Low Resource Dependency Parsing: Cross-lingual Parameter Sharing in a Neural Network Parser

Long Duong,[12] **Trevor Cohn,**[1] **Steven Bird,**[1] and **Paul Cook**[3]
[1]Department of Computing and Information Systems, University of Melbourne
[2]National ICT Australia, Victoria Research Laboratory
[3]Faculty of Computer Science, University of New Brunswick
lduong@student.unimelb.edu.au {t.cohn,sbird}@unimelb.edu.au paul.cook@unb.ca

Abstract

Training a high-accuracy dependency parser requires a large treebank. However, these are costly and time-consuming to build. We propose a learning method that needs less data, based on the observation that there are underlying shared structures across languages. We exploit cues from a different source language in order to guide the learning process. Our model saves at least half of the annotation effort to reach the same accuracy compared with using the purely supervised method.

1 Introduction

Dependency parsing is a crucial component of many natural language processing systems, for tasks such as text classification (Özgür and Güngör, 2010), statistical machine translation (Xu et al., 2009), relation extraction (Bunescu and Mooney, 2005), and question answering (Cui et al., 2005). Supervised approaches to dependency parsing have been successful for languages where relatively large treebanks are available (McDonald et al., 2005). However, for many languages, annotated treebanks are not available. They are costly to create, requiring careful design, testing and subsequent refinement of annotation guidelines, along with assessment and management of annotator quality (Böhmová et al., 2001). The Universal Treebank Annotation Guidelines aim at providing unified annotation for many languages enabling cross-lingual comparison (Nivre et al., 2015). This project provides a starting point for developing a treebank for resource-poor languages. However, a mature parser requires a large treebank for training, and this is still extremely costly to create. Instead, we present a method that exploits shared structure across languages to achieve a more accurate parser. Structural information from the source resource-rich language is incorporated as a prior in the supervised training of a resource-poor target language parser using a small treebank. When compared with a supervised model, the gain is as high as 8.7%[1] on average when trained on just 1,000 tokens. As we add more training data, the gains persist, though they are more modest. Even at 15,000 tokens we observe a 2.9% improvement.

There are two main approaches for building dependency parsers for resource-poor languages: delexicalized parsing and projection (Täckström et al., 2013). The delexicalized approach was proposed by Zeman et al. (2008). A parser is built without any lexical features, and trained on a treebank in a resource-rich source language. It is then applied directly to parse sentences in the target resource-poor languages. Delexicalized parsing relies on the fact that identical part-of-speech (POS) inventories are highly informative of dependency relations, enough to make up for cross-lingual syntactic divergence.

In contrast, projection approaches use parallel data to project source language dependency relations to the target language (Hwa et al., 2005). McDonald et al. (2011) and Ma and Xia (2014) exploit both delexicalized parsing and parallel data. They use parallel data to constrain the model which is usually initialized by the English delexicalized parser.

In summary, existing work generally starts with a delexicalized parser and uses parallel data to improve it. In this paper, we start with a source language parser and refine it with help from dependency annotations instead of parallel data. This choice means our method can be applied in cases where linguists are dependency-annotating small amounts of field data, such as in Karuk, a nearly-extinct language of Northwest California (Garrett et al., 2013).

[1]We use absolute values herein.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 845–850,
Beijing, China, July 26-31, 2015.

