Frustratingly Easy Cross-Lingual Transfer for Transition-Based Dependency Parsing

Ophélie Lacroix[1], **Lauriane Aufrant**[1,2], **Guillaume Wisniewski**[1] and **François Yvon**[1]
[1]LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay
[2]DGA, 60 boulevard du Général Martial Valin, F-75509 Paris
{ophelie.lacroix, lauriane.aufrant, guillaume.wisniewski, francois.yvon}@limsi.fr

Abstract

In this paper, we present a straightforward strategy for transferring dependency parsers across languages. The proposed method learns a parser from partially annotated data obtained through the projection of annotations across unambiguous word alignments. It does not rely on any modeling of the reliability of dependency and/or alignment links and is therefore easy to implement and parameter free. Experiments on six languages show that our method is at par with recent algorithmically demanding methods, at a much cheaper computational cost. It can thus serve as a fair baseline for transferring dependencies across languages with the use of parallel corpora.

1 Introduction

Cross-lingual learning techniques enable to transfer useful supervision information from well-resourced to under-resourced languages, helping the development of NLP tools for a large number of languages. In this work, we present a simple method for transferring dependency parsers between languages.

Two main strategies have been considered to transfer syntactic annotations: (a) direct model transfer and (b) annotation transfer. The first approach assumes a common representation between the source and target languages (e.g. at the level of PoS tags), which enables to train a model on source data and to use it to parse target sentences. The performance of 'pure' delexicalized dependency transfer can be significantly improved using additional techniques such as self-training (Zeman and Resnik, 2008), smart data selection (Søgaard, 2011), relexicalization and/or multi-source model transfer (Cohen et al., 2011; Naseem et al., 2012; Täckström et al., 2013). The second approach (transfer of annotations) requires parallel sentences, in which word alignments are used to infer target syntactic structures from source dependencies. The main difficulty here is to cope with cases of non-isomorphism between the source and target structures as well as with the noise in source annotations and in alignments. Turning source trees into target trees indeed may require to filter poor alignments and to apply various heuristic transformation rules, such as the ones introduced in Hwa et al. (2005), later improved in Tiedemann (2014).

In this study, we consider a simple, yet effective approach to transfer annotations, which entirely dispenses from the transfer rules of Hwa et al. (2005), the sharp filtering of partially annotated trees (Tiedemann, 2014), the inclusion of fake root dependencies for unattached words (Spreyer and Kuhn, 2009), or the multi-step process of Rasooli and Collins (2015). Our proposal is, in fact, quite as straightforward (apart from the use of parallel texts) as the delexicalized transfer method of McDonald et al. (2013) while achieving performances that surpass this state-of-the-art method by a wide margin, and competing with recent algorithmically costly methods: it globally outperforms the scores of (Ma and Xia, 2014) and even achieves the same performance as (2015) for 1 language out of 5. It can thus be used as a fair and simple baseline when evaluating new transfer methodologies.

Proceedings of NAACL-HLT 2016, pages 1058–1063,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

Our method relies on the observation (Section 2) that transition-based dependency parsers using the dynamic oracle strategy can be trained from partially annotated trees (in which some words may not have a governor) *using exactly the same algorithm that is used to train from fully annotated tree*. As explained in Section 3, this observation allows us to design a simple transfer strategy that, first, (partially) projects syntactic annotations from a source language onto a target language via unambiguous word alignments and, second, learns a dependency parser from these partially annotated target data. We then apply this strategy for six language pairs.

2 Training Dependency Parsers on Partially Annotated Data

2.1 Training with a Dynamic Oracle

We consider a transition-based dependency parser based on the arc-eager algorithm (Nivre, 2003): this parser builds a dependency tree incrementally by performing a sequence of *actions*. At each step of the parsing process, a classifier scores each possible action and the highest scoring one is applied.

Training relies on the dynamic oracle of Goldberg and Nivre (2012): for each sentence, a parse tree is built incrementally; at each step, if the predicted action creates an erroneous dependency (or, equivalently, prevents the creation of a gold dependency), a weight vector is updated, according to the perceptron rule. The set of all 'correct' actions is built considering the (potentially wrong) predicted tree and the gold action is defined as the correct action with the highest model score.

