Abstract

We present $KL_{cpos}^3$, a language similarity measure based on Kullback-Leibler divergence of coarse part-of-speech tag trigram distributions in tagged corpora. It has been designed for multilingual delexicalized parsing, both for source treebank selection in single-source parser transfer, and for source treebank weighting in multi-source transfer. In the selection task, $KL_{cpos}^3$ identifies the best source treebank in 8 out of 18 cases. In the weighting task, it brings +4.5% UAS absolute, compared to unweighted parse tree combination.

1 Introduction

The approach of delexicalized dependency parser transfer is to train a parser on a treebank for a source language ($src$), using only non-lexical features, most notably part-of-speech (POS) tags, and to apply that parser to POS-tagged sentences of a target language ($tgt$) to obtain dependency parse trees. Delexicalized transfer yields worse results than a supervised lexicalized parser trained on a target language treebank. However, for languages with no treebanks available, it may be useful to obtain at least a lower-quality parse tree for tasks such as information retrieval.

Usually, multiple source treebanks are available, and it is non-trivial to select the best one for a given target language. As a solution, we present a language similarity measure based on KL divergence (Kullback and Leibler, 1951) of distributions of coarse POS tag trigrams in POS-tagged corpora, which we call $KL_{cpos}^3$. The measure has been designed and tuned specifically for multilingual delexicalized parser transfer, and it often succeeds in selecting the best source treebank in a single-source setting, as well as in appropriately weighting the source treebanks by similarity to the target language in a multi-source parse tree combination approach.

2 Related Work

Delexicalized parser transfer was conceived by Zeman and Resnik (2008), who also introduced two important preprocessing steps – mapping treebank-specific POS tagsets to a common set using Interset (Zeman, 2008), and harmonizing treebank annotation styles into a common style, which later developed into the HamleDT harmonized treebank collection (Zeman et al., 2012).

McDonald et al. (2011) applied delexicalized transfer in a setting with multiple source treebanks available, finding that the problem of selecting the best source treebank without access to a target language treebank for evaluation is non-trivial. They combined all source treebanks by concatenating them but noted that this yields worse results than using only the best source treebank.

An alternative is the (monolingual) parse tree combination method of Sagae and Lavie (2006), who apply several independent parsers to the input sentence and combine the resulting parse trees using a maximum spanning tree algorithm. Surdeanu and Manning (2010) enrich tree combination with weighting, assigning each parser a weight based on its Unlabelled Attachment Score (UAS). In our work, we introduce an extension of this method to a crosslingual setting by combining parsers for different languages and using source-target language similarity to weight them.

Several authors (Naseem et al., 2012; Søgaard and Wulff, 2012; Täckström et al., 2013b) employed WALS (Dryer and Haspelmath, 2013) to estimate source-target language similarity for delexicalized transfer, focusing on genealogy distance and word-order features. Søgaard and Wulff (2012) also introduced weighting into the treebank concatenation approach, using a POS n-gram model trained on a target-language corpus...
to weight source sentences in a weighted perceptron learning scenario (Cavallanti et al., 2010). KL divergence (Kullback and Leibler, 1951) of POS tag distributions, as well as several other measures, was used by Plank and Van Noord (2011) to estimate monolingual domain similarity.

As is quite common in parsing papers, including those dealing with semi-supervised and unsupervised parsing, we use gold POS tags in all our experiments. This enables us to evaluate the effectiveness of our parsing method alone, not influenced by errors stemming from the POS tagging. Based on the published results, it seems to be considerably easier to induce POS tags than syntactic structure for under-resourced languages, as there are several high-performance weakly-supervised POS taggers. Das and Petrov (2011) report an average accuracy of 83% using word-aligned texts, compared to 97% reached by a supervised tagger. Täckström et al. (2013a) further improve this to 89% by leveraging Wiktionary. For some languages, there are even less resources available; Agić et al. (2015b) were able to reach accuracies around 70% by using partial or full Bible translation. Our methods could thus be applied even in a more realistic scenario, where gold POS tags are not available for the target text, by using a weakly-supervised POS tagger. We intend to evaluate the performance of our approach in such a setting in future.

