TeIAS | Tehran Institute for Advanced Studies

بنام خداوند جان و خرد

Multilingual Learning

Mohammad Taher Pilehvar

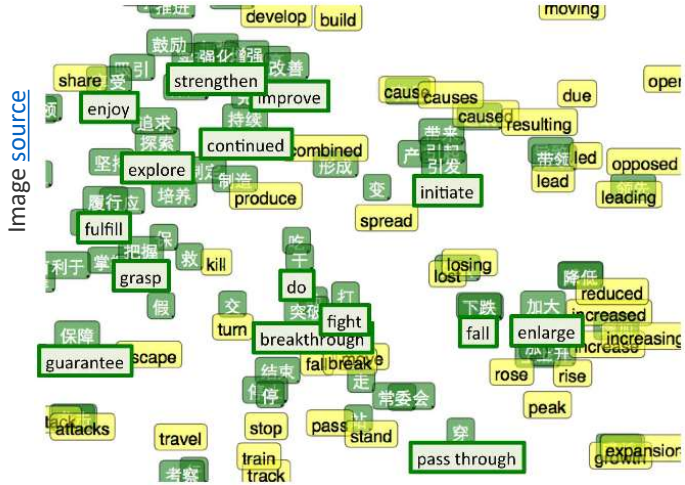


Image source

Natural Language Processing 1400
https://teias-courses.github.io/nlp00/

Most materials of these slides are taken from Graham Neubig, Carnegie Mellon University, Language Technologies Institute, CS11-747 course

Most languages are resource-poor

https://meta.wikimedia.org/wiki/List_of_Wikipedias

1M+ articles: 18

100K articles: 52 (Farsi has around 900K articles)

10K articles: 89

1K+ articles: 123

100+ articles: 30

Not enough monolingual data for many languages; even less annotated data for NMT, sequence labeling, etc.

Multilingual learning

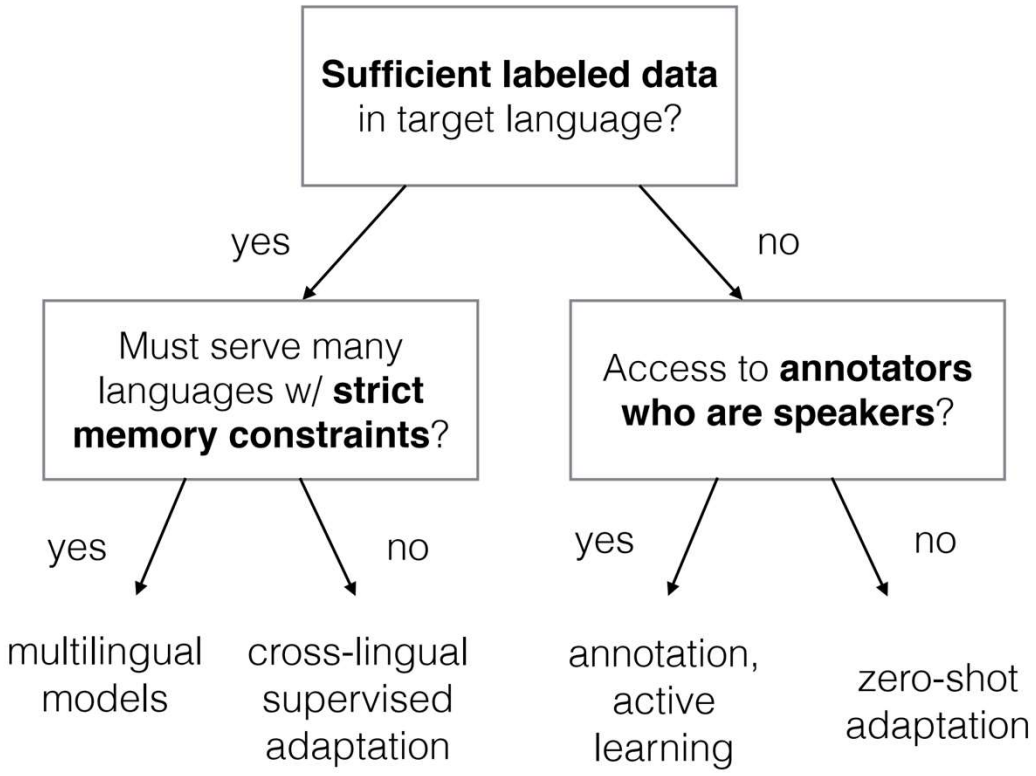
Ideally, we would like to have models that can process input texts from multiple languages

Why?

- **Transfer Learning**
  Improve accuracy on lower resource languages by transferring knowledge from higher-resource languages

- **Memory Savings**
  Use one model for all languages, instead of one for each

High-level Multilingual Learning Flowchart

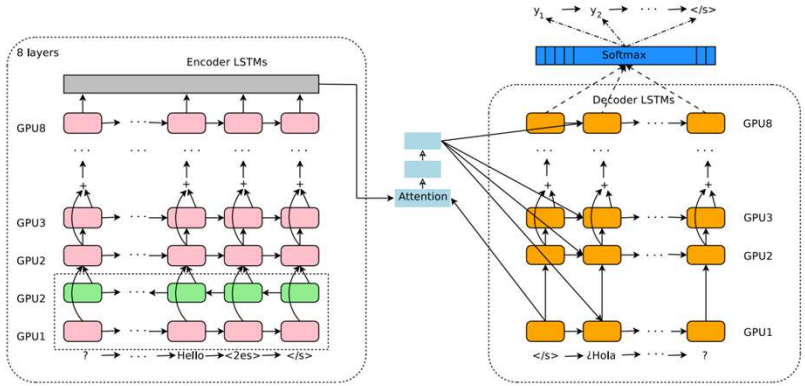


Multilingual learning

Multilingual seq2seq models

Google's Multilingual NMT (Johnson et al, 2017)

- It is possible to learn a single model that handles several languages
- Even as simple as adding a tag about the target language for generation



<fr> this is an example → ceci est un exemple

<ja> this is an example → これは例です

Multilingual seq2seq models

Multilingual BERT (mBERT)

- It is possible to learn a single model that handles several languages
- Or even just processing different input languages using the same network (We and Dredze, 2019)

ceci est un exemple
これは例です

Difficulties in Fully Multilingual Learning

- For a fixed sized model, the per-language capacity decreases as we increase the number of languages (Conneau et al, 2019)

- Increasing the number of low-resource languages –> decrease in the quality of high-resource language translations (Aharoni et al, 2019)