It is crucial to notice that the training algorithm is an error-correction learning procedure that solely depends on its ability to detect when an action choice will result in an error: when no error is detected, the construction of the parse tree continues according to the model prediction. Consequently, this training procedure can also be used, unchanged, to train a dependency parser from partially annotated data: when no supervision information is available (no reference dependency is known), all actions are considered as correct; in this case, the predicted action is one of the correct actions, the weight vector is not updated, and the training process goes on.

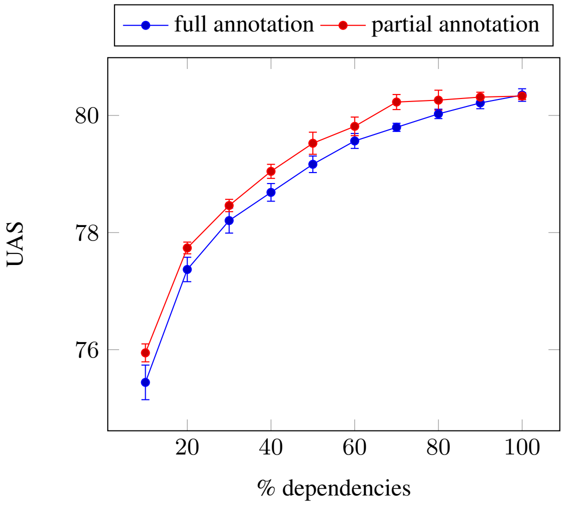

Figure 1: UAS achieved by a parser trained on $n\%$ of the dependencies on German.

This observation can be readily generalized to dependency parsers using a beam search procedure.[1] For the experiments in Section 3, we use a beam-search version of the parser trained with an early-update strategy (Collins and Roark, 2004).

2.2 Experiments on Artificial Datasets

We first carry out a control experiment on datasets in which dependencies have been artificially removed to show that learning from partially annotated data is possible. We compare the performance achieved by a parser trained on $n\%$ of the sentences of the train set with the performance of a parser trained on the whole train set, but in which only $n\%$ of the dependencies of each sentence are known. In both conditions, the total number of dependencies considered during training is roughly the same. Figure 1 plots the parsing performance for German, evaluated by the UAS, with respect to the percentage of dependencies that were kept. To avoid any bias, the reported scores have been averaged over 10 runs. Similar results are observed for 5 other languages of the Universal Dependency Treebank[2] (UDT) (McDonald et al., 2013).

Overall, these results show that learning a parser from partially annotated data is possible. Two other

[1] See (Aufrant and Wisniewski, 2016) for a detailed explanation.

[2] See Section 3.2 for more details on datasets.

conclusions can also be drawn. First, it appears that the number of training examples can be reduced without significantly hurting the performance: removing half the training sentences only reduces the UAS by 1.2 absolute. Second, for a similar number of annotations (i.e. number of dependencies known), better results are achieved when more sentences are annotated, even if this annotation is only partial: in Figure 1, the UAS of a parser trained on partially annotated sentences is higher than the UAS of a parser trained from a subset of the training set.

Indeed, in a partial structure, information on unknown dependencies can be inferred from neighbouring dependencies because of the projectivity constraints. Therefore, the set of gold actions is sometimes smaller than the set of possible actions and an update can happen even if the dependency is unknown. For instance, when training a German dependency parser, 35,382 updates are performed when only 60% of the dependencies are known, to be compared with the 31,339 updates that take place when training on 60% of the fully annotated sentences.

3 Application to Dependency Transfer

In this section, we show how learning from partially annotated data can be used for cross-lingual dependency transfer. A partial projection strategy is first applied to infer partially annotated data for a target language from a full-parsed source data. The target annotations are then used to learn an effective parsing model for the target language.

