Køpsala

Transition-Based Graph Parsing via Efficient Training and Effective Encoding

Hershcovich, Daniel; De Lhoneux, Miryam; Kulmizev, Artur; Pejhan, Elham; Nivre, Joakim

Published in:
Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies

DOI:
10.18653/v1/2020.iwpt-1.25

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Køpsala: Transition-Based Graph Parsing via Efficient Training and Effective Encoding

Daniel Hershcovich♦ Miryam de Lhoneux♦ Artur Kulmizev♥
Elham Pejhan♦ Joakim Nivre♥
♦University of Copenhagen ♥Uppsala University
{dh,ml,ep}@di.ku.dk,
{artur.kulmizev,joakim.nivre}@lingfil.uu.se

Abstract

We present Køpsala, the Copenhagen-Uppsala system for the Enhanced Universal Dependencies Shared Task at IWPT 2020. Our system is a pipeline consisting of off-the-shelf models for everything but enhanced graph parsing, and for the latter, a transition-based graph parser adapted from Che et al. (2019). We train a single enhanced parser model per language, using gold sentence splitting and tokenization for training, and rely only on tokenized surface forms and multilingual BERT for encoding. While a bug introduced just before submission resulted in a severe drop in precision, its post-submission fix would bring us to 4th place in the official ranking, according to average ELAS. Our parser demonstrates that a unified pipeline is effective for both Meaning Representation Parsing and Enhanced Universal Dependencies.

1 Introduction

The IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies (Bouma et al., 2020) involves sentence segmentation, tokenization, lemmatization, part-of-speech tagging, morphological analysis, basic dependency parsing, and finally (for the first time) enhanced dependency parsing. The enhancements encode case information, elided predicates, and shared arguments due to conjunction, control, raising and relative clauses (see Figures 1 and 2).

In Universal Dependencies v2 (UD; Nivre et al., 2020), enhanced dependencies (ED) are a separate dependency graph than the basic dependency tree (BD). However, ED is almost a super-set of BD,¹ and so most previous approaches (Schuster and Manning, 2016; Nivre et al., 2018) have attempted to recover ED from BD using language-specific rules. On the other hand, Hershcovich et al. (2018) experimented with TUPA, a transition-based directed acyclic graph (DAG) parser originally designed for parsing UCCA (Abend and Rappaport, 2013), for supervised ED parsing. They converted ED to UCCA-like graphs and did not use pre-trained contextualized embeddings, yielding sub-optimal results. Taking a similar approach, we adapt a transition-based graph parser (Che et al., 2019) designed for Meaning Representation Parsing (Oepen et al., 2019), but parse ED directly and use BERT embeddings (Devlin et al., 2019).

The main contribution of our work is a transition system supporting the graph structures exhibited by ED, including null nodes (meaning this is not a strictly bilexical formalism), cycles and non-crossing graphs (§3.1), as Figure 4 demonstrates for the sentence from Figure 2. We parse ED completely separately from BD, demonstrating the applicability of a full graph parser, starting from only segmented and tokenized text, to ED. Our code is available at https://github.com/coastalcph/koepsala-parser.

2 Preprocessing

As the focus of this shared task is ED parsing, we rely on existing systems for preprocessing. Here, we consider two off-the-shelf pipelines: STANZA...
(Qi et al., 2020)\textsuperscript{2} and UDPIPE 1.2 (Straka and Straková, 2017; Straka et al., 2016),\textsuperscript{3} both of which have models pre-trained on UD v2.5 treebanks. We experiment with either pipeline during prediction to process the raw text files for the dev and test sets, eventually selecting UDPIPE for our primary submission. This process entails sentence segmentation, tokenization, lemmatization, part-of-speech tagging, morphological feature tagging, and BD parsing.\textsuperscript{4} For training our ED parser (§3), however, we use gold inputs for simplicity. We use the conllu Python package\textsuperscript{5} to read CoNLL-U files.

**Preprocessing model selection.** Since the dev and test data do not denote their source treebanks, we simply process the text using the pipeline model trained on the language’s largest treebank. To experiment with an alternative method, for languages with more than one treebank, we also train UDPIPE models on combined training treebanks. Table 1 shows the comparison of LAS on the combined dev set, for these models and for the models (pre-)trained on the language’s largest treebank. The results show that using the combined training sets does not lead to consistent improvements in terms of LAS, and so we continue using pre-trained treebank-specific preprocessing models henceforth.