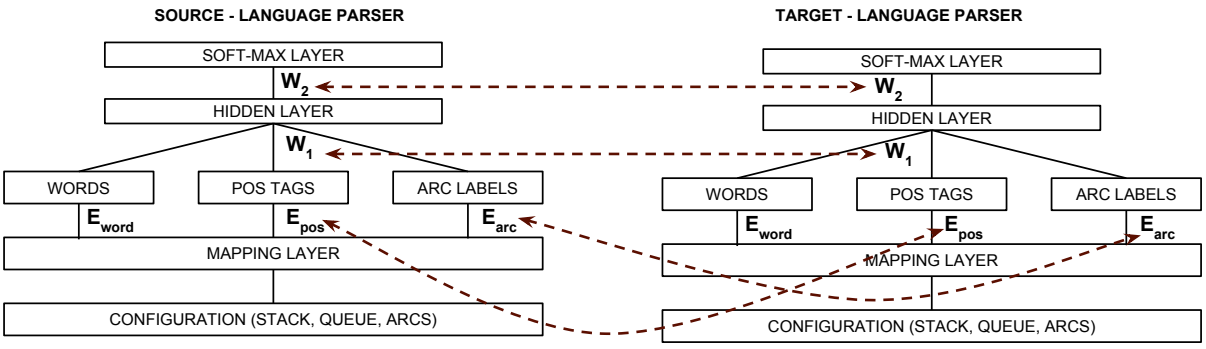

Figure 1: Neural Network Parser Architecture from Chen and Manning (2014) (left). Our model (left and right) with soft parameter sharing between the source and target language shown with dashed lines.

2 Supervised Neural Network Parser

In this section we review the parsing model which we use for both the source language and target language parsers. It is based on the work of Chen and Manning (2014). This parser can take advantage of target language monolingual data through word embeddings, data which is usually available for resource-poor languages. Chen and Manning's parser also achieved state-of-the-art monolingual parsing performance. They built a transition-based dependency parser (Nivre, 2006) using a neural-network. The neural network classifier decides which transition is applied for each configuration.

The architecture of the parser is illustrated in Figure 1 (left), where each layer is fully connected to the layer above. For each configuration, the selected list of words, POS tags and labels from the Stack, Queue and Arcs are extracted. Each word, POS or label is mapped to a low-dimension vector representation (embedding) through the Mapping Layer. This layer simply concatenates the embeddings which are then fed into a two-layer neural network classifier to predict the next parsing action. The set of parameters for the model is $E_{word}, E_{pos}, E_{labels}$ for the mapping layer, W_1 for the cubic hidden layer and W_2 for the softmax output layer.

3 Cross-lingual parser

Our model takes advantage of underlying structure shared between languages. Given the source language parsing structure as in Figure 1 (left), the set of parameters E_{word} will be different for the target language parser shown in Figure 1 (right) but we hypothesize that E_{pos}, E_{arc}, W_1 and W_2 can be shared as indicated with dashed lines. In particular we expect this to be the case when languages use the same POS tagset and arc label sets, as we presume herein. This assumption is motivated by the development of unified annotation for many languages (Nivre et al., 2015; Petrov et al., 2012; McDonald et al., 2013).

To allow parameter sharing between languages we could jointly train the parser on the source and target language simultaneously. However, we leave this for future work. Here we take an alternative approach, namely regularization in a similar vein to Duong et al. (2014). First we train a lexicalized neural network parser on the source resource-rich language (English), as described in Section 2. The learned parameters are $E^{en}_{word}, E^{en}_{pos}, E^{en}_{arc}, W^{en}_1, W^{en}_2$. Second, we incorporate English parameters as a prior for the target language training. This is straightforward when we use the same architecture, such as a neural network parser, for the target language. All we need to do is modify the learning objective function so that it includes the regularization part. However, we don't want to regularize the part related to E^{en}_{word} since it will be very different between source and target language. Letting $W_1 = (W_1^{word}, W_1^{pos}, W_1^{arc})$, the learning objective over training data $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^N$, becomes:[2]

$$\mathcal{L} = \sum_{i=1}^{N} \log P(y^{(i)}|x^{(i)}) - \frac{\lambda_1}{2}\Big[\|W_1^{pos} - W_1^{en:pos}\|_F^2 + \|W_1^{arc} - W_1^{en:arc}\|_F^2 + \|W_2 - W_2^{en}\|_F^2\Big] - \frac{\lambda_2}{2}\Big[\|E_{pos} - E^{en}_{pos}\|_F^2 + \|E_{arc} - E^{en}_{arc}\|_F^2\Big] \quad (1)$$

This is applicable where we use the same POS

[2] All other parameters, i.e. W_1^{word} and E_{word}, are regularized using a zero-mean Gaussian regularization term, with weight $\lambda = 10^{-8}$, as was done in the original paper.

