Multilinguality in Transformers Architectures

Edoardo Barba

barba@di.uniroma1.it

Sapienza University of Rome

SAPIENZA NLP

European Research Council

Established by the European Commission

Table of Contents

Table of Contents

Transfer Learning: What is it?

Definition from Wikipedia: “Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.”

Transfer Learning: What is it?

Definition from Wikipedia: “Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.”

You may have heard of it: Word2Vec, ELMo, BERT, BART, …

Transfer Learning: What is it?

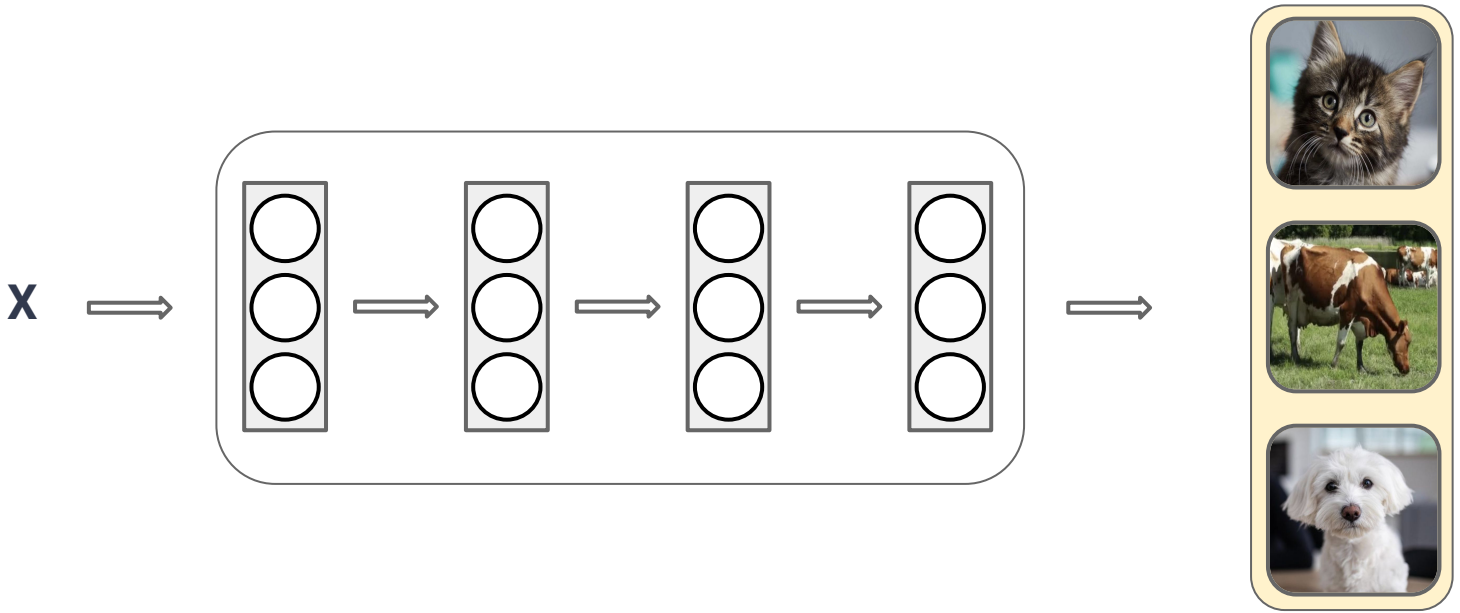
Definition from Wikipedia: “Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.”

You may have heard of it: Word2Vec, ELMo, BERT, BART, …

But how does it work?

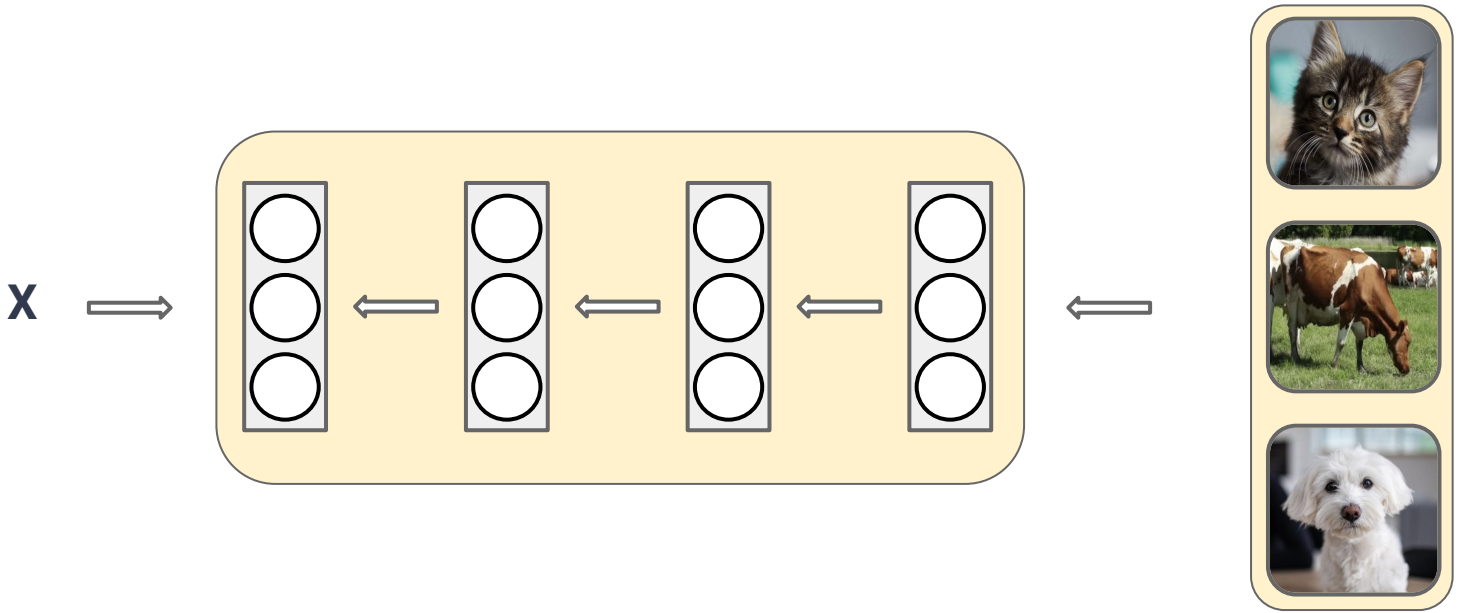
Transfer Learning: What is it?

Legend has it, the first research area where this problem was investigated is Computer Vision.



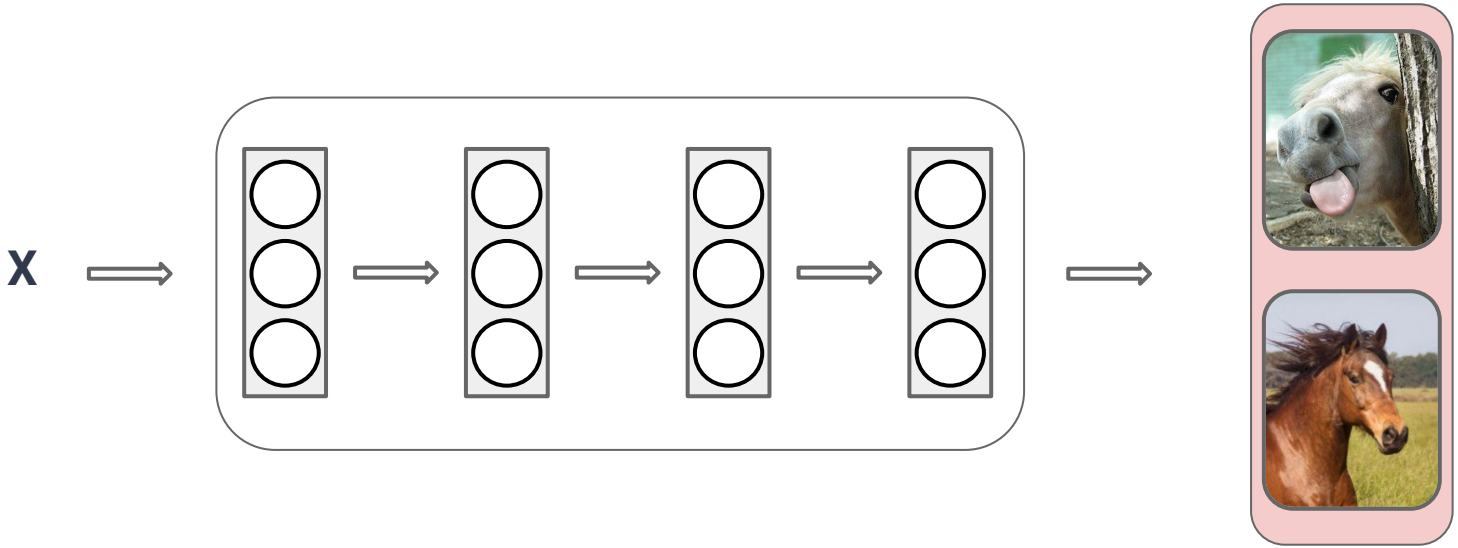
Transfer Learning: What is it?

Legend has it, the first research area where this problem was investigated is Computer Vision.



