Isomorphic Transfer of Syntactic Structures in Cross-Lingual NLP

Edoardo Maria Ponti*, Roi Reichart†, Anna Korhonen*, Ivan Vulić*
*Language Technology Lab, University of Cambridge †Technion, IIT
{ep490,alk23,iv250}@cam.ac.uk roiri@ie.technion.ac.il

1. Introduction

- **Transferring or sharing knowledge** among languages is a popular solution to mitigate resource scarcity and harness language-independent information in NLP.
- Their effectiveness is challenged by cross-lingual variation in morpho-syntactic structures. This results in anisomorphism between the nodes $V$ and $U$ of equivalent dependency trees: there exists no bijection $f(V) \rightarrow U$ such that adjacencies between corresponding nodes are preserved.
- Can we a) **measure** anisomorphism, b) use it to **select** compatible source languages for knowledge transfer, and c) **process** source dependency trees to tailor them and improve downstream tasks?

2a. Metrics: Jaccard Index

Language-wide anisomorphism is measured by the Jaccard index of two sets of morphological features (e.g., tense= Past) $M_S$ and $M_T$ occurring at least once in a treebank.

$$J(M_S, M_T) = \frac{|M_S \cap M_T|}{|M_S \cup M_T|}$$

2b. Metrics: Tree Edit Distance

Instance-level anisomorphism is estimated by the (average) tree edit distance between tree pairs $S$ and $T$ in a multi-parallel Bible corpus with the Zhang-Sashua algorithm [1] based on a mapping $M$.

$$\gamma(M,S,T) = \sum_{\epsilon \in M} \gamma(S_i \rightarrow T_j) + \gamma(S_i \rightarrow \epsilon) + \gamma(\epsilon \rightarrow T_j)$$

3. Processing Dependency Trees

We leverage the ZS operations (change, delete, add) to process trees. Thus we adapt the constructions (e.g., predicative possession) of a source tree to the strategies of a target language (as defined by WALS).

4. Data

- **Parsing**: a sample of 21 treebanks from from Universal Dependencies v1.4;
- **Neural Machine Translation**: a novel dataset created from the Open Subtitles 2016 corpus for Arabic-Dutch and Indonesian-Portuguese (3M sentences train / 5K test);
- **Sentence Similarity**: Sentence pairs annotated with a label ranging from 0 (dissimilarity) to 5 (equivalence). 9,709 train (in English from the STS benchmark) / 250 test (in Arabic from Task 1 of SemEval 2017).

5. Source Selection: Parsing

We perform delexicalised model transfer for **syntactic parsing** with an SVM (DeSR) and a neural network (Syntaxnet).

For each of the 7 target languages, we choose 3 source languages (highest, middle, and lowest) ranked according to the Jaccard Index.

6a. Task: Neural Machine Translation

We run a syntax-based NMT model in two settings: with and without the tree processing. we use an attentional encoder-decoder network that jointly learns to translate and align words, enriched with linguistic features (including syntax) [2].

<table>
<thead>
<tr>
<th></th>
<th>AR-NL</th>
<th>ID-PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>7.01</td>
<td>14.79</td>
</tr>
<tr>
<td>+ Syntax</td>
<td>14.40</td>
<td>23.70</td>
</tr>
<tr>
<td>+ Preprocessing</td>
<td>15.40</td>
<td>24.12</td>
</tr>
</tbody>
</table>

6b. Task: Sentence Similarity

We classify sentence similarity based on original and processed trees in a lexicalised transfer setting (through multilingual word embeddings).

The two sentences are encoded with a TreeLSTM, then concatenated, and finally fed to a multi-layer perceptron [3].

<table>
<thead>
<tr>
<th></th>
<th>Pearson</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono-lingual</td>
<td>77.9</td>
<td>0.94</td>
</tr>
<tr>
<td>Cross-lingual</td>
<td>44.7</td>
<td>1.82</td>
</tr>
<tr>
<td>+ Preprocessing</td>
<td>48.0</td>
<td>1.64</td>
</tr>
</tbody>
</table>

7. Conclusions

The results demonstrate that reducing anisomorphism leads to enhancements in performance:

- Savvy metrics reliably rank source languages by similarity (better than genealogy).
- Tree processing grants algorithms a better leverage on syntactic information, which is pivotal to several tasks, and make them more robust to cross-lingual variation.

Acknowledgements

This work is supported by the ERC Consolidator Grant LEXICAL (no 648099).

References