<table>
<thead>
<tr>
<th>Language</th>
<th>Czech</th>
<th>Dutch</th>
<th>Estonian</th>
<th>Polish</th>
</tr>
</thead>
<tbody>
<tr>
<td>combined</td>
<td>78.88</td>
<td>76.50</td>
<td>77.01</td>
<td>82.96</td>
</tr>
<tr>
<td>largest</td>
<td>83.97</td>
<td>74.97</td>
<td>77.61</td>
<td>82.59</td>
</tr>
</tbody>
</table>

Table 1: LAS on the combined dev set for UDPIPE models trained on the language’s combined training treebanks and the models trained on the language’s largest treebank. No consistent trend is observed.

### 3 Transition-Based Enhanced Dependency Parser

Our system is a transition-based graph parser, based on the HIT-SCIR system (Che et al., 2019), which achieved the highest average score across frameworks (AMR, EDS, UCCA, DM and PSD) in the CoNLL 2019 shared task on Meaning Representation Parsing (MRP; Oepen et al., 2019). It is written in the AllenNLP (Gardner et al., 2018) framework. For training efficiently, it employs stack LSTMs (Dyer et al., 2015), batching operations across sentences. For better encoding, HIT-SCIR fine-tuned BERT (Devlin et al., 2019) while training the parser.

A transition-based parser operates by manipulating a buffer (originally containing the input words provided by the preprocessor, see §2) and a stack (originally containing the root, i.e., word at index 0), to incrementally create the output dependency graph. At each point in the parsing process, a tran-
sition is selected out of a pre-defined set of possible transitions. A classifier is trained to predict the best transition to apply at each step, by mimicking an oracle during training (see §3.1).

HIT-SCIR used a different transition system per framework (AMR, EDS, UCCA; and one system for DM and PSD), according to the graph properties of each and based on existing framework-specific parsers (Liu et al., 2018; Buys and Blunsom, 2017; Hershcovich et al., 2017; Wang et al., 2018). We construct a transition system for ED using subsets of transitions from two of the HIT-SCIR systems: their system for DM and PSD, as well as their system for UCCA, with some further adaptations specific to ED graphs.

3.1 Transition System

Our system contains the following transitions: \{\texttt{Shift, Left-Edge}_l, \texttt{Right-Edge}_l, \texttt{Reduce-0}, \texttt{Reduce-1}, \texttt{Node, Swap and Finish}\}. The \texttt{Shift} transition pops the first element of the buffer and pushes it onto the stack. The \texttt{Left-Edge}_l and \texttt{Right-Edge}_l transitions add an arc\(^6\) between the two top items of the stack with label \(l\). We need two different \texttt{Reduce} transitions to pop the topmost and second topmost items of the stack, which we name \texttt{Reduce-0} and a \texttt{Reduce-1} respectively. This makes it possible to construct length-2 cycles, which ED allows (and most MRP frameworks do not). The \texttt{Node} transition inserts a null node as the first element of the buffer, needed to support null nodes. \texttt{Swap} moves the second-top node of the stack to the buffer, thus swapping the order between the two top nodes of the stack. This is necessary for handling crossing graphs (analogous to non-projective trees). Finally, \texttt{Finish} terminates the transition sequence. A formal definition of the transition set is shown in Figure 3.

Separate $\texttt{Edge}$ transitions exist for each edge label. Labels containing coordination or case suffixes (such as \texttt{nmod:van}) are treated as any other label and are not split, resulting in a large number of transitions for some languages, shown in Table 2.

\texttt{Node} transitions, on the other hand, do not select any label or features, since null nodes are only evaluated with respect to their incoming and outgoing edges. All other information is ignored, and thus not predicted by the parser: predicted null nodes are thus only placeholders.

\textbf{Constraints.} In addition to the modified transition set, we change the constraints for some transitions according to the required graph structure.

Since \texttt{Left-Edge}_l and \texttt{Right-Edge}_l transitions do not reduce the dependent, we need to ensure that we do not draw the same arc twice. For this reason, these transitions are not allowed if there is already an arc with label \(l\) between the two nodes. We also disallow to add an arc with the root as dependent.

To ensure every node gets attached to at least one head, we disallow the \texttt{Reduce-0} and \texttt{Reduce-1} transitions for nodes that do not have a head yet. We also disallow reducing the root.

For the \texttt{Swap} transition, we maintain the \textit{generated order} of each node, assigned when the node is shifted (for words) or created (for null nodes). To prevent infinite loops during inference, we only allow swapping nodes whose order in the stack is the same as their generation order.

To limit repeated actions, we arbitrarily constrain \texttt{Node} transitions such that there are no more null nodes than words (although a lower limit would suffice), and \texttt{Edge} transitions to limit the

<table>
<thead>
<tr>
<th>Language</th>
<th>Total</th>
<th>$\texttt{Edge}$</th>
<th>w/ suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>402</td>
<td>395</td>
<td>345</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>197</td>
<td>191</td>
<td>137</td>
</tr>
<tr>
<td>Czech</td>
<td>768</td>
<td>761</td>
<td>702</td>
</tr>
<tr>
<td>Dutch</td>
<td>393</td>
<td>386</td>
<td>336</td>
</tr>
<tr>
<td>English</td>
<td>300</td>
<td>293</td>
<td>232</td>
</tr>
<tr>
<td>Estonian</td>
<td>445</td>
<td>438</td>
<td>381</td>
</tr>
<tr>
<td>Finnish</td>
<td>266</td>
<td>259</td>
<td>210</td>
</tr>
<tr>
<td>French</td>
<td>112</td>
<td>106</td>
<td>59</td>
</tr>
<tr>
<td>Italian</td>
<td>281</td>
<td>274</td>
<td>216</td>
</tr>
<tr>
<td>Latvian</td>
<td>238</td>
<td>232</td>
<td>161</td>
</tr>
<tr>
<td>Lithuanian</td>
<td>323</td>
<td>317</td>
<td>263</td>
</tr>
<tr>
<td>Polish</td>
<td>676</td>
<td>669</td>
<td>615</td>
</tr>
<tr>
<td>Russian</td>
<td>944</td>
<td>938</td>
<td>861</td>
</tr>
<tr>
<td>Slovak</td>
<td>266</td>
<td>259</td>
<td>204</td>
</tr>
<tr>
<td>Swedish</td>
<td>209</td>
<td>202</td>
<td>153</td>
</tr>
<tr>
<td>Tamil</td>
<td>146</td>
<td>140</td>
<td>103</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>290</td>
<td>283</td>
<td>225</td>
</tr>
</tbody>
</table>

Table 2: Number of transitions for each language.

\(^6\)For consistency, we keep the transition nomenclature using “\texttt{Edge}”, although they create directed dependency \textit{arcs}. Note that in analogous transitions in some transition systems, such as \texttt{Arc-Eager} (Nivre, 2003), the dependent of the transition is also popped out of the stack as part of either of these two transitions. Here, since dependents can have multiple heads and can have arcs with multiple labels, we stick to the \texttt{Edge} action and use our two \texttt{Reduce} transitions to pop elements of the stack when necessary.
number of heads per node to 7.\footnote{While the observed number of heads per node in the data goes up to 36, in the training data there is only a small minority of cases where a node has more than 7 heads.}

**Finish** is only allowed when the buffer is empty and the stack only contains the root. If no valid transition is available, the sequence is terminated prematurely by applying the **Finish** transition, regardless of the **Finish** constraints.

**Oracle.** We use a static oracle similar to HIT-SCIR (a single “gold” transition sequence is given during training, which the parser is forced to follow), but develop one for our transition system.

The oracle deterministically chooses the transition to take given the current configuration. Let $s_1$ and $s_0$ be the two top items of the stack and $b$ the first item of the buffer (if these are defined in the current configuration). If the buffer is empty and the stack only contains the root, take a **Finish** transition. Otherwise, if there is an arc between $s_1$ and $s_0$ with label $l$ that has not yet been constructed, take the necessary **Right-Edge** or **Left-Edge** action. Otherwise, if $s_0$ has a node dependent, take a **Node** transition. Otherwise, if $s_0$ has all its heads and dependents, take **Reduce-0**, if $s_1$ has all its heads and dependents, take **Reduce-1**. Otherwise, if $s_1$ and $s_0$ are in their generated order and $s_0$ has a head or a dependent in the stack that is not $s_1$, take a **Swap**. Otherwise **Shift**. Figure 4 shows an example transition sequence.

### 3.2 Classifier

The parser uses BERT (Devlin et al., 2019) for token representation. While Che et al. (2019) used pre-trained English model (wwm-cased_L-24_H-102416), we replaced it with a pre-trained multilingual one (multi_cased_L-12_H-76812).\footnote{https://github.com/google-research/bert/blob/master/multilingual.md trained on 104 languages, including all 17 languages participating in the shared task. As done by Che et al. (2019), we use the bert-pretrained text field embedder from AllenNLP, which extracts the first word-piece of each token, applying a scalar mix on all layers of transformer.}

The transition classifier is a stack-LSTM (Dyer et al., 2015) with only BERT embedding features for words, as well as a scalar feature denoting the ratio between the number of (null) nodes and the number of words (Hershcovich et al., 2017), as in HIT-SCIR. We do not fine-tune BERT due to memory limitations, though fine-tuning would likely result in improved performance.

#### 3.3 Postprocessing

The enhanced graphs are required to be connected, i.e., every node must be reachable from the root.\footnote{This is enforced by the task organizers by running validate.py --level 2 --lang ud on the system predictions before evaluation.}

While the transition constraints ensure that every node has a head, there may be unconnected cycles at the end of the parse, resulting in invalid graphs. To fix the problem, at the end of the parse, we iteratively find the unconnected node with the most descendants, and attach it to the predicate (the first dependent of the root) with an orphan-labeled arc. In addition to unconnected cycles, this resolves the problem of prematurely terminated transition sequences due to no valid transition being available according to the constraints: instead of resulting in partially-constructed graphs, headless nodes are similarly attached with an orphan-labeled arc to the predicate, if it exists, or otherwise to the root.

**Parsing tragedy.** Our postprocessing procedure to attach unconnected subgraphs had a bug at the time of submission, where many nodes were incorrectly identified as unconnected and thus un-
necessarily attached to the predicate/root. While this still yielded valid graphs, precision dropped precipitously from before the introduction of the postprocessing procedure. Due to the late stage in the evaluation period at which we made this change, we failed to properly monitor our development scores and could not identify the cause for the drop in time, resulting in low official scores. However, after submission we identified the bug and fixed it, improving our parser’s accuracy back to the range we had observed during development.

3.4 Training

For training the ED parser we do not simply train it on the largest treebank per language, but rather train it on the concatenated training treebanks per language. In preliminary experiments, this did lead to improvements in terms of combined dev ELAS over treebank-specific models, contrary to our findings in BD parsing for preprocessing (§2). We train our models on an NVIDIA P100 GPU with a batch size of 8. All other hyperparameters can be found in the configuration files in the repository.¹¹

Training until convergence took 1h30 (for Tamil, ¹⁰https://github.com/coastalcph/koepsala-parser/blob/master/config/transition_eud.jsonnet
¹¹https://github.com/coastalcph/koepsala-parser/blob/master/config/transition_eud.jsonnet
the smallest treebank) to up to 2 days (for Arabic, which contains many long sentences). Prediction on the dev set took between 4 minutes (for Tamil) and 55 minutes (for Czech), ranging from 117 words/second (7 sentences/second, for Tamil) to 1300 words/second (81 sentences/second, for Czech), including the model loading time.

3.5 Baselines

In addition to providing validation scores for our trained parsers, we consider three competitive baselines, as provided by the task organizers:

- **B1**: gold standard dependency trees copied as enhanced graphs. Though this can be technically considered an upper bound, as gold tree information is provided, it should nonetheless provide some idea of how much of the enhanced graph can be derived from the dependency tree.

- **B2**: predicted trees yielded by UDpipe models trained on UD v2.5 (using the largest treebank where applicable), copied as enhanced graphs. This is more representative than B1 of realistic parsing scenarios, which rely on predictions.

- **B3**: similar to B2, but applying the Stanford enhancer post-hoc over the predicted trees. Scores for Finnish and Latvian were not provided.

4 Results

Table 3 displays our results on the per-language (not per-treebank) test partitions of the shared task data. As explained in §2, for languages with multiple training treebanks available (Czech, Estonian, Dutch, Polish), we preprocessed the raw text of each treebank using the pipeline trained on the largest treebank available for that language (e.g. alpino for Dutch). Also, aforementioned in §3.4, we then trained our parsers on the concatenation of each language’s treebanks, so that we could parse at the language level (as opposed to treebank). Though we observed scant differences between the two preprocessing pipelines, it was UDPIPE that produced fewer validation errors. As such, we adopted it as the main preprocessor for our official submission.