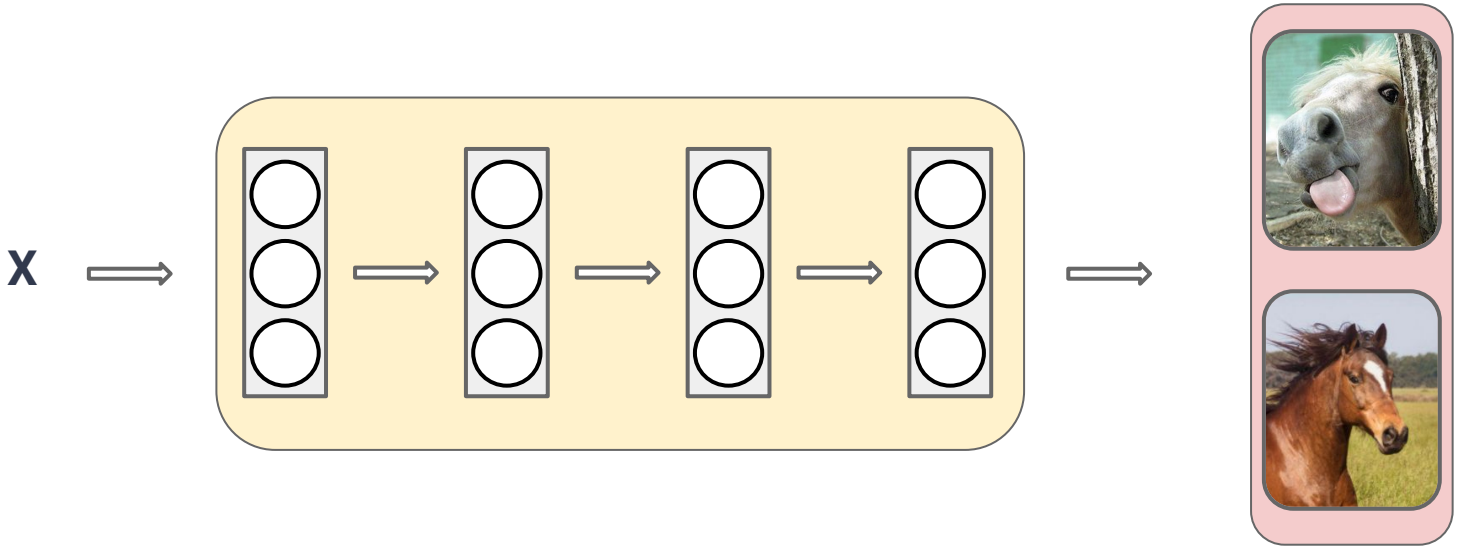
Transfer Learning: What is it?

Legend has it, the first research area where this problem was investigated is Computer Vision.



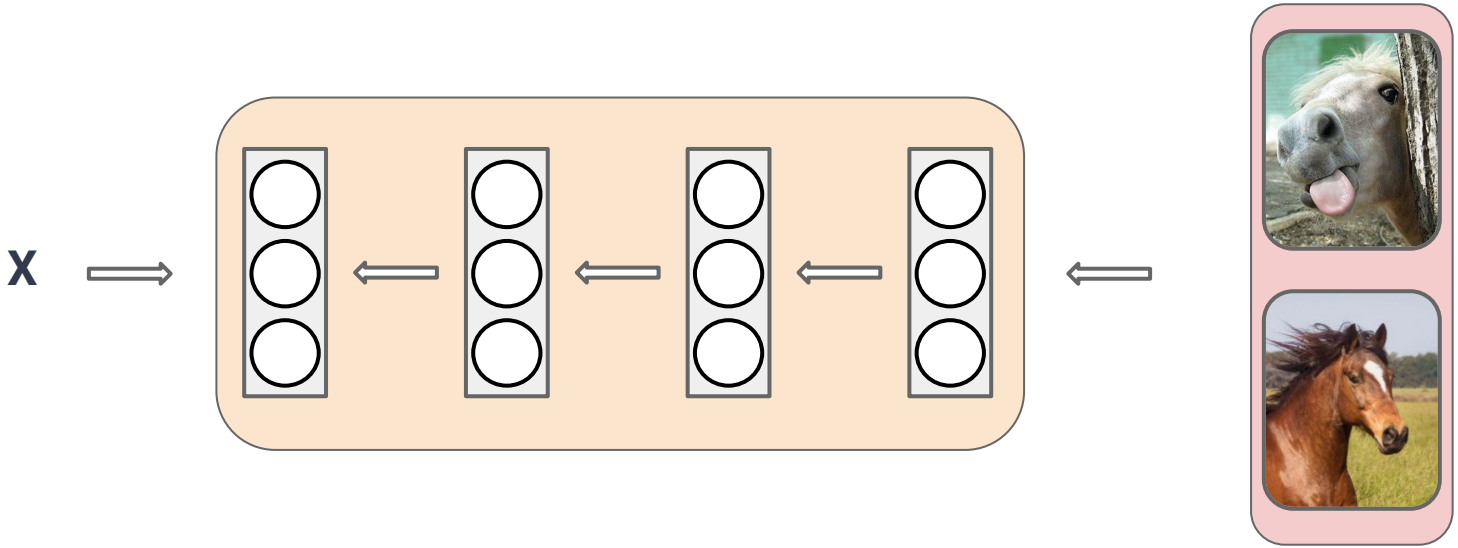
Transfer Learning: What is it?

Legend has it, the first research area where this problem was investigated is Computer Vision.



Transfer Learning: What is it?

Legend has it, the first research area where this problem was investigated is Computer Vision.



Transfer Learning: What is it?

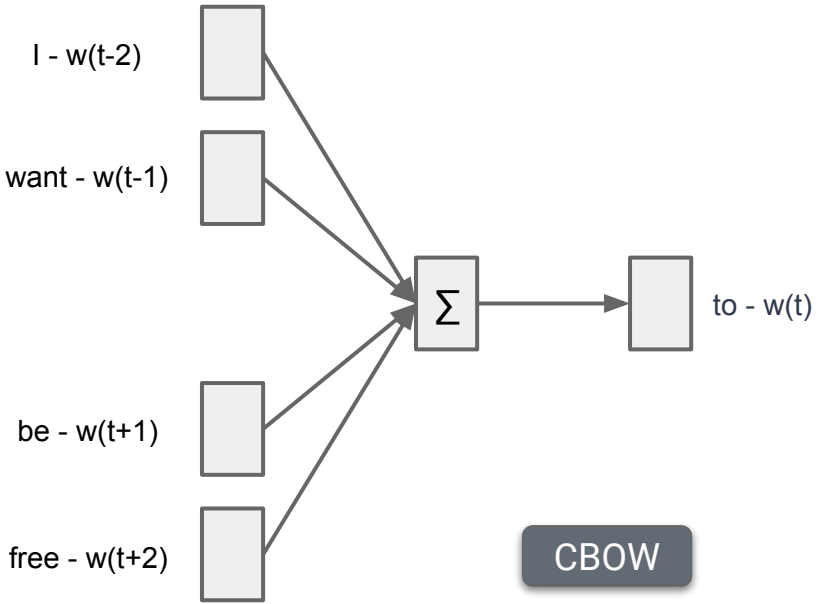
- **Time**: model pretrained on similar task usually guarantees faster convergence
- **Data**: models trained on huge amounts of data can be used to mitigate the lack of data for other tasks
- **Performances**: initializing the model weights with a pretrained model can boost the final performances

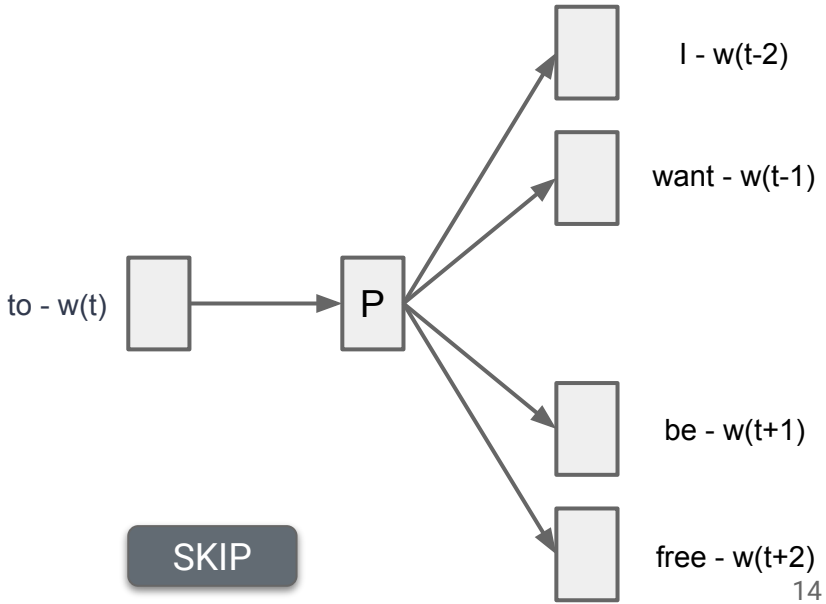
The brain work in this way...

