Relational Graph Attention Networks for Syntax Encoding in Zero-shot Cross-lingual Semantic Role Labeling

Anonymous ACL submission

Abstract

Recent models in cross-lingual semantic role labeling (SRL) rely heavily on BiLSTMs, a derivation of RNNs, as their main encoders. However, a previous study in dependency parsing has shown that RNN-based cross-lingual models are ineffective in distant languages. Therefore, we propose graph neural networks (GNNs) built on dependency trees to replace BiLSTMs' role as the encoder for cross-lingual models. We hypothesize that encoding sentences based on their dependency trees helps cross-lingual SRL models achieve better generalization. Through a simple encoder-decoder architecture, we compare various GNNs, i.e., gated graph convolutional networks (GGCNs), graph attention networks (GATs), two-attention relational GATs (2ATT-GATs), and modified self-attention networks from Transformer (SATs). We focus on a zero-shot setting and evaluate the models in 23 languages available in Universal Proposition Bank. The evaluation shows that 2ATT-GATs outperform other GNNs. Moreover, comparisons against BiLSTM-based models show that 2ATT-GATs are more effective for building cross-lingual SRL models, especially in languages with different word orders.

1 Introduction

Semantic role labeling (SRL) is a task to determine the predicates of a sentence and argument roles for each corresponding predicate, as shown in Figure 1. SRL supports many natural language processing (NLP) tasks, e.g., information extraction (Christensen et al., 2010), abstractive summarization (Khan et al., 2015), and machine translation (Rapp, 2022). However, SRL resource availability is low, hindering the performance of other NLP tasks in diverse languages. Cross-lingual SRL models try to solve this problem by training the models in resource-rich languages and transferring the models to resource-poor languages.

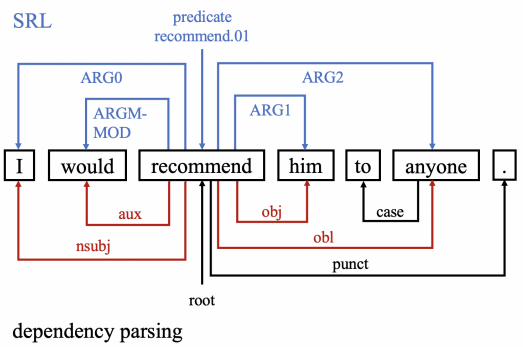

Figure 1: An example of the SRL task (top) and the dependency parsing task (bottom) applied to a sentence taken from UPB. The red color indicates path intersections in both tasks.

A study in cross-lingual dependency parsing (Ahmad et al., 2019) proves that recurrent neural network (RNN)-based cross-lingual models are sensitive to word orders, making them unable to transfer effectively to distant languages. Despite the developments in cross-lingual SRL, recent models still heavily rely on BiLSTMs, the derivation of RNNs, as their encoders, e.g., Fei et al. (2020), Cai and Lapata (2020), and Conia et al. (2021).

We propose to apply GNNs over universal dependency trees provided by Universal Dependencies (UD) (Nivre et al., 2016b, 2020) as the encoder for cross-lingual models. We hypothesize that encoding sentences based on their dependency trees makes cross-lingual SRL models generalize better. We provide two main reasons: (1) Many predicate-argument paths and argument roles in SRL intersect with dependency paths and dependency relations in dependency parsing (Marcheggiani and Titov, 2017), as shown in Figure 1. (2) Universal dependency tree, representing a sentence's grammatical structure in a language-universal scheme, is a more generalized representation across languages than word sequences.

We experiment on various GNNs as an encoder to extract language-universal information from

dependency trees, including GGCNs (Marcheggiani and Titov, 2017), graph attention networks (GATs) (Veličković et al., 2017), modified relational GATs (Wang et al., 2020) (2ATT-GATs), and modified self-attention networks from Transformer (Shaw et al., 2018) (SATs). We apply GGCNs and 2ATT-GATs since they have been proven useful to encode dependency trees in monolingual SRL (Marcheggiani and Titov, 2017) and sentiment analysis (Wang et al., 2020), respectively. We compare 2ATT-GATs with GATs that treat the dependency tree as unlabeled. In addition, we employ SATs as it is taken from self-attention in the popular Transformer (Vaswani et al., 2017). Furthermore, we compare the best GNN-based model with BiLSTMs-based models to show the effectiveness of the GNN-based model in cross-lingual SRL.

SRL consists of four steps, i.e., predicate detection, predicate sense disambiguation, argument detection, and argument labeling. Following previous work in cross-lingual SRL (Fei et al., 2020), we focus on argument detection and argument labeling in the dependency-based SRL. We conduct experiments in a zero-shot setting to find the most transferable networks across languages.

We train and evaluate the models in seven and 23 languages provided by Universal Proposition Bank (UPB) v1.0 and UPB v2.0, respectively. Throughout the paper, we show that in cross-lingual SRL:

1. 2ATT-GATs outperform other GNN-based SRL models, indicating that 2ATT-GATs transfer more effectively across languages than other GNNs.

