Multi-Source Cross-Lingual Constituency Parsing

Hour Kaing[†‡], **Chenchen Ding**[†], **Katsuhito Sudoh**[‡], **Masao Utiyama**[†],
Eiichiro Sumita[†], **Satoshi Nakamura**[‡]
[†]National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
[‡]Nara Institute of Science and Technology,
8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan

Abstract

Pretrained multilingual language models have become a key part of cross-lingual transfer for many natural language processing tasks, even those without bilingual information. This work further investigates the cross-lingual transfer ability of these models for constituency parsing and focuses on multi-source transfer. Addressing structure and label set diversity problems, we propose the integration of typological features into the parsing model and treebank normalization. We trained the model on eight languages with diverse structures and use transfer parsing for an additional six low-resource languages. The experimental results show that the treebank normalization is essential for cross-lingual transfer performance and the typological features introduce further improvement. As a result, our approach improves the baseline F1 of multi-source transfer by 5 on average.

1 Introduction

Recent pretrained multilingual language models have become a key step in cross-lingual transfer for many natural language processing tasks such as name entity recognition, part-of-speech tagging, natural language inference, and dependency parsing (Wu and Dredze, 2019). These models are desirable in research on cross-lingual transfer because bilingual information is not required.

Cross-lingual transfer is when a trained model for a source language is applied to a target (unseen) language. There are two transfer scenarios, single-source and multi-source transfer. For single-source transfer, each time, the model is trained on only one source language. In this scenario, multiple models are available for cross-lingual transfer in practice. Additional model selection is necessary for single-source transfer because cross-lingual transfer relies on language isomorphism. For multi-source transfer, to leverage all existing resources, treebanks of multiple languages are combined to train a multilingual parser that can be later used for any unseen language. In this work, we study the multi-source transfer for sophisticated structure prediction, i.e., constituency parsing. Our work will serve as a benchmark for cross-lingual constituency parsing using pretrained multilingual language model.

For constituency parsing, training a multilingual parser has two main issues that must be considered. First, the source languages can produce diverse word orders—for instance, different *subject-verb-object* or *noun-adjective* orders. These language properties can be simply identified using existing typology databases, e.g., The World Atlas of Language Structures (WALS) or Syntactic Structures of the World's Languages (SSWL). It is intuitive that these language properties can be used to guide a multilingual parser to share corresponding model parameters among similar languages (Naseem et al., 2012; Ammar et al., 2016; Scholivet et al., 2019; Üstün et al., 2020). For cross-lingual transfer, the typological features could hurt performance (Ammar et al., 2016), and an effective integration technique is required (Üstün et al., 2020). Inspired by this, we investigate the usefulness of typological features for cross-lingual constituency parsing and propose a training strategy to generalize the cross-lingual capability of the model using smooth sampling and random dropout.

The second issue is that even though constituency structure is universal, the design of a label set is language specific. For dependency structures, this problem has inspired the creation of the Universal Dependency project (Nivre et al., 2016). The syntactic label sets of constituency structure vary across languages—for instance, very few labels are shared and even labels for the same syntactic category may be different across languages. This increases the complexity of multi-source transfer. Therefore, we propose normalization of the con-

Proceedings of the 18th International Conference on Natural Language Processing, pages 341–346
Silchar, India. December 16 - 19, 2021.

