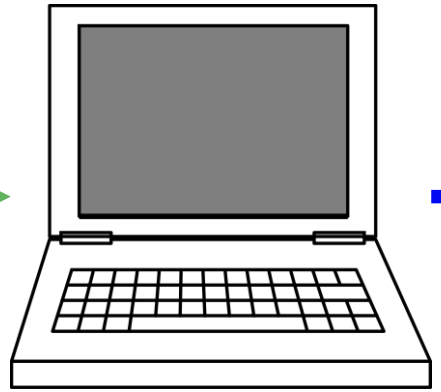
Recent Advances in Transferable Representation Learning

Techniques for Facilitating Cross-Lingual/Domain Transfer
A Case Study in Dependency Parsing

AAAI 2020 Tutorial

Muhao Chen, **Kai-Wei Chang**, Dan Roth

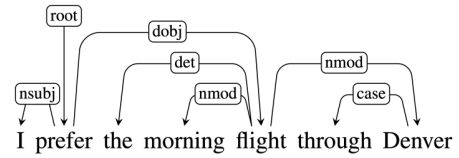
Cross-lingual Dependency Parsing

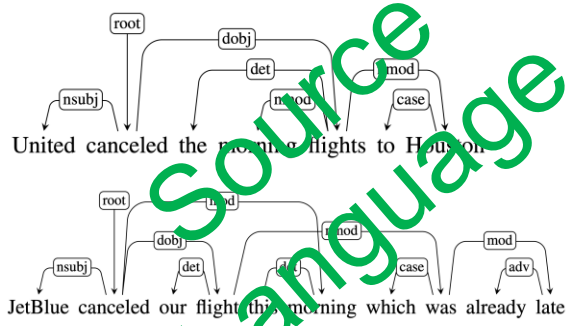


Train

Parse

English Treebank





…

Source Language

French Corpus

Aujourd'hui j'ai rencontré un accident

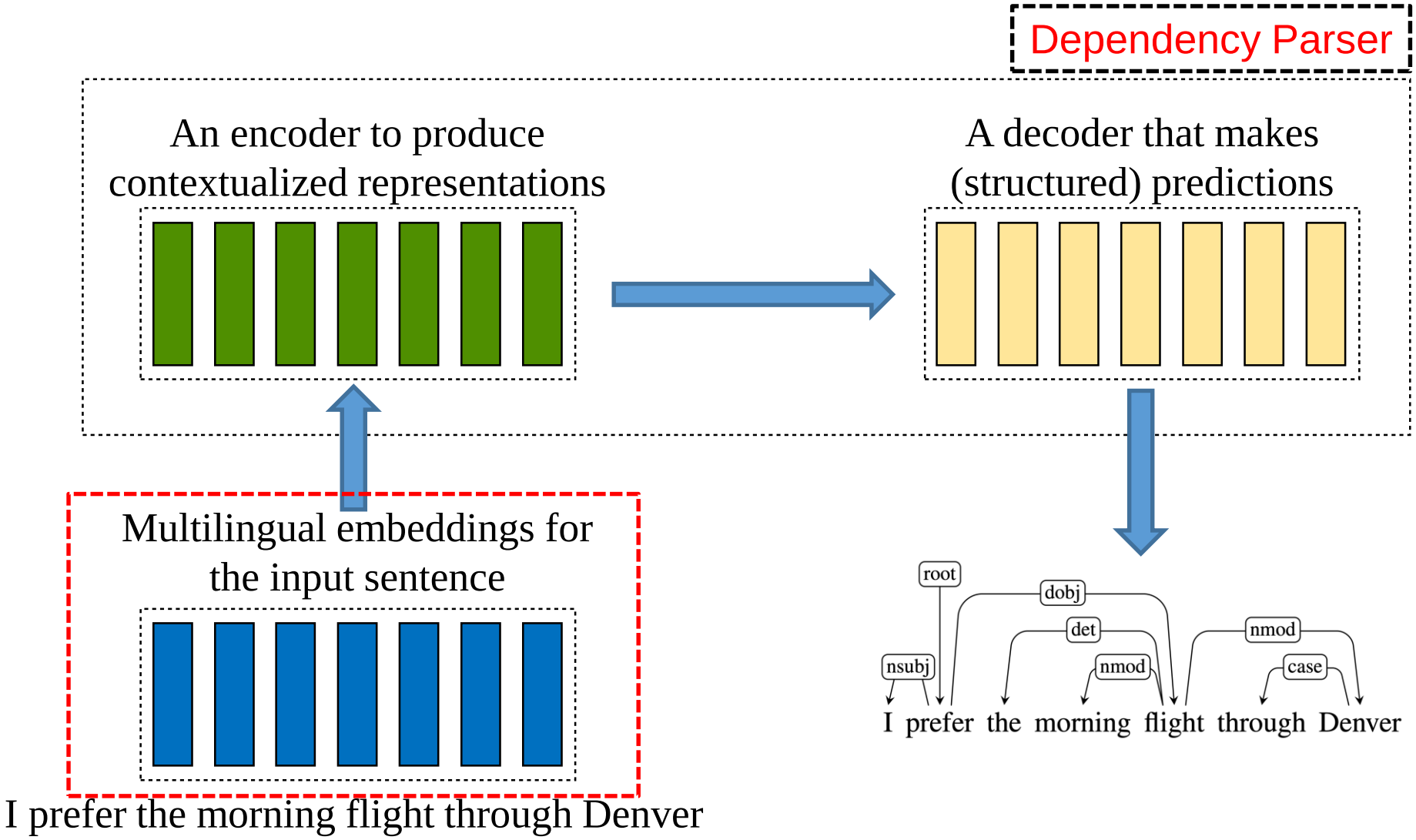
J'ai besoin de prendre le vol

Je ne pouvais pas déjeuner aujourd'hui à cause d'une réunion

…

Target Language