	Train	Dev	Test	Total
cs	1173.3	159.3	173.9	1506.5
de	269.6	12.4	16.6	298.6
en	204.6	25.1	25.1	254.8
es	382.4	41.7	8.5	432.6
fi	162.7	9.2	9.1	181.0
fr	354.7	38.9	7.1	400.7
ga	16.7	3.2	3.8	23.7
hu	20.8	3.0	2.7	26.5
it	194.1	10.5	10.2	214.8
sv	66.6	9.8	20.4	96.8

Table 1: Number of tokens ($\times$ 1,000) for each language in the Universal Dependency Treebank collection.

tagset and arc label annotation for the source and target language. The same POS tagset is required so that the source language parser has similar structure with the target language parser. The requirement of same arc label annotation is mainly needed for evaluation using the Labelled Attachment Score (LAS).[3] We fit two separate regularization sensitivity parameters, λ_1 and λ_2, since they correspond to different parts of the model. λ_1 is used for the shared (universal) part, while λ_2 is used for the language specific parts. Together λ_1 and λ_2 control the contribution of the source language parser towards the target resource-poor model. In the extreme case where λ_1 and λ_2 are large, the target model parameters are tied to the source model, except for the word embeddings E_{word}. In the opposite case, where they are small, the target language parser is similar to the purely supervised model. We expect that the best values fall between these extremes. We use stochastic gradient descent to optimize this objective function with respect to $W_1, W_2, E_{word}, E_{pos}, E_{arc}$.

4 Experiments

In this part we want to see how much our cross-lingual model helps to improve the supervised model, for various data sizes.

4.1 Dataset

We experimented with the Universal Dependency Treebank collection V1.0 (Nivre et al., 2015) which contains treebanks for 10 languages.[4] These treebanks have many desirable properties for our model: the dependency types and coarse POS are the same across languages. This removes the need for mapping the source and target language tagsets to a common tagset. Moreover, the dependency types are also common across languages allowing LAS evaluation. Table 1 shows the dataset size of each language in the collection. Some languages have over $400k$ tokens such as *cs, fr* and *es*, meanwhile, *hu* and *ga* have only around $25k$ tokens.

4.2 Monolingual Word Embeddings

We initialize the target language word embeddings E_{word} of our neural network cross-lingual model with pre-trained embeddings. This is an advantage since we can incorporate monolingual data which is usually available for resource-poor languages. We collect monolingual data for each language from the Machine Translation Workshop (WMT) data,[5] Europarl (Koehn, 2005) and EU Bookshop Corpus (Skadiņš et al., 2014). The size of monolingual data also varies significantly. There are languages such as English and German with more than 400 million words, whereas, Irish only has 4 million. We use the skip-gram model from `word2vec` to induce 50-dimension word embeddings (Mikolov et al., 2013).

4.3 Coarse vs Fine-Grain POS

Our model uses the source language parser as the prior for the target language parser. The requirement is that the source and target should use the same POS tagset. It is clear that information will be lost when using the coarser shared-POS tagset. Here, we simply want to quantify this loss. We run the supervised neural network parser on the coarse-grained Universal POS (UPOS) tagset, and the language-specific fine-grained POS tagset for languages where both are available in the Universal Dependency Treebank.[6] Table 2 shows the average LAS for coarse- and fine-grained POS tagsets with various data sizes. For the smaller dataset, using the coarse-grained POS tagset performed better. Even when we used all the data, the coarse-grained POS tagset still performed reasonably well, approaching the performance obtained using the fine-grained POS tagset. Thus, the choice of the coarse-grained Universal POS tagset

[3] However, same arc-label set also informs some information about the structure.