3 Delexicalized Parser Transfer

Throughout this work, we use MSTperl (Rosa, 2015b), an implementation of the unlabelled single-best MSTParser of McDonald et al. (2005b), with first-order features and non-projective parsing, trained using 3 iterations of MIRA (Crammer and Singer, 2003).¹

Our delexicalized feature set is based on the set of McDonald et al. (2003a) with lexical features removed. It consists of combinations of signed edge length (distance of head and parent, bucketed for values above 4 and for values above 10) with POS tag of the head, dependent, their neighbours, and all nodes between them.² We use the Universal POS Tagset (UPOS) of Petrov et al. (2012).

3.1 Single-source Delexicalized Transfer

In the single-source parser transfer, the delexicalized parser is trained on a single source treebank, and applied to the target corpus. The problem thus reduces to selecting a source treebank that will lead to a high performance on the target language.

3.2 Multi-source Delexicalized Transfer

In our work, we extend the monolingual parse tree combination method to a multi-source crosslingual delexicalized parser transfer setting:

1. Train a delexicalized parser on each source treebank.
2. Apply each of the parsers to the target sentence, obtaining a set of parse trees.
3. Construct a weighted directed graph as a complete graph over all tokens of the target sentence, where each edge is assigned a score equal to the number of parse trees in which it appears (each parse tree contributes by either 0 or 1 to the edge score). In the weighted variant of the method, the contribution of each parse tree is multiplied by its weight.
4. Find the final dependency parse tree as the maximum spanning tree over the graph, using the algorithm of Chu and Liu (1965) and Edmonds (1967).

4 KL_{cpos}³ Language Similarity

We introduce KL_{cpos}³, a language similarity measure based on distributions of coarse POS tags in source and target POS-tagged corpora. This is motivated by the fact that POS tags constitute a key feature for delexicalized parsing.

The distributions are estimated as frequencies of UPOS trigrams in the treebank training sections:

\[
f(cpos_{i-1}, cpos_i, cpos_{i+1}) = \frac{\text{count}(cpos_{i-1}, cpos_i, cpos_{i+1})}{\sum_{a,b,c} \text{count}(cpos_a, cpos_b, cpos_c)};
\]

we use a special value for cpos_{i-1} or cpos_{i+1} if cpos_i appears at sentence beginning or end.

We then apply the Kullback-Leibler divergence

1¹Note that while our approach does not depend in principle on the actual parser used, our results and conclusions may not be valid for other parsers.

2²The feature set, as well as scripts and configuration files for the presented experiments, are available in (Rosa, 2015a).

³Bigrams and tetragrams performed comparably on the weighting task, but worse on the selection task. Using more fine-grained POS tags led to worse results as fine-grained features tend to be less shared across languages.
\[ D_{KL}(tgt||src) \text{ to compute language similarity:}  \\
KL_{cpos^3}(tgt, src) = \\
\sum_{cpos^3 \in tgt} f_{tgt}(cpos^3) \cdot \log \frac{f_{tgt}(cpos^3)}{f_{src}(cpos^3)}, \tag{2} \]

where \( cpos^3 \) is a coarse POS tag trigram. For the KL divergence to be well-defined, we set the source count of each unseen trigram to 1.

### 4.1 \( KL_{cpos^3} \) for Source Selection

For the single-source parser transfer, we compute \( KL_{cpos^3} \) distance of the target corpus to each of the source treebanks and choose the closest source treebank to use for the transfer.

### 4.2 \( KL_{cpos^3}^{-4} \) for Source Weighting

To convert \( KL_{cpos^3} \) from a negative measure of language similarity to a positive source parser weight for the multi-source tree combination method, we take the fourth power of its inverted value.\(^5\) The parse tree produced by each source parser is then weighted by \( KL_{cpos^3}^{-4}(tgt, src) \).

### 5 Dataset

We carry out our experiments using HamleDT 2.0 of Rosa et al. (2014), a collection of 30 treebanks converted into Universal Stanford Dependencies (de Marneffe et al., 2014), with POS tags converted into UPOS; we use gold-standard POS tags in all experiments. We use the treebank training sections for parser training and language similarity estimation, and the test sections for evaluation.\(^6\)

### 5.1 Tuning

To avoid overfitting the exact definition of \( KL_{cpos^3} \) and \( KL_{cpos^3}^{-4} \) to the 30 treebanks, we used only 12 development treebanks for hyperparameter tuning: ar, bg, ca, el, es, et, fa, fi, hi, hu, it, ja.\(^7\)

Table 1 contains evaluation of several language similarity measures considered in the tuning phase, applied to weighted multi-source transfer and evaluated using average UAS on the development set. We evaluated KL divergences computed in both directions, as well as Jenses-Shannon divergence (Lee, 2001) and cosine similarity. Based on the results, \( KL_{cpos^3}^{-4} \) was selected, as it performed best in all aspects.

Once the hyperparameters were fixed, we applied the parser transfer methods to the full set of 30 treebanks; our final evaluation is based on the results on the 18 test treebanks as targets.

### 5.2 Other datasets

Additionally, we also report preliminary results on the Prague style conversion of HamleDT, which loosely follows the style of the Prague Dependency Treebank of Böhmová et al. (2003), and on the subset of CoNLL 2006 and 2007 shared tasks (Buchholz and Marsi, 2006; Nilsson et al., 2007) that was used by McDonald et al. (2011).\(^8\)

### 6 Evaluation

#### 6.1 Results

Table 2 contains the results of our methods both on the test languages and the development languages.

---

\(^4\)The KL divergence is non-symmetric; \( D_{KL}(P||Q) \) expresses the amount of information lost when a distribution \( Q \) is used to approximate the true distribution \( P \). Thus, in our setting, we use \( D_{KL}(tgt||src) \), as we try to minimize the loss of using a \( src \) parser as an approximation of a \( tgt \) parser.

\(^5\)A high value of the exponent strongly promotes the most similar source language, giving minimal power to the other languages, which is good if there is a very similar source language. A low value enables combining information from a larger number of source languages. We chose a compromise value of 4 based on performance on the development data.

\(^6\)Contrary to the motivation, we do not evaluate our method on truly underresourced languages, since automatic intrinsic evaluation is not possible on languages without treebanks. Still, e.g., Bengali and Telugu can be considered low-resourced, since their treebanks are very small.

\(^7\)We tuned the choice of the similarity measure, POS n-gram length, and the way of turning \( KL_{cpos^3} \) into \( KL_{cpos^3}^{-4} \). To tune our method to perform well in many different situations, we chose the development set to contain both smaller and larger treebanks, a pair of very close languages (ca, es), a very solitary language (ja), multiple members of several language families (Uralic, Romance), and both primarily left-branching (bg, el) and right-branching (ar, ja) languages.

\(^8\)The CoNLL subset is: da, de, el, en, es, it, nl, pt, sv.
For each target language, we used all remaining 29 source languages for training (in the single-source method, only one of them is selected and applied).

Our baseline is the treebank concatenation method of McDonald et al. (2011), i.e., a single delexicalized parser trained on the concatenation of the 29 source treebanks.

As an upper bound,9 we report the results of the oracle single-source delexicalized transfer: for each target language, the oracle source parser is the one that achieves the highest UAS on the target treebank test section.10 For space reasons, we do not include results of a higher upper bound of a supervised delexicalized parser (trained on the target treebank), which has an average UAS of 68.5%. It was not surpassed by our methods for any target language, although it was reached for Telugu, and approached within 5% for Czech and Latin.

### 6.2 Discussion

The results show that $KL_{cpos}$ performs well both in the selection task and in the weighting task, as both the single-source and the weighted multi-source transfer methods outperform the unweighted tree combination on average, as well as the treebank concatenation baseline. In 8 of 18 cases, $KL_{cpos}$ is able to correctly identify the oracle source treebank for the single-source approach. In two of these cases, weighted tree combination further improves upon the result of the single-source transfer, i.e., surpasses the oracle; in the remaining 6 cases, it performs identically to the single-source method. This proves $KL_{cpos}$ to be a successful language similarity measure for delexicalized parser transfer, and the weighted multi-source transfer to be a better performing approach than the single-source transfer.

The weighted tree combination is better than its unweighted variant only for half of the target languages, but it is more stable, as indicated by its lower standard deviation, and achieves an average UAS higher by 4.5% absolute. The unweighted tree combination, as well as the treebank concatenation, perform especially poorly for English, German, Tamil, and Turkish, which are rich in determiners, unlike the rest of the treebanks;11 there-

<table>
<thead>
<tr>
<th>Tgt lang</th>
<th>TB conc</th>
<th>Oracle del trans</th>
<th>Single-src KL</th>
<th>Multi-src KL</th>
<th>$w_1$</th>
<th>$w_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bn</td>
<td>61.9</td>
<td>te 66.7</td>
<td>66.7</td>
<td>63.2</td>
<td>66.7</td>
<td></td>
</tr>
<tr>
<td>cs</td>
<td>60.5</td>
<td>sk 65.8</td>
<td>65.8</td>
<td>60.4</td>
<td>65.8</td>
<td></td>
</tr>
<tr>
<td>da</td>
<td>56.2</td>
<td>en 55.4</td>
<td>54.4</td>
<td>53.9</td>
<td>54.0</td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>12.6</td>
<td>en 55.8</td>
<td>57.6</td>
<td>55.8</td>
<td>57.6</td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>12.3</td>
<td>de 42.6</td>
<td>41.6</td>
<td>41.6</td>
<td>41.6</td>
<td></td>
</tr>
<tr>
<td>eu</td>
<td>41.2</td>
<td>da 42.1</td>
<td>40.8</td>
<td>40.8</td>
<td>40.8</td>
<td></td>
</tr>
<tr>
<td>grc</td>
<td>43.2</td>
<td>et 42.2</td>
<td>44.7</td>
<td>44.7</td>
<td>44.7</td>
<td></td>
</tr>
<tr>
<td>la</td>
<td>38.1</td>
<td>grc 40.3</td>
<td>40.3</td>
<td>40.3</td>
<td>40.3</td>
<td></td>
</tr>
<tr>
<td>nl</td>
<td>55.0</td>
<td>da 57.9</td>
<td>56.2</td>
<td>56.2</td>
<td>56.2</td>
<td></td>
</tr>
<tr>
<td>pt</td>
<td>62.8</td>
<td>en 64.2</td>
<td>62.7</td>
<td>62.7</td>
<td>62.7</td>
<td></td>
</tr>
<tr>
<td>ro</td>
<td>44.2</td>
<td>it 66.4</td>
<td>51.2</td>
<td>51.2</td>
<td>51.2</td>
<td></td>
</tr>
<tr>
<td>ru</td>
<td>55.5</td>
<td>sk 57.7</td>
<td>58.7</td>
<td>58.7</td>
<td>58.7</td>
<td></td>
</tr>
<tr>
<td>sl</td>
<td>32.2</td>
<td>sl 61.7</td>
<td>59.6</td>
<td>59.6</td>
<td>59.6</td>
<td></td>
</tr>
<tr>
<td>sv</td>
<td>45.9</td>
<td>sk 53.9</td>
<td>57.3</td>
<td>57.3</td>
<td>57.3</td>
<td></td>
</tr>
<tr>
<td>tr</td>
<td>67.5</td>
<td>bn 77.4</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td></td>
</tr>
<tr>
<td>te</td>
<td>18.8</td>
<td>te 40.3</td>
<td>41.1</td>
<td>41.1</td>
<td>41.1</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>44.5</td>
<td>-</td>
<td>48.0</td>
<td>48.0</td>
<td>48.0</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>16.9</td>
<td>-</td>
<td>13.1</td>
<td>13.1</td>
<td>13.1</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Evaluation using UAS on test target treebanks (upper part of the table) and development target treebanks (lower part).

For each target language, all 29 remaining non-target treebanks were used for training the parsers. The best score among our transfer methods is marked in bold; the baseline single-source method of McDonald et al. (2011), i.e., a single source language with lowest $KL_{cpos}$ distance to the target language (language bold if identical to oracle).

Legend:
- **Tgt lang** = Target treebank language.
- **TB conc** = Treebank concatenation.
- **Oracle del trans** = Single-source delexicalized transfer using the oracle source language.
- **Single-src** = Single-source delexicalized transfer using source language with lowest $KL_{cpos}$ distance to the target language (language bold if identical to oracle).
- **Multi-src** = Multi-source delexicalized transfer, unweighted ($\times 1$) and $KL_{cpos}$ weighted ($\times w$).
- **Test, Dev, All, SD** = Average on test/development/all, and its standard sample deviation.
- **PRG, CoNLL** = Preliminary results (average UAS) on Prague conversion of HamleDT, and on subset of CoNLL used by McDonald et al. (2011).
fore, determiners are parsed rather randomly.\textsuperscript{12} In the weighted methods, this is not the case any-
more, as for a determiner-rich target language, determiner-rich source languages are given a high weight.

For target languages for which $KL_{cpos^3}$ of the closest source language was lower or equal to its average value of 0.7, the oracle treebank was ident-
ified in 7 cases out of 12 and a different but com-
petitive one in 2 cases; when higher than 0.7, an appropriate treebank was only chosen in 1 case out of 6. When $KL_{cpos^3}$ failed to identify the oracle, weighted tree combination was always better or equal to single-source transfer but mostly worse than unweighted tree combination. This shows that for distant languages, $KL_{cpos^3}$ does not perform as good as for close languages.

We believe that taking multiple characteristics of the languages into account would improve the results on distant languages. A good approach might be to use an empirical measure, such as $KL_{cpos^3}$, combined with supervised information from other sources, such as WALS. Alternatively, a backoff approach, i.e. combining $KL_{cpos^3}$ with e.g. $KL_{cpos^2}$, might help to tackle the issue.

Still, for target languages dissimilar to any source language, a better similarity measure will not help much, as even the oracle results are usually poor. More fine-grained resource combination methods are probably needed there, such as selectively ignoring word order, or using different sets of weights based on POS of the dependent node.

6.3 Evaluation on Other Datasets

In (Rosa, 2015c), we show that the accuracies ob-
tained when parsing HamleDT treebanks in the Universal Stanford Dependencies annotation style are significantly lower than when using the Prague style. Preliminary experiments using the Prague style conversion of HamleDT generally show our methods to be effective even on that dataset, although the performance of $KL_{cpos^3}$ is lower in source selection – it achieves lower UAS than un-
weighted tree combination, and only identifies the oracle source treebank in 30% cases. This may be due to us having used only the Stanfordized tree-
banks for tuning the exact definition of the mea-
sure.

\textsuperscript{12}UAS of determiner attachment tends to be lower than 5\%, which is several times less than for any other POS.

Preliminary trials on the subset of CoNLL used by McDonald et al. (2011) indicated that our methods do not perform well on this dataset. The best results by far are achieved by the unweighted combina-
tion, i.e., it is best not to use $KL_{cpos^3}$ at all on this dataset. We believe this to be a deficiency of the dataset rather than of our methods – it is rather small, and there is low diversity in the languages involved, most of them being either Germanic or Romanic. The HamleDT dataset is larger and more diverse, and we believe it to correspond better to the real-life motivation for our methods, thus providing a more trustworthy evaluation.

In the near future, we intend to reevaluate our methods using the Universal Dependencies tree-
bank collection (Nivre et al., 2015; Agić et al., 2015a), which currently contains 18 languages of various types and seems to be steadily growing. A potential benefit of this collection is the fact that the annotation style harmonization seems to be done with more care and in a more principled way than in HamleDT, presumably leading to a higher quality of the dataset.

7 Conclusion

We presented $KL_{cpos^3}$, an efficient language sim-
ilarity measure designed for delexicalized depen-
dency parser transfer. We evaluated it on a large set of treebanks, and showed that it performs well in selecting the source treebank for single-source transfer, as well as in weighting the source tree-
banks in multi-source parse tree combination.

Our method achieves good results when applied to similar languages, but its performance drops for distant languages. In future, we plan to explore combinations of $KL_{cpos^3}$ with other language sim-
ilarity measures, so that similarity of distant lan-
guages is estimated more reliably.

In this work, we only used the unlabelled first-
order MSTParser. We intend to also employ other parsers in future, possibly in combination, and in a labelled as well as unlabelled setting.

Acknowledgments

This research was supported by the grants GAUK 1572314, SVV 260 224, and FP7-ICT-2013-10-
610516 (QTLeeP). This work has been using lan-
guage resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Re-
public (project LM2010013).
References


Željko Agić, Dirk Hovy, and Anders Søgaard. 2015b. If all you have is a bit of the bible: Learning POS taggers for truly low-resource languages. In The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015). Hrvatska znanstvena bibliografija i MZOS-Svibor.