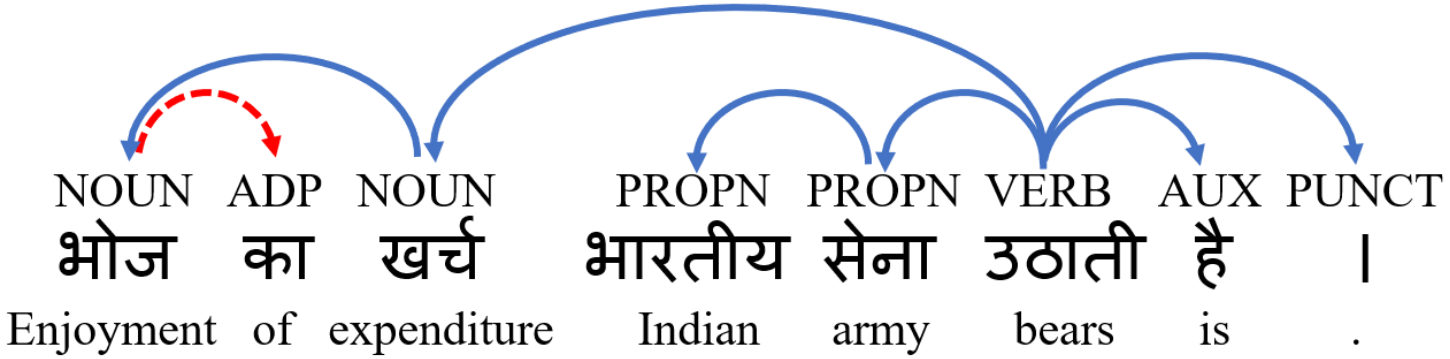
Dependency Parsing



Challenges for Cross-Lingual Transfer

Different languages have different properties

(e.g., word order)



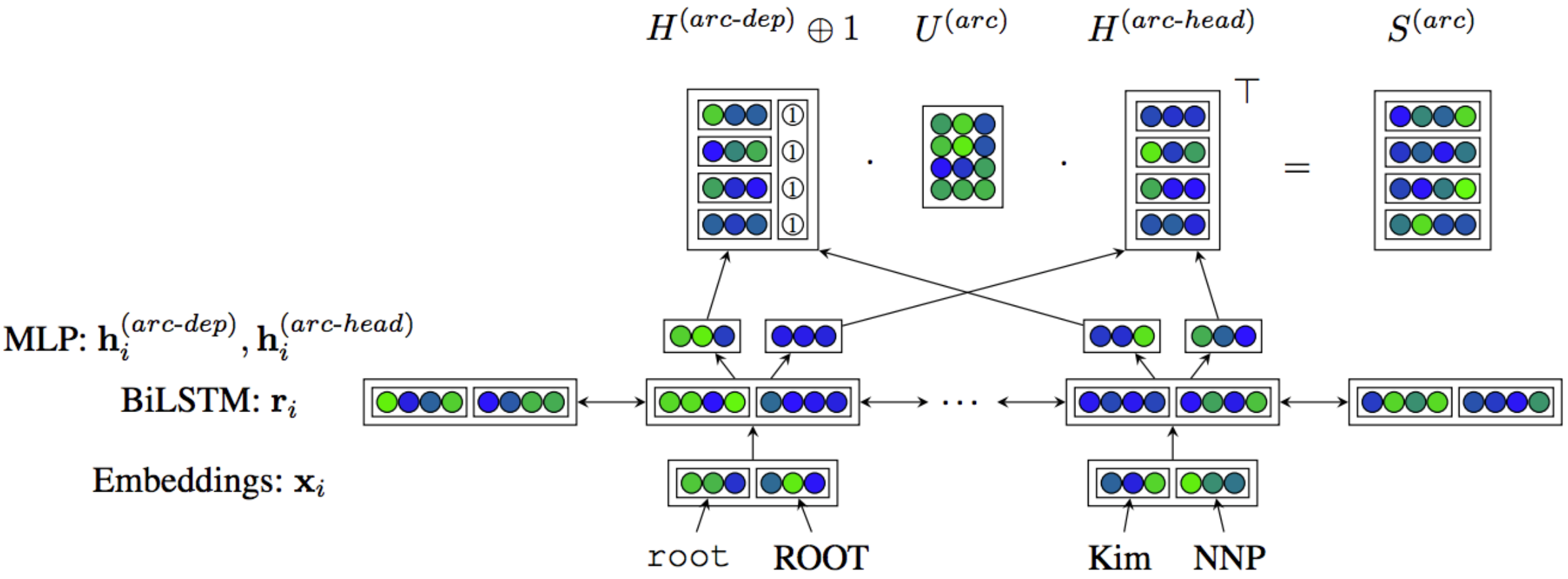
Improve transfer learning across languages

(Learning language-agnostic representation)

How to Perform Better Cross-Lingual Transfer?

- Examine and verify our hypothesis on cross-lingual dependency parsing
  - UD annotation for over 70 languages
  - Parser is a low-level task that reflects the problems
- Remove language-specific knowledge (e.g., word order) from encoder
- Add language-specific knowledge (weak supervision) to decoder

