Multilingual End-to-end Dependency Parsing with Linguistic typology knowledge

Chinmay Choudhary
National University of Ireland
c.choudhary1@nuigalway.ie

Dr. Colm O'riordan
National University of Ireland
colm.oriordan@nuigalway.ie

Abstract

We evaluate a *Multilingual End-to-end BERT based Dependency Parser* which parses an input sentence by directly predicting the relative head-position for each word within it. Our model is a Cross-lingual dependency parser which is trained on a diverse polyglot corpus of high-resource source languages, and is applied on a low-resource target language.
To make model more robust to typological variations between source and target languages, and to facilitate the cross-lingual transferring, we utilized the Linguistic typology knowledge, available in typological databases **WALS** and **URIEL**. We induce such typology knowledge within our model through an auxiliary task within Multi-task Learning framework.

1 Introduction

Linguistic typology is the classification of human languages according to their syntactic, phonological and semantic features. There are numerous available typological databases such as WALS (Haspelmath, 2009), SSWL (Collins and Kayne, 2009), LAPSyd (Maddieson et al., 2013), ValPal (Hartmann and Bradley Taylor, 2013), AUTOTYP (Bickel et al., 2017), APCLS (Michaelis and Magnus Huber, 2013) etc. These databases provide taxonomies of typological features and their possible values, as well as the respective feature values for most of the world's languages.
Linguistic typology existed as an independent research domain since long (Greenberg, 1963; Comrie, 1989; Nichols, 1992) but recently it has been used along with *Cross-lingual/Multi-lingual* NLP (Ponti et al., 2018; Wang and Eisner, 2017; Agić, 2017; Bender, 2016; O'Horan et al., 2016) to address the issue of data-sparsity in low-resource languages.
However all the popular typological databases suffer from a major shortcoming of limited coverage. In fact, values of many important typological features for most languages (specially less documented ones) are missing in these databases. This sparked a line of research on automatic acquisition of such missing typology knowledge. Many researchers (Malaviya et al., 2017; Bjerva and Augenstein, 2018; Bjerva et al., 2019; Bjerva and Augenstein, 2017; Östling and Tiedemann, 2016) indeed successfully used Multi-lingual NLP and ML techniques to predict these missing feature values. Thus Multilingual NLP and Language typology feature prediction are very closely related tasks which would complement each other. Based on this intuition, we propose a model that performs both Multilingual NLP and Linguistic typology feature prediction tasks simultaneously, in a multi-tasking setup.
Multi-task Learning (MTL) (Ruder, 2017) is neural network framework which involves performing of two or more tasks simultaneously leading to knowledge/parameter sharing. These tasks are closely related thus complement each other leading to improved performance on all of them. Even in scenarios where we primarily care about a single task, using a closely related task as an auxiliary task for MTL can be useful (Caruana, 1998; Zhang et al., 2014; Liu et al., 2015; Girshick, 2015; Arik et al., 2017).
In this work, we use *Linguistic Typology* feature prediction task as auxiliary task for *End-to-end Cross-lingual Dependency Parsing*. Hence, we make following contributions.

1. We evaluated the performance an *End-to-end BERT Based Parser* which can parse a sentence by directly predicting relative head-position tag for each word within input sentence. This is inspired by (Li et al., 2018) which is an *End-to-end Seq2seq Dependency Parser*. We evaluated the performance of this BERT based End-to-end parser in both monolingual and cross-lingual/multilingual setups (using mBERT). We will refer to this model

as *Base E2E BERT parser* in this paper.

2. We added the auxiliary task of Linguistic typology prediction to our *Base E2E BERT parser* to observe the change in performence under different settings. We will refer to this model as *Multitasking E2E BERT Parser* in this paper.

2 Related Work

Cross-lingual *Model-transfer* approaches to Dependency Parsing such as (McDonald et al., 2011; Cohen et al., 2011; Duong et al., 2015; Guo et al., 2016; Vilares et al., 2015; Falenska and Çetinoğlu, 2017; Mulcaire et al., 2019; Vania et al., 2019; Shareghi et al., 2019) involve training a model on high-resource languages and subsequently adapting it to low-resource languages.
Participants of CoNLL 2017 shared-task (Daniel et al., 2017) and CoNLL 2018 shared task (Zeman et al., 2018) also provide numerous approaches to dependency parsing of low-resource languages.
Some approaches such as (Naseem et al., 2012; Täckström et al., 2013; Barzilay and Zhang, 2015; Wang and Eisner, 2016a; Rasooli and Collins, 2017; Ammar, 2016; Wang and Eisner, 2016b) used typological information to facilitate cross-lingual transfer. All these approaches directly feed the linguistic typology features into the model whereas we induce the linguistic typology knowledge through Multitask learning.
Inducing typology knowledge through MTL rather than directly feeding it along with word-embeddings have following advantages.