How to mitigate?
Better data balancing, better parameter sharing

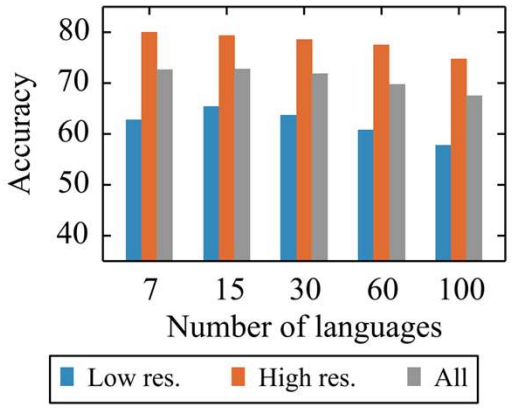
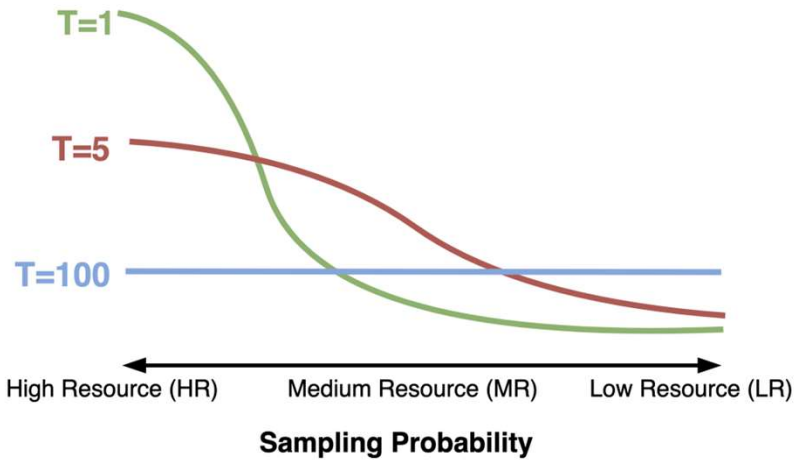


Figure 2: The transfer-interference trade-off: Low-resource languages benefit from scaling to more languages, until dilution (interference) kicks in and degrades overall performance.

Heuristic Sampling of Data

- Massively Multilingual Neural Machine Translation in the Wild (Arivazhgan et al, 2019)

- Sample data based on dataset size scaled by a temperature term



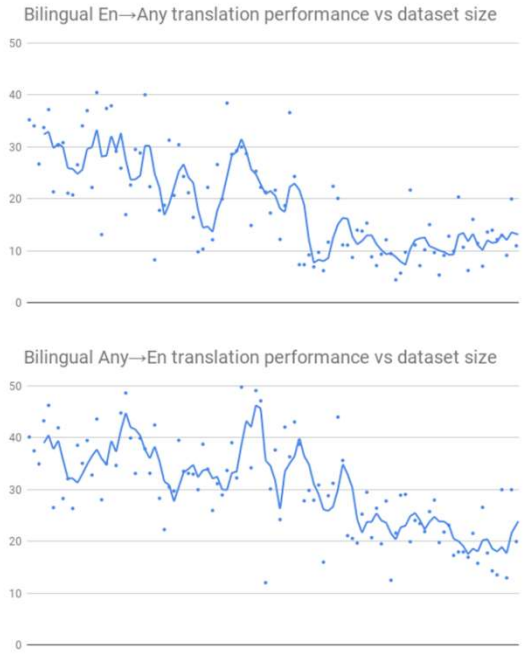


Figure 2: Quality (measured by BLEU) of individual bilingual models on all 204 supervised language pairs, measured in terms of BLEU (y-axes). Languages are arranged in decreasing order of available training data from left to right on the x-axes (pair ids not shown for clarity). Top plot reports BLEU scores for translating from English to any of the other 102 languages. Bottom plot reports BLEU scores for translating from any of the other 102 languages to English. Performance on individual language pairs is reported using dots and a trailing average is used to show the trend.

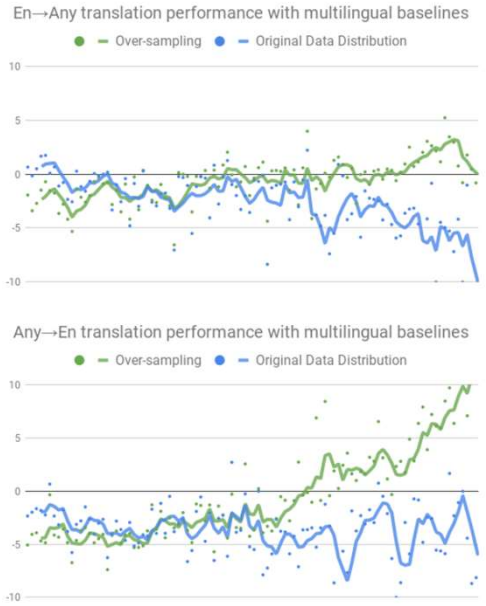


Figure 3: Effect of sampling strategy on the performance of multilingual models. From left to right, languages are arranged in decreasing order of available training data. While the multilingual models are trained to translate both directions, *Any→En* and *En→Any*, performance for each of these directions is depicted in separate plots to highlight differences. Results are reported relative to those of the bilingual baselines (2). Performance on individual language pairs is reported using dots and a trailing average is used to show the trend. The colors correspond to the following sampling strategies: (i) Blue: original data distribution, (ii) Green: equal sampling from all language pairs. Best viewed in color.

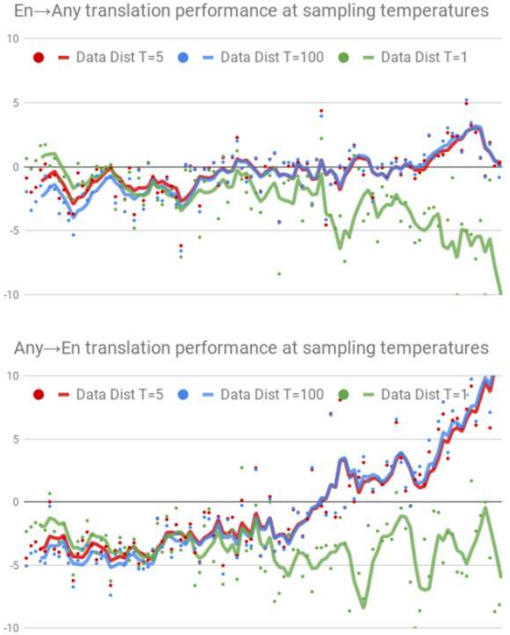


Figure 5: Effect of varying the sampling temperature on the performance of multilingual models. From left to right, languages are arranged in decreasing order of available training data. Results are reported relative to those of the bilingual baselines (2). Performance on individual language pairs is reported using dots and a trailing average is used to show the trend. The colors correspond to the following sampling strategies: (i) Green: True data distribution ($T = 1$) (ii) Blue: Equal sampling from all language pairs ($T = 100$) (iii) Red: Intermediate distribution ($T = 5$). Best viewed in color.