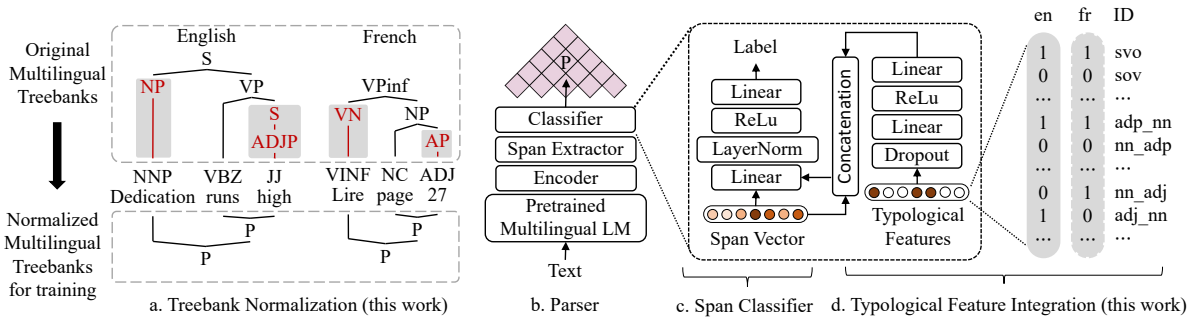

Figure 1: Overall architecture of our parser. The multilingual treebanks are normalized (a) before training the parser (b). A span classifier (c) is also integrated with a feature extractor (d) for binary typological vectors, as shown in the right-most example.

stituency treebanks to universalize the multilingual parsing model.

The contributions of this paper are summarized as follows: 1) typological feature integration for model generalization on unseen languages (Section 4.1), and 2) treebank normalization is proposed to reduce the complexity of cross-lingual structural prediction (Section 4.2).

2 Related Work

In multi-source transfer, task-specific knowledge of multiple source languages is combined and jointly transferred to an unseen or zero-shot language. This combination can be categorized according to three levels (Das and Sarkar, 2020), that is, the level of treebanks (McDonald et al., 2011; Ammar et al., 2016; Scholivet et al., 2019; Üstün et al., 2020), model parameters (Cohen et al., 2011; Søgaard and Wulff, 2012), or parse outputs (Rosa and Žabokrtský, 2015; Agić, 2017). This work focuses on treebank level, that is, treebank concatenation and, unlike previous studies, we study a more sophisticated structure, constituency treebanks, which simultaneously contain diverse syntactic labels across multiple source languages.

Typological features are a valuable resource for multi-source transfer where source languages have diverse structures, and they have been used specifically for sharing the parameters of non-neural (Naseem et al., 2012; Täckström et al., 2013; Zhang and Barzilay, 2015) and neural (Ammar et al., 2016; Scholivet et al., 2019; Üstün et al., 2020) models. Following the same motivation, we also investigate the usefulness of typological features for a multilingual constituent parser and propose a training strategy that generalizes the model for zero-shot languages. Specifically, we integrate typological features into the self-attentive constituency parser (Kitaev and Klein, 2018).

Our work is similar to that of Kitaev et al. (2019) who investigated the multilingualism of the self-attentive constituency parser (Kitaev and Klein, 2018) using the pretrained multilingual language model. However, our work differs from theirs such that we focus on zero-shot performance. In addition, we propose to normalize the concatenated treebanks and integrate typological features for better zero-shot performance. We also extend the sampling technique that Kitaev et al. (2019) use by constraining the minimum size of each treebank.

3 The Self-Attentive Parser

The basis of our model (Fig. 1b) follows the self-attention based encoder–decoder architecture of Kitaev and Klein (2018). Specifically, the encoder consists of word embedding and self-attention layers to produce the contextual presentation for each word. At the decoder side, all possible spans are extracted and each span (i, j) is represented by a hidden vector $v_{i,j}$ that is constructed by subtracting the representations associated with the start and end of the span. Then, each span (i, j) is assigned a labeling score $s(i, j, \cdot)$ by an MLP span classifier as

$$s(i, j, \cdot) = W_2 g(f(W_1 v_{i,j} + c_1)) + c_2, \quad (1)$$

where W_* and c_* are the weight and bias, respectively; f and g are the layer normalization and ReLU ("Re"ctified "L"inear "U"nit) activation function, respectively, as shown in Figure 1c. For each sentence, the constituency structure T is represented by a set of labeled spans $\{(i_t, j_t, l_t) : t = 1, \ldots, |T|\}$ where l is a label. Therefore, the score of T is

$$s(T) = \sum_{(i,j,l) \in T} s(i, j, l). \quad (2)$$

At test time, the optimal structure can be obtained using a CKY-style inference algorithm. For training, the model is optimized using a max-margin objective function, the details of which can be found in Kitaev and Klein (2018). In addition, the parser's hyperparameters are unchanged from Kitaev and Klein (2018).

To perform cross-lingual parsing, an external pre-trained multilingual language model must be used and simply take the place of the word embedding layer. Because the model is trained on sub-words, only the last sub-word unit of the corresponding token is used to represent a word. In this experiment, we use a recent multilingual language model, i.e. XLM-RoBERTa-Large (Conneau et al., 2020).