Background: Deep Biaffine Parser

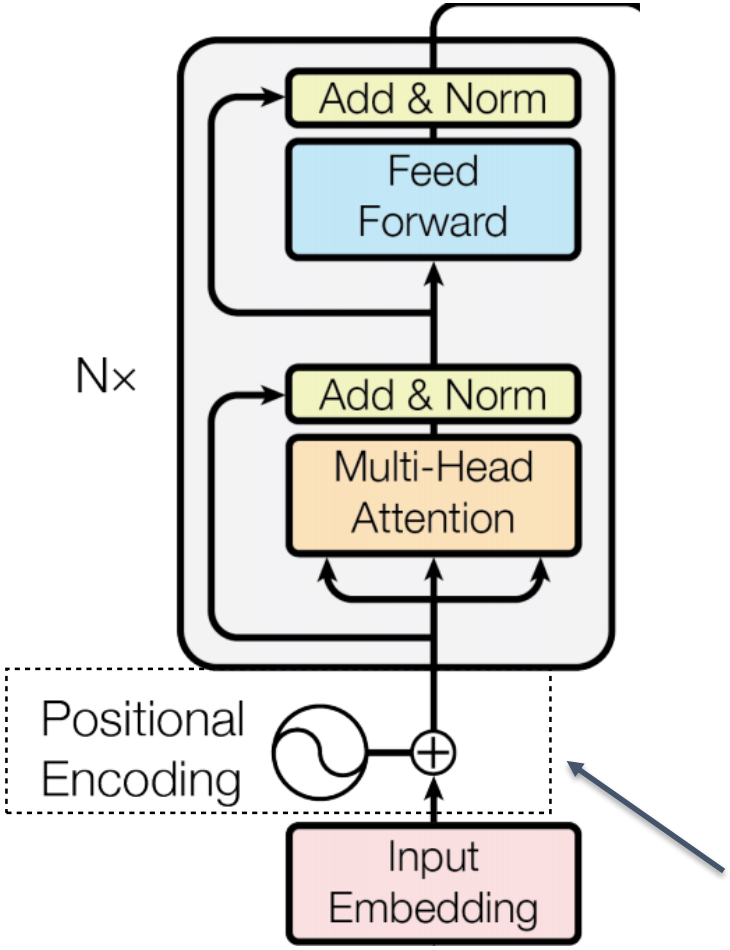


- **Graph-based** parser
- Encoder: RNN (Order-sensitive); Decoder: Graph (Order-free)

Dozat and Manning (ICLR2017)

Remove Word Order information [Ahamd+ 19]

Multi-Head Self-Attention with Relative Position



Flexible positional encoding (order-free)

- In the original paper:

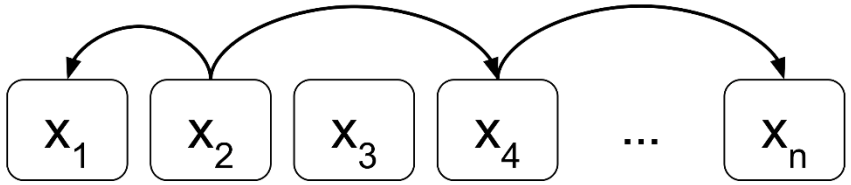
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Vaswani et. al. (NIPS 2017)

- Encoder absolute distance

$a^V_{2,1}=w^V_{-1}$ $a^V_{2,4}=w^V_2$ $a^V_{4,n}=w^V_k$

$a^K_{2,1}=w^K_{-1}$ $a^K_{2,4}=w^K_2$ $a^K_{4,n}=w^K_k$



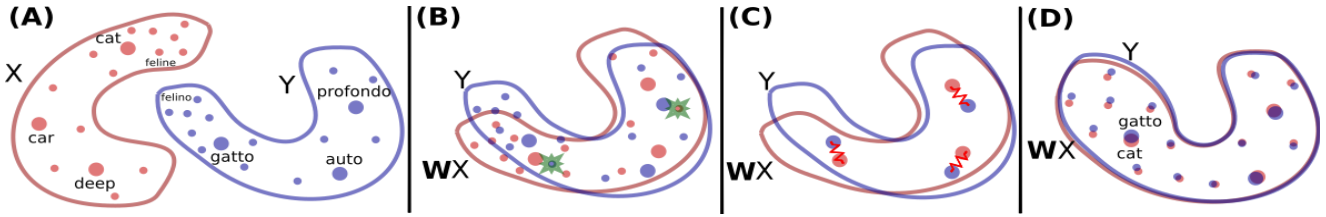
Shaw et. al. (NAACL2018)

Kai-Wei Chang (http://kwchang.net/talks/genderbias/)

Architectures for Cross-lingual Parser

- Embedding

Facebook MUSE



Conneau et. al. ICLR2018

- Encoders
  - ☐ BiLSTMs (order-sensitive) v.s.
  - ☐ Multi-Head Self-Attention with Absolute Relative Positional Encoding (order-free)
- Decoders
  - ☐ Pointer Network (order-sensitive) v.s.
  - ☐ BiAffine Attention (order-free)