Adapters

MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer (Pfeiffer et al, 2020)

Standard cross-lingual transfer:

1) Fine-tune it on labelled data of a downstream task in a source language
2) Apply it directly to perform inference in a target language

Downside:

Multilingual initialization balances many languages. It is thus not suited to excel at a specific language at inference time.

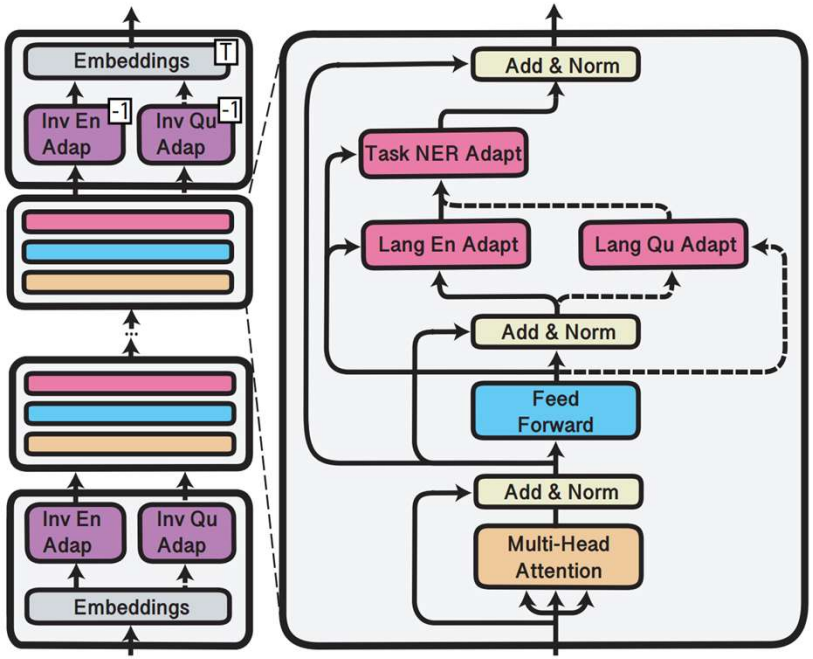
Adapters

MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer (Pfeiffer et al, 2020)

Language adapters

Learn language-specific transformations

- Trained for MLM on a language (to make the multilingual model more suitable for that language)
- Kept fixed during task-specific training
- Zero-shot: simply replace the source language adapter with its target language component (in the figure, English to Quechua)



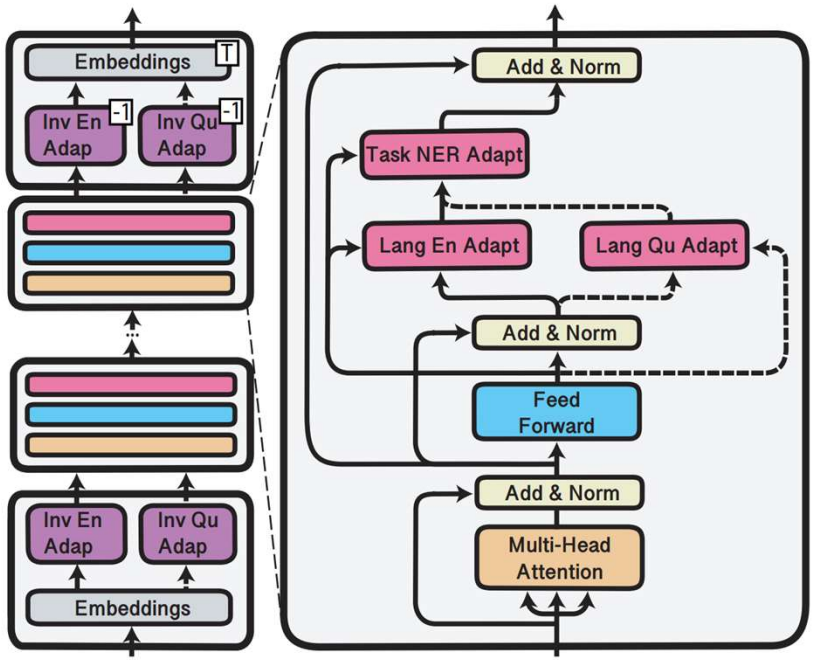
Adapters

MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer (Pfeiffer et al, 2020)

Task adapters

Learn task-specific knowledge across languages

- Task adapters are the only parameters that are updated when training on a downstream task (e.g., NER)



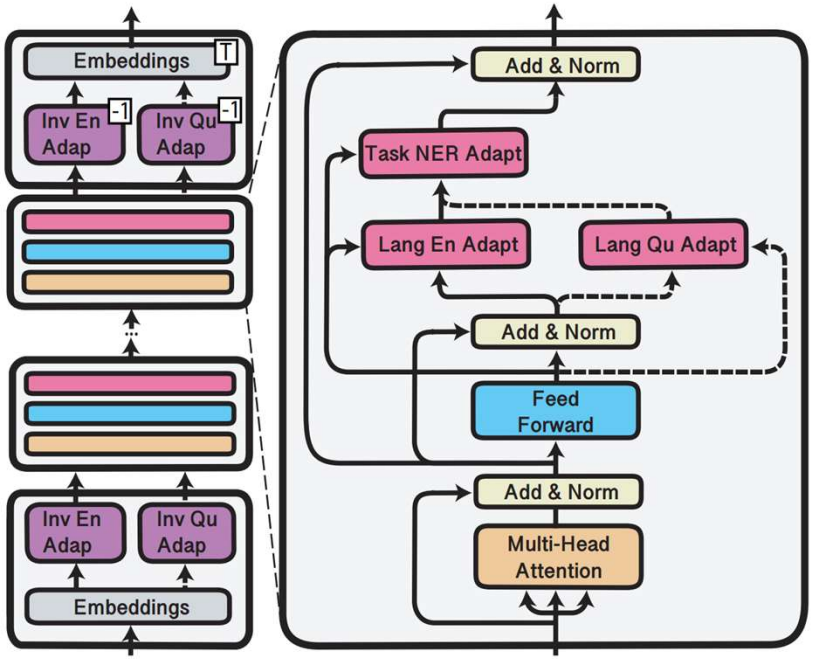
Adapters

MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer (Pfeiffer et al, 2020)

Invertible adapters

Capture token-level language-specific transformations

- The majority of the "parameter budget" of pretrained multilingual models is spent on token embeddings of the shared multilingual vocabulary (mBERT 110M parameters, 110K*768 token parameters).
- Mitigate this mismatch between multilingual and target language vocabulary.
- Trained with language adapters using MLM.

