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

English Treebank

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
Dependency Parsing

An encoder to produce contextualized representations

Multilingual embeddings for the input sentence

A decoder that makes (structured) predictions

I prefer the morning flight through Denver

Dependency Parser
Different languages have different properties (e.g., word order)

Enjoyment of expenditure Indian army bears is.

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)
Multi-Head Self-Attention with Relative Position

- In the original paper:
  \[
  PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{model}}) \\
  PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{model}})
  \]
  
  Vaswani et. al. (NIPS 2017)

- Encoder absolute distance

  \[
  a_{2,1}^v = w_{-1}^v \\
  a_{2,4}^v = w_2^v \\
  a_{4,n}^v = w_k^v \\
  a_{2,1}^k = w_{-1}^k \\
  a_{2,4}^k = w_2^k \\
  a_{4,n}^k = w_k^k
  \]

  Shaw et. al. (NAACL2018)

Remove Word Order information [Ahamd+ 19]

Kai-Wei Chang (http://kwchang.net/talks/genderbias/)
Architectures for Cross-lingual Parser

- **Embedding**
  - Facebook MUSE

- **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)

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

Zero-shot Transfer UAS Results (Except for English)

Distances to English increase, Transfer performances decrease.

Selected Transfer Results of Different Architectures

- RNN-Stack
- SelfAtt-Stack
- RNN-Graph
- SelfAtt-Graph

(order-sensitive) RNN-Stack, SelfAtt-Stack, RNN-Graph, SelfAtt-Graph (order-free)

Languages:
- English
- Swedish
- Spanish
- Croatian
- Hindi
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

Train

English Treebank

I prefer the morning flight through Denver

United canceled the morning flights to Houston.

JetBlue canceled our flight which was already late.

Parse

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.

Russian Corpus

У меня сильная головная боль.
Они прекрасно проводят время вместе.
Они скоро поженятся.

Source Language

Auxiliary Language

Target Language
Adversarial Learning

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 Multilingual embedding
Cross-lingual transfer with Multilingual 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
# 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

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Authors</th>
<th>Area</th>
<th>Languages</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Search</td>
<td>Search</td>
<td></td>
<td>any</td>
<td></td>
</tr>
<tr>
<td>1A</td>
<td>Consonant Inventories</td>
<td>Ian Maddieson</td>
<td>Phonology</td>
<td>563</td>
<td>Values</td>
</tr>
<tr>
<td>2A</td>
<td>Vowel Quality Inventories</td>
<td>Ian Maddieson</td>
<td>Phonology</td>
<td>564</td>
<td>Values</td>
</tr>
</tbody>
</table>
Corpus-Statistics Constraints

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

- Compiling from WALS features:
  - Dominant order $\Rightarrow$ 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

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

\[
Y^* = \arg \max_Y \text{score}(X, Y) \quad \text{s.t.} \quad |\text{Ratio}_{predicted} - \text{Ratio}_{given}| \leq \delta
\]

- Reuse model inference through Lagrangian relaxation

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

Lagrangian Relaxation

\[
\max_{y_i} \sum_i s(y_i, \text{sentence}_i)
\]

s. t. Corpus–Statistics Constraints

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