3.1 Partial Projection of Dependencies

Using sentence-aligned bitexts associating an automatically parsed text in a resource-rich language with its translation in target language, dependencies can readily be projected via alignment links, yielding 'cheap', albeit noisy, supervision data. The main difficulties with the projection arise with many-to-many links and un-aligned tokens. Hwa et al. (2005) have proposed several specific heuristics to deal with the different kinds of alignments and project a full dependency tree. However, this solution comes at the expense of deleting words or creating fake dependencies in the target sentence, which may introduce unreliable annotations in the target data.

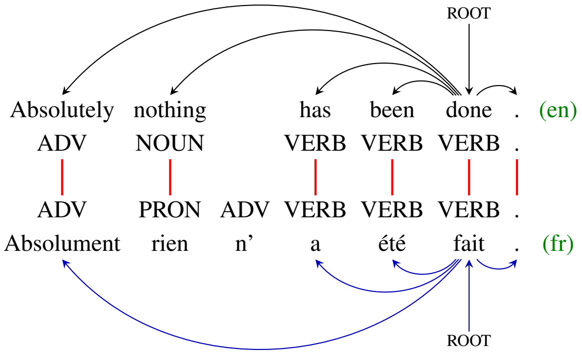

Figure 2: Partial dependency projection from English to French. Only English dependencies compatible with `1:1` alignments, and for which the POS of the aligned words are consistent, are transferred to French.

In this work, we advocate another approach and show that it is simpler and more effective to ignore unattached words and many-to-many alignments: we claim that training a parser from a corpus of high-quality annotated (albeit partially) data will result in better parsing performances than a parser trained from fully-annotated but noisy data.

In practice, parallel sentences are aligned in both directions with `Giza++` (Och and Ney, 2003) and these alignments are merged with the intersection heuristic. This heuristic only selects `1:1` alignment links that occur in the two directional alignments and, intuitively, contains only reliable alignment points, as they have been predicted by two independent models. Note that we do not try to model the reliability of dependency and/or alignment links, making our approach easy to implement and parameter free.

We additionally consider three simple heuristics to filter the transferred annotations and improve their precision: we first remove from the training set target sentences containing non-projective dependencies,[3] as well as sentences for which less than 80% of the words are attached. The latter case indeed corresponds to parallel sentences with few alignment links that are often not perfect translation of each other. Finally, following Rasooli and Collins (2015), we ignore all alignment links that associate words

[3] As shown in the work of (Mareček, 2011), sentences containing non-projective dependencies often results in low-quality projected dependency structures.

with different PoS tags. As shown in Figure 2, for each pair of aligned sentences, only the dependencies for which both the head and the dependent are each aligned to exactly one word (PoS-consistent) are projected.[4]

This approach finally produces an automatically annotated corpus for the target language that contains mostly accurate annotations, even if the dependency structure is incomplete.

3.2 Datasets and Experimental Setup

All our experiments are carried out on six languages[5] of the Universal Dependency Treebank Project: German, English, Spanish, French, Italian and Swedish. We considered as parallel corpora a subset of the Europarl corpus (Koehn, 2005) that have exactly the same English sentences, collecting $1,231,216$ parallel sentences for the 6 language pairs.

For training the target partial data, we used our own implementation of the arc-eager dependency parser with a dynamic oracle, using the features described in (Zhang and Nivre, 2011), with a beam size of 8. The beam-search strategy is used for training (20 iterations) and decoding.

3.3 Dependency Transfer Experiments

For each language pair, the source dataset (Europarl) is PoS-tagged and parsed using the transition-based version of the MateParser (Bohnet and Nivre, 2012), trained on the UDT corpus with a beam size of 40.[6] Dependencies are then (partially) projected onto the target side of the corpus and filtered using the method described above. As reported in Table 2, after filtering, the number of sentences in the train set varies between $15,191$ for German and $52,554$ for Swedish and the percentage of tokens receiving a dependency varies from 88.15% for French to 90.84% for German.

Our parser is then trained on the resulting partially annotated dataset and its performance evaluated on the target UDT test set by the Unlabeled Attachment Score, UAS (excluding punctuation). Gold PoS were used for evaluating in order to make results of our method comparable with state-of-art methods.