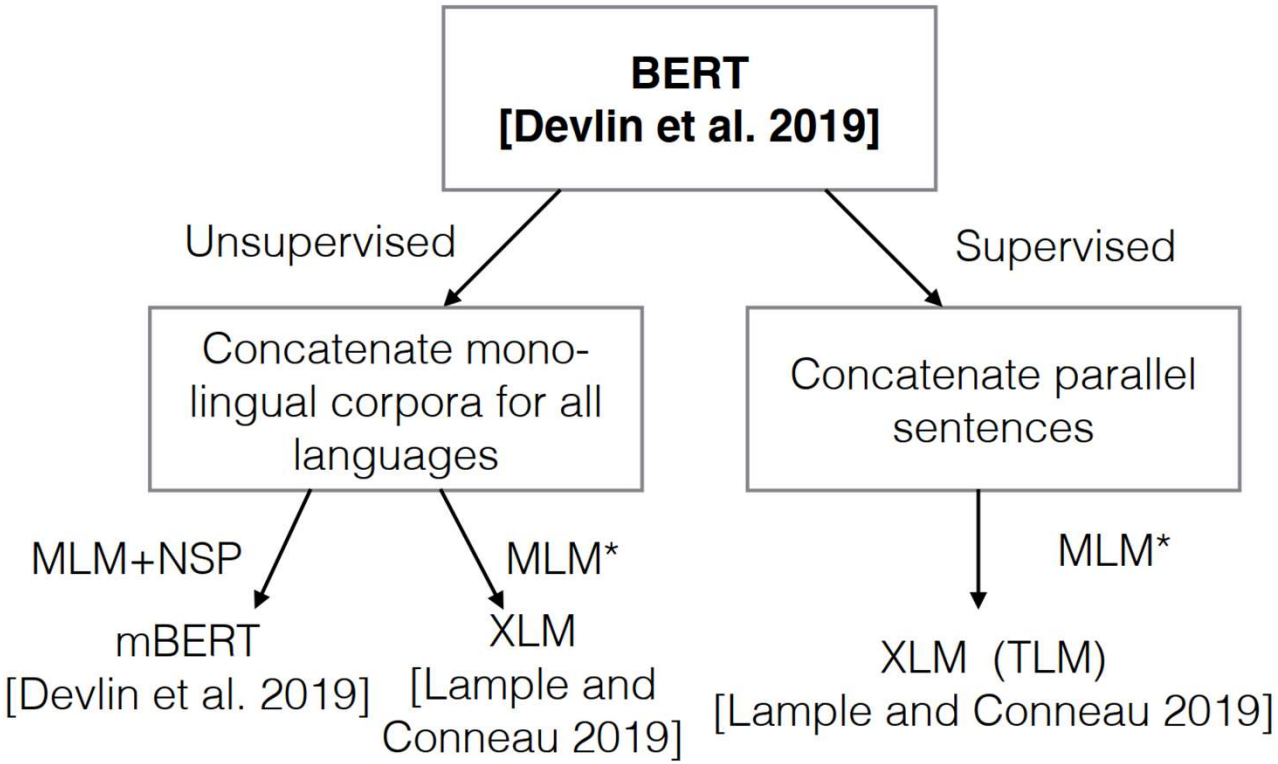


Multilingual Pre-trained Models

Multi-lingual pre-training

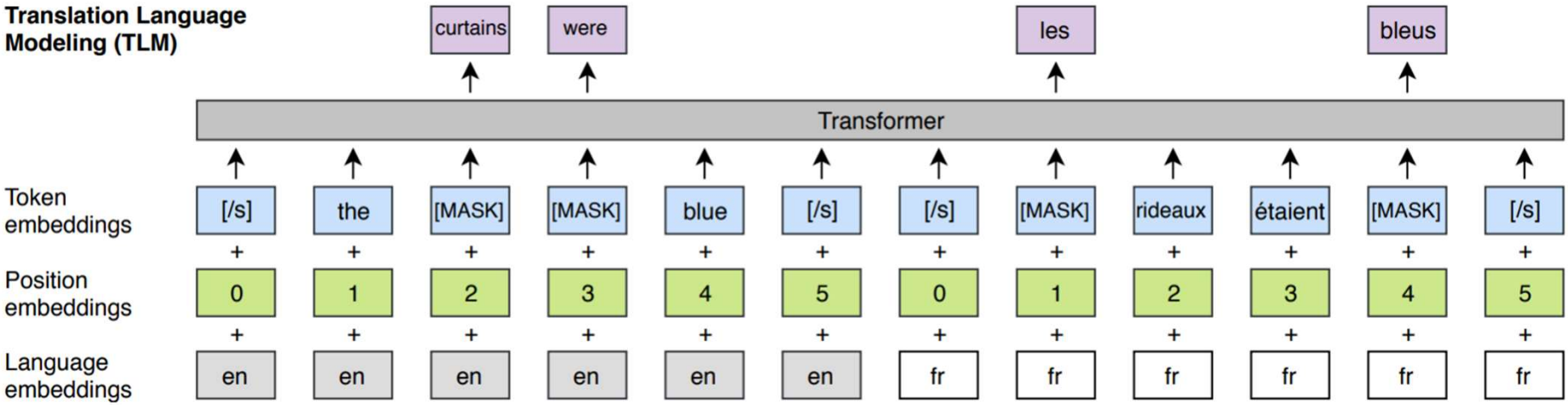
- Language model pre-training has shown to be effective for many NLP tasks, eg. BERT.
- BERT uses masked language model (MLM) and next sentence prediction (NSP) objective.
- Models such as mBERT, XLM, XLM-R extend BERT for multi-lingual pre-training.

