On Difficulties of Cross-Lingual Transfer with Order Differences: A Case Study on Dependency Parsing

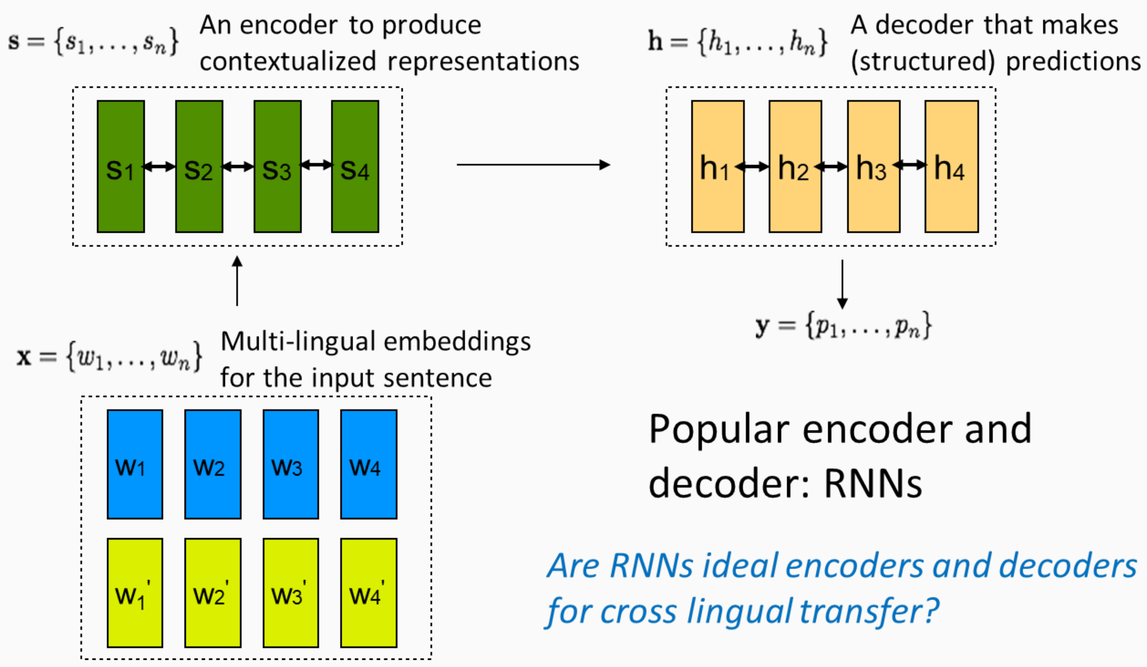
Wasi Uddin Ahmad[1*], Zhisong Zhang[2*], Xuezhe Ma[2],
Eduard Hovy[2], Kai-Wei Chang[1], Nanyun Peng[3]

[1]University of California, Los Angeles, [2]Carnegie Mellon University,
[3]University of Southern California

* This work was supported in part by NSF IIS-1760523.

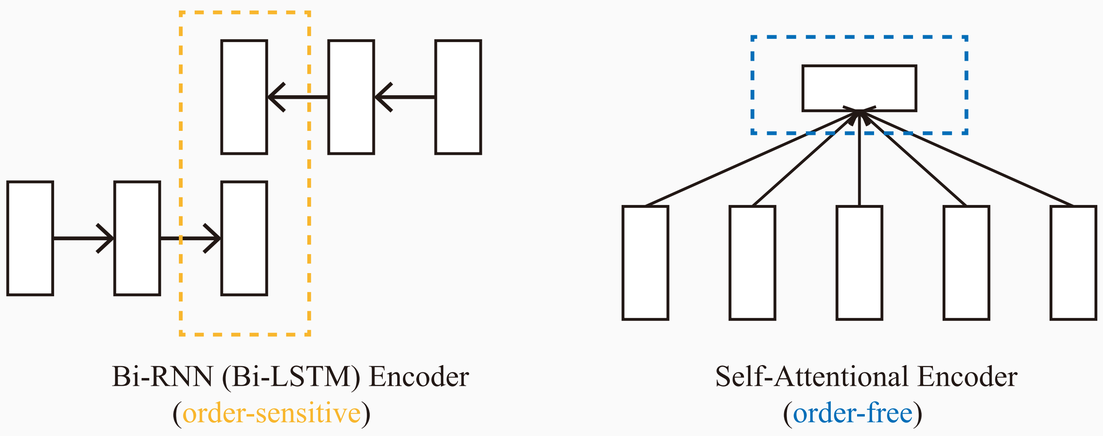
Introduction

- We investigate neural model architectures for **cross-lingual transfer**.
- Since different languages may have **different word orders**, **an order-agnostic model** may perform better generally.
- Our hypothesis is that a model that overfits less to language-specific order information can generally perform better when transferred to other languages.
- We use zero-resource cross-lingual **dependency parsing** as a testbed and conduct evaluations on 31 languages across a broad spectrum of language families.