1. The model can also be applied to low-resource languages for which many typology feature values are unknown/missing.

2. The auxiliary task should help to improve the performance on the main dependency parsing task as well, since it would make the model give special emphasis on the syntactic typology (specially word-order typology) of language being parsed while predicting the dependency relations.

3 Base End-to-end BERT Parser

This section elaborates the details of our *End2End BERT based Dependency Parser* which directly predicts the relative head position tag of each word within input sentence.

Given a sentence of length T, its dependency parse-tree can be represented as a sequence of T relative head-position tags as demonstrated in figure 1a.
Figure 2a depicts the architecture of our baseline model. The depicted architecture comprises of three components namely *BERT Encoder*, *Output Network* and *Tree-decoder* described as section 3.1, 3.2 and 3.3.

3.1 BERT Encoder

It is a BERT based network which takes as input, the entire sentence as sequence of tokens. The model outputs $d-1$ dimensional word-embeddings for all words within the input sentence. Thus for a sentence of length T, it would output matrix $E \in R^{T*(d-1)}$.
We use WordPiece tokenizer (Wu et al., 2016) to tokenize input sentence and extract embeddings. For each word within input sentence, we use the BERT output corresponding to the first wordpiece of it as its embedding, ignoring the rest.

3.1.1 POS tag information

We add pos-tag information in our parser by appending index of pos-tag of each word, to the encodings outputted by BERT encoder as evident in figure 2b. Thus matrix $\hat{E}$ is derived from E through equation 1 .

$$\hat{E} = E; [t_1; t_2; ....; t_T] \tag{1}$$

Here t_i is POS-tag index of i^{th} word. $\hat{E} \in R^{T*d}$

3.2 Output Network

Its a simple feed-forward network with *softmax* activation. The network takes-in embedding matrix from the BERT encoder and outputs the probabilities of all possible relative head position tags at each word by applying equation 1.

$$Pr = softmax(\hat{E} * W + b) \tag{2}$$

Here W,b are weights and biases. $Pr \in R^{T*N}$ where N is the number of valid relative head-position tags.
For the sentence of length T, set of all possible relative head position tags S_T is given as

$$S_T = [L_1, L_2, ..., L_T, R_1, R_2, ..., R_{T-1}, \\ < root >, < EOS >]$$

Here $< root >$ and $< EOS >$ are tags to be assigned to $< s >$ and $< /s >$ tokens at the begin

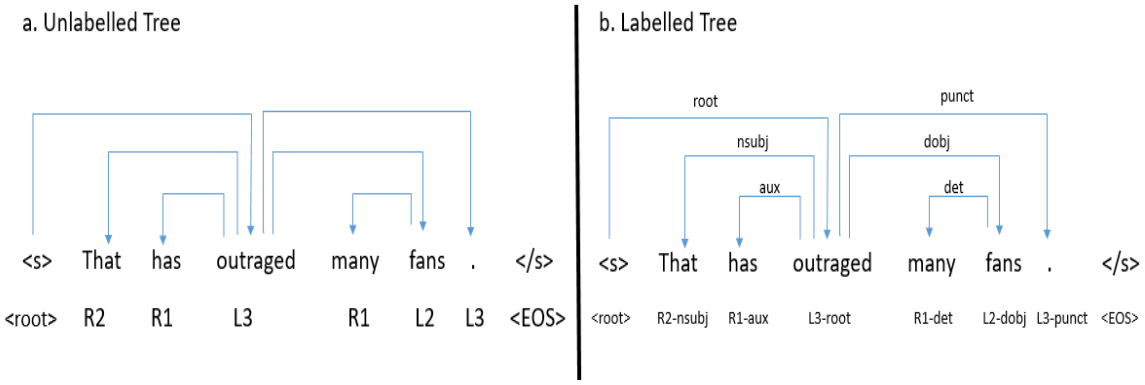

Figure 1: Examples of dependency parse tree being represented as relative head-position tag sequence

and end of the input sentence as shown in figure 1a.