2. 2ATT-GATs perform better than BiLSTM-based SRL models, even when built on inaccurate dependency trees, especially in target languages with different word orders than the source language.

2 Background

2.1 Universal Proposition Bank

Universal Proposition Bank (UPB) is a dataset containing SRL annotations across languages. UPB v1.0 (Akbik et al., 2015, 2016) provides SRL annotations for nine treebanks and eight languages, including English and other seven languages shown at the left side of Table 1. UPB v2.0 (Jindal et al., 2022) provides SRL annotations for 43 treebanks and 23 languages, shown in Table 1. UPB is annotated semi-automatically through filtered annotation projection and bootstrap training (Akbik et al., 2015). UPB v2.0 has significantly improved over UPB v1.0 regarding SRL annotation quality, language scope, and availability of span-based SRL annotations (Jindal et al., 2022).

v1.0 and v2.0	**v2.0**	
Chinese (ZH)	Czech (CS)	Dutch (NL)
Finnish (FI)	Greek (EL)	Polish (PL)
Italian (IT)	Korean (KO)	Telugu (TE)
Spanish (ES)	Romanian (RO)	Indonesian (ID)
French (FR)	Hindi (HI)	Japanese (JA)
German (DE)	Marathi (MR)	Russian (RU)
Portuguese (PT)	Tamil (TA)	Ukrainian (UK)
	Hungarian (HU)	Vietnamese (VI)

Table 1: List of target languages available in UPB v1.0 and UPB v2.0.

2.2 Universal Dependencies

Universal Dependencies is a dataset that contains consistent syntactic annotations across languages, i.e., part-of-speech (POS) tags, morphological features, and syntactic dependencies. UD v1 (Nivre et al., 2016b) and UD v2 (Nivre et al., 2020) have different annotation schemes[1] in word segmentation, pos tags, morphological features, and syntactic relations. UD v1 and UD v2 have 40 and 37 universal dependencies relations, respectively. UPB v1.0 and UPB v2.0 that we use throughout the experiments are annotated based on UD v1.4 (Nivre et al., 2016a) and UD v2.9 (Zeman et al., 2021), respectively.

2.3 Dependency-based Semantic Role Labeling

Dependency-based SRL labels the argument heads for each predicate in a sentence based on its dependency tree. For example, in Figure 1, the phrase "to anyone" is the argument "ARG2" of the predicate "recommend". Based on the dependency tree, "anyone" is the head of the phrase "to anyone". Therefore, dependency-based SRL annotates the edge that connects "recommend" and "anyone" with "ARG2" label.

2.4 Related Works

Existing models in cross-lingual SRL rely on BiLSTMs as their main encoders despite the findings in Ahmad et al. (2019), which prove that RNN-based models perform ineffectively in distant languages for dependency parsing task. Fei

[1] https://universaldependencies.org/v2/summary.html

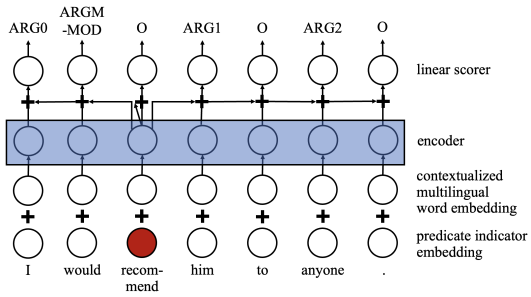

Figure 2: The architecture applied to a sentence with "recommend" as the predicate.

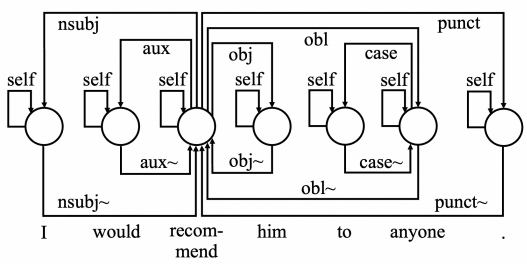

Figure 3: Dependency graph of a sentence converted from its dependency tree.

et al. (2020) propose parameter generation network (PGN)-BiLSTMs to build a cross-lingual SRL model. Cai and Lapata (2020) propose BiLSTM-based models as semantic role labeler and compressor in their architecture. Conia et al. (2021) propose BiLSTM-based universal sentence encoder and BiLSTM-based universal predicate-argument encoder to encode predicate-related and predicate-argument information.

GNNs have been used to encode dependency trees in monolingual SRL and aspect-based sentiment analysis (ABSA). In monolingual SRL, Marcheggiani and Titov (2017) employ GGCNs on top of BiLSTMs to incorporate dependency trees as graphs. In ABSA, Wang et al. (2020) and Jiang et al. (2021) employ relational GATs (R-GATs) and attention-based relational GCNs (ARGCNs), respectively. They apply GNNs on top of modified dependency trees to establish direct connections between aspects and their corresponding words.