4 Proposed Methods

4.1 Typological Feature Integration

A typology database is a valuable resource that represents various aspects of languages. Recent `lang2vec` (Littell et al., 2017) provides an interface to represent languages as binary vectors of typological features. Inspired by the recent work of Üstün et al. (2020), we also integrate typological features f (TF) into our model to guide the multilingual model's sharing of the structural knowledge among similar languages, as Figure 1d shows. We use simple feature concatenation to integrate typological features into the span classifier. Like Üstün et al. (2020), we embed binary typological vectors using two linear layers and a ReLU activation function g, and further apply random dropout over the binary typological vectors as

$$f' = M_2 g(M_1 dropout(f) + z_1) + z_2. \quad (3)$$

We then concatenate f' with each span vector, $v_{i,j}$, which modifies Equation 1 as

$$s(i, j, \cdot) = W_2 g(f(W_1[v_{i,j}, f'] + c_1)) + c_2. \quad (4)$$

Dropout is applied directly to the binary features because, during training, typological features only vary with respect to the number of source languages, and each feature is only helpful in the context of other features, which is known as co-adaptation (Hinton et al., 2012). Therefore, for a zero-shot language, without dropout, the model would not be able to extract individual features in a new feature context, which can be prevented using simple random dropout (Hinton et al., 2012). Like Hinton et al. (2012), we drop 50% of the features during training.

The number of multilingual treebanks commonly differs, and high-resource languages tend to be over-represented during training. Similar to the exponential smoothing in Kitaev et al. (2019), at each epoch, we sample d^a examples from each language, where d is the size of each language treebank and a is a hyperparameter. Unlike Kitaev et al. (2019), we use $a = 0.95$ because the size of each treebank is not as large as the unlabeled corpora. We also constrain the smoothed number of examples as $d^a > m$, where m is the smallest treebank size in the source-language pool. We call this approach "smooth sampling."

For the typological features, we combine the syntax features of WALS (Dryer and Haspelmath, 2013) or SSWL (Collins and Kayne, 2011)[1]. We only select the relevant features such as 81A, 82A, 83A, 85A, 87A, 88A, 89A, 90A, 144A, and other unknown ID features such as *subject_b/a_object*[2], *possessor_b/a_noun*, *degree_word_b/a_adjective*, and *subordinator_word_b/a_clause*. In addition, we exclude the morphological features, which contain the word *prefix* or *suffix*, and the missing features of any source language. For zero-shot languages, the missing features are set to zero. After that, we further automatically remove unnecessary features that are repeated for all source languages. Like Ustun et al. (Üstün et al., 2020), we set the hidden and output layer of our TF to 10 and 32, respectively.

4.2 Treebank Normalization

Another obvious issue of constituency treebanks is the difference in their syntactic labels. We observed that high-resource languages tend to have more diverse labels, whereas low-resource languages use a much smaller label set; for instance, Myanmar and Khmer have five and six labels, respectively, whereas English has 26. Moreover, label symbols for each treebank are very language specific; for example, French and English, which have large label sets, only share two labels.

Therefore, we propose treebank normalization (TN) as the preprocessing step in our approach. Specifically, we first remove any non-terminal span that has length or number of children less than two. In other words, they are any span $(i, j) \in T$ where $j - i < 2$. After that, we mask the labels of all the remaining non-terminal spans with an unified symbol, e.g., "P" as in the example in Figure 1a.

[1] These features can be obtained using `lang2vec` by passing a `syntax_wals+syntax_sswl.` argument.

[2] b/a denote "before or after".

Code	Language	Train	Valid	Test
de	German	40,472	5,000	5,000
en	English	39,832	1,700	2,416
ko	Korean	23,010	2,066	2,287
my	Myanmar	18,088	1,000	1,018
zh	Chinese	17,544	352	348
ja	Japanese	17,204	953	931
ar	Arabic	15,762	1,985	1,959
fr	French	14,759	1,235	2,541
km	Khmer	8,788	510	654
hu	Hungarian	8,146	1,051	1,009
eu	Basque	7,577	948	946
pl	Polish	6,578	821	822
sv	Swedish	5,000	494	666
he	Hebrew	5,000	500	716

Table 1: Data statistics. The numbers refer to numbers of sentences where upper languages are high-resource languages and lower for low-resource languages.