		# sentences		
source		**en**		**multi**
filter		100%	80%	80%
target	**de**	7,346	15,191	70,905
	es	9,293	27,700	178,147
	fr	6,626	21,381	144,755
	it	7,353	21,204	160,864
	sv	20,550	52,554	175,201

Table 2: Number of sentences in projected and filtered target data.

The proposed method is compared to three transfer method baselines: the relexicalisation procedure of McDonald et al. (2011), the method of Ma and Xia (2014) for transferring cross-lingual knowledge using entropy regularization, and the recent density-driven approach of Rasooli and Collins (2015) exploiting partially annotated data. The results are first compared for cross-lingual transfer from English and second, applying a voting method[7] for transferring from multiple sources. Note, however, that a direct comparison with these results is not completely fair as systems were not trained with the same exact conditions (less features, lower beam size, etc). As a baseline for comparing parsers, we also report the scores achieved by Rasooli and Collins (2015) and by our method on fully projected sentences ('en-100%').

3.4 Results

Table 1 reports the results of the various transfer methods. Our method achieves significantly better results than the relexicalisation procedure of McDonald et al. (2011) (up to +8.33 in Spanish) and outperforms the method of Ma and Xia (2014) for 3 languages (from +0.91 (fr) to +2.86 (sv)) and equalizes it for one (it). Finally, for Swedish, it achieves performance that are on a par with that of Rasooli

[4] To account for the root dependency, we consider that both the source and target sentences contain an additional ROOT token that is always aligned.

[5] These are the languages that are both in Europarl and UDT.

[6] Here are the supervised scores obtained with the MateParser (predicted PoS-tags) on the source languages: 92.4 (en), 80.4 (de), 83.1 (es), 83.8 (fr), 84.2 (it) and 85.7 (sv).

[7] The voting method chooses, for each token of a sentence, the most frequent head among the projected heads from the various source languages if it does not impede the projectivity of the resulted tree (otherwise the next most frequent head is chosen). The most frequent "head" may be null. Finally, the sentence may be partially annotated.

		M11	MX14		RC15			this work		sup.
source		(en)	(en)	(en)	(en-100%)	(multi)	(en)	(en-100%)	(multi)	
target	de	69.77	74.30	74.32	70.56	79.68	73.40	69.36	75.99	84.43
	es	68.72	75.53	78.17	75.69	80.86	77.05	73.98	78.94	85.51
	fr	73.13	76.53	79.91	77.03	82.72	77.44	75.89	80.80	85.81
	it	70.74	77.74	79.46	77.35	83.67	77.74	75.50	79.39	86.97
	sv	75.87	79.27	82.11	78.68	84.06	82.13	77.26	82.97	87.89

Table 1: Parsing quality (evaluated in UAS) of our method and previous works: M11 stands for McDonald et al. (2011), MX14 for Ma and Xia (2014), RC15 for Rasooli and Collins (2015) and 'sup' corresponds to the supervised scores. State-of-the-art scores are from (Rasooli and Collins, 2015).

and Collins (2015).

It therefore appears that, while being much simpler, the proposed approach achieves results very competitive with state-of-the-art methods at a much cheaper computational cost: our results have been obtained by training a single parser with a beam size of 8, while Ma and Xia (2014) use a parser with exact inference, the training and inference complexity of which is $\mathcal{O}(n^4)$ and the method of Rasooli and Collins (2015) requires the costly training of 4 different parsers each using a beam size of 64.

Results of Table 1 also show that Rasooli and Collins (2015) achieves better scores than our method when training on fully projected trees. This can be explained by the differing training conditions (as previously mentioned).[8] Finally, these results show the benefits of considering partial dependency trees and not only sentences for which a complete parse tree is transferred: a parser trained with partial dependencies improves the UAS up to 4.8 points.

4 Conclusion

In this paper, we have proposed and evaluated a very simple procedure to train a dependency parser with projected partial annotations. In fact, our training algorithm is virtually unchanged with respect to the fully supervised case. Yet, it has proved extremely effective when combined with an appropriate selection of the transferred annotations.[9]

Further improvements could be obtained using additional tricks, such as better data selection strategies or constrained parsing. Besides that, it is worth noting that our method is not only on a par with the method of Rasooli and Collins (2015) but could also be combined with it. Indeed, their first step can be substituted by our method. Since the latter outperforms the former, the combination of the two should improve their best final scores.