For training and evaluations, we always computed probabilities of all relative head-position tags within the tag-set for a sentence of length *Max* i.e. S_{Max} as the dimensions of model parameters should be fixed. Here *Max* is the length of largest sentence from all copra used during experiments.

In this paper we experimented with only Unlabeled Dependency Parsing however same architecture can be used for Labeled Dependency Parsing as well. In such case the output tags would comprise of relative head positions as well as relationship labels (eg: L2-nsubj). Hence, the set of all possible relative head position tags S would be much larger. Figure 1b depicts a labelled parse-tree being represented as sequence of head-position tags.

3.3 Tree-Decoder

This component decodes the most probable correct label sequence from Probabilities outputted by Output Network. The correct label sequence would satisfy following constraints.

1. Sequence should start with $< root >$ and end with $< EOS >$ tags. These tags should not appear anywhere else.

2. At each index (of word being labelled) the assigned label should be within the range of sentence. For eg: Word 'That' within sentence shown in figure 1a can not have tags L_2, L_3, L_4, L_5, L_6 and word '.' in the sentence can not have any right tags as these are outside the range of sentence.

3. Label sequence should not generate any cycles within dependency tree.

4. One of the words should have the head at $< root >$ token.

We used dynamic programming with beam-search to efficiently extract the most probable label sequence which satisfies the above listed constraints, out of all possible label sequences.

3.4 Multitasking End-to-end BERT Parser

Figure 2b demonstrates the architecture of our proposed model. The model is very similar to the *Base E2E BERT Parser* described in section 3 with one extra component namely *Linguistic typology predictor* which predicts the typology features of language being parsed. Thus model is Multi-tasking model with hard-parameter sharing (Ruder, 2017).

3.4.1 Linguistic typology predictor

It is a simple deep feed forward neural network which takes in the embedding generated by BERT Encoder for token $< /s >$ and outputs probabilities of values of binary syntactic typology features for the language being parsed as 1. Such features are provided by URIEL database (Littell et al., 2017). Let $\hat{N}$ be the number of syntactic typology features provided by URIEL database. The *Linguistic typology predictor* would then predict probability matrix $Pr_{ty} \in R^{\hat{N}}$ by applying equation 2.

$$Pr_{ty} = sigmoid(e_{</s>} * U + c) \quad (3)$$

Here $e_{</s>} \in R^d$ is embedding from BERT Encoder for $< /s >$ token. $U \in R^{d*\hat{N}}$ and $c \in R^{\hat{N}}$ are weights and biases respectively.

3.5 Training

We trained both *BERT Encoder* (fine-tuning of pre-trained BERT model) and *Output Network* components of *Base E2E BERT Parser* model jointly, by

