

# Cross-lingual NLP

Sara Stymne

Uppsala University  
Department of Linguistics and Philology

September 1, 2020

# What is a cross-lingual model

- Used to describe systems that involve more than one language
- Not one clear definition

# Typical NLP scenario

- You want to solve problem X for language Y
- You collect annotated data
- You apply some ML algorithm

# Typical NLP scenario

- You want to solve problem X for language Y
- You collect annotated data
- You apply some ML algorithm
- But:
  - There might not be any annotated data for X and Y
  - There might not even be much data at all for Y
  - There might be no pre-processing tools for Y

# Typical NLP scenario

- You want to solve problem X for language Y
- You collect annotated data
- You apply some ML algorithm
- But:
  - There might not be any annotated data for X and Y
  - There might not even be much data at all for Y
  - There might be no pre-processing tools for Y
  - You do not feel up to creating all these resources

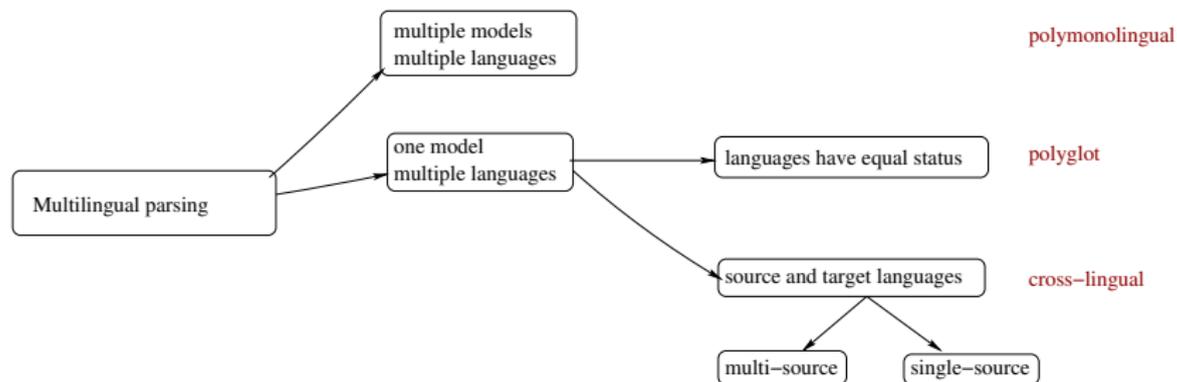
# Use other languages!

- Luckily, languages are related, and can have a lot in common!
- Maybe there is a language similar to Y which has data and resources
- Cross-lingual NLP: Use data/resources for one (or more) languages, to solve a problem for another language!

# Use other languages!

- Luckily, languages are related, and can have a lot in common!
- Maybe there is a language similar to Y which has data and resources
- Cross-lingual NLP: Use data/resources for one (or more) languages, to solve a problem for another language!
- Often used for low-resource languages
- But can also improve systems for medium/high resource languages

# Terminology suggestion for parsing



From Miryam de Lhoneux

# Focus for our group

- Polyglot:
  - Models that include several languages with equal status
- Cross-lingual:
  - Models that use one or more source languages and apply to a target language
  - No or little (annotated) data from the target language

# Not in focus

- Polymonolingual systems
  - Systems where one architecture is used for many languages, but where an individual model is trained for each language
- Machine translation
  - Except when machine translation systems are trained in a cross-lingual/polyglot manner

# Applications

- Multilingual systems can be trained for all type of applications
  - Tagging
  - Parsing
  - Machine translation
  - Lemmatization
  - Language modelling
  - Semantic role labelling
  - ...

# Resources used for cross-lingual systems

- Parallel corpora
- Bilingual lexicons/Tag dictionaries
- Typology, databases like WALS
- Language relatedness
- Target data (possibly tiny, noisy and/or incomplete)
- Cross-lingual word embeddings
- Multilingual large LMs

# Cross-lingual methods

- Annotation projection
- Translation of data
- Delexicalized transfer
- Parameter transfer
- Training guidance/soft constraints
- Joint learning
- Zero-shot learning with multilingual LMs
- ...

# Neural cross-lingual systems

- Neural models typically work well for cross-lingual models
- Cross-lingual systems can be viewed as multi-task systems
- Possible to share all or parts of an architecture
- Allows language representations as part of models
- Multilingual word and sentence representations:
  - Cross-lingual embeddings
  - Multilingual LMs

## Example: cross-lingual dependency parsing

- Work from our parsing group at UU (de Lhoneux, Nivre, Smith, Stymne)
- Neural dependency parser
- Add a treebank embedding to the representation of words
- The rest of the architecture is shared for all languages
- Train cross-lingual models for groups of mainly related languages

## Example: cross-lingual dependency parsing

- Work from our parsing group at UU (de Lhoneux, Nivre, Smith, Stymne)
- Neural dependency parser
- Add a treebank embedding to the representation of words
- The rest of the architecture is shared for all languages
- Train cross-lingual models for groups of mainly related languages
- This method also works monolingually when a language has many (diverse) treebanks

# Our BiLSTM-based parser

$X_{\text{the}}$

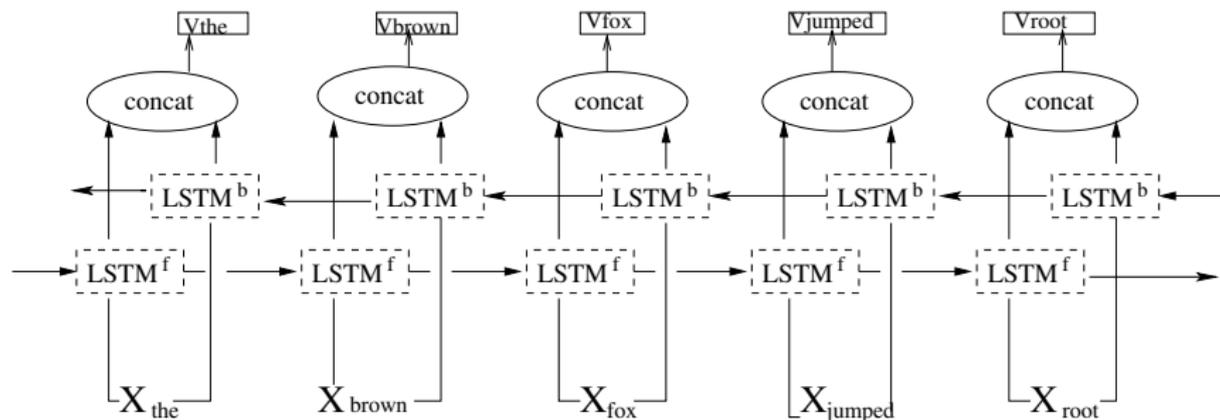
$X_{\text{brown}}$

$X_{\text{fox}}$

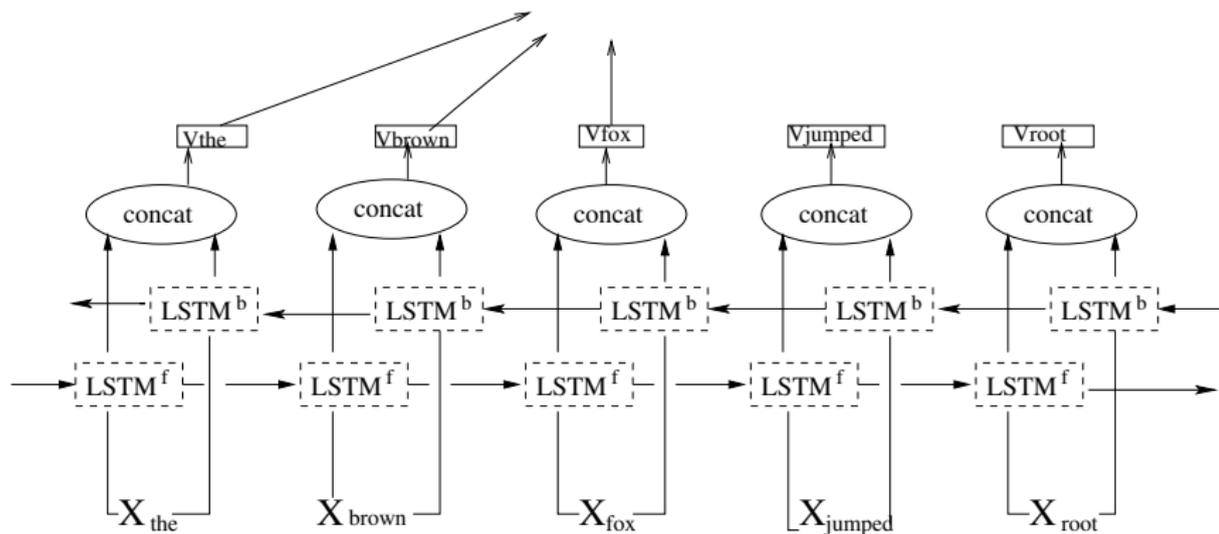
$X_{\text{jumped}}$

$X_{\text{root}}$

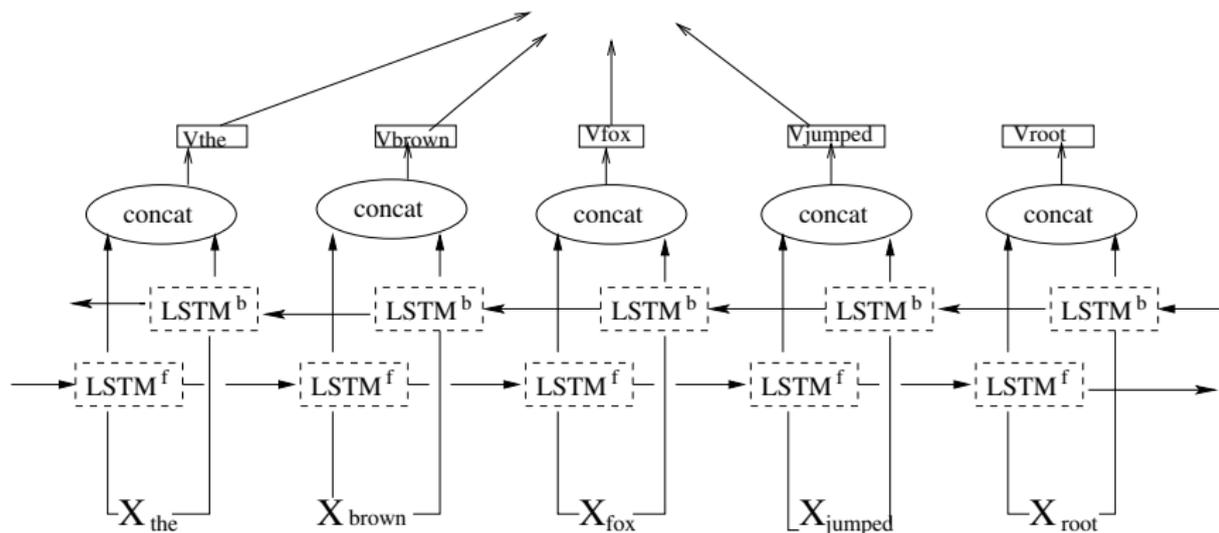
# Our BiLSTM-based parser



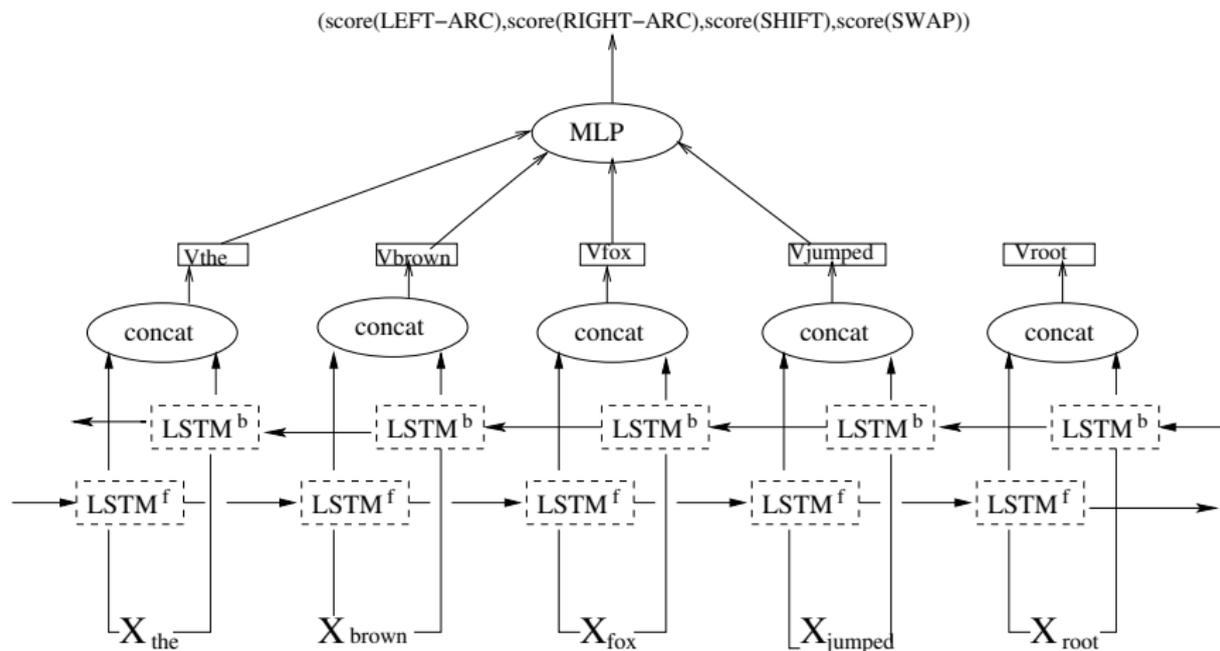
# Our BiLSTM-based parser



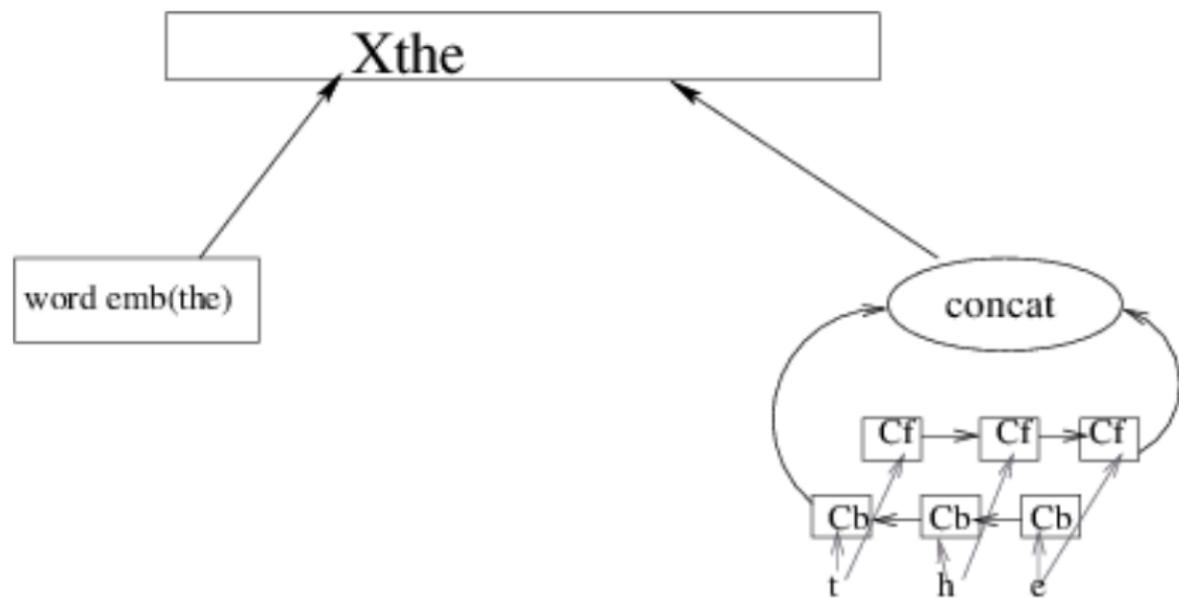
# Our BiLSTM-based parser



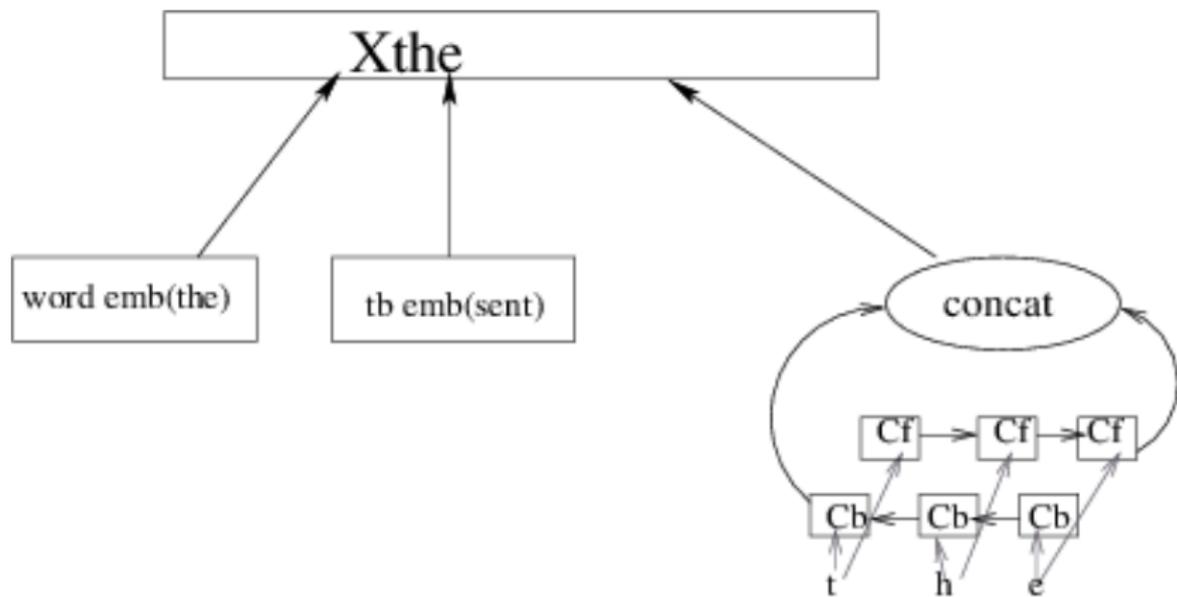
# Our BiLSTM-based parser



# Word representations



# Word representations + treebank embeddings



# Cross-lingual parsing: results

- Results at CoNLL 2018 shared task
- Comparison with a monolingual model
- Metric: LAS

Language(s)	Monolingual	X-ling	Diff
Kazakh	23.9	32.0	+8.1
Swedish	83.3	84.3	+1.0
German	75.2	75.5	+0.3
Low-resource	17.7	25.3	+7.6
All	70.7	72.3	+1.6

# Project suggestions

- All projects should involve more than one language