Model Overview

- Inputs: (aligned multilingual) Embeddings + Universal-POS.
- Encoder: BiLSTMs (order-sensitive) v.s. Multi-Head Self-Attention (order-free)



- Decoder: Stack Pointer (order-sensitive) v.s. Graph-based (order-free):
  The comparison of the two decoders is also "RNN v.s. Self-Att" in some way, although being more specific to the parsing problem.

Self-Attention with Relative Positional Repr.

- Relative positional Self-Attention: The input sequence of vectors $\mathbf{x} = (x_1, \ldots, x_n)$ are transformed into $\mathbf{z} = (z_1, \ldots, z_n)$, based on the self-attention mechanism:

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V + a_{ij}^V) \qquad \alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}} \qquad e_{ij} = \frac{x_i W^Q (x_j W^K + a_{ij}^K)^T}{\sqrt{d_z}}$$

Here, a_{ij}^V and a_{ij}^K are relative positional representations for the two position i and j.

- We adopt "Direction-Free" Relative Positional Representation:

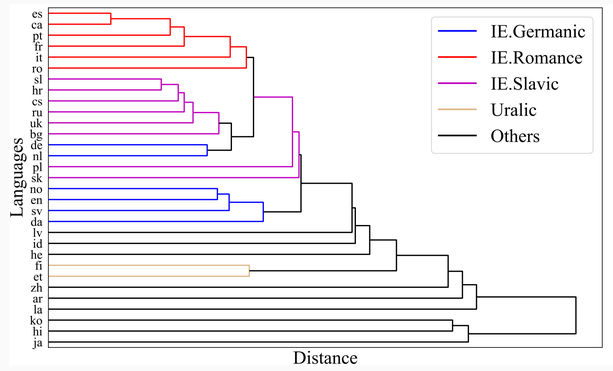
$$a_{ij}^K = w_{clip(|j-i|,k)}^K \qquad a_{ij}^V = w_{clip(|j-i|,k)}^V \qquad \text{clip}(x,k) = \min(|x|, k)$$

- Comparisons of different encoders (average transfer parsing performances), where our non-directional relative position representation ("SelfAtt-Relative") performs the best:

Model	UAS%	LAS%
SelfAtt-Relative (Ours)	64.57	54.14
SelfAtt-Relative+Dir	63.93	53.62
RNN	63.25	52.94
SelfAtt-Absolute	61.76	51.71
SelfAtt-NoPosi	28.18	21.45

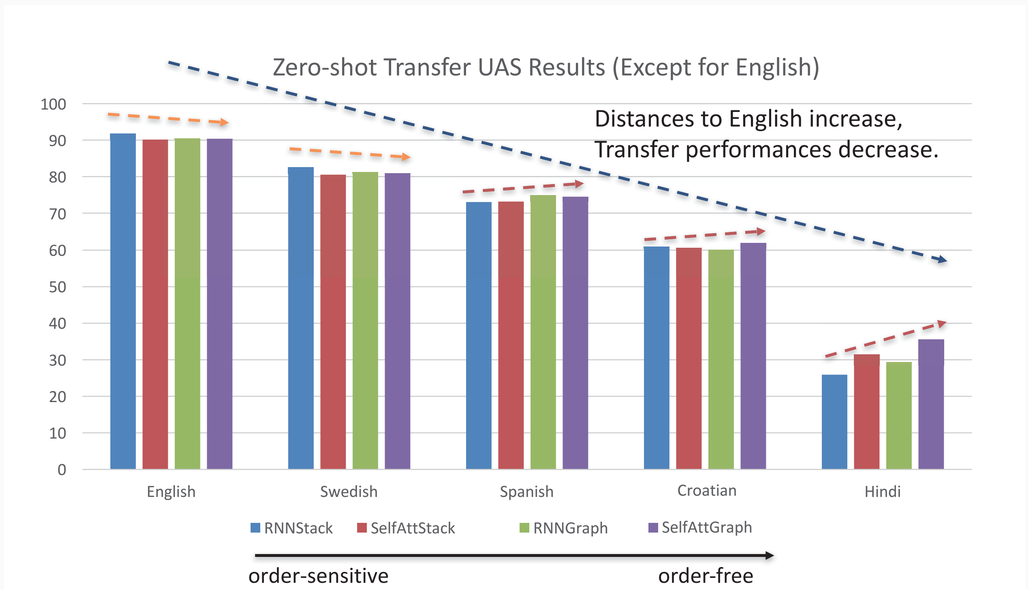
Experiment

- Language Clustering (left figure) and Overall Averaged Results (right table):