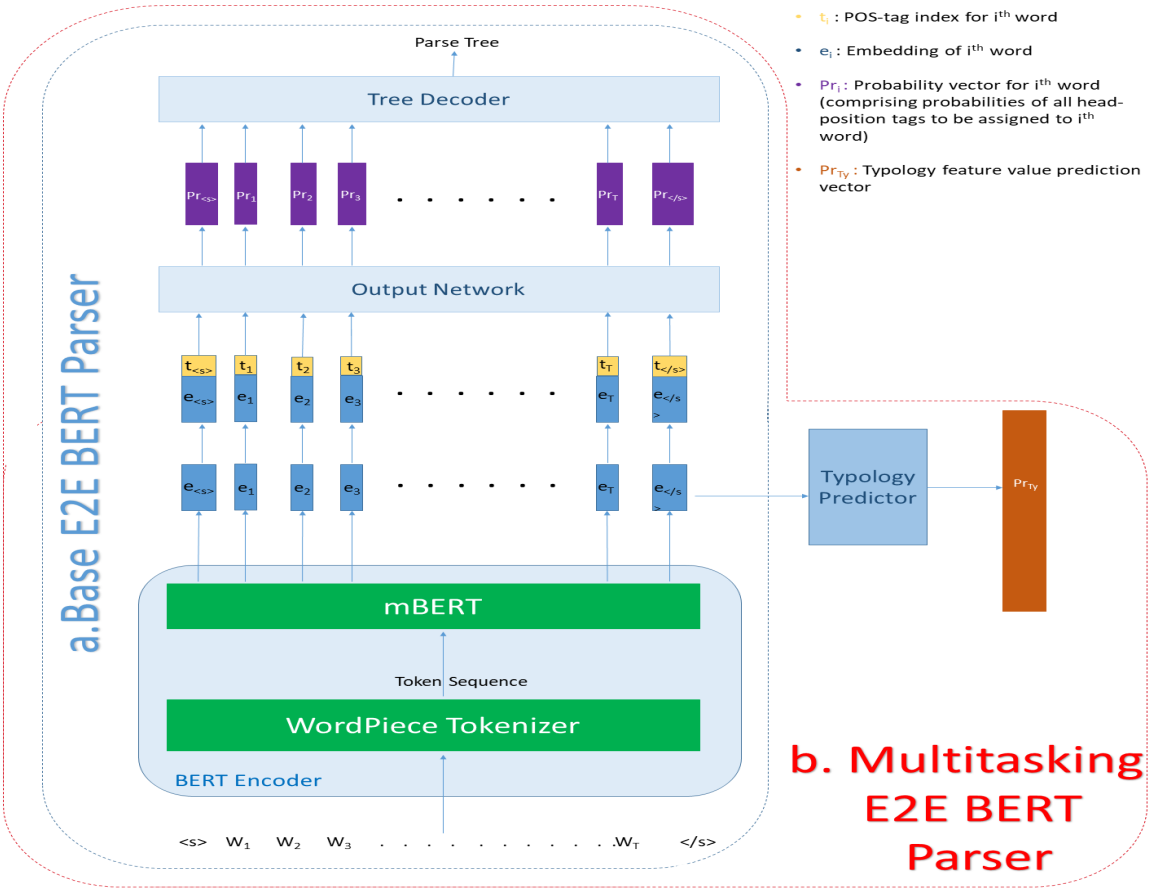

Figure 2: a. Base End-to-end BERT parser architecture. b.Multitasking End-to-end BERT parser architecture. Its an extension of Base End-to-end BERT parser architecture with one extra component namely *Typology Predictor*.

optimizing the cross-entropy loss (Gómez, 2018) between true relative head-position tags and probabilities outputted by the *Output Network*.

On the other hand, *Multitasking E2E BERT parser* is trained to perform tasks of *Prediction of relative heap-position tag sequence* and *Prediction of typology features* simultaneously through MTL, by optimizing the total-loss as the sum of cross-entropy loss over true head-position tag-sequence and the binary cross-entropy loss over true typology values.

Table 4 outlines values of hyper-parameters used during experimentation. These values are obtained by minimizing loss on *Validation* dataset for English language.

4 Experiments

4.1 Experimental setups

We evaluated the monolingual and multilingual variants of our proposed models within two distinct experimental setups namely *Monolingual* and *Cross-lingual* setups. These are described as sections 4.1.1 and 4.1.2 respectively.[1].

4.1.1 Monolingual Setup

In this setup we conducted experiments to evaluate the performance of fully monolingual variants of our proposed *Base E2E BERT Parses* and *Multitasking E2E BERT Parser*. In these settings we experimented in two languages namely *English* and *Chinese*. These monolingual variants use pre-trained monolingual English and Chinese BERT models provided by [].

For all experiments within this setup, we used the *Deep Biaffine Parser* (Dozat and Manning, 2016) as baseline. Its is a neural graph-based dependency parser which uses biaffine attention classifiers to predict the arcs and labels of the required parse-tree for an input sentence.

4.1.2 Cross-lingual setups

We conducted numerous experiments to evaluate the performance of Multilingual/Cross-lingual variants of our proposed *Base BERT Parses* and

[1]Source code at https://github.com/XXXXX

Experimental Settings	Source Languages	Target Languages
Monolingual	English, Chinese	English, Chinese
Cross-lingual with single source language	English	German, Croatian, Italian, Hindi, Chinese, Estonian, Vietnamese
Cross-lingual with multiple source languages	English, Urdu, French, Arabic, Japanese, Polish, Latvian, Tamil, Greek, Coptic, Kazakh, Turkish	German, Croatian, Italian, Hindi, Chinese, Estonian, Vietnamese

Table 1: Source and Target Languages used during experiments

Languages	Corpus
English	en_ewt-ud-train
Urdu	ur_udtb-ud-train
French	fr_ftb-ud-train
Arabic	ar_padt-ud-train
Japanese	ja_gsd-ud-train
Polish	pl_pdb-ud-train
Latvian	la_ittb-ud-train
Tamil	ta_ttb-ud-train
Greek	el_gdt-ud-train
Coptic	cop_scriptorium-ud-train
Kazak	kk_ktb-ud-train
Turkish	tr_imst-ud-train

Table 2: Copra for source languages listed in table 1 used during experiments. All copra are part of Universal Dependencies dataset.