It is apparent in Table 3 that the unconnected graph issue (described in §3.3) severely affected our official submission to the shared task (observed in the official column). After diagnosing and fixing this problem, we observed an improvement of 14.1 ELAS, which is consistent with our scores on the treebanks’ development sets. With this in mind, our (fixed) parser tends to perform in a generally stable fashion across languages, with an average ELAS of 76.48 and standard deviation of 6.86. Among our highest scoring languages are Bulgarian, French, and Italian—the former two of which are corroborated by the strong B1 baseline. Indeed, Tamil is the notable exception among all results, with 56.85 ELAS. We surmise that treebank size is the biggest factor in this degradation of performance, as Tamil has, by far, the smallest treebank at 400 sentences. As such, our parser has comparatively fewer graph samples to train on than it would for some other languages.

When comparing against the organizer-provided baselines, we see a strong improvement in using our system over both B2 and B3 systems. This is encouraging, as it demonstrates the benefit of parsing enhanced dependency graphs directly, rather
than relying on predicted trees to accurately relay the enhanced structure \((B1)\) or employing a heuristic-driven post-processor to derive it \((B2)\). Furthermore, though the organizers consider \(B1\) as an indirect upper bound due to the gold-standard tokenization and dependency structure employed therein, we can nonetheless observe an advantage in using our parser for some languages. These are Arabic (+2.16 ELAS), Finnish (+3.32), Italian (+4.46), and Ukranian (+0.86). Again, this is promising, given that our parser does not rely on any tree structure in order to parse graphs.

### 4.1 Pre-processing Analysis

Since the test data was provided in a raw, untokkenized format, we were interested in the extent of accuracy loss we might observe when relying on off-the-shelf pre-processors. Table 4 displays these results over the development data. It is clear that when we employ predicted segmentation, etc. using either \textsc{stanza} or \textsc{udpipe} pipelines, we observe a slight degradation in accuracy, as compared to the gold data. Omitting Czech, Estonian, Dutch, and Polish (which had several associated treebanks), all other languages degrade by an average of 2.00 ELAS for \textsc{stanza} and 2.32 for \textsc{udpipe}. Though one typically expects such a degradation when evaluating with predicted segmentation, we did observe some unreasonably large gaps in accuracy: namely for Arabic \((-4.02, -8.32\) for \textsc{stanza} and \textsc{udpipe}, respectively) and Tamil \((-12.19, -8.59\). The latter can likely be explained via its small training set, which undoubtedly affects all components of the preprocessing pipeline.

When we examine the scores for all multi-treebank languages, we do not notice a large difference between gold and predicted tokenization—which we expect to be different across treebanks. Here, we simply choose the one trained on the largest treebank \((\text{PicTree for Czech, EDT for Estonian, Alpino for Dutch, and LFG for Polish})\), as we consider this a simple yet reliable heuristic. However, when generating predictions for the smaller treebanks using the bigger treebank’s preprocessing model, we only notice a notable drop in accuracy for Dutch \((-2.15, -2.54\) for \textsc{stanza} and \textsc{udpipe}, respectively). This indicates that there are likely major differences in the treebanks’ domains or how they are respectively segmented or annotated. In general, however, the differences between gold and predicted tokenization is surprisingly not as large as we expected.

### 5 Conclusion

In this paper, we have described the IWPT 2020 Shared Task submission by the Copenhagen-Uppsala team, consisting of graphs predicted by a transition-based neural dependency graph parser with pre-trained multilingual contextualized embeddings. While not ranked among the top submission according to the official scores, the parser architecture proved effective for the type of dependency graphs exhibited by ED, and after fixing a critial bug we found the scores to improve dramatically and agree with the scores we had observed during development.

We expect that with more resources for BERT fine-tuning, hyperparameter tuning, language-specific pre-trained representations and careful pre-and post-processing, our parser will be a competitive system in this task. However, our contribution is a transition system that can directly handle ED, in a unified transition-based parsing system.
Acknowledgments

We thank the anonymous reviewers for their helpful comments. ML is funded by a Google Focused Research Award. We acknowledge the computational resources provided by CSC in Helsinki and Sigma2 in Oslo through NeIC-NLPL (www.nlpl.eu).

References