Multi-lingual pre-training



Multi-lingual MLM

- Also called translation language modeling
- XLM (Lample and Conneau, 2019)
- XLM-R: 100 languages, based on RoBERTa, 2TB+ of CommonCrawl data

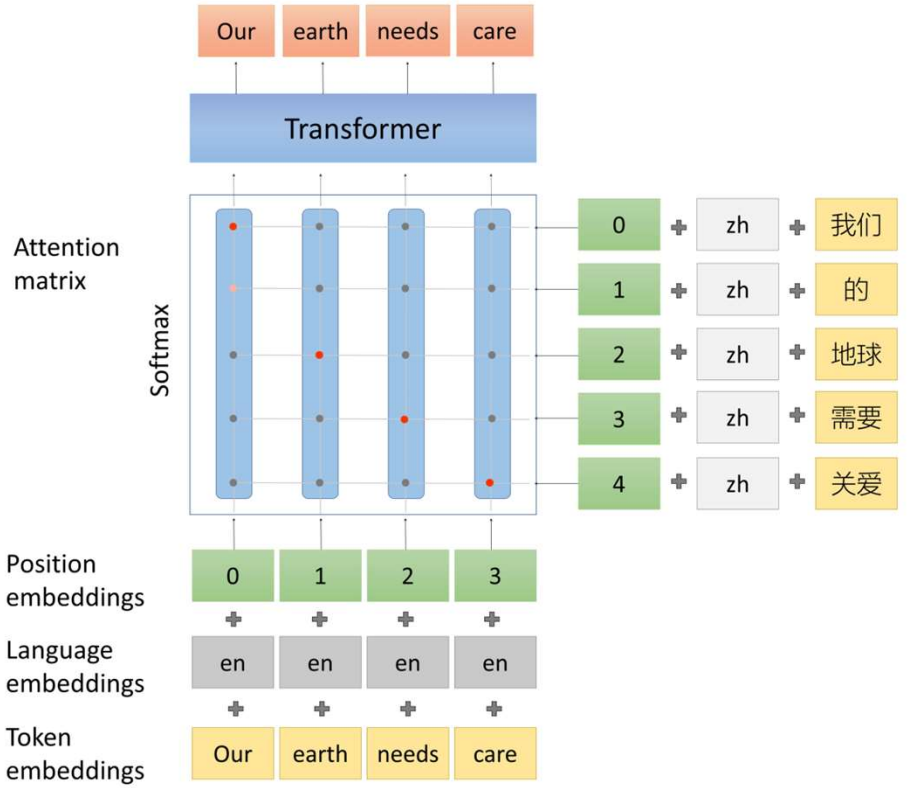


More explicit alignment objectives

Unicoder (Huang et al., 2019)

Cross-lingual word recovery
Learn the underlying alignment between two languages

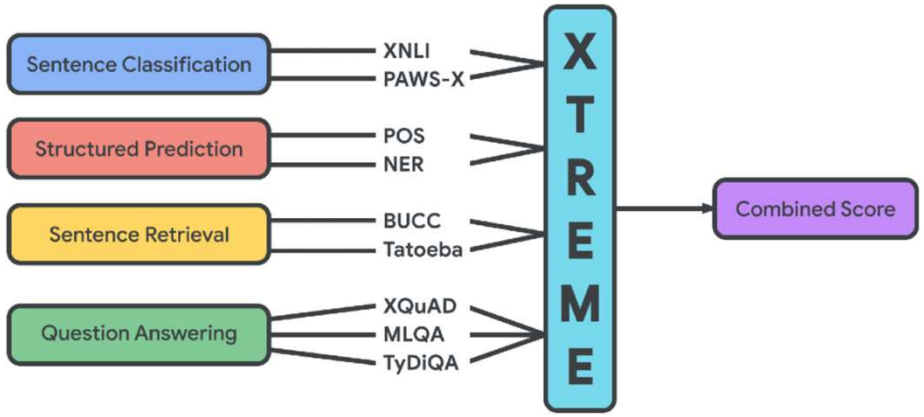
- Trained on sentence pairs
- It can be trained by recovering all words (given that input is not seen directly)



Evaluation of multilingual representations

Large-scale benchmarks that cover many tasks

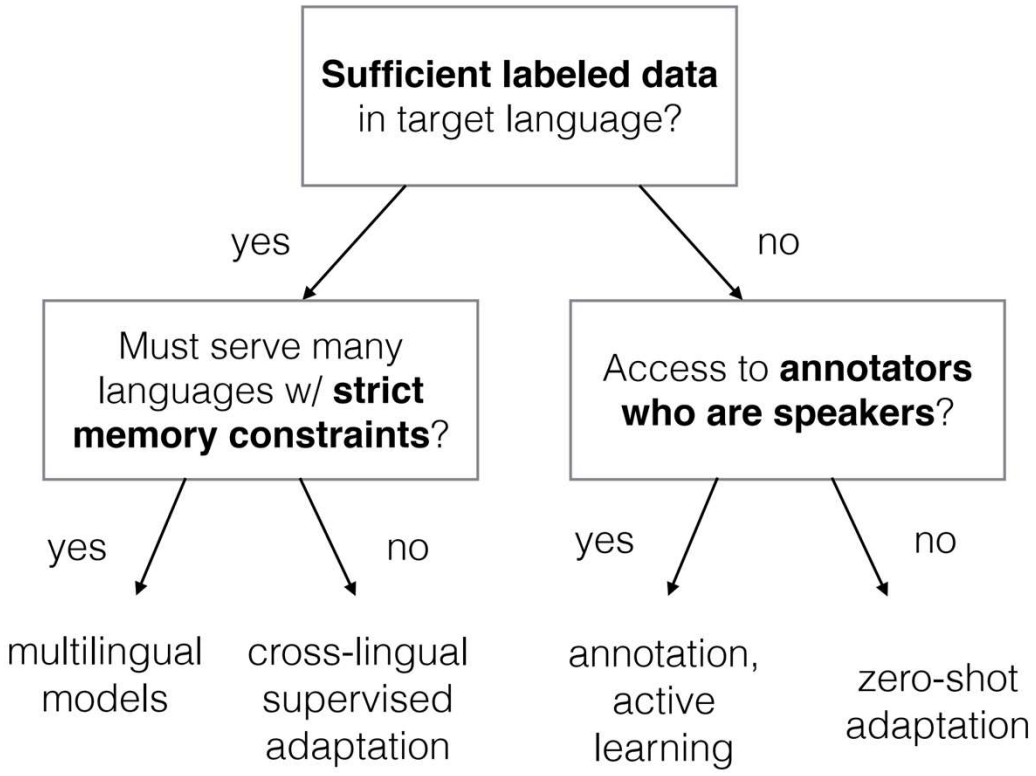
XTREME: 40 languages, 9 tasks (Hu et al., 2020)



XGLUE: less typologically diverse but contains generation (Liang et al. 2020)