Languages	Corpus	Dev Corpus*
German	de_hdt-ud-test	de_hdt-ud-dev
Croatian	hr_set-ud-test	hr_set-ud-dev
Italian	it_isdt-ud-test	it_isdt-ud-dev
Hindi	hi_hdtb-ud-test	hi_hdtb-ud-dev
Chinese	zh_gsd-ud-test	zh_gsd-ud-dev
Estonian	et_edt-ud-test	et_edt-ud-dev
Vietnamese	vi_vtb-ud-test	vi_vtb-ud-dev

Table 3: Copra for target languages listed in table 1 used during experiments. All copra are part of Universal Dependencies dataset. * A small subset of sentences are sampled from these copra to be added to the source copra in *Few-shot* scenarios

Multitasking E2E BERT Parser models in cross-lingual settings. These Multilingual variants use pre-trained Multilingual BERT (mBERT) (Wu and Dredze, 2019) model which is trained on data from Wikipedia in 104 languages.
We evaluated the Multilingual variants of our models under following two Cross-lingual setups.

1. *Cross-lingual with single source language (CL-Single)*: In this setup, all the parsers are trained in single source language English, but tested on a diverse range of target languages

2. *Cross-lingual with multiple source languages (CL-Poly)*: In this setup, all the parsers are trained on diverse polygot corpus and tested on a diverse range of target languages. There is no overlap between source and target language sets.

Furthermore, the experiments within *Cross-lingual with single source language (CL-Single)* and *Cross-lingual with multiple source languages (CL-Poly)* setups are conducted under both *Few-shot* and *Zero-shot* learning scenarios.
Within *Zero-shot* learning scenario the training corpus does not contain any sentence in the target language on which the model is being evaluated. On the other hand, within *Few-shot* learning scenario the training corpus consists of few sentences in the target language on which the model is being evaluated, along with other source language sentences (covering over 80% the corpus). In Cross-lingual setups we used Graph-based mBERT parser by (Wu and Dredze, 2019) as baseline. It is a multilingual parser that uses same architecture as (Dozat and Manning, 2016) except the LSTM encoder which is replaced by mBERT.

4.2 Languages

Table 1 lists various source and target language used in each of the experimental settings. In *CL-Poly* setup, we trained our models on joint polygot corpus of all twelve source languages listed in Table 2. All these twelve languages belong to distinct

Hyper-parameter	Value
d	768
Dropout prob.	0.01
Bach-size	32
Number of steps per epoch	Size of training corpus / 32
Epochs	50
BERT dimensions	cased_L-12_H-768_A-12

Table 4: Hyper-parameters

linguistic families thus making the corpus typologicaly diverse.
For all experiments, the training corpus size is always fixed to 30,000 sentences. The joint polygot corpus to train *CL-Poly* is created by randomly sampling 2500 sentences from the training coprus for each of the 12 source languages listed in Table 1, concatenating them as one treebank and randomly shuffling the order.
Our *Cross-lingual* models are tested on seven target languages, belonging to distinct linguistic families. Three of these seven languages namely *Chinese*, *Estonian* and *Ammheric* belong to a linguistic family which is distinct from language families of all the source languages listed in Table 2. Thus performance on these languages indicate true robustness of the evaluated models to typological variations between source and target languages.

4.3 Treebank and Typology datasets

Tables 2 and 3 list the treebank copra for each of the languages listed in Table 1, used during experiments. All these copra are downloaded from Universal Dependencies[2] .
For Linguistic typology feature prediction auxiliary tasks we used Linguistic typology feature values provided by URIEL database (Littell et al., 2017). URIEL database is a collection of binary features extracted from multiple typological, phylogenetic, and geographical databases such as WALS (Haspelmath, 2009), PHOIBLE (Moran and Richard Wright, 2014), Ethnologue (M. Paul Lewis and Fennig, 2015) and Glottolog (Harald Hammarstrom and Bank, 2015). URIEL database can be accessed through Pyton PyPi library called *lang2vec*[3]. Library also allows users to access only a subset of all binary features as well.