Table of Contents

Transfer Learning in NLP

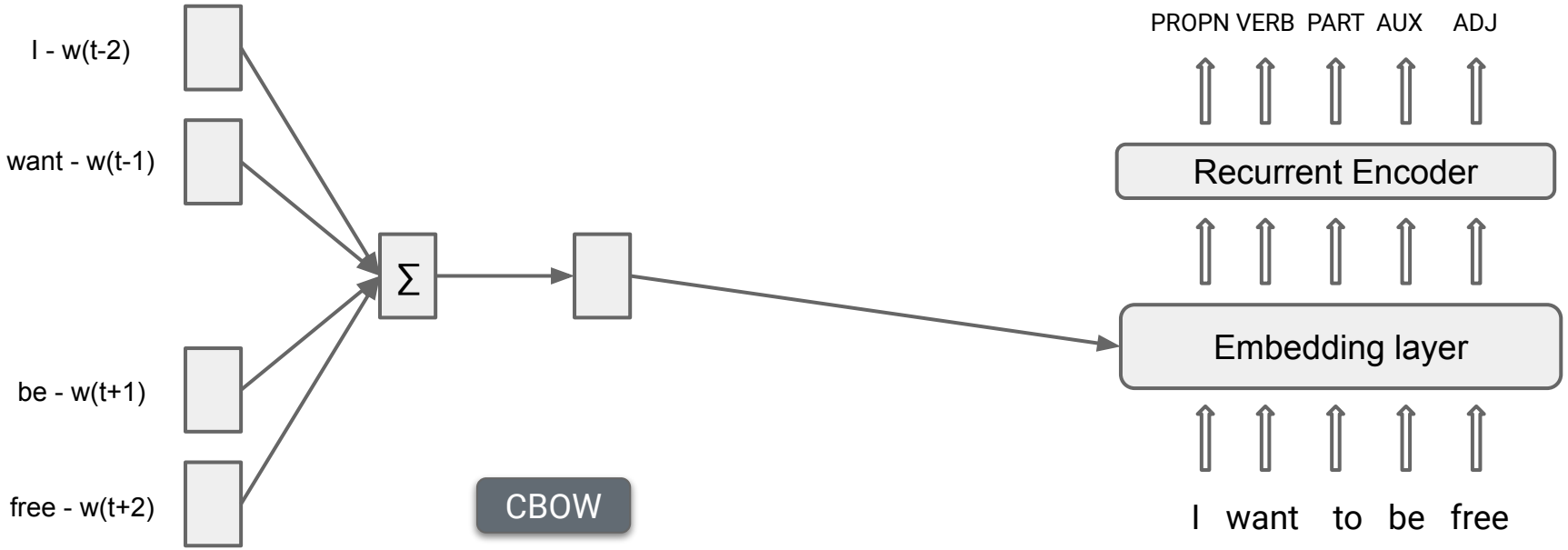
In Natural Language Processing, we **pretrain words representation**!





Transfer Learning in NLP

In Natural Language Processing, we **pretrain words representation**!



Transfer Learning in NLP

In Natural Language Processing, we **pretrain words representation**!

- But words are **ambiguous**!

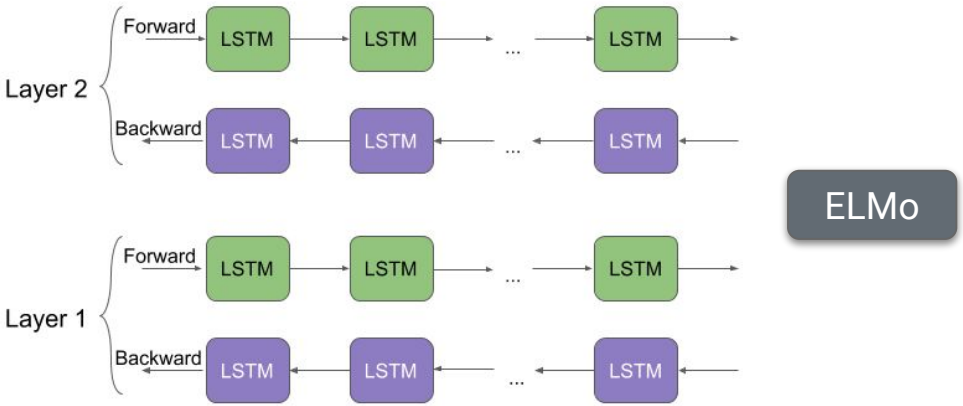
Transfer Learning in NLP

In Natural Language Processing, we **pretrain words representation**!

- But words are **ambiguous**!
- And the meaning of a word varies depending on the context in which it appears…
  - *The boss **fired** his secretary today* *(terminate the employment)*
  - *The neurons **fired** fast* *(generate an electrical impulse)*

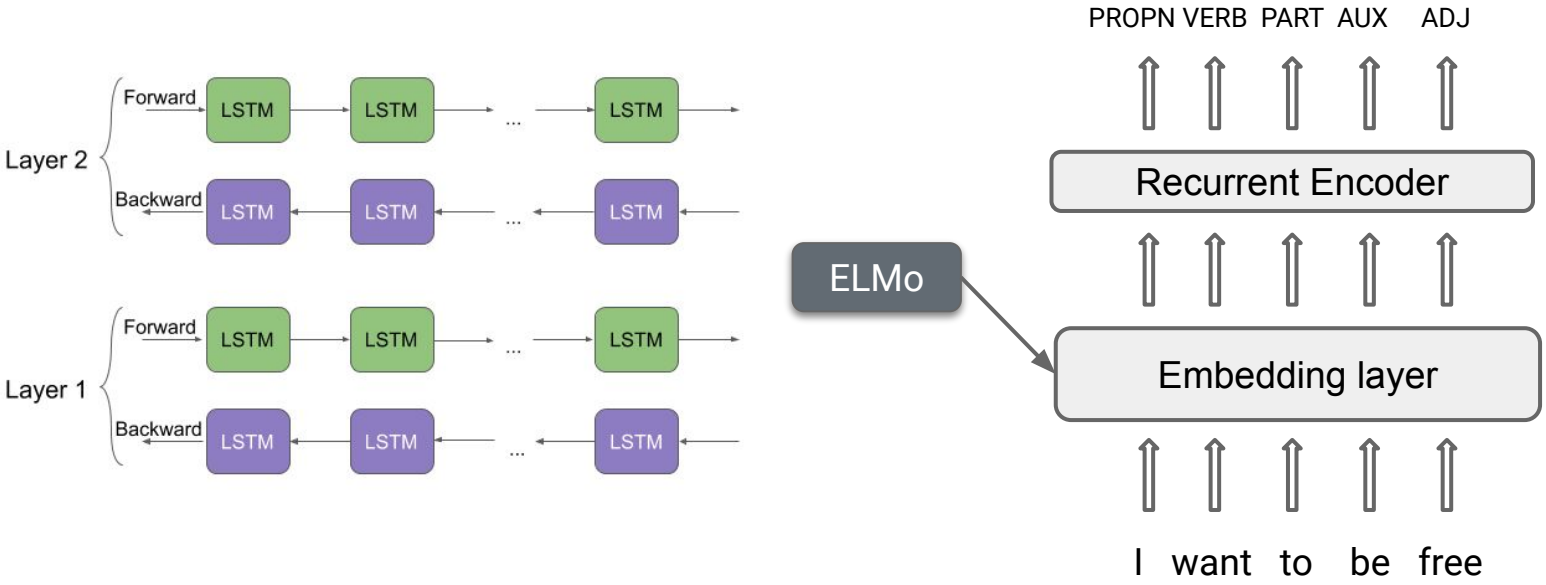
Transfer Learning in NLP

In Natural Language Processing, we **pretrain words representation**!



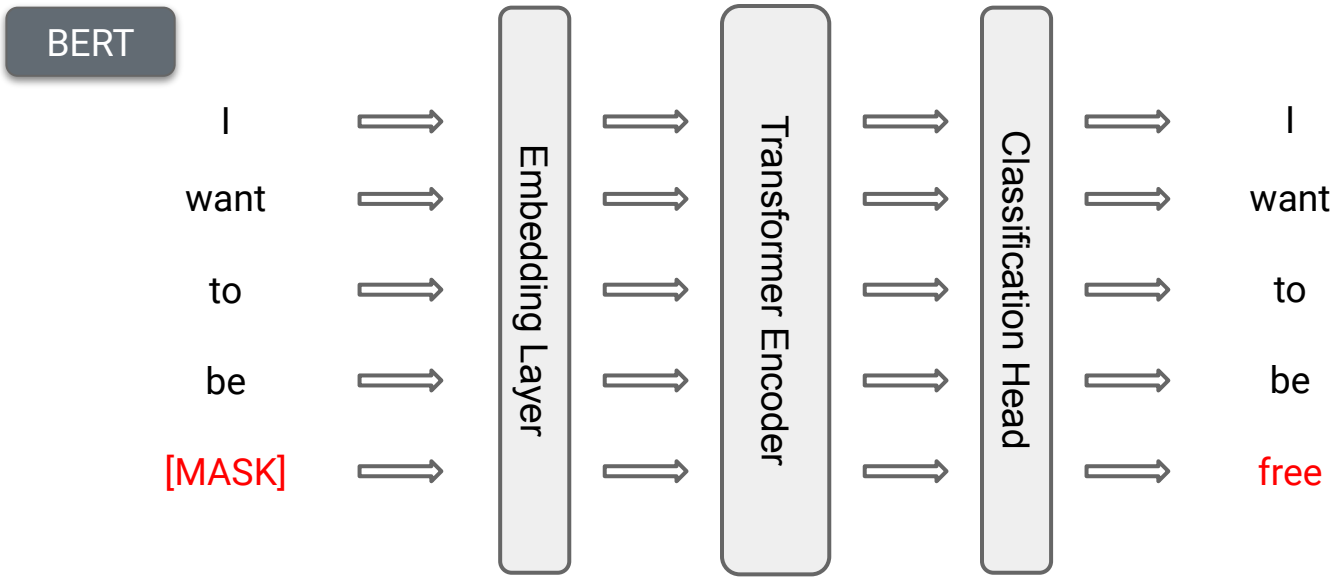
Transfer Learning in NLP

In Natural Language Processing, we **pretrain words representation**!



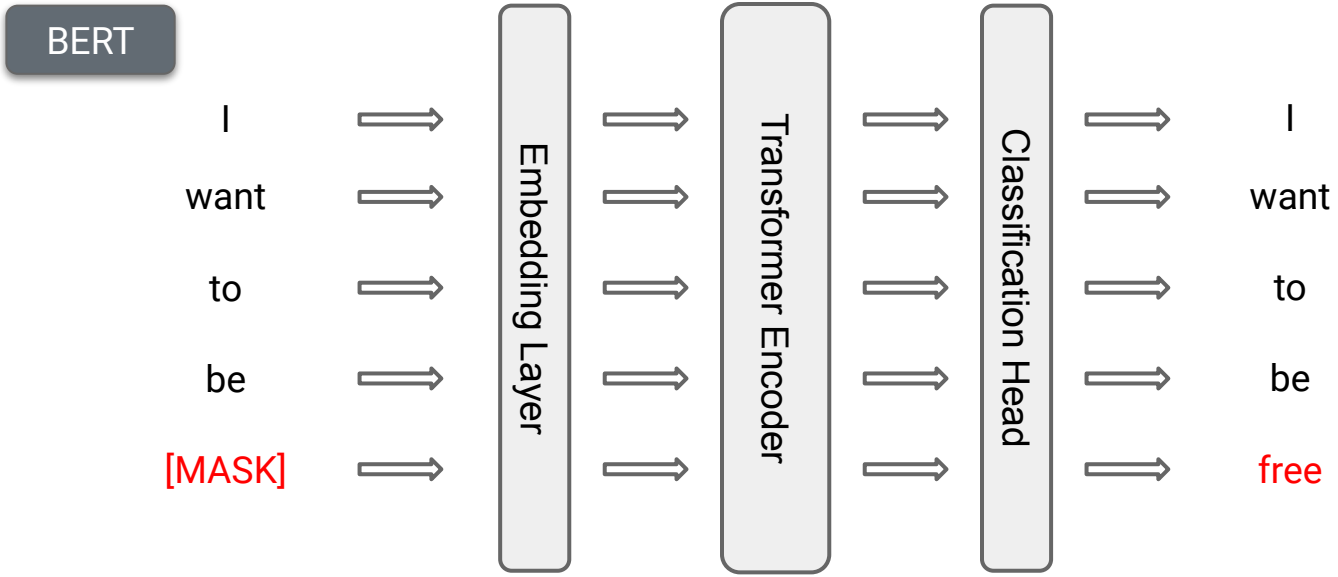
Transfer Learning in NLP

In Natural Language Processing, we **pretrain words representation**!



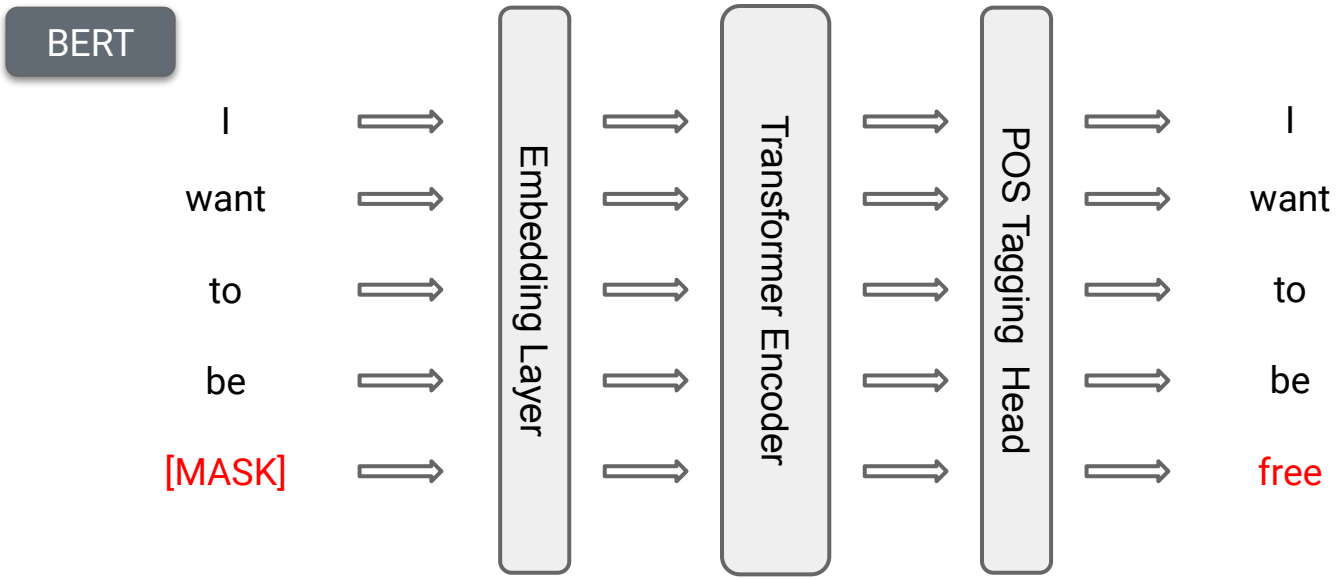
Transfer Learning in NLP

In Natural Language Processing, we **pretrain words representation**!



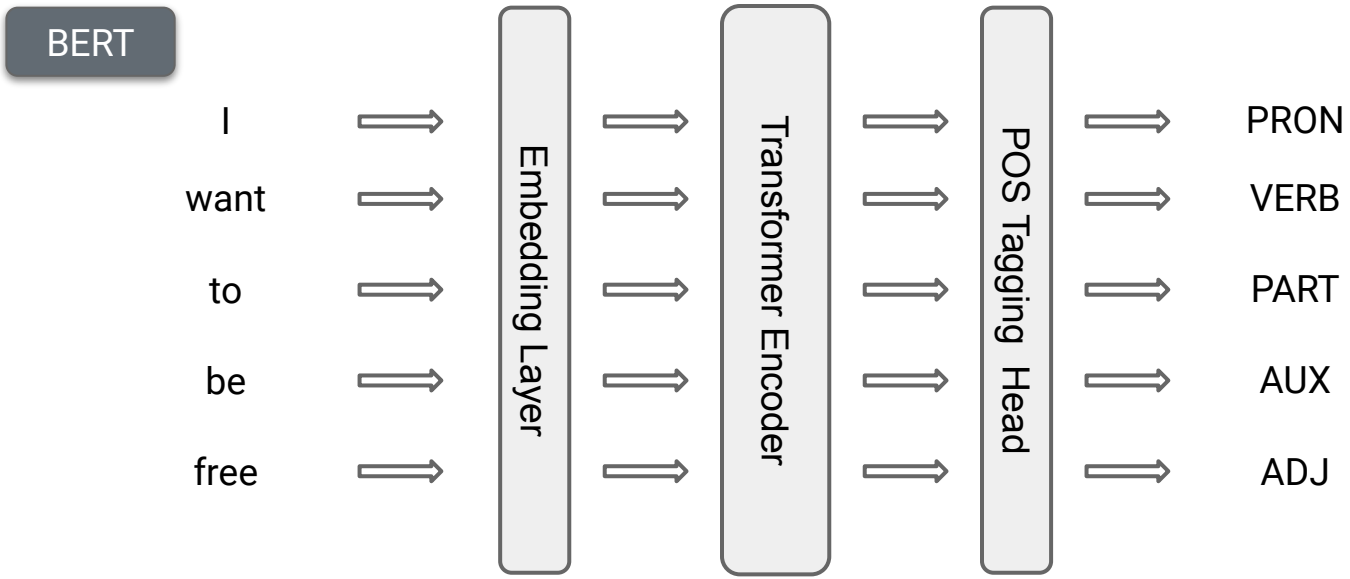
Transfer Learning in NLP

In Natural Language Processing, we **pretrain words representation**!



Transfer Learning in NLP

In Natural Language Processing, we **pretrain words representation**!



Transfer Learning in NLP

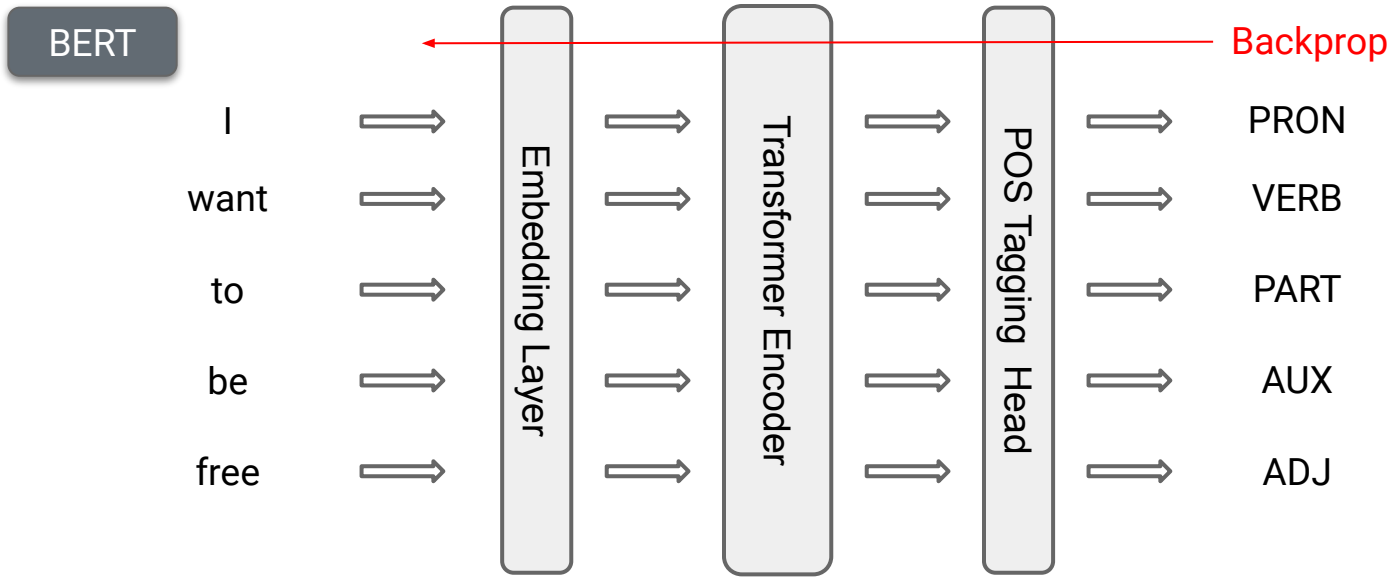
Two main ways to adapt pretrained contextualized embeddings to other tasks:

- **Fine Tuning**: while adapting the pretrained model to the new task, we update **both** the weights of the new layers and the pretrained model

- **Feature Based**: while adapting the pretrained model to the new task, we update **only** the new layers

Transfer Learning in NLP

Fine Tuning: let the gradient flow!



Transfer Learning in NLP

Feature Based: freeze!

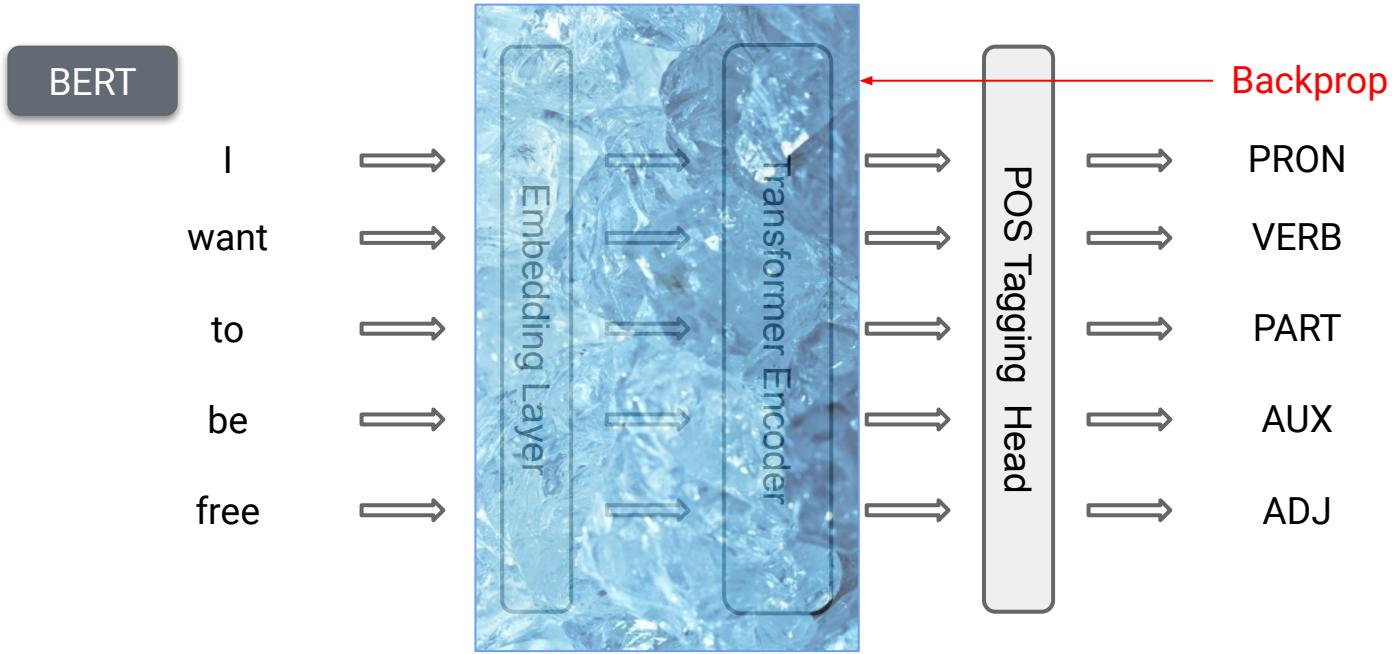
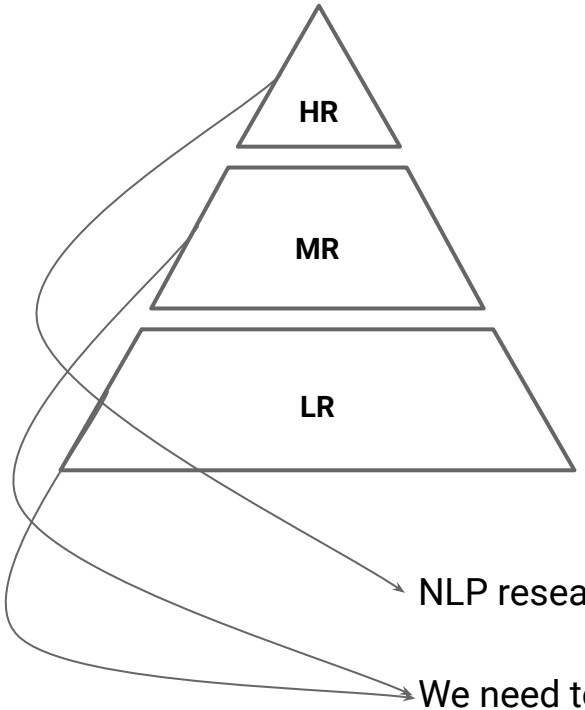


Table of Contents

Multilingual Transfer Learning

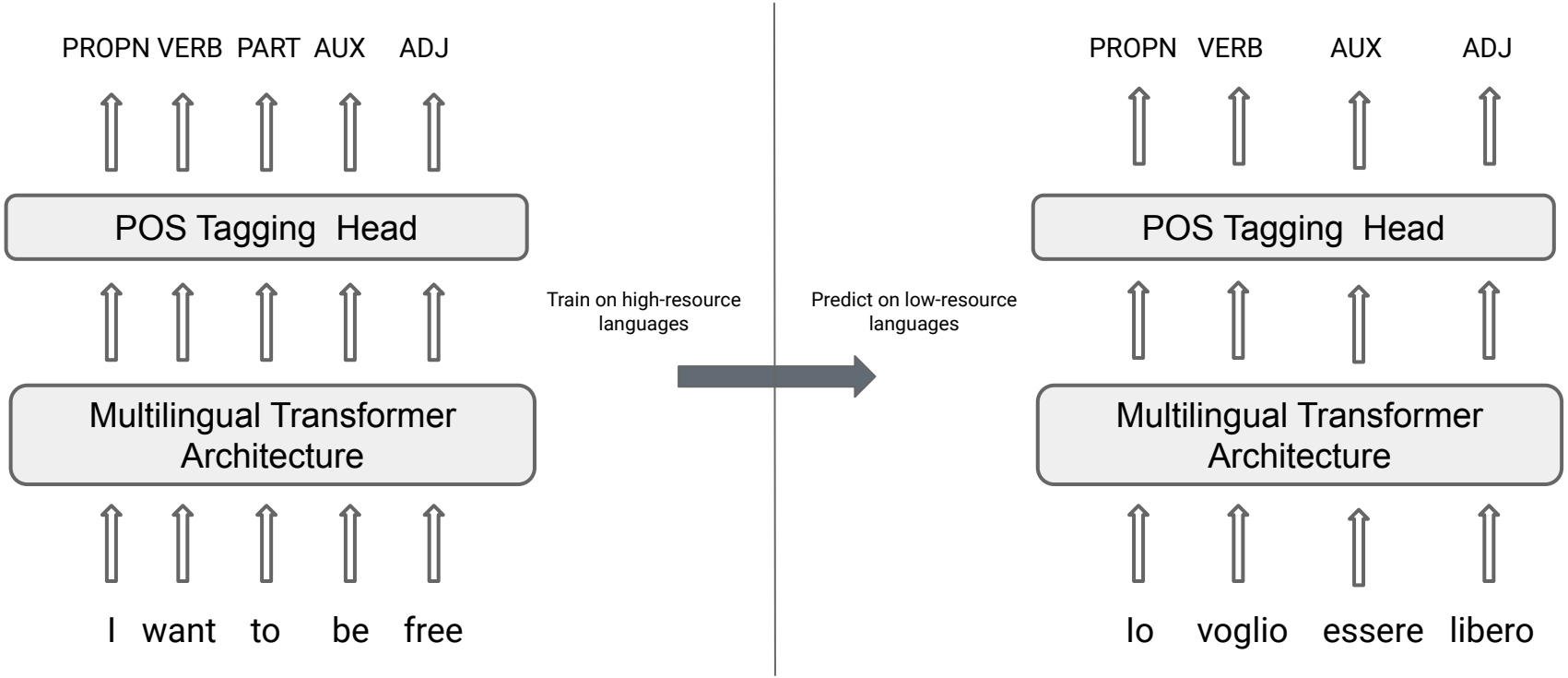


High-resource: 100s of millions of documents online, large labelled datasets, large Wikipedia

Medium-resource: Few labelled data, millions to 100,000s of online documents, medium-sized Wikipedia

Low-resource: No labelled data, scarce data online, small or non Wikipedia

Multilingual Transfer Learning



Multilingual Transfer Learning

- **Zero-shot learning**
- **Hardest transfer learning setting**
- Overall idea
  - Train multilingual encoders on high-resource languages
  - Test the model on **unseen** low-resource languages

Multilingual Transfer Learning

By definition,

- Low-resource languages have **scarce to none labelled data**
- Obviously, this affects test data as well
- **How do we evaluate them?**

Multilingual Transfer Learning

- Enter **XNLI,** Cross-lingual Natural Language Inference
- Simply put, an evaluation corpus spanning over 15 languages
- Widely accepted as an **adequate proxy** of cross-lingual transfer

Premise:	Hypothesis:	
He shot him with a gun	*He has never used a gun*	entailment contradiction ⟵ neutral

Multilingual Transfer Learning

So, how well do current models behave?

