Multilingual Grammar Induction with Continuous Language Identification*

Wenjuan Han[†], **Ge Wang**[†], **Yong Jiang**[‡], **Kewei Tu**[†]
[†]School of Information Science and Technology, ShanghaiTech University, Shanghai, China
[‡]Alibaba Group
{hanwj,wangge,tukw}@shanghaitech.edu.cn
yongjiang.jy@alibaba-inc.com

Abstract

The key to multilingual grammar induction is to couple grammar parameters of different languages together by exploiting the similarity between languages. Previous work relies on linguistic phylogenetic knowledge to specify similarity between languages. In this work, we propose a novel universal grammar induction approach that represents language identities with continuous vectors and employs a neural network to predict grammar parameters based on the representation. Without any prior linguistic phylogenetic knowledge, we automatically capture similarity between languages with the vector representations and softly tie the grammar parameters of different languages. In our experiments, we apply our approach to 15 languages across 8 language families and subfamilies in the Universal Dependency Treebank dataset, and we observe substantial performance gain on average over monolingual and multilingual baselines.

1 Introduction

Human languages bear striking resemblance at the syntactic level in spite of their diversity on the surface, as many studies have revealed (Greenberg, 1963; Hawkins, 2014). This fact provides the basis for multilingual grammar induction which tries to simultaneously induce grammars of multiple languages. Intuitively, one can couple grammar parameters of different languages with similar typology and learn them simultaneously. However, the lacking of measures of language similarity prevents this idea from being further exploited in practice.

Previous work in multilingual grammar induction either does not consider language similarity measures (Iwata et al., 2010) or models language similarity based on linguistic phylogeny (Berg-Kirkpatrick and Klein, 2010). The phylogenetic knowledge, however, could be misleading in measuring language similarity. For example, English and German are both Germanic languages, but English exhibits dominant Subject-Verb-Object (SVO) word order while German does not.

In this paper, we propose a novel approach to multilingual grammar induction. Our induction model represents language identities as continuous vectors (i.e., language embeddings) and employs a neural network to predict the grammar parameters of each language based on its embedding. The neural network parameters are universally shared across languages, which softly tie the grammar parameters of different languages. The language embeddings and the neural network parameters are trained with a standard grammar induction objective without any guidance from prior linguistic phylogenetic knowledge. We also introduce an auxiliary language identification task in which we predict the language identities of input sentences using the language embeddings.

We evaluate our approach on corpora of 15 languages across 8 language families and subfamilies. We observe that our approach achieves substantial performance gain on average over monolingual and multilingual baselines.

2 Dependency Model with Valence and Other Related Works

Dependency Model with Valence (DMV) (Klein and Manning, 2004) is the best known generative model for dependency grammar induction. The DMV generates a sentence and its dependency tree following three types of grammar rules (ATTACH, DECISION and ROOT). It firstly samples a token c from the ROOT distribution $\mathrm{P}_{\mathrm{ROOT}}(c)$

*The first and second authors contributed equally. The third author contributed to this work when at ShanghaiTech University. The fourth author is the corresponding author.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5728–5733,
Hong Kong, China, November 3–7, 2019.

and then recursively decides whether to generate a new child token and what child token to generate by sampling from the DECISION and ATTACH distributions $\mathrm{P}_{\text{DECISION}}(dec|h, dir, val)$ and $\mathrm{P}_{\text{ATTACH}}(c|h, dir)$, where dir is a binary variable representing the direction of generation (left or right), val is a binary variable representing whether the current head token already has a child in the direction dir or not, dec is a binary variable deciding whether to continue generation in the current direction, c is the child token and h is the head token.

Almost all previous methods of multilingual grammar induction are based on DMV. Their focus is on designing various priors to couple DMV parameters across languages: Cohen and Smith (2009) propose a logistic normal prior while Berg-Kirkpatrick and Klein (2010) design a hierarchical Gaussian prior according to linguistic phylogeny.

The usage of continuous language embeddings has been explored in other tasks. For example, Ammar et al. (2016) and de Lhoneux et al. (2018) apply language embeddings in supervised multilingual dependency parsing.

3 Approach

We perform unlexicalized grammar induction in which a sentence is represented as a sequence of part-of-speech (POS) tags. We assume that all the languages share the same set of POS tags.

