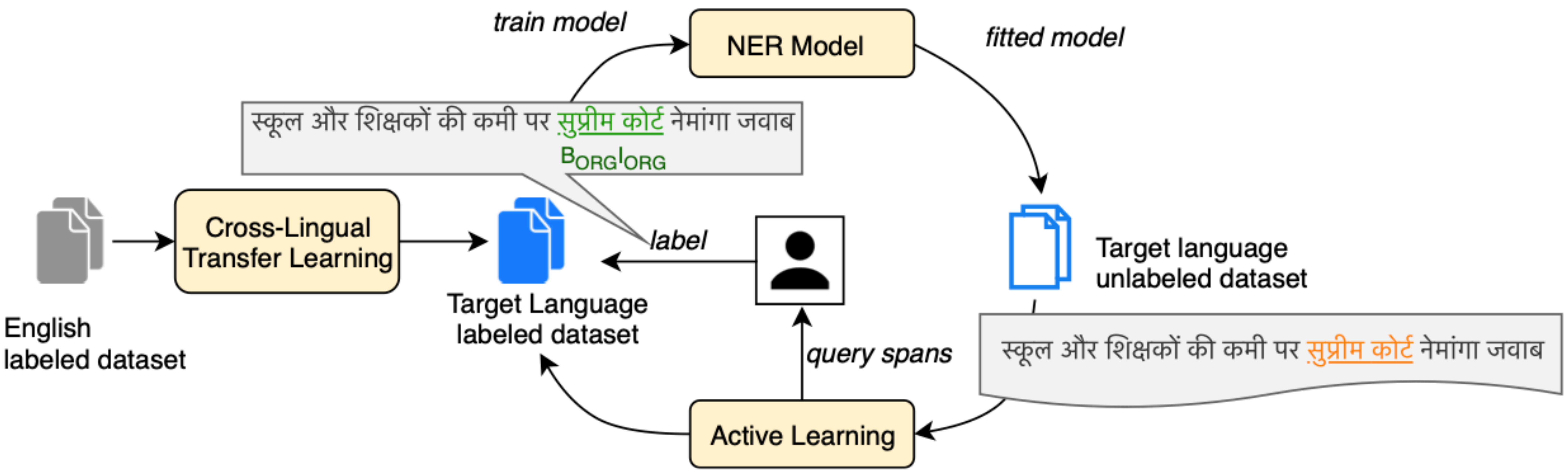
- Select phrases that are infrequent in parallel data (uncertain), but frequent in monolingual data (representative)

Bloodgood, Michael, and Chris Callison-Burch. "Bucking the trend: Large-scale cost-focused active learning for statistical machine translation." *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010.

Active Learning+Cross-lingual Transfer

[Chaudhary+ 19]

- Train a cross-lingual model, gradually improve via monolingual annotation



- Select examples where the cross-lingual model *has uncertain predictions*

- Using both cross-lingual and active supervision improves significantly over using just one

Chaudhary, Aditi, et al. "A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers." *arXiv preprint arXiv:1908.08983* (2019).

Conclusion

The Low-resource NLP Toolbox

- Lots of paired data <X,Y>
  -> **supervised learning**
- Lots of source or target data X or Y
  -> **monolingual pre-training, back-translation**
- Paired data in another, similar language <X',Y> or <X,Y'>
  -> **multilingual training, transfer**
- Can ask speakers to do a little work to generate data
  -> **active learning**

Use any tool available to you!

Thank you to sponsors:
(views expressed here do not reflect views of the US government)