[2]https://universaldependencies.org/
[3]https://pypi.org/project/lang2vec/

Model	en	zh
Deep Biaffine Network	93.77	93.77
Base E2E BERT Parser	93.00	93.77
Multitasking E2E BERT parser	93.13	93.77

Table 5: Unlabeled Attachment Scores (UAS) achieved in *Monolingual* experimental settings.

For the experiments within this paper, we used only syntactic binary features generated from WALS database (categorised as *Syntax-WALS* within URIEL database).

4.3.1 Missing Typology

As with most typology databases, URIEL also comprises of several missing values of features for many languages. These missing values are indicated as '–' in typology vector provided by URIEL (rather than having values 0 or 1). A typology feature can also have value as '–' for a well-documented language if that feature has no dominant value observed within the respective language
These missing features pose a problem during training of *Multitasking BERT Parser* as there are no true-values for these to optimize loss with. We address this issue through masking technique (Vaswani et al., 2017). We masked the missing typology features and train only on available ones for each source language.

4.3.2 Short tree-bank copra

For each experiment under *Few-shot learning* scenario, we extracted a small set of target language sentences (on which model is being evaluated), to be added to the source training corpus before training.
We extracted this subset by randomly sampling sentences from the *dev* corpus of the respective target-language tree-bank dataset until the token-size becomes approximately equal to 3000. This is inspired by (Ammar et al., 2016) who used same yardstick to evaluate their *Multi-lingual Dependency Parser (MALOPA)*.

5 Results and Inference

Tables 5 outlines Unlabeled Attachment Score (UAS) achieved by the baseline *Deep Biaffine Parser* as well as our *Base E2E BERT Parser* and

	CL-Single				CL-Poly			
	mBERT	**Base E2E**	**Multi E2E**	**Aux task***	**mBERT**	**Base E2E**	**Multi E2E**	**Aux task***
zh	43.32	42.98	41.74	0.01	66.81	66.52	65.35	0.28
hr	72.49	72.07	70.91	0.07	75.28	75.01	74.05	0.14
et	71.05	70.69	69.72	0.05	67.2	66.8	65.67	0.26
de	78.07	77.68	76.67	0.04	78.85	78.54	77.33	0.21
hi	44.83	44.42	43.18	0.11	74.68	74.4	73.32	0.22
it	86.63	86.32	85.23	0.04	77.77	77.4	76.3	0.21
vi	40.74	40.34	39.25	0.08	66.89	66.56	65.45	0.24

Table 6: Unlabeled Attachment Scores (UAS) achieved in both Cross-lingual settings under *Zero-shot* scenario. *F1 values achieved on the auxiliary task of linguistic typology prediction (excluding missing values)

	CL-Single				CL-Poly			
	mBERT	**Base E2E**	**Multi E2E**	**Aux task***	**mBERT**	**Base E2E**	**Multi E2E**	**Aux task***
zh	44.04	43.69	44.29	0.57	67.68	67.37	68.19	0.76
hr	73.38	73.0	73.46	0.6	75.93	75.58	76.28	0.68
et	71.89	71.5	71.96	0.56	67.91	67.55	68.45	0.78
de	78.8	78.47	79.08	0.57	79.74	79.45	80.25	0.71
hi	45.63	45.33	45.91	0.61	75.59	75.16	76.13	0.62
it	87.44	87.12	87.63	0.61	78.51	78.14	78.98	0.66
vi	41.44	41.16	41.62	0.61	67.68	67.41	68.37	0.75

Table 7: Unlabeled Attachment Scores (UAS) achieved in both Cross-lingual settings under *Few-shot* scenario. *F1 values achieved on the auxiliary task of linguistic typology prediction (excluding missing values)

Multitasking E2E BERT Parser in monolingual settings, on both English and Chinese.

Tables 6 and 7 outline Unlabeled Attachment Scores (UAS) obtained under the *Few-shot* and the *Zero-shot* learning scenarios respectively. Results indicate that in both *Monolingual* and *Cross-lingual settings*, our *Base E2E BERT parser* performed at par with the baseline *Deep Biaffine Parser* (Dozat and Manning, 2016) and *Graph-based mBERT parser* (Wu and Dredze, 2019) models respectively, despite being much simpler in design as its end-to-end.

5.1 Effect of Polygot Training

It is evident from results that in *CL-Single* setup under both *Few-shot* and *Zero-shot* scenarios, all the evaluated mBERT based cross-lingual models (baseline and proposed models) perform better on target languages which are genealogically or geographically closer to the source language English. Thus high performance is observed for the European languages **de**, **et**, **it** and **hr**, whereas performance drop significantly on Asian languages **zh**, **hi** and **vi** as these are both genealogically and geographically apart from English.