[4] Czech (cs), German (de), English (en), Spanish (es), Finnish (fi), French (fr), Irish (ga), Hungarian (hu), Italian (it), Swedish (sv)

[5] `http://www.statmt.org/wmt14/`

[6] Czech, English, Finnish, Irish, Italian, and Swedish

Tokens	Coarse UPOS	Fine POS
1k	46.8	42.3
3k	54.3	52.4
5k	56.9	55.8
10k	59.9	59.8
15k	61.5	61.4
All	74.7	75.2

Table 2: Average LAS for supervised learning using the modified version of the Universal POS tagset and the fine-grained POS tagset across various training data sizes.

instead of the original POS tagset is relevant, given that we assume there will only be a small treebank in the target language. Moreover, even when we have a bigger treebank, using the UPOS tagset does not hurt the performance much.[7]

4.4 Tuning regularization sensitivity

As shown in equation 1, λ_1 and λ_2 control the contribution of the source language parser toward the target language parser. We tune these parameters separately using development data. Firstly, we tune λ_1 by fixing $\lambda_2 = 0.1$. The reason for choosing such a large value of 0.1 is that we expect the POS and arc label embeddings to be fairly similar across languages. Figure 2 shows the average LAS for all 9 languages (except English) on different data sizes using different values of λ_1. We observed that $\lambda_1 = 0.001$ gives the optimum value on the development data consistently across different data sizes. We compare the performance at two extreme values of λ_1. For small data size, at 1k tokens, $\lambda_1 = 100$ is better than when $\lambda_1 = 10^{-8}$. This shows that when trained using a small data set, the source language parser is more accurate than the supervised model. However, at 3k tokens, the supervised model is starting to perform better.

We now fix $\lambda_1 = 0.001$ to tune λ_2 in the same range as λ_1. However, the average LAS didn't change much for different values of λ_2. It appears that λ_2 has very little effect on parsing accuracy. This is understandable since λ_2 affects only a small number of parameters (POS and arc embeddings). Thus, we choose $\lambda_1 = 0.001$ and $\lambda_2 = 0.1$ for our experiments.

[7]This is because UPOS generalizes better, and when aggregating with lexical information, it has similar distinguishing power compared with the fine-grained POS tagset.

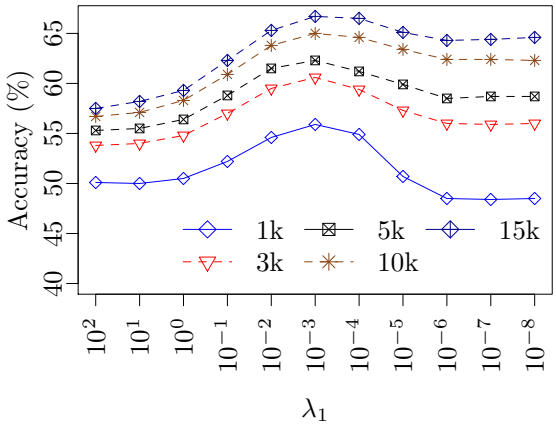

Figure 2: Sensitivity of regularization parameter λ_1 against the average LAS measured on all 9 languages (except English) on the development set for various data sizes (tokens)

4.5 Learning Curve

We choose English as our source language to build different target parsers for each language in the Universal Dependency Treebank collection. We train the supervised neural network parser as mentioned in Section 2 on the Universal Dependency English treebank using UPOS tagset. The UAS and LAS for the English parser is 85.2% and 82.9% respectively, when evaluated on the English test set. We use the English parser as the prior for our cross-lingual model, as described in Section 3. Figure 3 shows the learning curve for both the supervised neural network parser and our cross-lingual model with respect to our implemention of McDonald et al.'s (2011) delexicalized parser, i.e. their basic model which uses no parallel data and no target language supervision. Overall, both the supervised model and the cross-lingual model are much better than this baseline. For small data sizes, our cross-lingual model is superior when compared with the supervised model, giving as much as an 8.7% improvement. This improvement lessens as the size of training data increases. This is to be expected, because the supervised model becomes stronger as the size of training data increases, while the contribution of the source language parser is reduced. However, at 15k tokens we still observed a 2.9% average improvement, demonstrating the robustness of our cross-lingual model. Using our model also reduced the standard deviation ranges on each data point from 12% to 7%.