As a result, our label classifier is simplified to only detect the span as a span or non-span.

5 Experiments

5.1 Setups

The evaluation was performed on 14 languages: English from the Penn Treebank (Marcus et al., 1993); Chinese from the Chinese Penn Treebank 5.1 (Xue et al., 2005); Japanese, Khmer, and Myanmar (my) from the Asian Language Treebank (Riza et al., 2016); and Arabic, Basque, French, German, Hebrew, Hungarian, Korean, Polish, and Swedish from the SPMRL 2013 shared task (Seddah et al., 2014). The standard splits of each treebank were applied to prepare the training, validation, and test datasets.

We grouped the languages into high- and low-resource (zero-shot) languages based on their amount of data; those with fewer than 10k samples were treated as low-resource languages (Khmer, Hungarian, Basque, Polish, Swedish, and Hebrew). We trained a multilingual model on the high-resource languages and evaluated the cross-lingual parsing on the low-resource languages.

Note that Khmer and Myanmar scripts have no word boundaries, so we simply use their gold segmented long token[3] for this experiment. We observe that XLM-RoBERTa-Large's tokenizer produces reasonable sub-words for Khmer and Myanmar's long tokens, even when the tokenizer was trained using SentencePiece for these two languages. Table 1 presents detailed data statistics for each language.

[3]Khmer and Myanmar written scripts can be segmented into morphemes (short tokens) or at compound level (long tokens) (Ding et al., 2018)

Lang.	S_{best}	S_{dist}	M_{base}	TN_{ours}	TN_{ours} + TF_{ours}
km	70.0	70.0	55.5	69.0	**71.8**
hg	64.7	31.2	68.6	73.9	**74.7**
eu	33.2	27.2	27.3	34.7	**35.8**
pl	**72.8**	**72.8**	65.6	67.9	68.3
sv	**74.8**	**74.8**	67.8	73.1	73.8
he	77.1	71.3	80.5	81.6	**82.2**
avg	64.5	55.5	61.9	66.2	**67.0**

Table 2: Main unlabeled F1 results. The best F1 for each row is highlighted in bold text.

For comparison, we trained two baselines, single- and multi-source models. For the single-source model, we trained parser for each high-resource language and then selected the best model based on its parsing accuracy on the oracle test set (S_{best}) of the low-resource language or used the precomputed syntactic distance (Littell et al., 2017) (S_{dist}). For the same-value syntactic distance, we further weighted each source language based on the size of its corresponding training data. For the multi-source baseline, a multilingual parser (M_{base}) was trained on concatenated treebanks without treebank normalization or typological features.

Because the label sets of each treebank differ and calculating the accuracy of label prediction was difficult, we calculated the unlabeled F1 measure to evaluate cross-lingual performance. All following F1 values refer to the unlabeled F1 for simplicity. We also removed unnecessary spans such as sentence-level and length-of-one spans.

5.2 Results

As shown in Table 2, the performance of single-source transfer was very high, especially when the best source language can be accurately detected. Unfortunately, the precomputed syntactic distance is not enough to choose the best source language; in the results, it failed in three out of six cases. The alternative to source selection is to train a multilingual parser. Interestingly, even the straightforward treebank concatenation M_{base} has a competitive performance when compared with single-source transfer. The results further show that treebank normalization is essential when training a multilingual

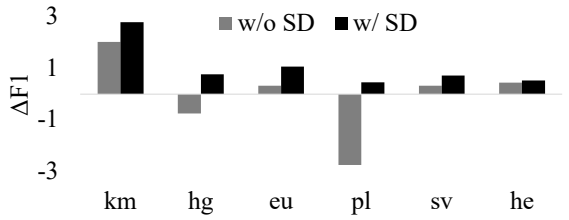

Figure 2: Improvements in the F1 of TN+TF over TN model with or without SD. SD refers to **S**mooth sampling and random **D**ropout.

constituency parser for zero-shot languages, where the improvement over M_{base} is 4.3 in average F1. This result suggests that reducing the complexity of the structure improves cross-lingual performance.