On the other hand, in *CL-Poly* setup, these models show almost uniform performance across all target languages. However even in *Cl-Poly* setup, the models achieved comparatively lower UAS on languages **zh**, **et** and **vi** than on other target languages, as these languages belong to a language family which is distinct from language families of all source languages listed in table 2 (section 4.2). Since **zh**, **et** and **vi** are fully unknown languages in both *CL-Single* and *CL-Poly*, the performance on these languages indicate the cross-lingual transfer ability of the evaluated mBERT based dependency parsing models.

It is evident from results outlined in Tables 6 and 7 that both baseline and our proposed End-to-end parsing models show very strong improvement in performance on languages **zh**, **et** and **vi** when trained on mixed polygot corpus as compared to when trained on single source language copra.

Thus it can be inferred that Cross-lingual transfering ability of an mBERT based multilingual dependency parser, to a distinct and unseen target

language increases significantly as a result of polygot training, as polygot training allows the model to generalise better over a diverse set of languages.

5.2 Effect of Auxiliary task

Tables 2, 3 and 4 also outline the F1-scores achieved by our *Multitasking E2E BERT parser* model on the auxiliary task of predicting linguistic-typology features in Monolingual settings as well as both *Cross-lingual with single source language* and *Cross-lingual with multiple source languages* under both *Zero-shot* and *Few-shot* scenarios.

5.2.1 Effect in Monolingual setting

Results in Table 1 show that within Monolingual setup, our *Multitasking E2E BERT parser* showed marginal improvement over *Base E2E BERT parser* for both English and Chinese. In-fact the monolingual variant of our *Multitasking E2E BERT parser* outperformed the baseline *Deep Biaffine Parser* (Dozat and Manning, 2016) for both English and Chinese.
Hence it can be inferred that in Monolingual settings, the auxiliary task of predicting linguistic typology features does lead to improvement in parsing performance indeed, as it enables the model the model to emphasize on syntactic typology of language being parsed (specifically word-order features) while predicting the dependency relations within the sentence.

5.2.2 Effect in Cross-lingual settings

Under the *Zero-shot learning* scenario, our *Multitasking E2E BERT parser* under-performed *Base E2E BERT parser* in both *CL-Single* and *CL-Poly* settings for all target languages.
On the other hand under *Few-shot learning* scenario, our *Multitasking E2E BERT parser* showed improvement in performance for all target languages, in both *CL-Single* and *CL-Poly* settings.
Within *CL-Poly* setting under *Few-shot learning* scenario, our *Multitasking E2E BERT parser* shows an average improvement of 4.6% in UAS across all target languages over *Base E2E BERT parser*. This is much higher than average improvement of 1.93% shown by our*Multitasking E2E BERT parser* over *Base E2E BERT parser* within *CL-Single* settings under *Few-shot learning* scenario.
Based on these trends it can be inferred that the auxiliary task does not help the model to improve the cross-lingual transfer parsing in an unseen language (which are not the part of training corpus). However the task does enable the model to better learn to distinctively parse in each of the languages on which it is trained, even if the training corpus consists of only few sentence in the language.
Further the improvement is higher in *CL-Poly* settings than *CL-Single* settings as the model generalizes better on the auxiliary task due to polygot training.

6 Conclusion and Future Work

In this paper we evaluated the performance of our proposed *End-to-end BERT Based Dependency Parser* which can parse a sentence by directly predicting relative head-position tag for each word within input sentence. Subsequently we added the auxiliary task of Linguistic typology prediction to our *Base E2E BERT parser* to observe the change in performance under different settings.
Our results show that adding such auxiliary task leads to improvement in performance of *Base E2E BERT Parser* within Cross-lingual settings under *Few-shot* learning scenario whereas no improvement is observed within the *Zero-shot* learning scenario.

References

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10.

Waleed Ammar. 2016. *Towards a Universal Analyzer of Natural Languages*. Ph.D. thesis, Ph. D. thesis, Google Research.

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.

Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. 2017. Deep voice: Real-time neural text-to-speech. *arXiv preprint arXiv:1702.07825*.

Regina Barzilay and Yuan Zhang. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. Association for Computational Linguistics.

Emily M Bender. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660.