3.1 Multilingual Grammar Model

Our multilingual grammar model adopts the NDMV (Jiang et al., 2016), a monolingual model, as the basic component. NDMV predicts grammar rule probabilities using neural networks. In our model, we add a continuous vector representation of the language identity l (i.e., a language embedding) as an additional input to the neural networks in NDMV. Specifically, to predict an ATTACH rule probability $\mathrm{P}_{\text{ATTACH}}(c|h, dir, val, l)$, we use a multilayer neural network that takes the embeddings of the head token h, valence val and language identity l as input, uses a weight matrix W_{dir} specific to the direction dir in the first layer, and uses a weight matrix W_c consisting of all the child POS tag vectors in the softmax output layer. The neural network structure is shown in the left part of Figure 1. We predict the DECISION rule probabilities in a similar way. We record the number of ROOT rule probabilities instead of predicting them

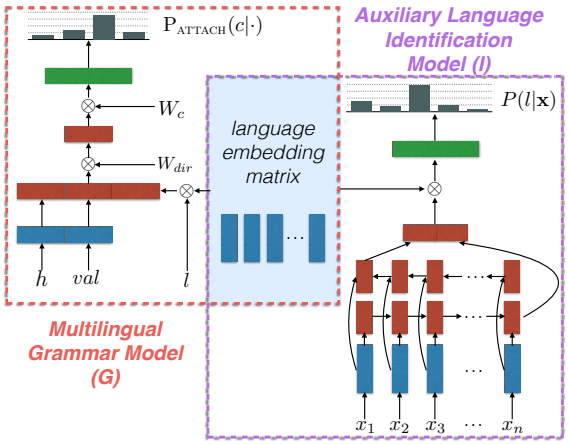

Figure 1: Model Architecture. The language embedding matrix contains the embeddings of all the languages. l is a one-hot vector and $\otimes$ means matrix multiplication. Blue bars represent the embeddings of input symbols; brown bars represent the hidden states of the neural networks; green bars represent the logits which are the inputs to the Softmax layers.

since there are only a small number of such rules. The language embeddings are part of the model parameters and are trained simultaneously with all the other parameters. We hope that after training, similar languages would have similar embeddings and therefore similar grammar rule probabilities.

3.2 Auxiliary Task

To improve the learning of the language embeddings, we introduce an auxiliary language identification task: given an input sentence represented by a sequence of POS tags $\{x_1, x_2, \ldots, x_n\}$, predict its language. We use a standard Bidirectional Long Short-Term Memory (Bi-LSTM) to encode the input sentence, and then a multilayer perceptron to classify the sentence into one of the languages. The weight matrix of the output layer of the multilayer perceptron contains the embeddings of all the languages. The model structure is shown in the right part of Figure 1.

3.3 Training

Denote the set of model parameters as $\mathbf{\Theta}$, the set of languages as $\mathbf{L}$, the set of grammars of different languages as $\mathbf{G} = \{\mathbf{G}_l, l \in \mathbf{L}\}$, and the training data as $\mathcal{D} = \{\mathbf{x}^{(i)}, l^{(i)}\}_{i=1}^{N}$ where $\mathbf{x}^{(i)}$ is the i-th training sentence and $l^{(i)}$ is its language identity. Our training objective function $\mathcal{L}(\mathbf{\Theta})$ combines two conditional probabilities for each training sentence $\mathbf{x}^{(i)}$: $P(\mathbf{x}^{(i)}|\mathbf{G}_{l^{(i)}})$, the probability of the training sentence $\mathbf{x}^{(i)}$ being generated from the corresponding grammar $\mathbf{G}_{l^{(i)}}$; and $P(l^{(i)}|\mathbf{x}^{(i)})$,

the probability of correct language identification of $\mathbf{x}^{(i)}$.

$$\mathcal{L}(\mathbf{\Theta}) = \sum_{(\mathbf{x},l)\in\mathcal{D}} \Big(\log P_{\mathbf{\Theta}}(\mathbf{x}|\mathbf{G}_l) + \lambda \log P_{\mathbf{\Theta}}(l|\mathbf{x}) \Big)$$

where λ is a hyper-parameter and is set to 1 by default. We follow the approach in NDMV to optimize the first term and use Adam (Kingma and Ba, 2014) to optimize the second term.

Note that the language identification model is only used during training to improve the learning of language embeddings. During testing, we run the multilingual grammar model to predict grammar rule probabilities without the need to invoke the language identification model.

4 Experiment

4.1 Setup

We selected 15 languages across 8 language families and subfamilies to ensure diversity. To enable comparisons with previous state-of-the-art approaches (Jiang et al., 2017; Li et al., 2019), we conducted our experiments on UD Treebank 1.4. For each language, we show its language family and the training corpus size in Table 1. We trained our method on the training sentences with length ≤ 15 and tested our method on the testing sentences with length ≤ 40 after removing all punctuations. Since we are doing unsupervised learning, gold dependency trees were not used during training. We use the directed dependency accuracy (DDA, the percentage of words in the testing dataset which are assigned the correct head, same to the unlabeled attachment score normally used in supervised parsing) as the evaluation metric and report the average DDA of 5 runs for each experiment. All the parameters of neural networks including language embeddings were randomly initialized and trained with learning rate 0.001, mini-batch size 1000 and epoch 50. The dimension of the head token embedding and the child token embedding is set to 10. The shape of the weight matrix W_{dir} is 20×10. The dimension of the valence embedding and language identity embedding is set to 5. For the auxiliary language identification task, we use a Bi-LSTM with hidden vector dimension of 10.[1]

[1]Our code is available at https://github.com/WinnieHAN/mndmv.git.

Language	UD Treebank	Language Family	Corpus Size
ET	Estonian	Finnic	11404
FI	Finnish	Finnic	9648
NL	Dutch	Germanic	8783
EN	English	Germanic	7674
DE	German	Germanic	7447
NO	Norwegian	Germanic	10017
GRC	Ancient_Greek	Hellenic	9387
HI	Hindi	Indo-Irian	4997
JA	Japanese	Janponic	7441
FR	French	Romance	4976
IT	Italian	Romance	6492
LA	Latin-ITTB	Romance	10136
BG	Bulgarian	Slavonic	6507
SL	Slovenian	Slavonic	3800
EU	Basque	Vasconic	4271

Table 1: Languages and treebanks used in our experiments.

CODE	MONOLINGUAL		MULTILINGUAL			
	DMV	**NDMV**	**DMV**	**NDMV**	**G**	**G+I**
ET	51.8	52.9	43.1	45.3	56.0	**56.4**
FI	31.8	27.6	39.1	40.0	**50.7**	49.3
NL	42.4	35.6	46.5	47.8	50.4	**50.6**
EN	51.8	**53.7**	47.7	50.8	51.7	52.7
DE	52.8	50.4	55.5	57.2	59.6	**61.4**
NO	58.9	59.2	55.7	58.8	61.0	**61.3**
GRC	40.4	37.7	41.1	40.8	**46.8**	46.2
HI	52.6	**53.9**	29.2	31.1	47.4	46.8
JA	39.8	37.1	27.8	29.6	43.4	**44.2**
FR	58.8	38.1	59.6	59.4	58.4	**60.1**
IT	60.8	63.6	**66.7**	66.4	64.4	65.9
LA	32.6	36.3	39.8	42.0	**45.1**	45.0
BG	58.9	61.8	65.9	69.4	**71.3**	**71.3**
SL	**70.7**	67.5	62.1	63.3	68.3	68.6
EU	42.1	45.5	45.7	45.2	**54.2**	53.6
Avg	49.7	48.1	48.4	49.8	55.3	**55.6**

Table 2: DDA of monolingual and multilingual approaches. Each language is indicated by its ISO 639 code. G: our multilingual grammar model. I: our auxiliary language identification task. Ave: Average DDA over 15 languages.

4.2 Results

We first compare our method with two baseline methods, DMV and NDMV, which are similar to our method[2]. The baseline methods are experimented in both monolingual and multilingual settings. For the monolingual setting we trained the baseline models on each language independently. For the multilingual setting we trained them on the combined training data of all the 15 languages and tested on one of the languages. Table 2 shows the experimental results. It can be seen that our multilingual grammar model (G) performs better on average than all the baselines. The improvement be-

[2]We re-implemented the DMV and the NDMV. We set the ATTACH valence and DECISION valence to 2 and used root constraints, similar to previous work (Gimpel and Smith, 2012; Bisk and Hockenmaier, 2013; Noji et al., 2016).

comes more significant when our model is jointly trained with the auxiliary language identification task (G+I). Note that our approach performs worse than the monolingual baseline on some languages, and we speculate that it is partly caused by data imbalance. In particular, the worst-performing Hindi language has only 4997 training sentences, much smaller than the average 7532. It would be interesting to make training more balanced by assigning weights to training samples of different languages, which we leave for future work.

To measure the statistical significance of the advantage of our method, we performed the nonparametric Friedman's test to support/reject the claim (null hypothesis): there is no difference between the G+I model and the NDMV model in a multilingual setting. Based on the above sample data, the P-value 7.8911×10^{-4} would result in rejection of the claim at the 0.05 significance level, thus showing the significance in our performance gain.

In Table 3 we compare our method with recent state-of-the-art approaches on the UD Treebank dataset: Convex-MST (Grave and Elhadad, 2015), LC-DMV (Noji et al., 2016) and D-J (Jiang et al., 2017). For the three approaches we use the results reported by Jiang et al. (2017). Our G+I model performs better than Convex-MST and LC-DMV on average, even though additional priors and delicate biases are integrated into the two methods (e.g, the universal linguistic prior for Convex-MST and the limited center-embedding for LC-DMV). Our method also slightly outperforms D-J on average, even though D-J combines Convex-MST and LC-DMV and therefore utilizes even more linguistic prior knowledge.

CODE	Convex MST	LC-DMV	D-J	G	G+I
ET	49.4	31.8	44.0	56.0	**56.4**
FI	44.7	26.9	43.5	**50.7**	49.3
NL	45.3	34.1	43.5	50.4	**50.6**
EN	54.0	56.0	**60.1**	51.7	52.7
DE	51.4	50.5	55.7	59.6	**61.4**
NO	55.3	45.5	60.8	61.0	**61.3**
GRC	43.4	33.1	44.9	**46.8**	46.2
HI	56.8	54.2	**60.0**	47.4	46.8
JA	44.8	43.8	**45.8**	43.4	44.2
FR	**62.0**	48.6	57.0	58.4	60.1
IT	69.1	**71.1**	70.3	64.4	65.9
LA	38.8	38.6	42.2	**45.1**	45.0
BG	61.6	62.4	**73.8**	71.3	71.3
SL	54.0	49.5	69.6	68.3	**68.6**
EU	50.0	45.4	**55.7**	54.2	53.6
Avg	52.0	46.1	55.1	55.3	**55.6**

Table 3: Comparison of the recent state-of-the-art approaches and G/G+I. Avg: Average DDA over 15 languages.

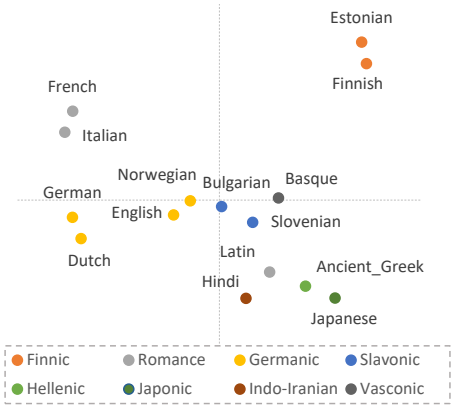

Figure 2: Visualization of the language embeddings.

5 Analysis

5.1 Visualization of Language Embeddings

One of our main expectations is that our approach can automatically learn language embeddings that capture similarities in typology between different languages. In order to verify our expectation, we collected the learned language embeddings and visualized them on a 2D plane using the t-SNE algorithm (Van der Maaten and Hinton, 2008).

Figure 2 shows the visualization result. It can be seen that in most cases languages in the same language family are close to each other. For example, Finnish is close to Estonian (Finnic languages) and Slovenian is close to Bulgarian (Slavonic languages). It is also interesting to note that some languages, such as English and Norwegian in the Germanic family and Latin in the Romance family, are closer to languages outside of their families than to their family siblings. We attribute this phenomenon to the difference in typology among some in-family languages: the flexible word order in German and Dutch is not shared by English and Norwegian; Latin, on the other hand, seems to share more common typological features with its classical counterpart, ancient Greek, than with its modern phylogenetic relatives in the Romance family. Such differences cannot be inferred from linguistic phylogeny, but our language embeddings have undoubtedly captured them.

5.2 In-family vs. Cross-family

In order to further examine the effectiveness of our model in coupling grammar parameters between languages regardless of their language families, we design an additional experiment on a bilingual

	ET	FI	NO	NL	EN	DE	LA	IT	FR	BG	SL	GRC	JA	HI	EU
ET		-1.1	1.9	2.1	0.3	0.9	3.1	5.0	5.6	3.6	-1.2	0.9	29.6	3.2	13.8
FI	-0.2		3.7	1.4	2.6	4.8	2.5	-3.4	0.7	9.5	6.4	2.0	10.3	2.0	2.0
NO	8.8	10.0		-3.2	0.4	1.0	2.7	-1.7	1.4	1.2	2.1	1.8	11.7	9.4	10.4
NL	5.3	-1.7	2.5		0.5	1.3	2.2	1.0	0.2	4.3	1.0	-1.0	0.2	4.6	4.8
EN	0.9	7.6	2.0	0.7		1.6	0.1	1.2	0.6	0.2	-1.0	2.3	18.3	1.2	0.6
DE	-0.8	5.4	0.7	2.1	1.9		1.9	0.9	1.8	2.6	-0.4	2.0	1.8	17.5	8.6
LA	7.0	0.8	0.3	-5.2	-3.6	4.0		-2.0	-5.4	-3.5	-1.6	-2.2	12.3	22.8	3.7
IT	0.7	11.4	-0.3	0.4	2.8	-2.0	-1.3		-0.1	2.2	1.1	6.2	11.3	20.1	10.0
FR	0.9	-1.8	-1.1	0.6	1.4	0.2	0.8	-0.4		5.2	-0.5	3.7	-0.2	21.1	9.1
BG	9.0	13.3	1.2	-2.0	-0.9	2.3	0.1	-1.5	-4.1		1.3	0.9	18.2	25.8	12.7
SL	5.2	11.7	1.3	-4.3	1.0	-4.1	-0.2	-0.6	0.4	2.3		4.5	2.6	21.3	15.2
GRC	4.1	-0.6	0.6	-4.8	4.1	1.8	2.4	3.5	-0.3	2.2	4.6		10.4	12.4	11.4
JA	11.6	0.0	-7.1	-4.4	5.4	-3.3	5.9	-5.6	3.9	0.2	8.0	7.6		0.0	2.1
HI	5.9	-1.2	5.6	1.7	1.4	5.9	3.1	9.0	2.1	7.6	4.6	2.0	0.3		-8.4
EU	16.0	-1.5	3.9	2.7	-0.6	1.7	2.0	1.1	8.0	-0.6	4.0	6.0	-0.2	-2.1	

Figure 3: Each cell shows the difference between the DDA of our model and that of DMV evaluated on the test data of the column language. Positive numbers are shown in red and negative numbers in blue. Languages are grouped by their families.

setup. Specifically, for each pair of languages, we tested our approach and the baseline of training DMV on the combined training set. Since in this setting DMV is blind to the language identity, we expect that it would perform poorly if the two languages come from different families and hence are very likely to have large difference. On the other hand, our model would not be as sensitive to the difference between the two languages. In Figure 3, we report the difference between the DDA of our model and that of DMV. It shows that the advantage of our model over DMV is more significant for cross-family language pairs than for in-family language pairs, which verifies our expectation.

6 Conclusion

In this paper, we incorporate continuous language identity representations into multilingual grammar induction, which softly tie grammar parameters from different languages, resulting in substantial performance gain over various baseline methods. Analysis of the language embeddings suggests that our approach may capture information about language similarity beyond linguistic phylogenetic knowledge.

While in this work we follow previous work and perform unlexicalized parsing, the proposed model can be extended for lexicalized parsing by replacing POS tag embeddings with cross-lingual word embeddings, which we leave for future work.

Acknowledgments

This work was supported by the Major Program of Science and Technology Commission Shanghai Municipal (17JC1404102).

References

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

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1288–1297. Association for Computational Linguistics.

Yonatan Bisk and Julia Hockenmaier. 2013. An hdp model for inducing combinatory categorial grammars. *Transactions of the Association for Computational Linguistics*, 1:75–88.

Shay B Cohen and Noah A Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 74–82. Association for Computational Linguistics.

Kevin Gimpel and Noah A Smith. 2012. Concavity and initialization for unsupervised dependency parsing. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 577–581. Association for Computational Linguistics.

Edouard Grave and Noémie Elhadad. 2015. A convex and feature-rich discriminative approach to dependency grammar induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1375–1384.

Joseph H Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.

John A Hawkins. 2014. *Cross-linguistic variation and efficiency*. OUP Oxford.

Tomoharu Iwata, Daichi Mochihashi, and Hiroshi Sawada. 2010. Learning common grammar from multilingual corpus. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 184–188. Association for Computational Linguistics.

Yong Jiang, Wenjuan Han, and Kewei Tu. 2016. Unsupervised neural dependency parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 763–771, Austin, Texas. Association for Computational Linguistics.

Yong Jiang, Wenjuan Han, and Kewei Tu. 2017. Combining generative and discriminative approaches to unsupervised dependency parsing via dual decomposition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1689–1694, Copenhagen, Denmark. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4992–4997. Association for Computational Linguistics.

Bowen Li, Jianpeng Cheng, Yang Liu, and Frank Keller. 2019. Dependency grammar induction with a neural variational transition-based parser. In *AAAI 2019*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85.

Hiroshi Noji, Yusuke Miyao, and Mark Johnson. 2016. Using left-corner parsing to encode universal structural constraints in grammar induction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 33–43.