- You can focus on essentially any application

# Project suggestions

- All projects should involve more than one language
- You can focus on essentially any application
- Some possibilities (CLP = cross-lingual/polyglot)
  - Come up with a new CLP method or an extension of an existing CLP method for a specific task
  - Extend CLP work to a new application/language
  - Perform an in-depth evaluation study of some CLP method
  - Compare different CLP methods or resources
  - Explore which languages to choose and/or how to mix languages for a/several target language(s)
  - Address issues with inconsistent tag sets/annotations across languages
  - ...

- Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. (2020) *A Call for More Rigor in Unsupervised Cross-lingual Learning*. ACL
- Shijie Wu and Mark Dredze. (2019) *Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT*. EMNLP-IJCNLP.
- David Yarowsky, Grace Ngai, and Richard Wicentowski. (2001) *Inducing multilingual text analysis tools via robust projection across aligned corpora*. HLT.
- Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. (2018) *82 Treebanks, 34 Models: Universal Dependency Parsing with Multi-Treebank Model*. CoNLL.
- Dan Kondratyuk, Milan Straka. (2019) *75 Languages, 1 Model: Parsing Universal Dependencies Universally*. EMNLP-IJCNLP.
- Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord (2020) *UDapter: Language Adaptation for Truly Universal Dependency Parsing*. EMNLP.
- Aditi Chaudhary, Jiateng Xie, Zaid Sheikh, Graham Neubig, Jaime Carbonell. (2019) *A Little Annotation does a Lot of Good: A Study in Bootstrapping Low-resource Named Entity Recognizers*. EMNLP-IJCNLP.
- Yu-Hsiang Lin et al. (2019) *Choosing Transfer Languages for Cross-Lingual Learning*. ACL.
- Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, Kristina Toutanova (2021) *Revisiting the Primacy of English in Zero-shot Cross-lingual Transfer*. arXiv.

Questions?