In our future work, we intend to study how this training strategy behaves for other transition-based systems or, more generally, for other NLP scenarios using partially annotated data.

Acknowledgments

This work has been partly funded by a DGA-RAPID project under grant agreement N.°1429060465 (Papyrus) and the French *Direction générale de l'armement*. We thank the reviewers for their accurate comments and suggestions.

References

Lauriane Aufrant and Guillaume Wisniewski. 2016. PanParser: a Modular Implementation for Efficient Transition-Based Dependency Parsing. Technical report, LIMSI, March.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea, July.

Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised Structure Prediction with Non-Parallel Multilingual Guidance. In *Proceedings of EMNLP 2011, the Conference on Empirical Methods in Natural Language Processing*, pages 50–61, Edinburgh, Scotland, UK., July.

[8] Indeed, the supervised scores achieved by Rasooli and Collins (2015) are 1.03 higher than ours.

[9] The external parameters used for filtering and training were selected according to the results of several experiments. The impact of these parameters are examined in Lacroix et al. (2016).

Michael Collins and Brian Roark. 2004. Incremental Parsing with the Perceptron Algorithm. In *Proceedings of ACL 2004, the 42nd Annual Meeting on Association for Computational Linguistics*, page 111. Association for Computational Linguistics.

Yoav Goldberg and Joakim Nivre. 2012. A Dynamic Oracle for Arc-Eager Dependency Parsing. In *Proceedings of COLING 2012, the International Conference on Computational Linguistics*, pages 959–976, Bombay, India.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection accross Parallel Texts. *Natural language engineering*, 11:311–325.

Philipp Koehn. 2005. Europarl: A parallel corpus for Statistical Machine Translation. In *2nd Workshop on EBMT of MT-Summit X*, pages 79–86, Phuket, Thailand.

Ophélie Lacroix, Guillaume Wisnewski, and François Yvon. 2016. Cross-lingual Dependency Transfer: What Matters? Assessing the Impact of Pre- and Post-processing. In *Proceedings of the NAACL-16 Workshop on Multilingual and Crosslingual Methods in NLP*, MLCL 2016, San Diego, CA, USA. Association for Computational Linguistics.

Xuezhe Ma and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1337–1348, Baltimore, Maryland, June.

David Mareček. 2011. Combining Diverse Word-Alignment Symmetrizations Improves Dependency Tree Projection. In *Computational Linguistics and Intelligent Text Processing*, pages 144–154. Springer.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source Transfer of Delexicalized Dependency Parsers. In *Proceedings of EMNLP 2011, the Conference on Empirical Methods in Natural Language Processing*, pages 62–72.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL 2013, the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 629–637.

Joakim Nivre. 2003. An Efficient Algorithm for Projective Dependency Parsing. In *Proceedings of IWPT 2003, the 8th International Workshop on Parsing Technologies*, Nancy, France.

Franz Joseph Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51.

Mohammad Sadegh Rasooli and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338, Lisbon, Portugal, September. Association for Computational Linguistics.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of ACL 2011, the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 682–686, Portland, Oregon, USA, June.

Kathrin Spreyer and Jonas Kuhn. 2009. Data-Driven Dependency Parsing of New Languages Using Incomplete and Noisy Training Data. In *Proceedings of CoNLL 2009, the Thirteenth Conference on Computational Natural Language Learning*, pages 12–20, Boulder, Colorado, June.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target Language Adaptation of Discriminative Transfer Parsers. In *Proceedings of ACL 2013, the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, Atlanta, Georgia.

Jörg Tiedemann. 2014. Rediscovering Annotation Projection for Cross-Lingual Parser Induction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1854–1864, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42, Hyderabad, India, January. Asian Federation of Natural Language Processing.

Yue Zhang and Joakim Nivre. 2011. Transition-based Dependency Parsing with Rich Non-local Features. In *Proceedings of ACL 2011, the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193, Portland, Oregon, USA, June. Association for Computational Linguistics.