Balthasar Bickel, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Fernando Hildebrandt, Kristine, and John B Lowe. 2017. The autotyp

typological databases. *Version 0.1. 0. Online: https://github.com/autotyp/autotyp-data/tree/0.1.0*.

Johannes Bjerva and Isabelle Augenstein. 2017. Tracking typological traits of uralic languages in distributed language representations. *arXiv preprint arXiv:1711.05468*.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. *arXiv preprint arXiv:1802.09375*.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

R Caruana. 1998. Multitask learning. autonomous agents and multi-agent systems.

Shay B Cohen, Dipanjan Das, and Noah A Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 50–61.

Chris Collins and Richard Kayne. 2009. Syntactic structures of the world's languages. http://sswl.railsplayground.net/..

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Zeman Daniel, Popel Martin, Straka Milan, Hajic Jan, Nivre Joakim, Ginter Filip, Luotolahti Juhani, Pyysalo Sampo, Petrov Slav, Potthast Martin, et al. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, volume 1, pages 1–19. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850.

Agnieszka Falenska and Özlem Çetinoğlu. 2017. Lexicalized vs. delexicalized parsing in low-resource scenarios. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 18–24.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Raúl Gómez. 2018. Understanding categorical cross-entropy loss, binary cross-entropy loss, softmax loss, logistic loss, focal loss and all those confusing names. *URL: https://gombru. github. io/2018/05/23/cross_entropy_loss/(visited on 29/03/2019)*.

Joseph Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *In J. Greenberg, ed., Universals of Language. 73-113. Cambridge, MA*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Martin Haspelmath Harald Hammarstrom, Robert Forkel and Sebastian Bank. 2015. Glottolog 2.6.

Martin Haspelmath Hartmann, Iren and editors Bradley Taylor. 2013. Valency Patterns Leipzig.

Martin Haspelmath. 2009. *The typological database of the World Atlas of Language Structures*. Berlin: Walter de Gruyter.

Zuchao Li, Jiaxun Cai, Shexia He, and Hai Zhao. 2018. Seq2seq dependency parsing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3203–3214.

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 *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval.

Gary F. Simons M. Paul Lewis and Charles D. Fennig. 2015. Ethnologue: Languages of the World, Eighteenth edition.

Ian Maddieson, Sébastien Flavier, Egidio Marsico, Christophe Coupé, and François Pellegrino. 2013. Lapsyd: lyon-albuquerque phonological systems database. In *INTERSPEECH*, pages 3022–3026.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. *arXiv preprint arXiv:1707.09569*.

Ryan McDonald, Slav Petrov, and Keith B Hall. 2011. Multi-source transfer of delexicalized dependency parsers.

Philippe Maurer Martin Haspelmath Michaelis, Susanne Maria and editors Magnus Huber. 2013. Atlas of Pidgin and Creole Language Structures Online.

Daniel McCloy Moran, Steven and editors Richard Wright. 2014. PHOBIA Online.

Phoebe Mulcaire, Jungo Kasai, and Noah A Smith. 2019. Low-resource parsing with crosslingual contextualized representations. *arXiv preprint arXiv:1909.08744*.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. The Association for Computational Linguistics.

Johanna Nichols. 1992. *Linguistic diversity in space and time*. University of Chicago Press.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. *arXiv preprint arXiv:1610.03349*.

Robert Östling and Jörg Tiedemann. 2016. Continuous multilinguality with language vectors. *arXiv preprint arXiv:1612.07486*.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures in cross-lingual nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1531–1542.

Mohammad Sadegh Rasooli and Michael Collins. 2017. Cross-lingual syntactic transfer with limited resources. *Transactions of the Association for Computational Linguistics*, 5:279–293.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ehsan Shareghi, Yingzhen Li, Yi Zhu, Roi Reichart, and Anna Korhonen. 2019. Bayesian learning for neural dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3509–3519.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. *arXiv preprint arXiv:1909.02857*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

David Vilares, Carlos Gómez-Rodríguez, and Miguel A Alonso. 2015. One model, two languages: training bilingual parsers with harmonized treebanks. *arXiv preprint arXiv:1507.08449*.

Dingquan Wang and Jason Eisner. 2016a. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Dingquan Wang and Jason Eisner. 2016b. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Dingquan Wang and Jason Eisner. 2017. Fine-grained prediction of syntactic typology: Discovering latent structure with supervised learning. *Transactions of the Association for Computational Linguistics*, 5:147–161.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer.