CS 784: Computational Linguistics
Lecture 18: Computational Multilingualism (in 2025)

Freda Shi

School of Computer Science, University of Waterloo
fhs@uwaterloo.ca

March 20, 2025

Reading: Ponti et al. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*.

Pre-Class Survey

Do you think language is a **special** type of time series?

In other words, how do you like the following statement?

Suppose we have a perfect stock price prediction model, it would be a (nearly) perfect language model as well.

Computational Multilingualism before LLMs

Machine Translation

- **Bilingual lexicon induction**: recognize mutually translatable pairs without word-level annotation.
- **Word alignment**: align words in parallel corpora.
- **Syntax (phrase)-based models**: recognize mutually translatable phrases, translate, and reorder.
- **Neural machine translation**: sequence-to-sequence models for translation.

Computational Multilingualism before LLMs

Machine Translation

- **Bilingual lexicon induction**: recognize mutually translatable pairs without word-level annotation.
- **Word alignment**: align words in parallel corpora.
- **Syntax (phrase)-based models**: recognize mutually translatable phrases, translate, and reorder.
- **Neural machine translation**: sequence-to-sequence models for translation.

Cross-Lingual Information Retrieval

- Represent cross-lingual information digitally and retrieve it

Computational Multilingualism before LLMs

Machine Translation

- **Bilingual lexicon induction**: recognize mutually translatable pairs without word-level annotation.
- **Word alignment**: align words in parallel corpora.
- **Syntax (phrase)-based models**: recognize mutually translatable phrases, translate, and reorder.
- **Neural machine translation**: sequence-to-sequence models for translation.

Cross-Lingual Information Retrieval

- Represent cross-lingual information digitally and retrieve it

Cross-Lingual Transfer (of Neural Models)

- Pretraining on one language and fine-tuning on another
- Multilingual BERT

Outline

Concept: what is linguistic typology?

Outline

Concept: what is linguistic typology?

Research questions the survey attempts to answer:

1. Which NLP tasks and applications can benefit from typology?

Outline

Concept: what is linguistic typology?

Research questions the survey attempts to answer:

1. Which NLP tasks and applications can benefit from typology?
2. What are the advantages and limitations of currently available typological databases? Can data-driven inference of typological features offer an alternative source of information?

Outline

Concept: what is linguistic typology?

Research questions the survey attempts to answer:

1. Which NLP tasks and applications can benefit from typology?
2. What are the advantages and limitations of currently available typological databases? Can data-driven inference of typological features offer an alternative source of information?
3. Which methods have been proposed to incorporate typological information in NLP systems, and how should such information be encoded?

Outline

Concept: what is linguistic typology?

Research questions the survey attempts to answer:

1. Which NLP tasks and applications can benefit from typology?
2. What are the advantages and limitations of currently available typological databases? Can data-driven inference of typological features offer an alternative source of information?
3. Which methods have been proposed to incorporate typological information in NLP systems, and how should such information be encoded?
4. To what extent does the performance of typology-savvy methods surpass typology-agnostic baselines? How does typology compare with other criteria of language classification, such as genealogy?

Outline

Concept: what is linguistic typology?

Research questions the survey attempts to answer:

1. Which NLP tasks and applications can benefit from typology?
2. What are the advantages and limitations of currently available typological databases? Can data-driven inference of typological features offer an alternative source of information?
3. Which methods have been proposed to incorporate typological information in NLP systems, and how should such information be encoded?
4. To what extent does the performance of typology-savvy methods surpass typology-agnostic baselines? How does typology compare with other criteria of language classification, such as genealogy?
5. How can typology be harnessed for data selection, rule-based systems, and model interpretation?

Typology

Linguistic typology: the study of the variation among the world's languages through **systematic** comparison.

Typology

Linguistic typology: the study of the variation among the world's languages through **systematic** comparison.

- **Morphosyntactic typology**: word order, case marking, etc.

私は本を読みます

I read the book

Typology

Linguistic typology: the study of the variation among the world's languages through **systematic** comparison.

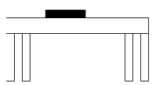
- **Morphosyntactic typology**: word order, case marking, etc.

私は本を読みます

I read the book

- **Semantic typology**: lexicalization patterns, etc.

[Feist and Gentner, 2003]


English	over	on	in
Spanish	sobre	en	en
Japanese	ue	ue	naka

The World Atlas of Language Structures (WALS)

The online database of typological features of languages `https://wals.info/`.

The English Bias

An interesting pragmatic phenomenon: whenever we don't mention the language, there is an R-based implicature that we are talking about English.

The English Bias

An interesting pragmatic phenomenon: whenever we don't mention the language, there is an R-based implicature that we are talking about English.

Things are turning better.

ACL papers with title containing "parsing":

[Source: Daniel Zeman, https://ufal.mff.cuni.cz/~zeman/langtech/npfl120/slides/npfl120-01-wals.pdf]

The English Bias

An interesting pragmatic phenomenon: whenever we don't mention the language, there is an R-based implicature that we are talking about English.

Things are turning better.

ACL papers with title containing "parsing":

- ACL 1998: 9 papers, 3 languages
  EN: 4, DE: 1, ES: 1, no evaluation on natural language: 1

[Source: Daniel Zeman, https://ufal.mff.cuni.cz/~zeman/langtech/npfl120/slides/npfl120-01-wals.pdf]

The English Bias

An interesting pragmatic phenomenon: whenever we don't mention the language, there is an R-based implicature that we are talking about English.

Things are turning better.

ACL papers with title containing "parsing":

- ACL 1998: 9 papers, 3 languages
  EN: 4, DE: 1, ES: 1, no evaluation on natural language: 1
- ACL 2007: 12 papers, 13 languages
  EN: 7, DE: 3, AR: 1, CS: 1, DA: 1, **EU**: 1, JA: 1, NL: 1, PT: 1, SL: 1, SV: 1, ZH: 1

[Source: Daniel Zeman, https://ufal.mff.cuni.cz/~zeman/langtech/npfl120/slides/npfl120-01-wals.pdf]

The English Bias

An interesting pragmatic phenomenon: whenever we don't mention the language, there is an R-based implicature that we are talking about English.

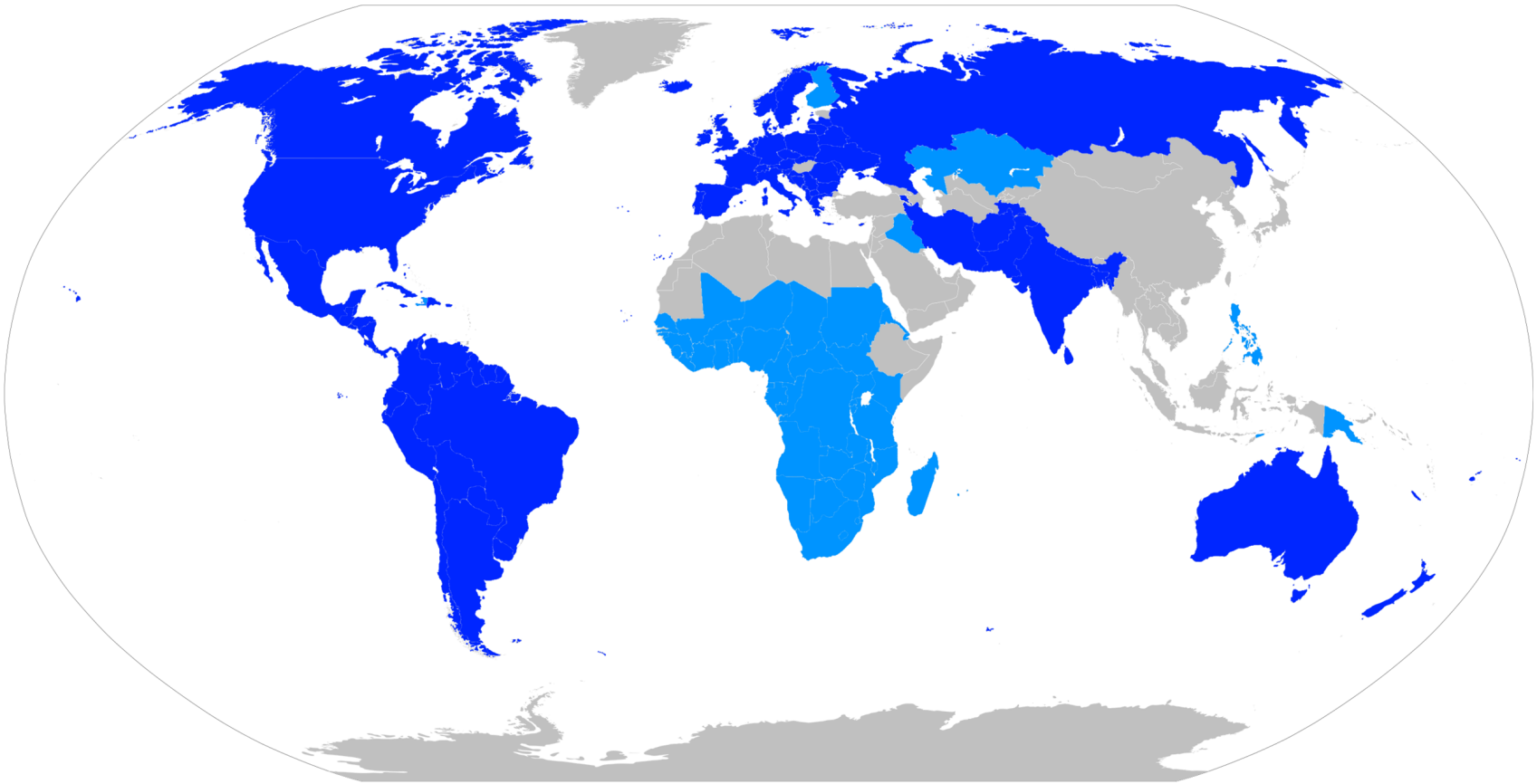
Things are turning better.

ACL papers with title containing "parsing":

- ACL 1998: 9 papers, 3 languages
  EN: 4, DE: 1, ES: 1, no evaluation on natural language: 1
- ACL 2007: 12 papers, 13 languages
  EN: 7, DE: 3, AR: 1, CS: 1, DA: 1, **EU**: 1, JA: 1, NL: 1, PT: 1, SL: 1, SV: 1, ZH: 1
- ACL 2016: 24 papers, 24 languages:
  EN: 18, DE: 6, ZH: 5, AR: 1, BG: 1, CA: 1, CS: 1, DA: 1, EL: 1, ES: 1, **EU**: 1, FR: 1, HE: 1, **HU**: 1, IT: 1, JA: 1, KO: 1, ML: 1, NL: 1, PL: 1, PT: 1, SL: 1, SV: 1, TR: 1

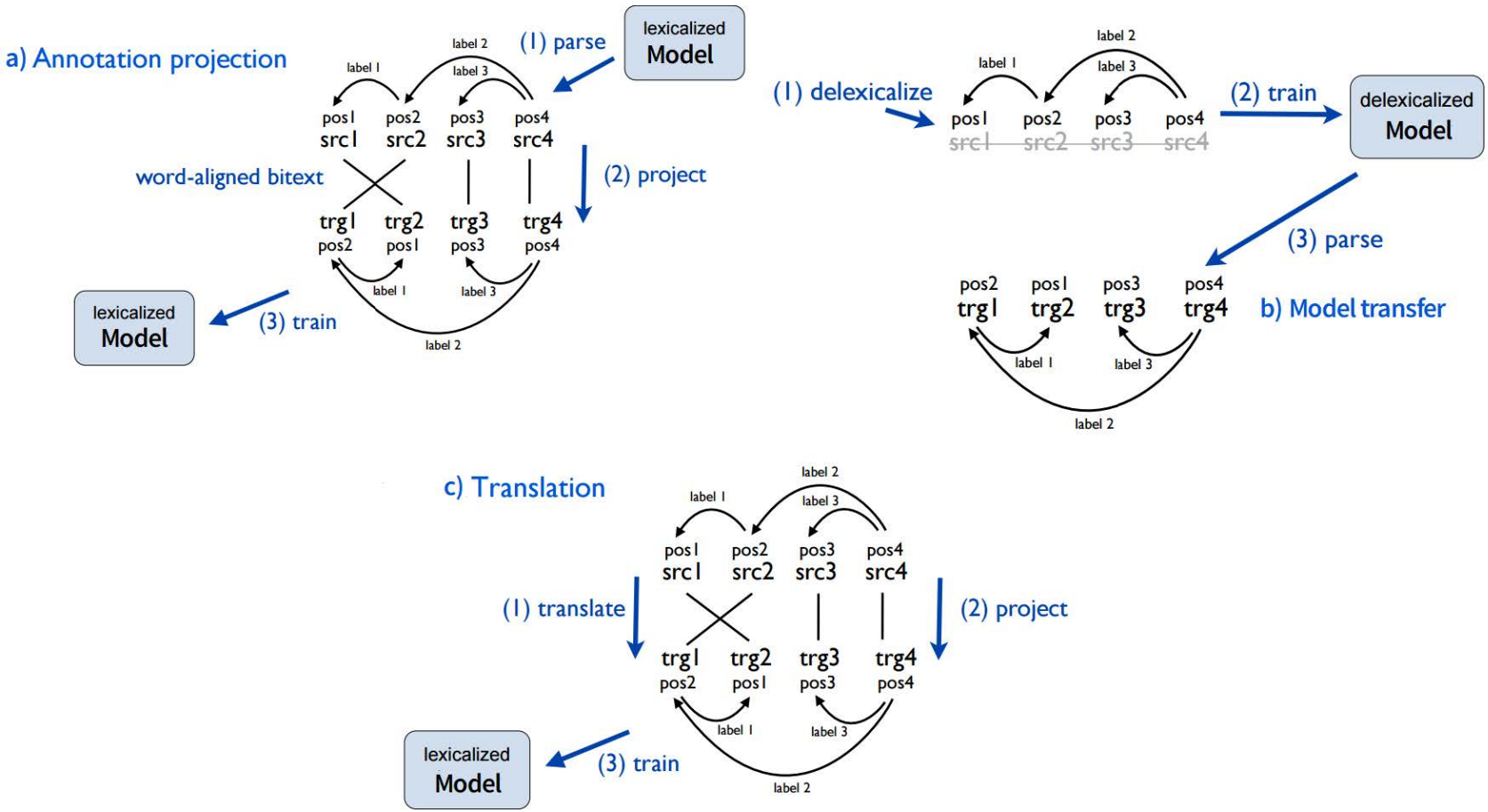
[Source: Daniel Zeman, https://ufal.mff.cuni.cz/~zeman/langtech/npfl120/slides/npfl120-01-wals.pdf]

The Indo-European Language Distribution



Word Alignment

Word alignment used to be the key of multilingual NLP.



Word Alignment: Algorithms

The IBM models propose probability-based generative processes for word alignment.

Word Alignment: Algorithms

The IBM models propose probability-based generative processes for word alignment.

It takes bitext (i.e., parallel sentences) as the training data, and infers the alignment between words in the two languages.

Word Alignment: Algorithms

The IBM models propose probability-based generative processes for word alignment.

It takes bitext (i.e., parallel sentences) as the training data, and infers the alignment between words in the two languages.

Key idea: if a word e in one language frequently co-occurs with a word f in its bitext counterpart (another language), they are likely to be aligned.

Word Alignment: Evaluation

The alignment error rate (AER):

$$\text{AER} = 1 - \frac{|P \cap A| + |P \cap R|}{|P| + |R|}$$

P :predicted alignment pairs $\langle s, t \rangle$

R :required alignment pairs

A :required and optional pairs

Word Alignment: Evaluation

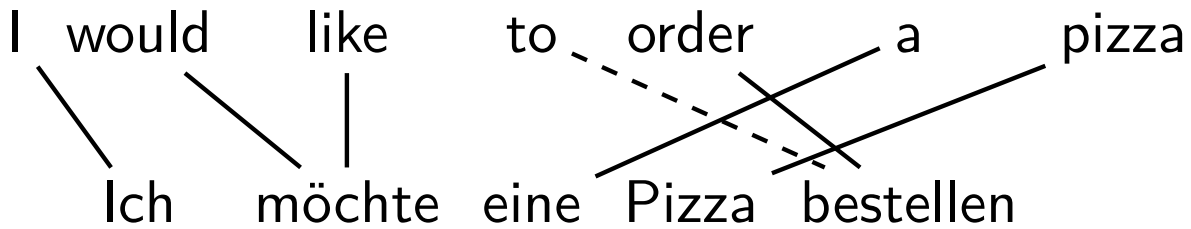
The alignment error rate (AER):

$$\text{AER} = 1 - \frac{|P \cap A| + |P \cap R|}{|P| + |R|}$$

P :predicted alignment pairs $\langle s, t \rangle$

R :required alignment pairs

A :required and optional pairs



Solid lines: required alignment pairs.
Dashed lines: optional alignment pairs.

Cross-Lingual Transfer

Suppose you have a pretrained model on English, how can you transfer it to another language X?

Cross-Lingual Transfer

Suppose you have a pretrained model on English, how can you transfer it to another language X?

- **Direct transfer**: pretend data X is English, and use the model directly.

Cross-Lingual Transfer

Suppose you have a pretrained model on English, how can you transfer it to another language X?

- **Direct transfer**: pretend data X is English, and use the model directly.
- **Data projection**: make some necessary modifications to the data in X to make it look like English, and use the model directly.

Cross-Lingual Transfer

Suppose you have a pretrained model on English, how can you transfer it to another language X?

- **Direct transfer**: pretend data X is English, and use the model directly.
- **Data projection**: make some necessary modifications to the data in X to make it look like English, and use the model directly.
- **Model adaptation**: make some necessary modifications to the model so that it can handle X.

Cross-Lingual Transfer

Suppose you have a pretrained model on English, how can you transfer it to another language X?

- **Direct transfer**: pretend data X is English, and use the model directly.
- **Data projection**: make some necessary modifications to the data in X to make it look like English, and use the model directly.
- **Model adaptation**: make some necessary modifications to the model so that it can handle X.

Cross-lingual word embeddings are usually helpful resources, as they capture some lexical semantic information.

Cross-Lingual Transfer

Suppose you have a pretrained model on English, how can you transfer it to another language X?

- **Direct transfer**: pretend data X is English, and use the model directly.
- **Data projection**: make some necessary modifications to the data in X to make it look like English, and use the model directly.
- **Model adaptation**: make some necessary modifications to the model so that it can handle X.

Cross-lingual word embeddings are usually helpful resources, as they capture some lexical semantic information.

Why?

Multilingual Pretraining

There are not much difference between monolingual and multilingual pretraining of BERT or GPT-2 architectures, except for the input data and the corresponding tokenization strategies.

Multilingual Pretraining

There are not much difference between monolingual and multilingual pretraining of BERT or GPT-2 architectures, except for the input data and the corresponding tokenization strategies.

Recap: Byte-based BPE.

That's great 👍

54 68 61 74 2019 73 **20** 67 72 65 61 74 **20** 1F44D

All in hexadecimal.

Multilingual Pretraining

There are not much difference between monolingual and multilingual pretraining of BERT or GPT-2 architectures, except for the input data and the corresponding tokenization strategies.

Recap: Byte-based BPE.

That's great 👍

54 68 61 74 2019 73 **20** 67 72 65 61 74 **20** 1F44D

All in hexadecimal.

Discussion: what are the potential issues?

Multilingual Pretraining

There are not much difference between monolingual and multilingual pretraining of BERT or GPT-2 architectures, except for the input data and the corresponding tokenization strategies.

Recap: Byte-based BPE.

That's great 👍

54 68 61 74 2019 73 **20** 67 72 65 61 74 **20** 1F44D

All in hexadecimal.

Discussion: what are the potential issues?

- Non-English languages generally take more bytes to represent.

Zero Cross-Lingual Transfer

Multilingual pretraining enables quite efficient (and sometimes "magically" zero-shot) cross-lingual transfer.

Zero Cross-Lingual Transfer

Multilingual pretraining enables quite efficient (and sometimes "magically" zero-shot) cross-lingual transfer.

Taking the row-wise and column-wise maximum over the multilingual BERT representation–based similarity matrix gives you much better performance than statistics-based models.

Jalili Sabet et al. 2020. *SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings*. Findings of EMNLP

Zero Cross-Lingual Transfer

Multilingual pretraining enables quite efficient (and sometimes "magically" zero-shot) cross-lingual transfer.

Taking the row-wise and column-wise maximum over the multilingual BERT representation–based similarity matrix gives you much better performance than statistics-based models.

Jalili Sabet et al. 2020. *SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings*. Findings of EMNLP

Schuster et al. (2019): pretrained multilingual models, such as mBERT, enables zero-shot cross-lingual dependency parsing.

The performance can be even better if some cross-lingual alignment is performed in the shared embedding space.

Zero Cross-Lingual Transfer

Multilingual pretraining enables quite efficient (and sometimes "magically" zero-shot) cross-lingual transfer.

Taking the row-wise and column-wise maximum over the multilingual BERT representation–based similarity matrix gives you much better performance than statistics-based models.

Jalili Sabet et al. 2020. *SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings*. Findings of EMNLP

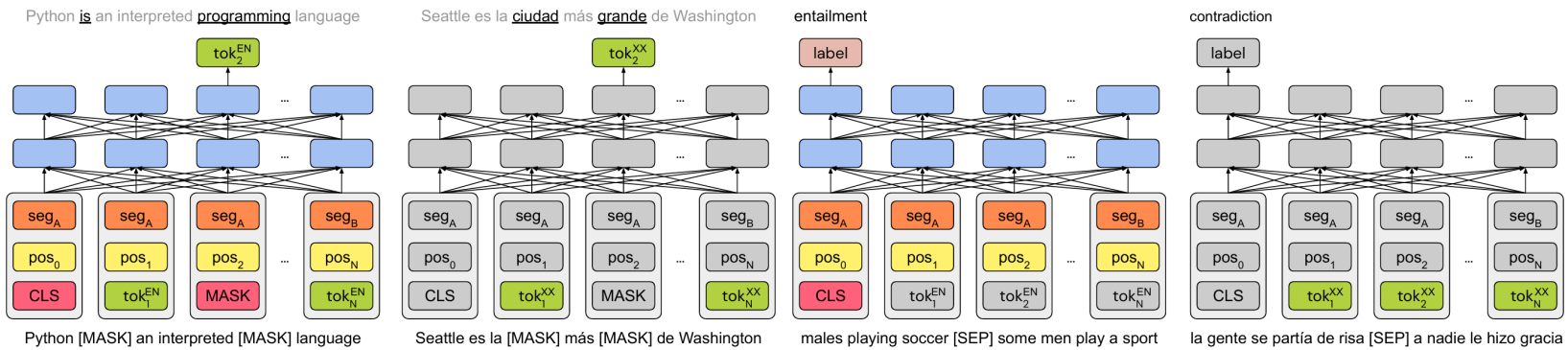
Schuster et al. (2019): pretrained multilingual models, such as mBERT, enables zero-shot cross-lingual dependency parsing.

The performance can be even better if some cross-lingual alignment is performed in the shared embedding space.

Open problem: why does zero-shot corss-lingual transfer work? Explain scientifically.

Parameter-Efficient Cross-Lingual Transfer

Key assumption: the Transformers have learned some language-agnostic processing protocols, so we can only fine-tune the non-contextualized (layer 0) word embeddings.

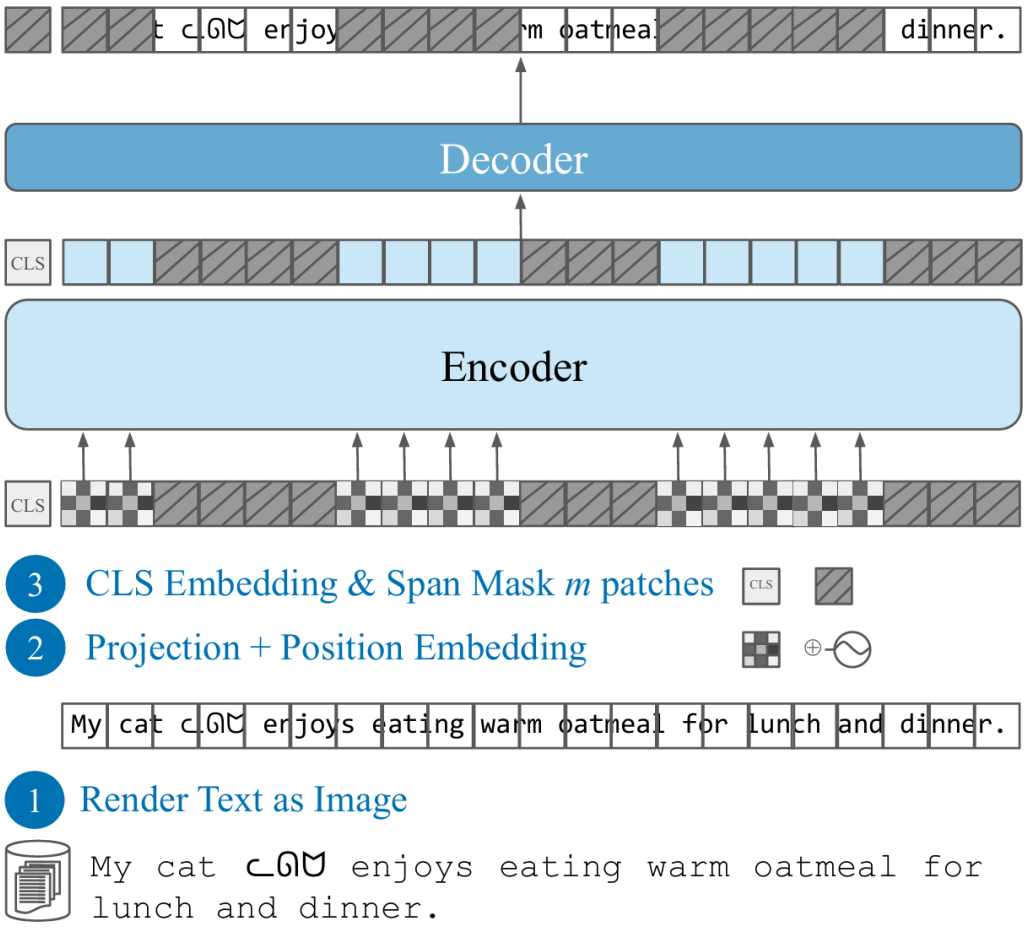


(a) English pre-training (b) L_2 embedding learning (c) English fine-tuning (d) Zero-shot transfer to L_2

[Artetxe et al. 2020. On the Cross-lingual Transferability of Monolingual Representations. ACL]

Pixel Language Models

Predict next pixel patch, instead of next word, in a sequence.



[Rust et al. 2023. Language Modelling with Pixels. ICLR]

Discussion 1

Q: Do we need typological features as language model inputs?

What experiments can (or will) help us answer this question?

Discussion 1

Q: Do we need typological features as language model inputs?

What experiments can (or will) help us answer this question?

- Train a language model from scratch with additional typological features as input, and compare its performance with the current model.

Discussion 1

Q: Do we need typological features as language model inputs?

What experiments can (or will) help us answer this question?

- Train a language model from scratch with additional typological features as input, and compare its performance with the current model.
  Possible issue(s): cost inefficient, or even intractable.
- Finetune a pretrained language model with additional typological features as input.

Discussion 1

Q: Do we need typological features as language model inputs?

What experiments can (or will) help us answer this question?

- Train a language model from scratch with additional typological features as input, and compare its performance with the current model.
  Possible issue(s): cost inefficient, or even intractable.
- Finetune a pretrained language model with additional typological features as input.
  Possible issue(s): cost inefficient, balance between the new and old data distribution.
- Probe if these features have been encoded in the model.

Discussion 1

Q: Do we need typological features as language model inputs?

What experiments can (or will) help us answer this question?

- Train a language model from scratch with additional typological features as input, and compare its performance with the current model.
  Possible issue(s): cost inefficient, or even intractable.
- Finetune a pretrained language model with additional typological features as input.
  Possible issue(s): cost inefficient, balance between the new and old data distribution.
- Probe if these features have been encoded in the model.
  Possible issue(s): this will not directly improve the model; there may be shortcuts in experimental designs.

Discussion 2

Q: What's next in computational multilingualism?

- How do language models learn and represent typological features?

Discussion 2

Q: What's next in computational multilingualism?

- How do language models learn and represent typological features?
- How can typological information be grounded in concrete real-world features?

Discussion 2

Q: What's next in computational multilingualism?

- How do language models learn and represent typological features?
- How can typological information be grounded in concrete real-world features?
- How can language models be human cognitive models, with awareness of cross-lingual variation?

Discussion 2

Q: What's next in computational multilingualism?

- How do language models learn and represent typological features?
- How can typological information be grounded in concrete real-world features?
- How can language models be human cognitive models, with awareness of cross-lingual variation?
- LLM assisted language documentation?

Any more ideas?

Next

Language models as human cognitive models