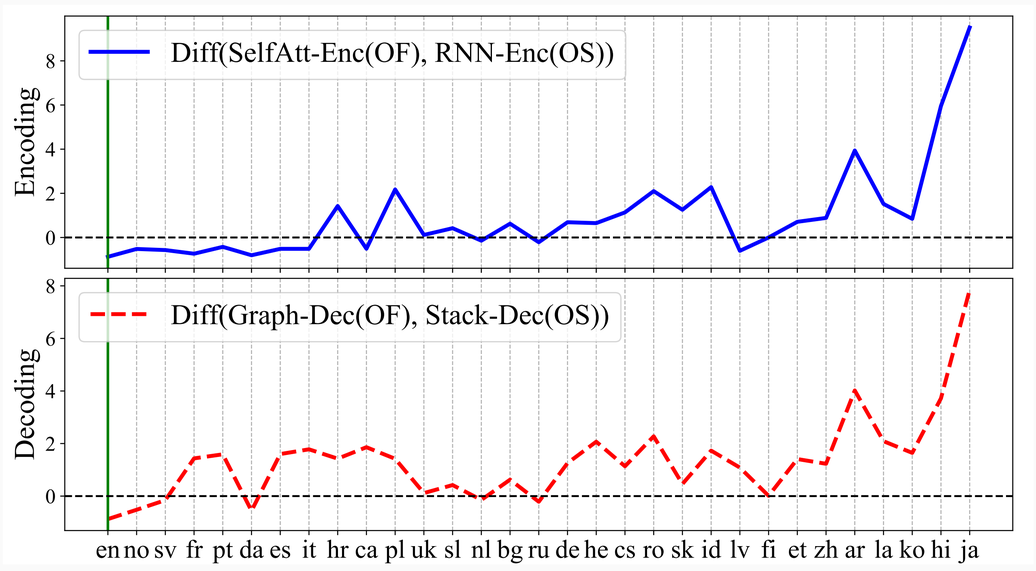


Model	UAS%	LAS%
SelfAtt-Graph	64.06	53.82
RNN-Graph	62.71	52.63
SelfAtt-Stack	62.22	52.00
RNN-Stack	62.37	51.89

- Performances on certain typical languages:



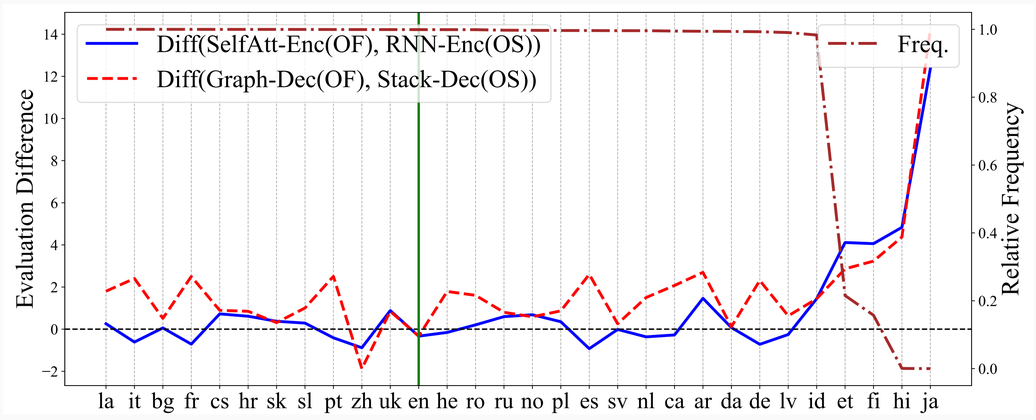
- Performance differences on all languages:



- Breakdown on Dependency Types

Analysis on specific dependency types. The blue and red curves and left y-axis represent the differences in evaluation scores, the brown curve and right y-axis represents the relative frequency of left-direction (modifier before head) on this type. The languages (x-axis) are sorted by this relative frequency from high to low:

- Adposition & Noun (ADP, NOUN, case):



- Adjective & Noun (ADJ, NOUN, amod):