Using our cross-lingual model can save the annotation effort that is required in order to reach

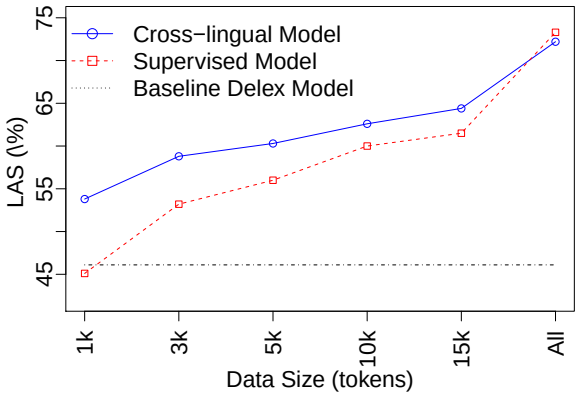

Figure 3: Learning curve for cross-lingual model and supervised model with respect to the baseline delexicalized parser from McDonald et al. (2011): the x-axis is the size of data (number of tokens); the y-axis is the average LAS measured on 9 languages (except English).

the same accuracy compared with the supervised model. For example, we only need 1k tokens in order to surpass the supervised model performance on 3k tokens, and we only need 5k tokens to match the supervised model trained on 10k tokens. The error rate reduction is from 15.8% down to 6.5% for training data sizes from 1k to 15k tokens. However, when we use all the training data, the supervised model is slightly better.

5 Conclusions

Thanks to the availability of the Universal Dependency Treebank, creating a treebank for a target resource-poor language has becoming easier. This fact motivates the work reported here, where we assume that only a tiny treebank is available in the target language. We tried to make the most out of the target language treebank by incorporating a source-language parser as a prior in learning a neural network parser. Our results show that we can achieve a more accurate parser using the same training data. In future work, we would like to investigate joint training on the source and target languages.

Acknowledgments

This work was supported by the University of Melbourne and National ICT Australia (NICTA). Trevor Cohn is the recipient of an Australian Research Council Future Fellowship (project number FT130101105).

References

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2001. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*, pages 103–127. Kluwer Academic Publishers.

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 724–731, Stroudsburg, PA, USA. ACL.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. ACL.

Hang Cui, Renxu Sun, Keya Li, Min-Yen Kan, and Tat-Seng Chua. 2005. Question answering passage retrieval using dependency relations. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 400–407, New York, NY, USA. ACM.

Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014. What can we get from 1000 tokens? a case study of multilingual pos tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 886–897, Doha, Qatar. ACL.

Andrew Garrett, Clare Sandy, Erik Maier, Line Mikkelsen, and Patrick Davidson. 2013. Developing the Karuk Treebank. Fieldwork Forum, Department of Linguistics, UC Berkeley.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11:311–325.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand.

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. Association for Computational Linguistics.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 91–98, Stroudsburg, PA, USA. ACL.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 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 the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. ACL.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C.j.c. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Hajič, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Maria Simi, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal dependencies 1.0.

Joakim Nivre. 2006. *Inductive Dependency Parsing (Text, Speech and Language Technology)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Levent Özgür and Tunga Güngör. 2010. Text classification with the support of pruned dependency patterns. *Pattern Recognition Letters*, 31(12):1598–1607.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, Atlanta, Georgia. ACL.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve smt for subject-object-verb languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253, Boulder, Colorado. ACL.

Daniel Zeman, Univerzita Karlova, and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *In IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42.