Multilingual Transfer Learning

So, how well do current models behave?

Well

Multilingual Transfer Learning

So, how well do current models behave?

Well, **kind of**

Multilingual Transfer Learning

So, how well do current models behave?

Well, **kind of**

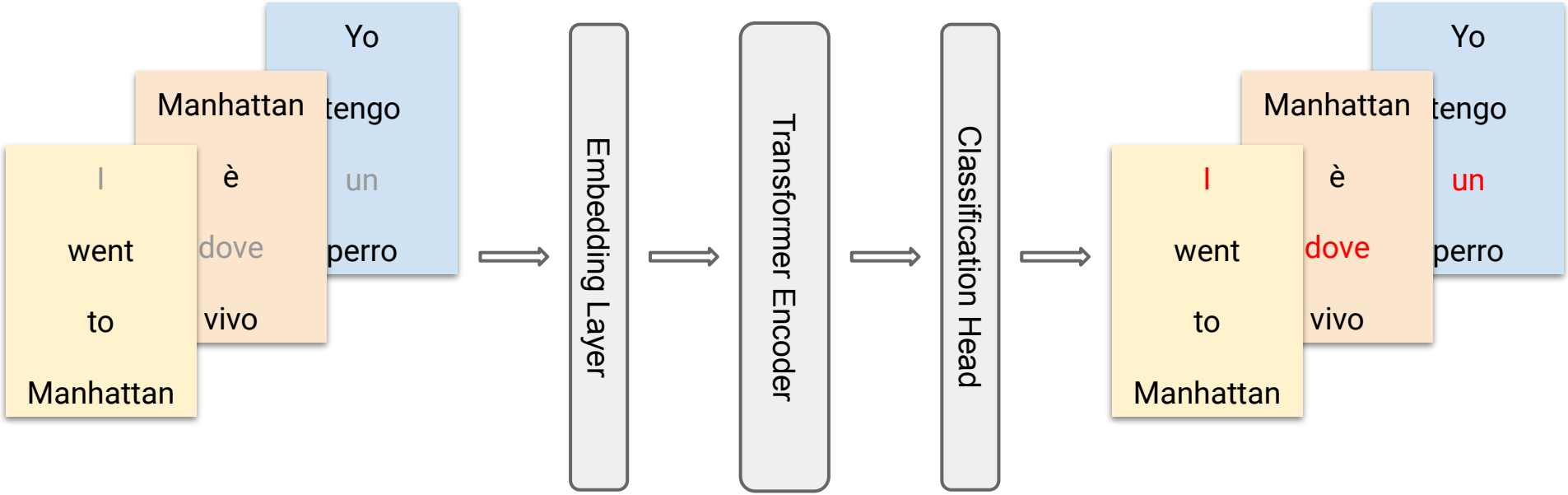
Model	EN	ES	ZH	AR
MBERT	81.4	74.9	70.1	70.4
XLM-TLM	85.0	78.9	76.5	73.1
XLMR	89.1	85.1	80.2	79.8

XNLI zero-shot results

Table of Contents

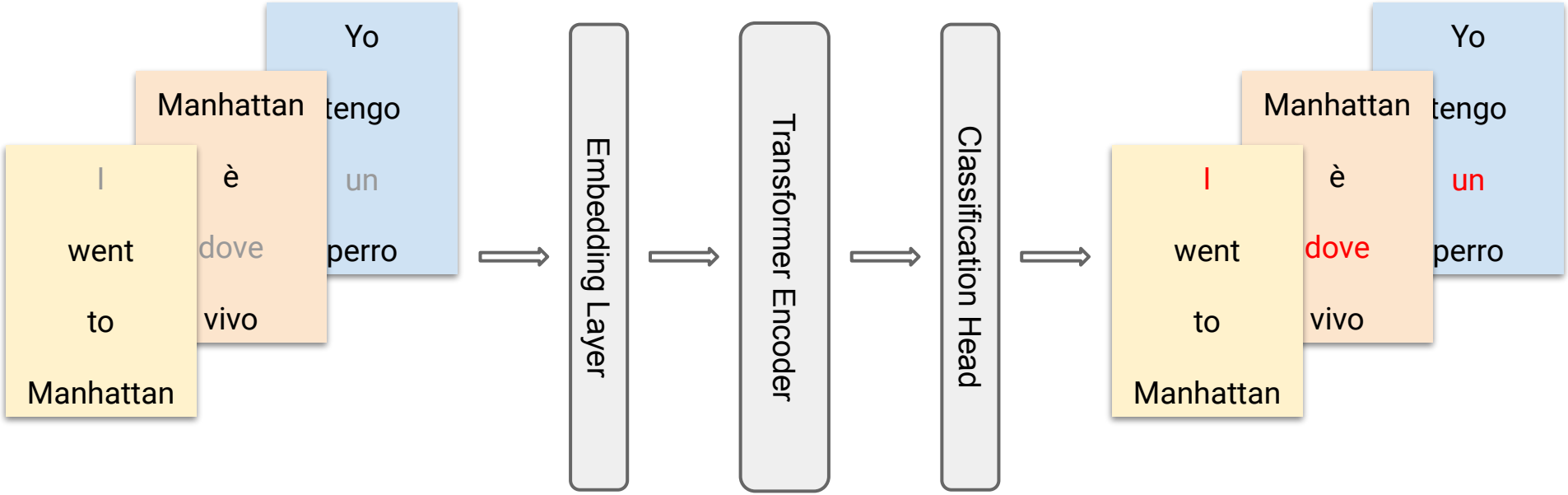
Multilingual Transfer Learning

Multilingual BERT training setting



Multilingual Transfer Learning

Why on Earth does it work?



Multilingual Transfer Learning

- It seems, even in this unsupervised setting, models tend to align common patterns in different languages, allowing cross-lingual share
- Two different interesting research points:
  - Which key features allow cross-lingual share between two different languages?
  - Which training settings fosters co-existence of multiple languages?

Bilingual Transfer

- [Wu et al. 2019] analyzed the first point extensively

- They performed a series of in-vivo evaluations of zero-shot performances, comparing **bilingual training settings**
  - *En-Fr*
  - *En-Ru*
  - *En-Zh*

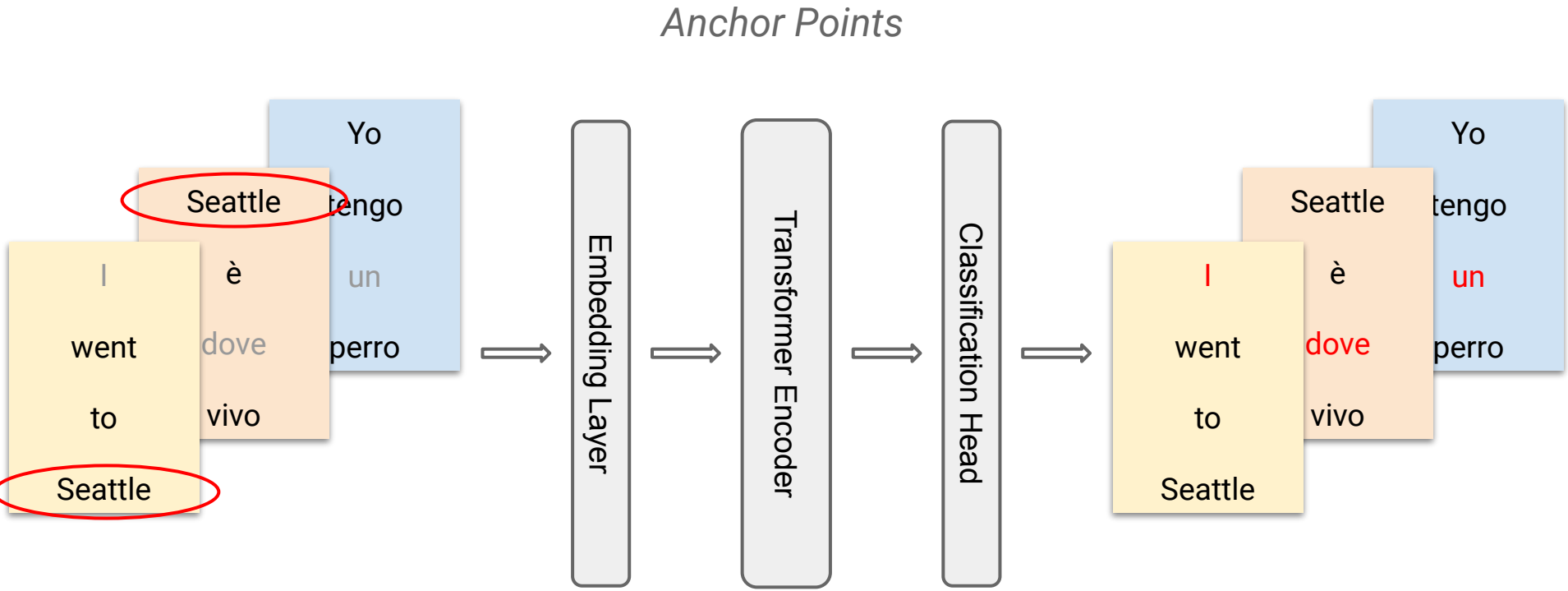
- They focused on three tasks: *XNLI*, *NER* and *Dependency Parsing*