Experiments

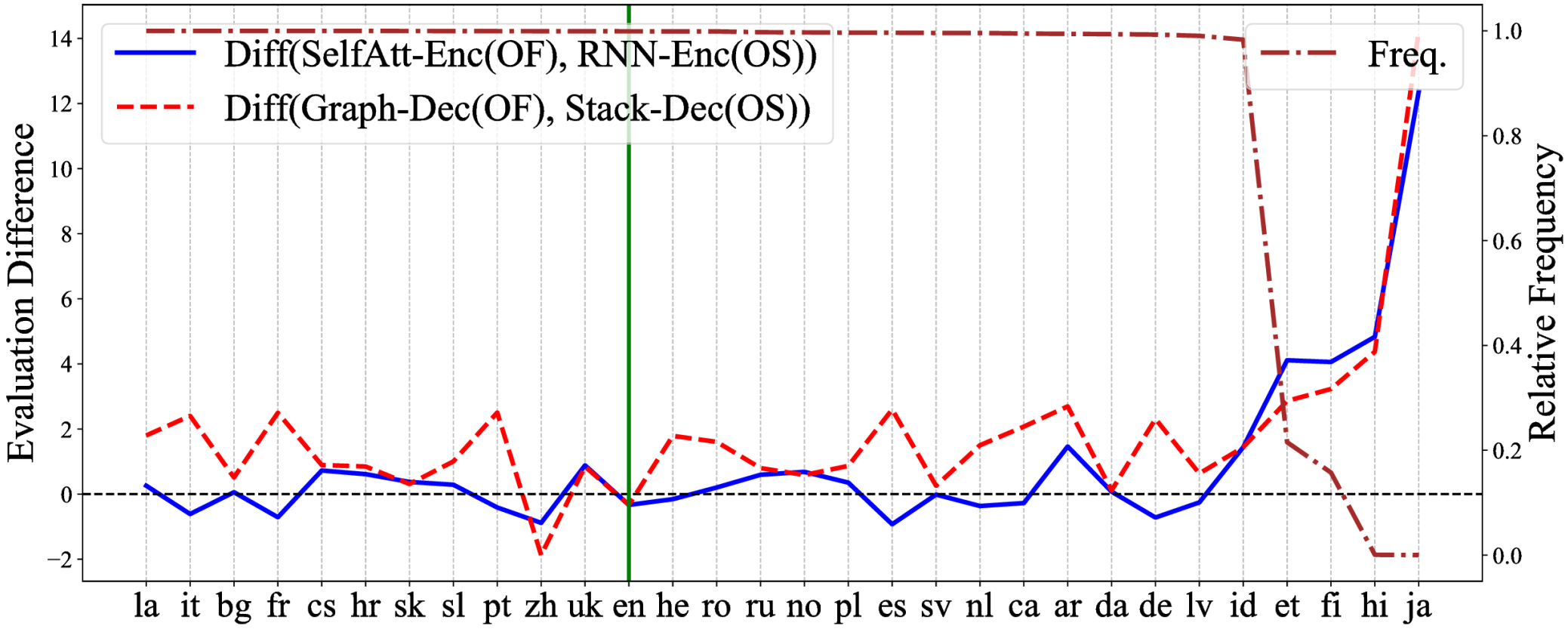
- Datasets:
  - UD (V2.2)
  - 31 languages, 12 families
- Setting:
  - Train/Dev on English
  - Directly predict on the rest 30 languages (zero-shot)

Language Families	Languages
Afro-Asiatic	Arabic (ar), Hebrew (he)
Austronesian	Indonesian (id)
IE.Baltic	Latvian (lv)
IE.Germanic	Danish (da), Dutch (nl), English (en), German (de), Norwegian (no), Swedish (sv)
IE.Indic	Hindi (hi)
IE.Latin	Latin (la)
IE.Romance	Catalan (ca), French (fr), Italian (it), Portuguese (pt), Romanian (ro), Spanish (es)
IE.Slavic	Bulgarian (bg), Croatian (hr), Czech (cs), Polish (pl), Russian (ru), Slovak (sk), Slovenian (sl), Ukrainian (uk)
Japanese	Japanese (ja)
Korean	Korean (ko)
Sino-Tibetan	Chinese (zh)
Uralic	Estonian (et), Finnish (fi)

Case Study – Adposition: Preposition v.s. postposition

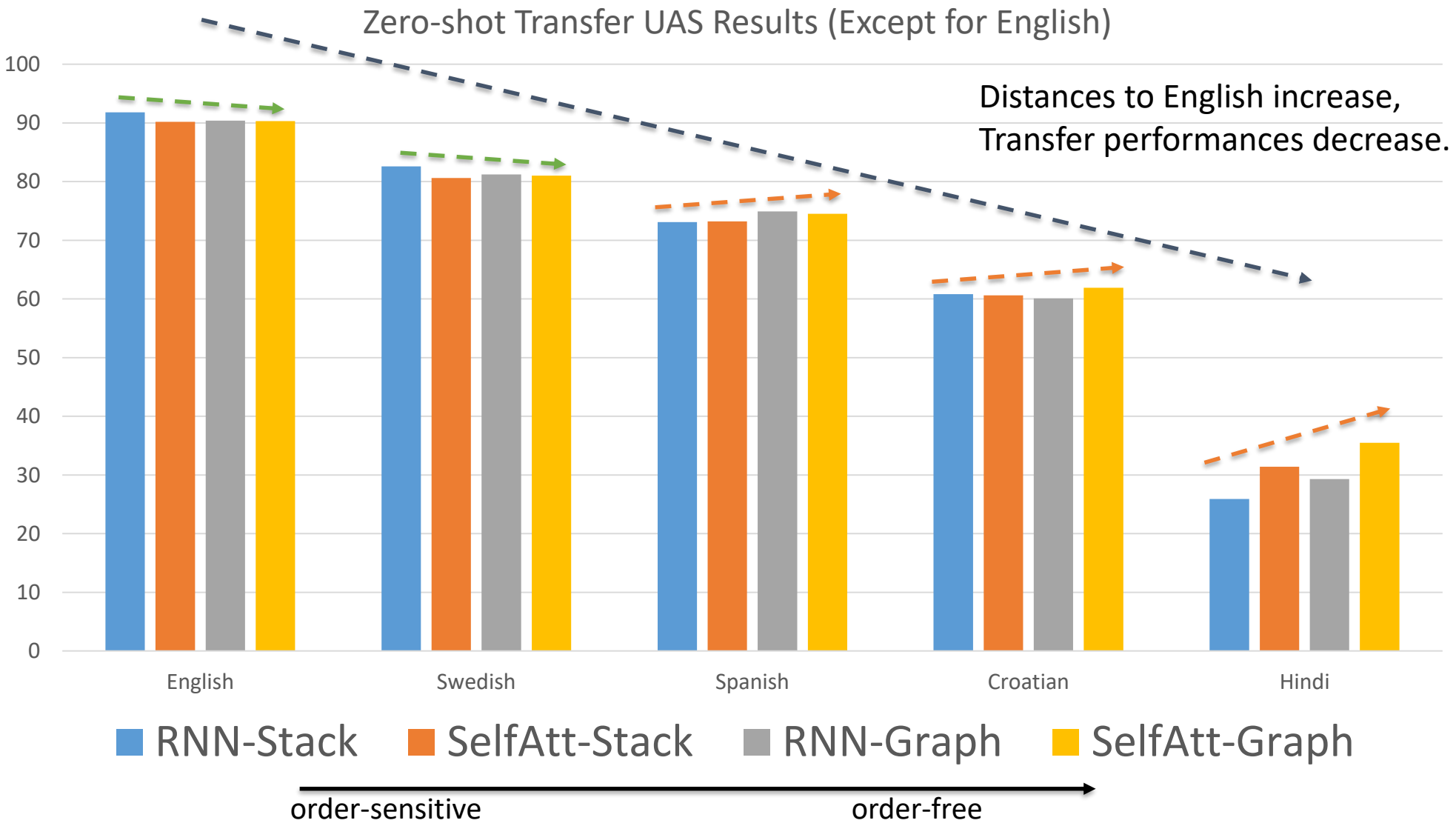
Postposition: I decided many years *ago to invent myself*