In addition to treebank normalization, our integration of typological features constantly improves cross-lingual performance. An analysis of Figure 2 further shows that the straightforward integration of typological feature yields smaller improvements or hurts the performance for some zero-shot languages, indicating the effectiveness of our smooth sampling and dropout, which generalize the typology-guided cross-lingual parser for zero-shot languages. We additionally observe that the combination of both smooth sampling and dropout is the best configuration for the cross-lingual parsing.

6 Conclusion

We demonstrated the strong ability of recent pretrained multilingual language models for cross-lingual constituency parsing. This result will serve as a new benchmark for future cross-lingual constituency parsing. Moreover, we found that our treebank normalization is crucial when training multilingual treebanks with diverse label sets. In addition, our typological feature integration with dropout and smooth sampling generalizes and improves the model for zero-shot languages. Because we integrated typological features into the span classifier using a simple concatenation approach, more advanced techniques—for instance, a parameter generator (Üstün et al., 2020)—with our dropout and smooth sampling could be studied in the future.

Additionally, our parser could make the applications that leverage structures possible for a wide range of languages without additional treebanks. For example, pseudo constituency structures that our parser generate could be used to apply the recurrent neural network grammar (Dyer et al., 2016; Kim et al., 2019) or the syntax-based neural machine translation (Ma et al., 2019) for many non-English languages. However, since the pseudo structures could be noisy or irrelevant to the model, selective or soft integration techniques should be considered (Chakrabarty et al., 2020).

References

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proc. of UDW*, pages 1–10.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2020. Improving low-resource nmt through relevance based linguistic features incorporation. In *Proc. of COLING*, pages 4263–4274.

Shay B Cohen, Dipanjan Das, and Noah A Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proc. of EMNLP*, pages 50–61.

Chris Collins and Richard Kayne. 2011. Syntactic structures of the world's languages.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. of ACL*, pages 8440–8451.

Ayan Das and Sudeshna Sarkar. 2020. A survey of the model transfer approaches to cross-lingual dependency parsing. *ACM Trans. Asian Low-Resour. Lang. Info. Process.*, 19(5):1–60.

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2018. Nova: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. *ACM Trans. Asian Low-Resour. Lang. Info. Process.*, 18(2):1–18.

Matthew S. Dryer and Martin Haspelmath. 2013. The world atlas of language structures online.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent neural network grammars. pages 199—-209.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580. [cs.NE].*

Yoon Kim, Chris Dyer, and Alexander M Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proc. of ACL*, pages 2369–2385.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proc. of ACL*, pages 3499–3505.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proc. of ACL*, pages 2676–2686.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proc. of EACL*, pages 8–14.

Chunpeng Ma, Akihiro Tamura, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Improving neural machine translation with neural syntactic distance. In *Proc. of NAACL*, pages 2032–2037.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Ryan McDonald, Slav Petrov, and Keith B Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proc. of EMNLP*, pages 62–72.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proc. of ACL*, page 629–637.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proc. of LREC*, pages 1659–1666.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Rapid Sun, Vichet Chea, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian language treebank. In *Proc. of O-COCOSDA*, pages 1–6.

Rudolf Rosa and Zdeněk Žabokrtský. 2015. KLcpos3-a language similarity measure for delexicalized parser transfer. In *Proc. of ACL-IJCNLP*, pages 243–249.

Manon Scholivet, Franck Dary, Alexis Nasr, Benoit Favre, and Carlos Ramisch. 2019. Typological features for multilingual delexicalised dependency parsing. In *Proc. of NAACL*, pages 3919–3930.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proc. of SPMRL*, pages 103–109.

Anders Søgaard and Julie Wulff. 2012. An empirical etudy of non-lexical extensions to delexicalized transfer. In *Proc. of COLING*, pages 1181–1190.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proc. of NAACL*, pages 1061–1071.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. Udapter: Language adaptation for truly universal dependency parsing. In *Proc. of EMNLP*, pages 2302–2315.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proc. of EMNLP-IJCNLP)*, pages 833–844.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proc. of EMNLP*, pages 1857–1867.