Bilingual Transfer

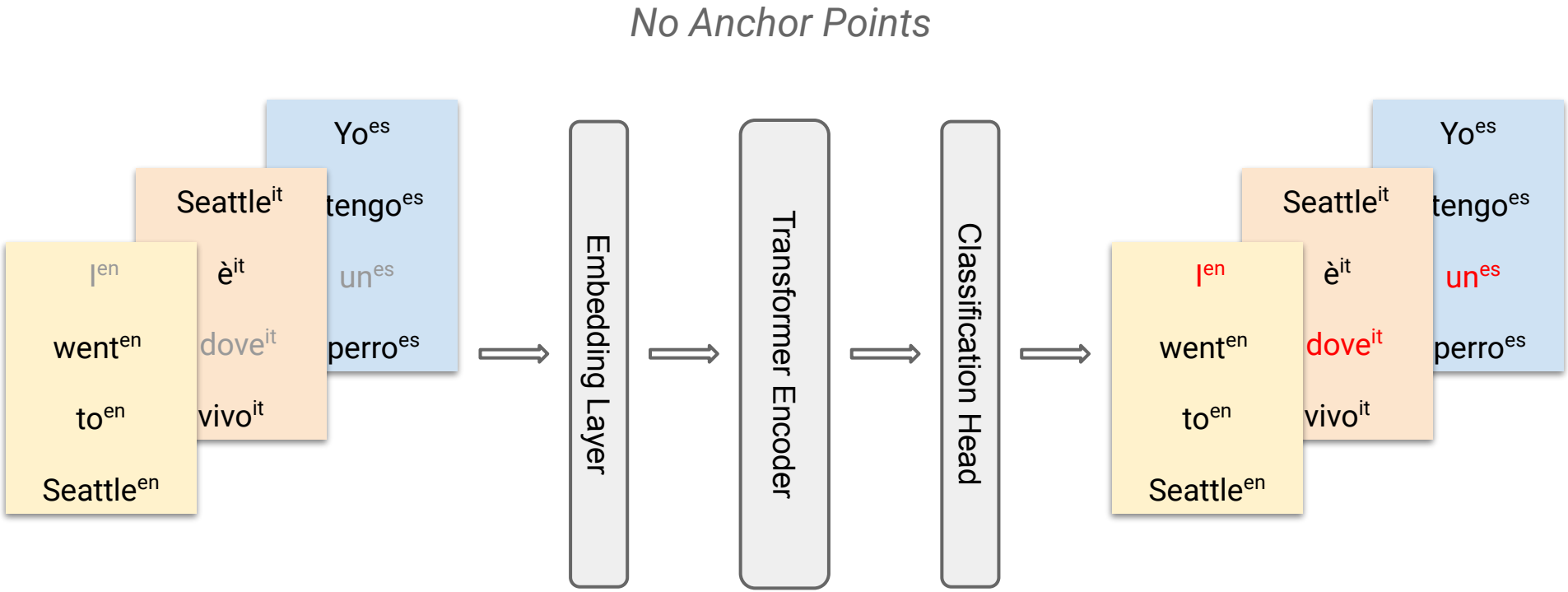
They reported interesting results on four training configurations:

- Anchor Points
- Domain similarity
- Parameter sharing
- Language similarity

Bilingual Transfer: Anchor Points



Bilingual Transfer: Anchor Points



Bilingual Transfer: Anchor Points

XNLI	FR	RU	ZH	AVG
Default Anchors	74.0	68.1	68.9	70.3
No Anchors	72.1	67.5	67.7	69.1

NER	FR	RU	ZH	AVG
Default Anchors	79.8	60.9	63.6	68.1
No Anchors	74.0	57.9	65.0	65.6

Parsing	FR	RU	ZH	AVG
Default Anchors	73.2	56.6	28.8	52.9
No Anchors	72.3	56.2	27.4	52.0

Bilingual Transfer: Domain Similarity

Domain Similarity

English

Target Language

English

Target Language

Bilingual Transfer: Domain Similarity

XNLI	FR	RU	ZH	AVG
Default	73.6	68.7	68.3	70.2
Wiki-CC	74.2	65.8	66.5	68.8

NER	FR	RU	ZH	AVG
Default	79.8	60.9	63.6	68.1
Wiki-CC	74.0	49.6	61.9	61.8

Parsing	FR	RU	ZH	AVG
Default	73.2	56.6	28.8	52.9
Wiki-CC	71.3	54.8	25.2	50.4

Bilingual Transfer: Parameter Sharing

- What if we don’t share all layers? That is, What if we **make some layers language-specific**?
- Configurations Explored:
  - Separated Embedding
  - Separated Embedding + L1-3
  - Separated Embedding + L1-6
  - Separated L1-3
  - Separated L1-6

Bilingual Transfer: Parameter Sharing

XNLI	FR	RU	ZH	AVG
Default	73.6	68.7	68.3	70.2
Sep Emb	72.7	63.6	60.8	65.7
Sep Emb+L1-3	69.2	61.7	56.4	62.4
Sep Emb+L1-6	51.6	35.8	34.4	40.6
Sep L1-3	72.4	65.0	63.1	66.8
Sep L1-6	61.9	43.6	37.4	47.6

Bilingual Transfer: Parameter Sharing

NER	FR	RU	ZH	AVG
Default	79.8	60.9	63.6	68.1
Sep Emb	75.5	57.5	59.0	64.0
Sep Emb+L1-3	73.8	46.8	53.5	58.0
Sep Emb+L1-6	56.5	5.4	1.0	21.0
Sep L1-3	74.0	53.3	60.8	62.7
Sep L1-6	61.2	23.7	3.1	29.3

Bilingual Transfer: Parameter Sharing

Parsing	FR	RU	ZH	AVG
Default	73.2	56.6	28.8	52.9
Sep Emb	71.7	54.0	27.5	51.1
Sep Emb+L1-3	68.2	53.6	23.9	48.6
Sep Emb+L1-6	50.9	6.4	1.5	19.6
Sep L1-3	69.7	54.1	26.4	50.1
Sep L1-6	61.7	31.6	12.0	35.1

Bilingual Transfer: Language Similarity

- The extent to which two languages are similar
- **Tricky to measure**: *word order, distribution of structural patterns, within language variance, etc.*
- Languages that have **common structural and functional features** are the one with the **highest scores** in cross lingual transfer

Bilingual Transfer: Language Similarity

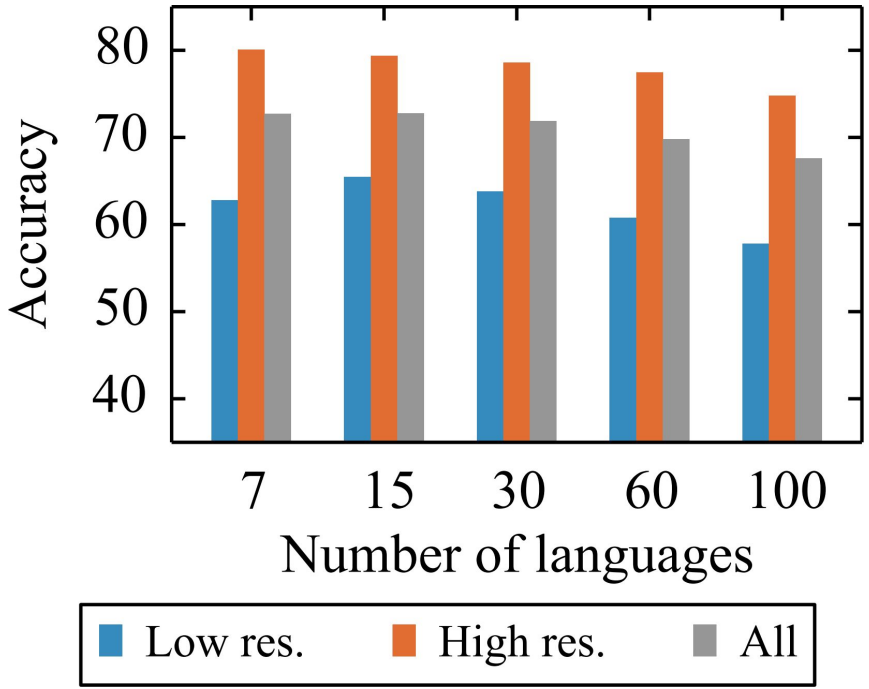
Model	FR	RU	ZH	AVG
XNLI (Acc.)	73.6	68.7	68.3	70.2
NER (F1)	79.8	60.9	63.6	68.1
Parsing (LAS)	73.2	56.6	28.8	52.9

Multilingual Transfer

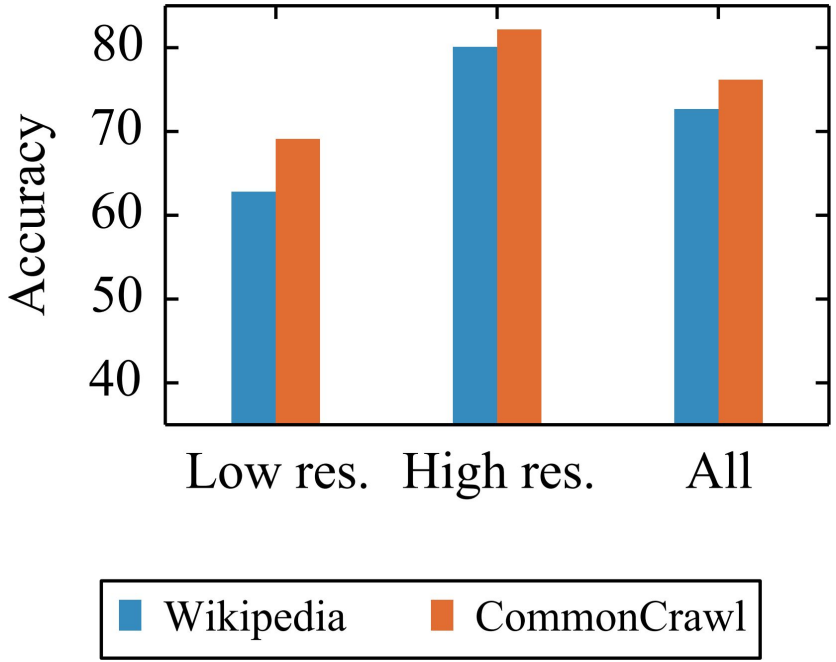
- [Conneau et al. 2020] highlights the key factors that allow multiple languages to coexist and share common patterns
- They trained different architectures on multiple set of languages, answering a few interesting questions
- Reaching state-of-the-art results on the XNLI benchmark

Multilingual Transfer: Number of Languages

What happens as we vary the number of languages we train upon?