Preposition: I decided many years *ago to invent myself*



The languages (x-axis) are sorted by this relative frequency from high to low

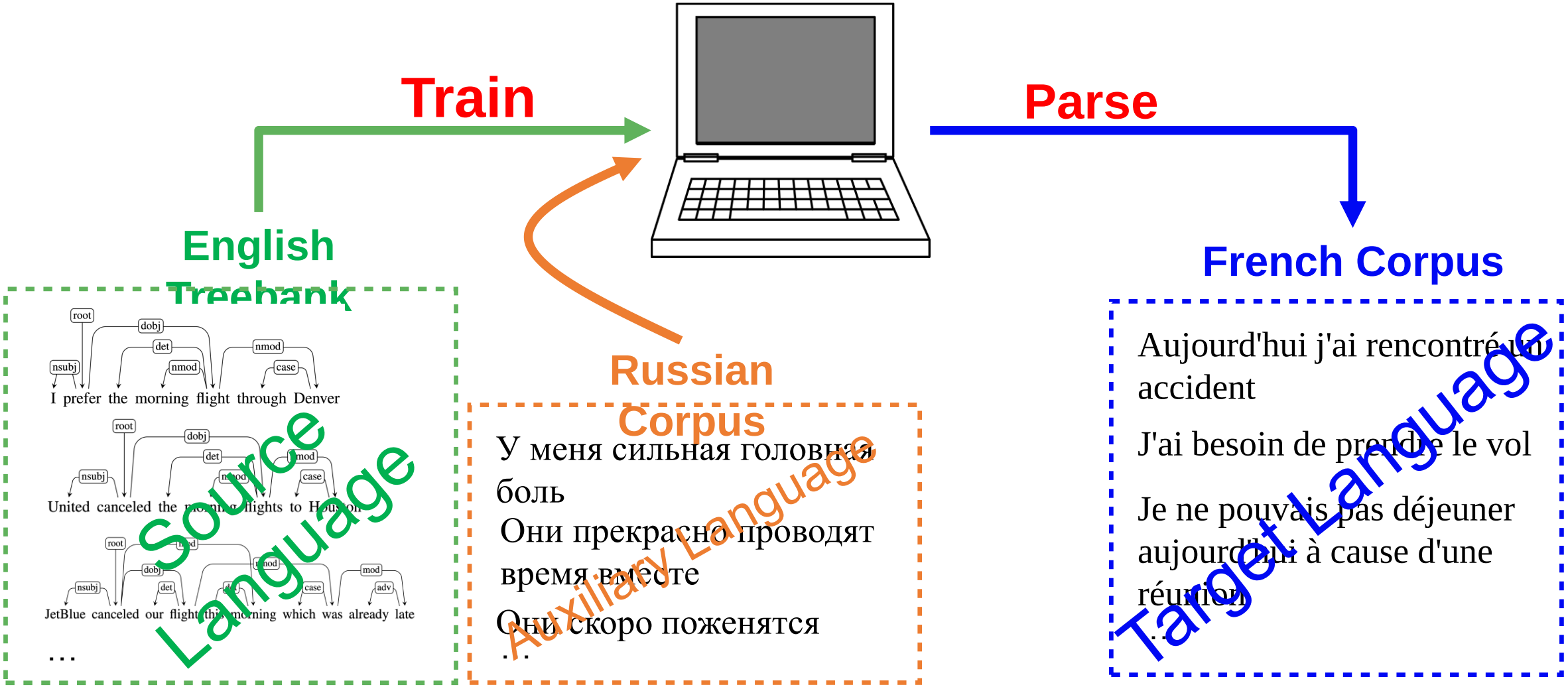
Selected Transfer Results of Different Architectures



How to Perform Better Cross-Lingual Transfer?

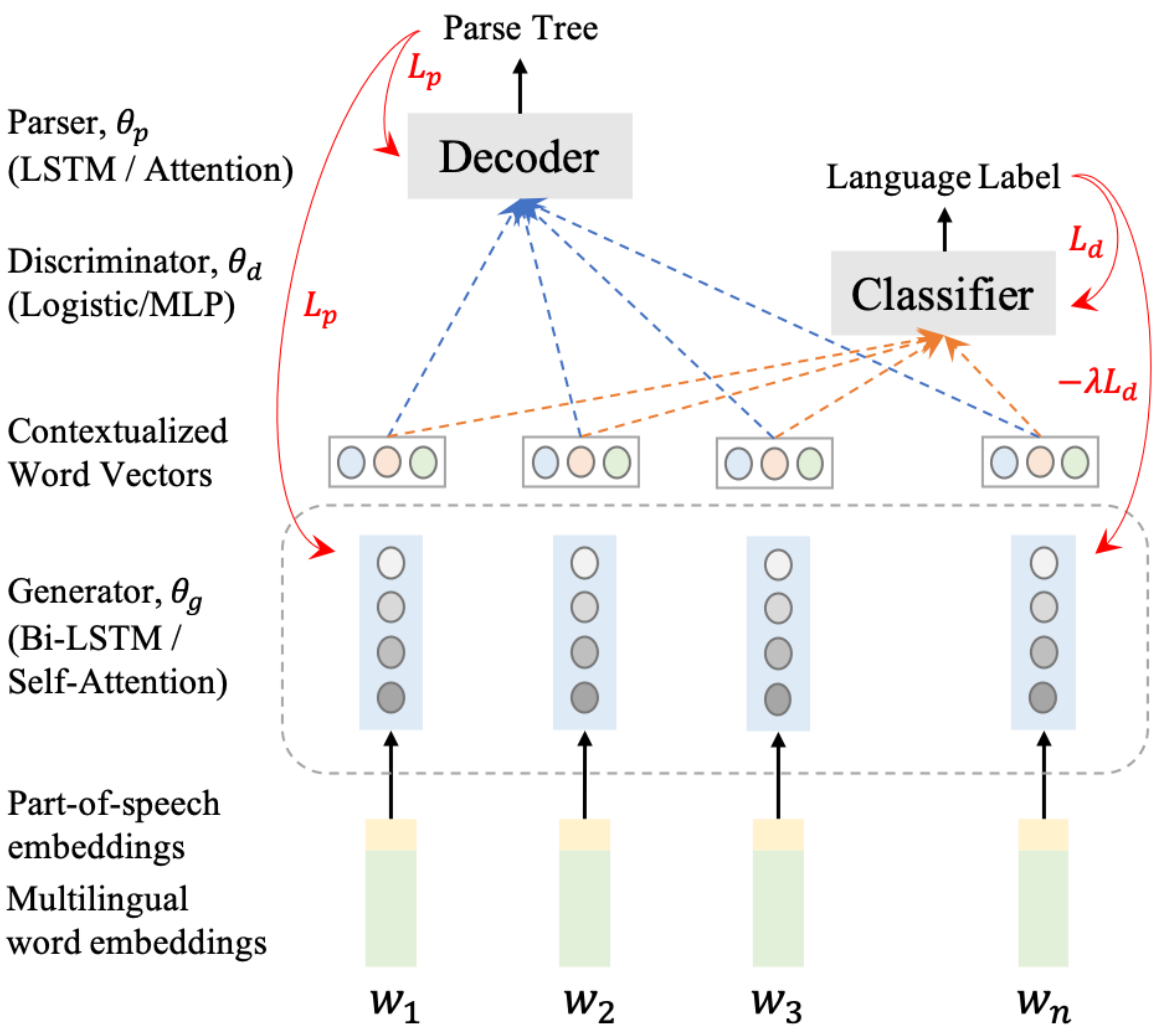
- Examine and verify our hypothesis on cross-lingual dependency parsing
  - UD annotation for over 70 languages
  - Parser is a low-level task that reflects the problems
- Remove language-specific knowledge (e.g., word order) from encoder
- Add language-specific knowledge (weak supervision) to decoder

Enhancing Shared Representations by Auxiliary Languages



Adversarial Learning

[Ahmad+19]



Similar ideas have been used in domain adaptation [Ganin+ 16] and other cross lingual tasks [Chen+18]

Experiment Setup

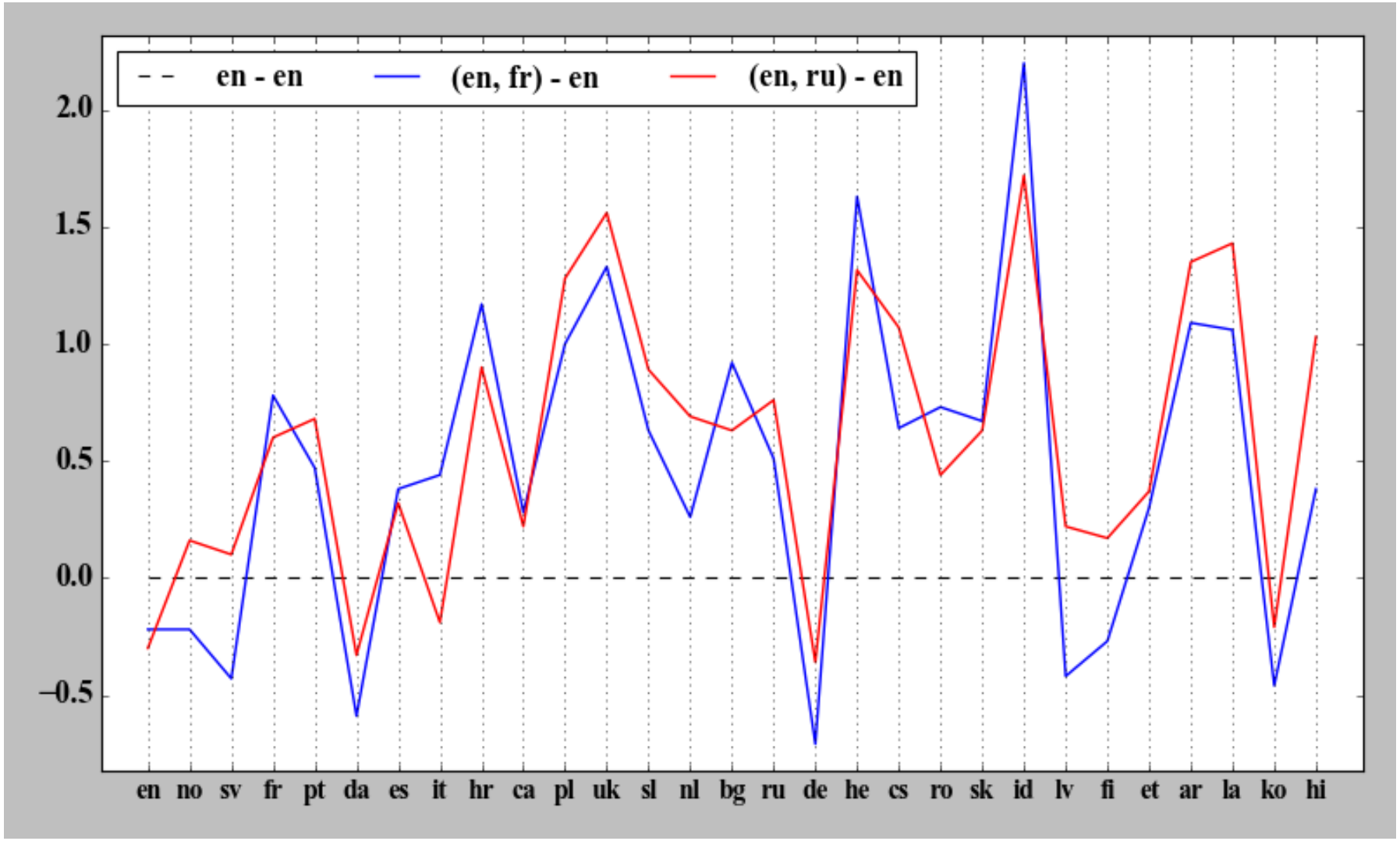
Embedding

- Token embeddings
  - Multilingual Embeddings (MUSE) [Smith et al., 2017, Bojanowski et al., 2017]
  - Multilingual BERT (M-BERT) [Devlin et al., 2017]
- Part-of-speech embeddings

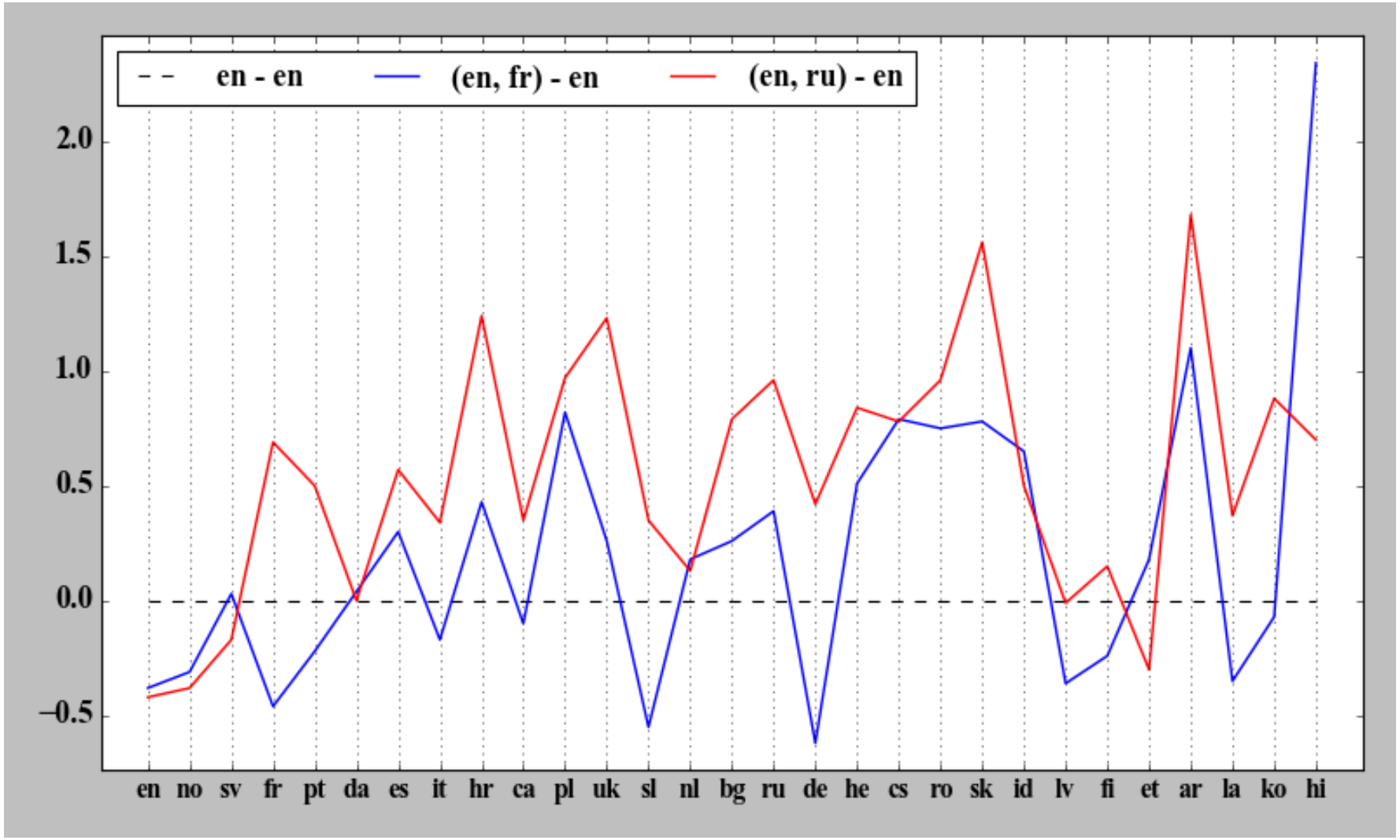
Parsers [Ahmad et al., 2019]

- Graph-based: Self-attentive-Graph
  - Multi-Head Self-Attention (**order-free**)
- Transition-based: RNN-StackPtr
  - BiLSTMs (**order-dependent**)

Cross-lingual transfer with Mulitilingual embedding



Cross-lingual transfer with Mulitilingual BERT



How to Perform Better Cross-Lingual Transfer?

- Examine and verify our hypothesis on cross-lingual dependency parsing
  - UD annotation for over 70 languages
  - Parser is a low-level task that reflects the problems
- Remove language-specific knowledge (e.g., word order) from encoder
- Add language-specific knowledge (weak supervision) to decoder

World ATLAS of Language Structures

THE WORLD ATLAS
OF LANGUAGE STRUCTURES
ONLINE

Home | Features | Chapters | Languages | References | Authors

Features

A feature is a structural property of language that describes one aspect of cross-linguistic diversity. A WALS feature has between 2 and 28 different values, shown by different colours on the maps. Most features correspond straightforwardly to chapters, but some chapters are about multiple features.

Showing 1 to 100 of 192 entries

← Previous | 1 | 2 | Next →

Id	Name	Authors	Area	Languages	Details
Search	Search		--any--	Search	
1A	Consonant Inventories	Ian Maddieson	Phonology	563	Values
2A	Vowel Quality Inventories	Ian Maddieson	Phonology	564	Values

Corpus-Statistics Constraints

[Tao+ 19]

- Consider constraints in the forms:
  - the ratio r of POS1 being on the left in POS1-POS2 arcs

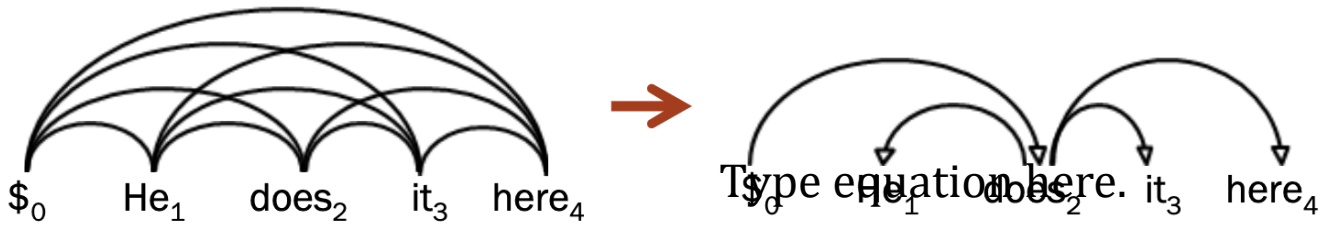
ADJ NN v.s. NN ADJ

- Compiling from WALS features:
  - Dominant order ⇒ 75% or more

- Add constraints when performing parsing

Parsing with Corpus-Level Constraints

- Parsing can be formulated as an integer linear programming problem
  - Finding the maximum spanning tree



$$Y^* = \arg\max_{Y \in \Phi(X)} score(X, Y)$$

Illustration is from https://slideplayer.com/slide/6623811/

- Corpus-level constraints added to guide inference
  - Doesn't require model retraining

$$Y^* = \arg\max_{Y} score(X, Y) \quad s.t. \quad \left| Ratio_{predicted} - Ratio_{given} \right| \leq \delta$$

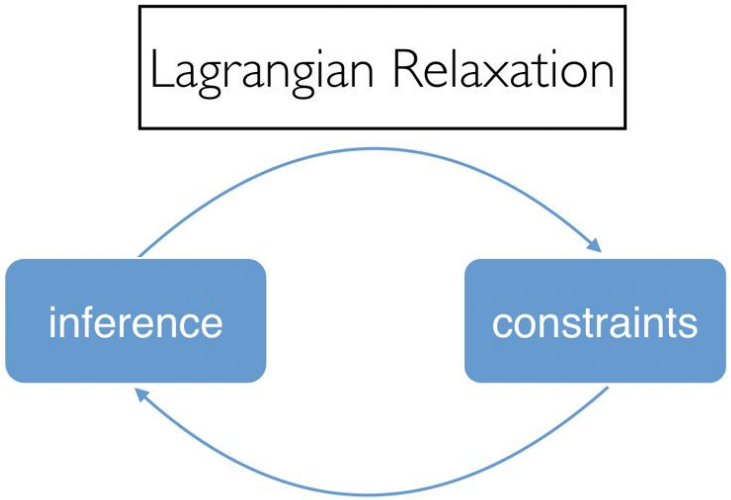
- Reuse model inference through Lagrangian relaxation

Constrained Inference

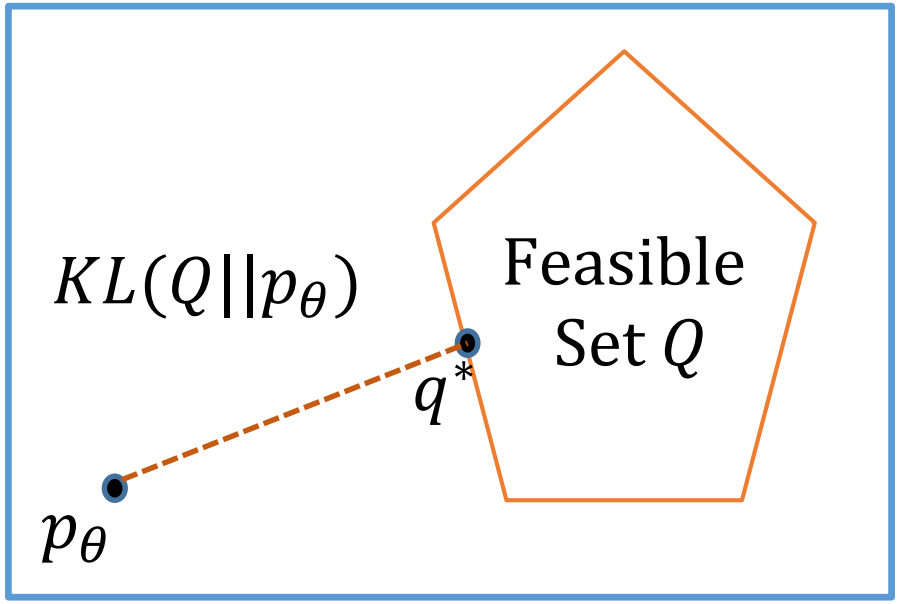
Lagrangian Relaxation

$$\max_{y_i} \sum_i s(y_i, sentence_i)$$

s. t. Corpus−Statistics Constraints



Posterior Regularization



LR, PR get improvements in 15, 17 out of 19 target languages from variant of language families, respectively