XTREME-R: harder version based on XTREME (Ruder et al. 2021)

High-level Multilingual Learning Flowchart



Cross-lingual Transfer Learning

Cross-lingual Transfer Learning

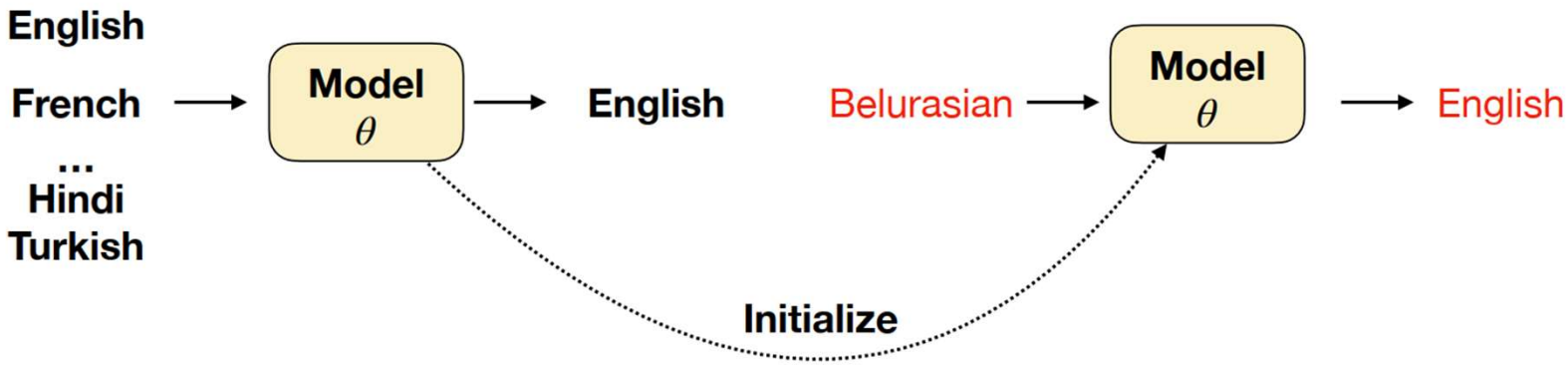
CLTL leverages data from one or more high-resource source languages

Popular strategies

- Multilingual learning (above)
- Pre-train and fine-tune
- Zero-shot transfer
- Annotation projection

Pre-train and fine-tune

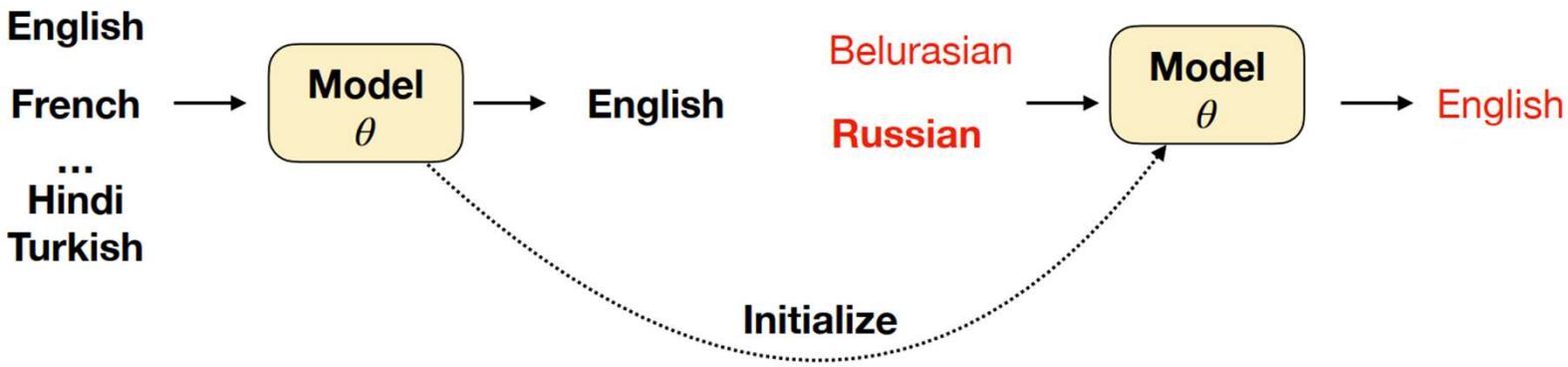
Rapid adaptation of Neural Machine Translation to New Languages (Neubig and Hu, 2018)



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

Similar language regularization

Rapid adaptation of Neural Machine Translation to New Languages (Neubig and Hu, 2018)



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

Zero-shot transfer

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

- 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!

Zero-shot transfer

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

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

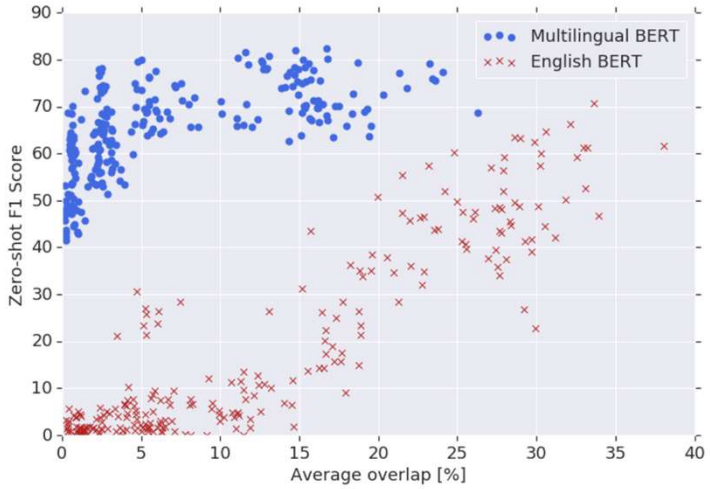
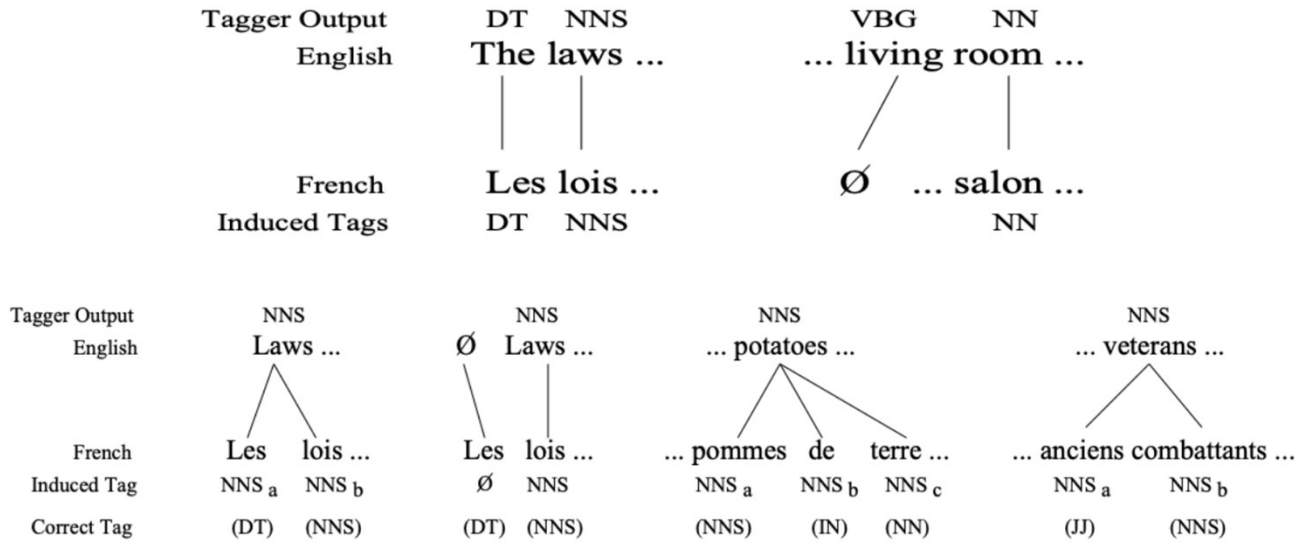


Figure 1: Zero-shot NER F1 score versus entity word piece overlap among 16 languages. While performance using EN-BERT depends directly on word piece overlap, M-BERT's performance is largely independent of overlap, indicating that it learns multilingual representations deeper than simple vocabulary memorization.

Annotation projection

Induce annotations in the target language using parallel data or bilingual dictionary (Yarowsky et al, 2001).



Which language to use?

When transferring from another language, it is ideal that it is:

- Similar to the target language
- Data-rich

Choosing Transfer Languages for Cross-Lingual Learning
(Lin et al, 2019)

	Method	MT	EL	POS	DEP
dataset	word overlap o_w	28.6	30.7	13.4	52.3
	subword overlap o_{sw}	29.2	–	–	–
	size ratio s_{tf}/s_{tk}	3.7	0.3	9.5	24.8
	type-token ratio d_{ttr}	2.5	–	7.4	6.4
ling. distance	genetic d_{gen}	24.2	50.9	14.8	32.0
	syntactic d_{syn}	14.8	46.4	4.1	22.9
	featural d_{fea}	10.1	47.5	5.7	13.9
	phonological d_{pho}	3.0	4.0	9.8	43.4
	inventory d_{inv}	8.5	41.3	2.4	23.5
	geographic d_{geo}	15.1	49.5	15.7	46.4
LANGRANK (all)		51.1	**63.0**	**28.9**	**65.0**
LANGRANK (dataset)		**53.7**	17.0	26.5	**65.0**
LANGRANK (URIEL)		32.6	58.1	16.6	59.6

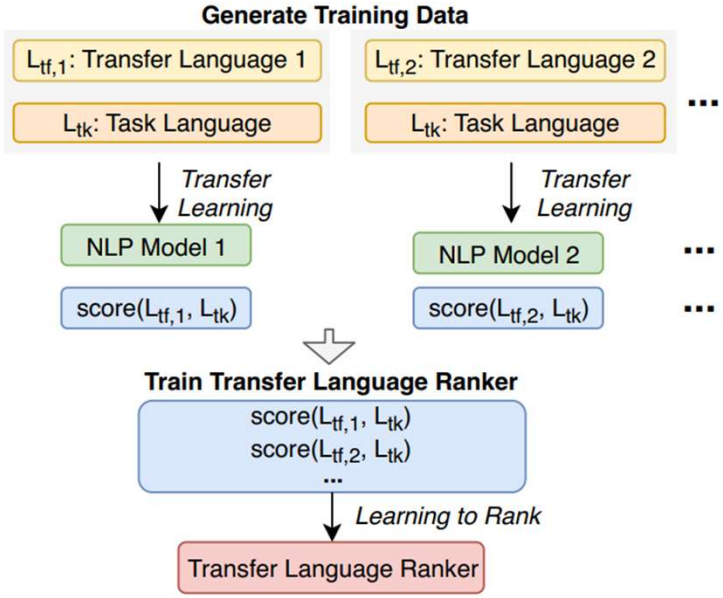
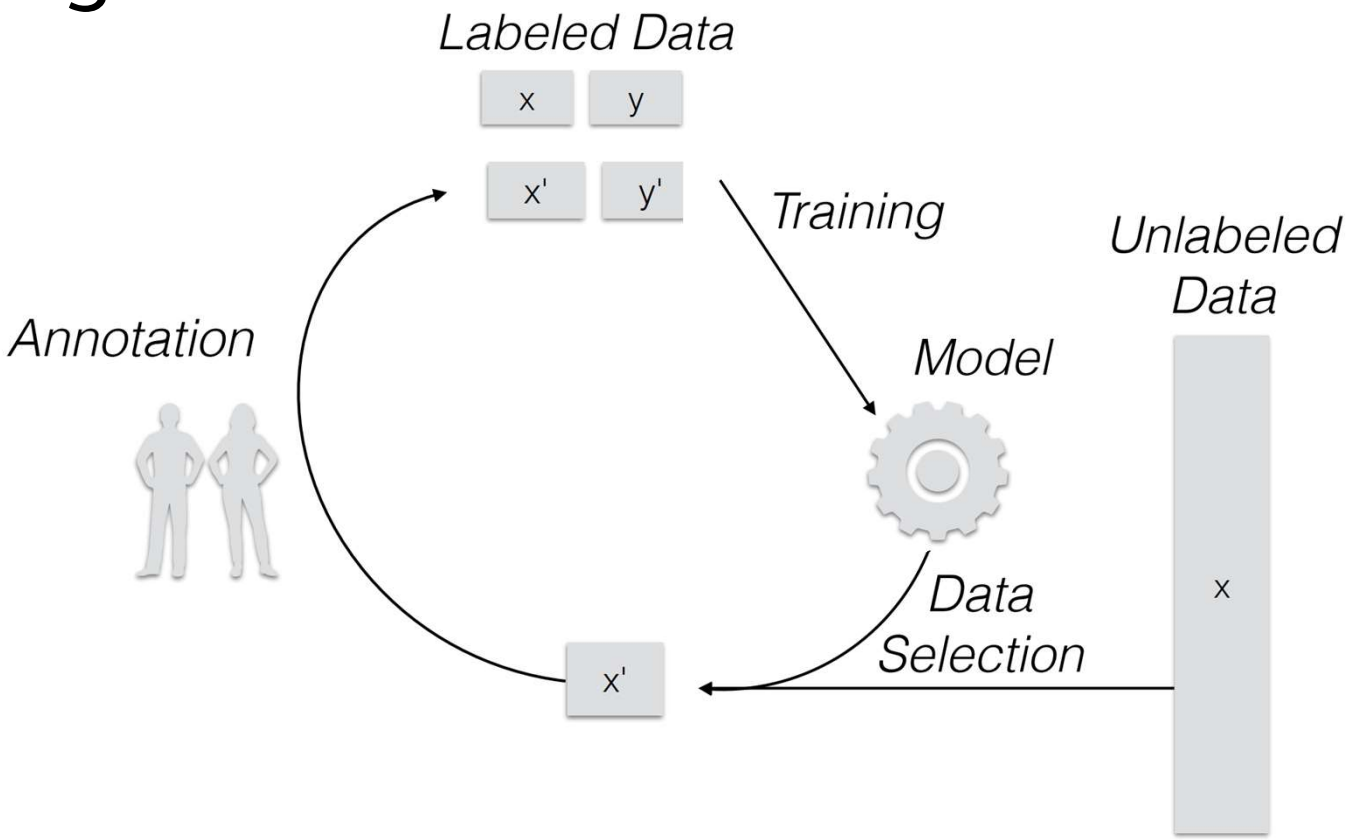


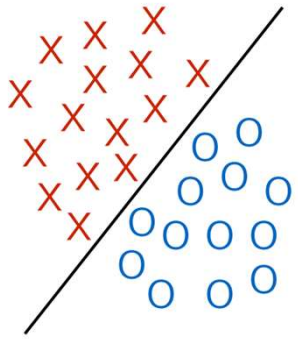
Figure 1: Workflow of learning to select the transfer languages for an NLP task: (1) train a set of NLP models with all available transfer languages and collect evaluation scores, (2) train a ranking model to predict the top transfer languages.

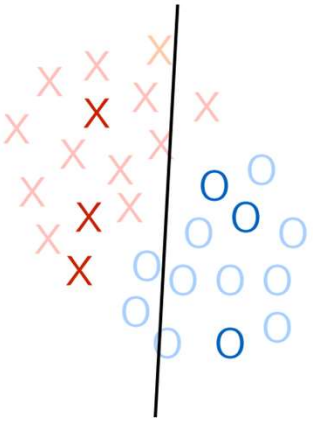
Creating New Data

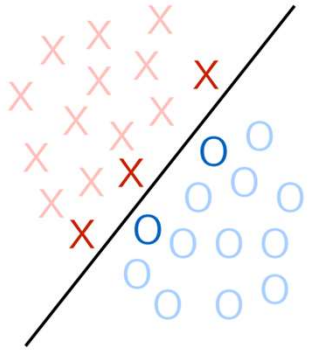
Active learning



Active learning







Active learning

- **Uncertainty:** we want data that are **hard** for our current models to handle

- **Representativeness:** we want data that are similar to the data that we are annotating

Uncertainty sampling criteria

Confidence: lower top-1 confidence = more uncertain

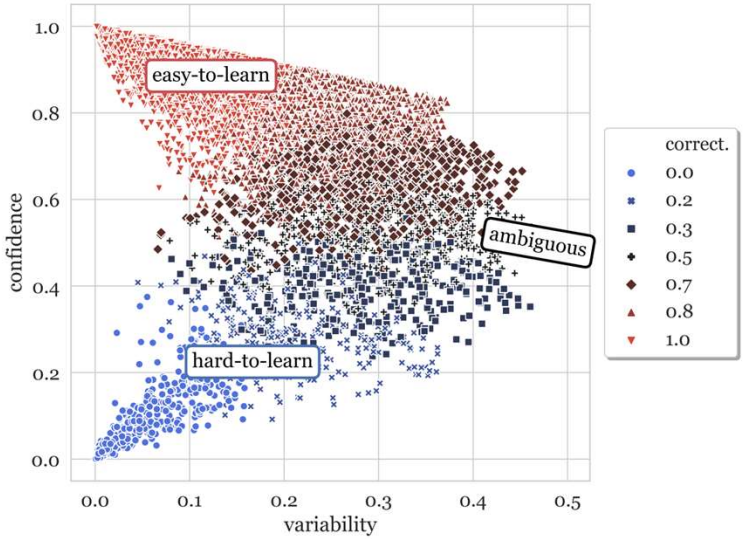
$$\hat{y} = \underset{y}{\operatorname{argmax}} \log P(y|x)$$

$$\text{top1}(x) = \log P(\hat{y}|x)$$

Margin: smaller difference between first and second candidates = more uncertain

$$\text{margin}(x) = \log P(\hat{y}|x) - \max_{y \neq \hat{y}} \log P(y|x)$$

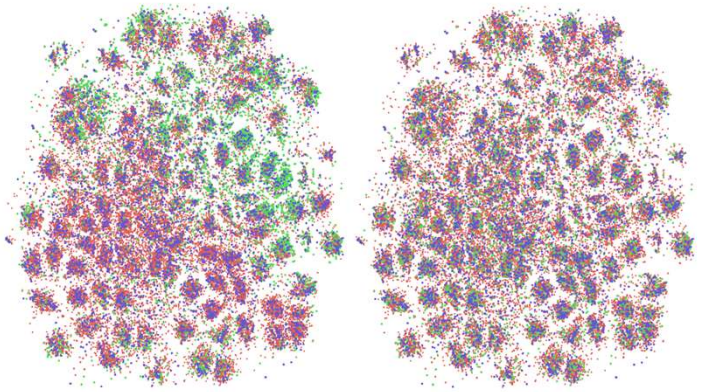
Cartography:



Representativeness

- Sener, Ozan, and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach.", ICLR 2018

- Based on similarity to other samples



(a) Uncertainty Oracle (b) Our Method

Figure 5: tSNE embeddings of the CIFAR dataset and behavior of uncertainty oracle as well as our method. For both methods, the initial labeled pool of 1000 images are shown in blue, 1000 images chosen to be labeled in green and remaining ones in red. Our algorithm results in queries evenly covering the space. On the other hand, samples chosen by uncertainty oracle fails to cover the large portion of the space.

Questions