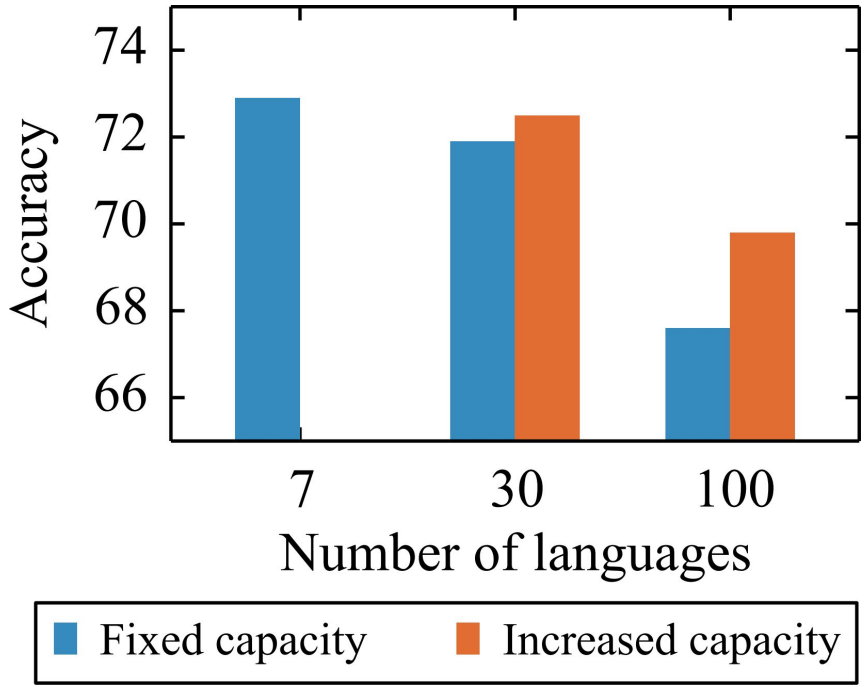
Multilingual Transfer: Training Corpora



How does the choice of the training corpora affect performances?

Multilingual Transfer Learning: Analyses

Does the model capacity mitigate the curse of multilinguality?

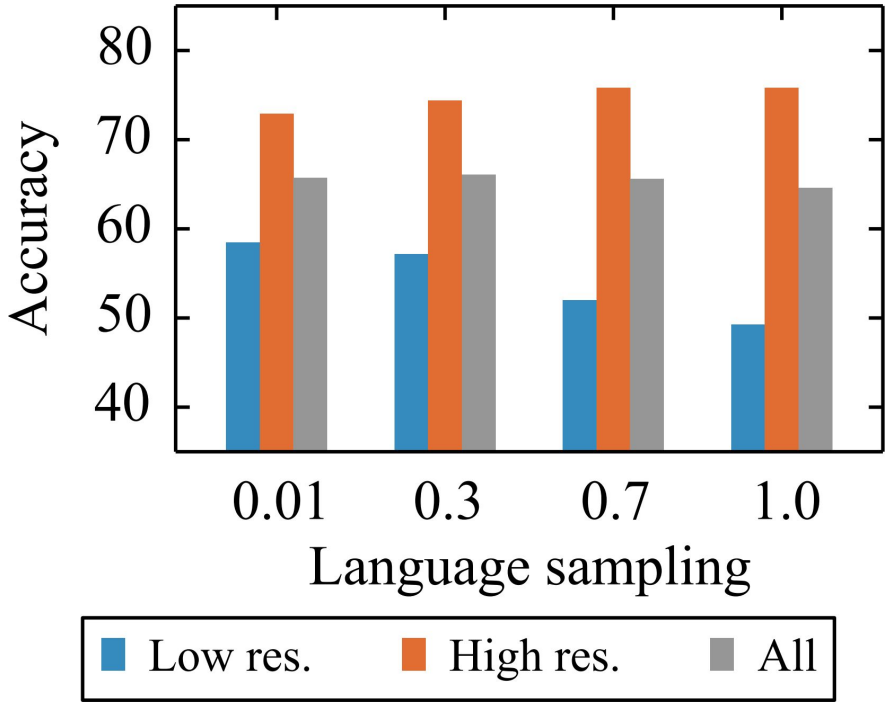


Multilingual Transfer Learning: Analyses

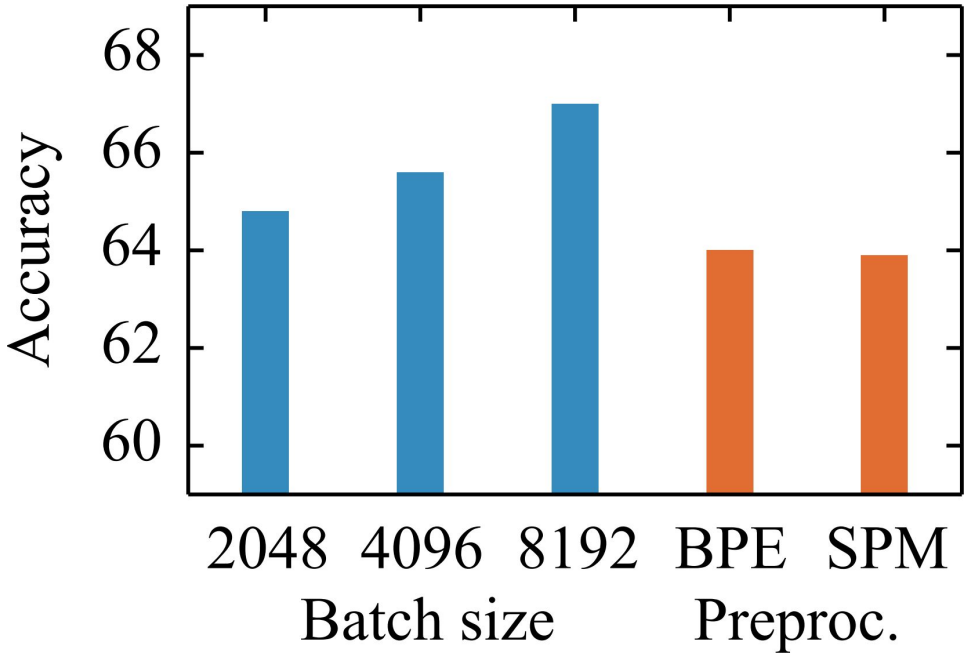
Language sampling

During the training phase the sequences of text provided to the model are chosen between the different languages with the following probability.

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}.$$



Multilingual Transfer: Batch Size



To what extent are higher batch sizes helpful?

Multilingual Transfer: Vocab Size

What about vocab sizes?

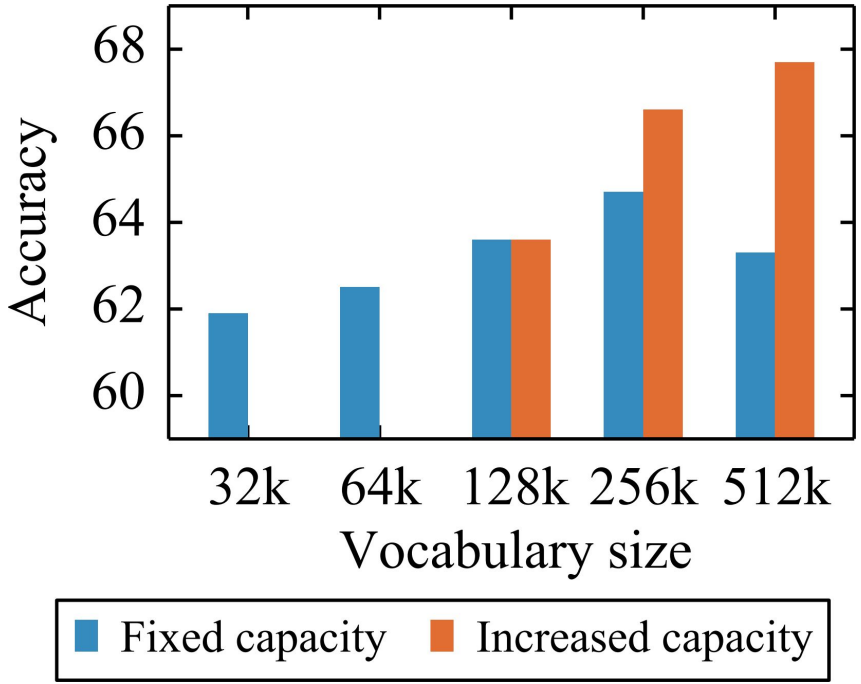


Table of Contents

Conclusions

- Linguistically speaking, **language similarity** is a key component to allow cross-lingual transfer
- From an architecture perspective, **parameter sharing** is crucial
- As for the hyperparameters, adequate choosing both the **model capacity** and the **number of languages** to train upon is essential