3 Model

We apply a common encoder-decoder architecture for comparing various GNN-based and BiLSTM-based cross-lingual SRL models. The architecture consists of an input layer (i.e., predicate indicator embedding and contextualized multilingual word embedding), an encoder, and a decoder (i.e., linear scorer), as shown in Figure 2.

3.1 Input Layer

For each word in a sentence, we concatenate predicate indicator embedding, $\overrightarrow{p_i}$, and contextualized multilingual word embedding, $\overrightarrow{c_i}$, to produce the final word representation, $\overrightarrow{h_i}$, as shown in Equation 1.

$$\overrightarrow{h_i} = [\overrightarrow{p_i}||\overrightarrow{c_i}] \quad (1)$$

Predicate indicator embedding is an embedding that represents whether a word is a predicate or not (Fei et al., 2020). We compare contextualized multilingual word embedding from two language models, i.e., multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020). Both mBERT and XLM-R produce embedding at a subword level according to WordPiece tokenization (Wu et al., 2016). We take the left-most subword as the representation for the corresponding word following Wang et al. (2019). To generate the word embedding, $\overrightarrow{c_i}$, we adopt the method proposed by existing cross-lingual SRL. Conia et al. (2021) concatenate the last four hidden layers obtained from corresponding pre-trained language models. Given the result of concatenation, $\overrightarrow{v_i}$, they apply a feed-forward neural network (FNN) and Swish activation function (Ramachandran et al., 2018), as shown in Equation 2. They also apply a dropout after the activation function.

$$\overrightarrow{c_i} = Swish(\mathbf{W}\overrightarrow{v_i} + \mathbf{b}) \quad (2)$$

3.2 Encoder

We experiment with various GNN-based and BiLSTM-based encoders. GNNs that we apply are graph attention networks (GATs), two-attention relational GATs (2ATT-GATs), gated graph convolutional networks (GGCNs), and modified self-attention networks from Transformer (SATs).

In GNN-based models, we encode a sentence by forming a dependency graph based on its dependency tree, which consists of dependency arcs and dependency relations. We follow the method proposed by Marcheggiani and Titov (2017). They convert a dependency tree to a graph by adding edges that flow in the opposite direction of the original dependency arcs and edges that flow from nodes to themselves. Figure 3 displays the dependency graph of the dependency tree at the bottom of Figure 1.

3.2.1 Graph Attention Networks

Given a graph that consists of nodes, i, and edges, e, GATs (Veličković et al., 2017) update each node representation, $\overrightarrow{h_i}$, according to its neighbor node representations, using multi-head attention mechanism that employs K heads. GATs utilize an attention weight, α, to measure the contribution of neighbor node representations when updating the corresponding node representation, h_i. The attention weight, α_{ij}, for the edge that connects node i and node j, is calculated by taking the dot-product between a weight vector, $\overrightarrow{\mathbf{a}}$, with the concatenation of linearly transformed $\overrightarrow{h_i}$ and $\overrightarrow{h_j}$. The result of the dot-product is passed to a LeakyReLU activation function, LR, and softmax function, SM, as shown in Equation 3.

$$\alpha_{ij}^k = \mathrm{SM}_j(\mathrm{LR}(\overrightarrow{\mathbf{a}}^k \cdot [\mathbf{W}^k\overrightarrow{h_i}||\mathbf{W}^k\overrightarrow{h_j}])) \quad (3)$$

3.2.2 Two-attention Relational Graph Attention Networks

Since GATs treat a graph as unlabeled (Veličković et al., 2017), Wang et al. (2020) modify GATs to use two attention weights. The first attention weight contains node representations, while the second includes dependency-type representations taken from dependency types that represent dependency arcs and relations in the dependency trees. Instead of using different equations for both attentions as in Wang et al. (2020), we find that applying the FNNs (Veličković et al., 2017) for calculating both attentions works best for our task. Therefore, we explain the modification in this section.

Equation 3 shows how we calculate the first attention weight, α. To calculate the second attention weight, β, we slightly modify Equation 3 to encode dependency-type representation, $\overrightarrow{r_{ij}}$, as shown in Equation 4.

$$\beta_{ij}^k = \mathrm{SM}_j(\mathrm{LR}(\overrightarrow{\mathbf{a}}^k \cdot \mathbf{W}^k\overrightarrow{r_{ij}})) \quad (4)$$

We obtain node representations from attention weight α and β, as shown in Equation 5 and Equation 6. $\mathcal{N}_i$ indicates the set of neighbor nodes of node i. $\overrightarrow{h_{\alpha,i}^l}$ and $\overrightarrow{h_{\beta,i}^l}$ are the node representations in layer l obtained from attention weight α and β, respectively.

$$\overrightarrow{h_{\alpha,i}^l} = \sum_{j\in\mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \overrightarrow{h_j^{l-1}} \quad (5)$$

$$\overrightarrow{h_{\beta,i}^l} = \sum_{j\in\mathcal{N}_i} \beta_{ij}^k \mathbf{W}^k \overrightarrow{h_j^{l-1}} \quad (6)$$

Finally, we calculate the node representation in layer l, $\overrightarrow{h_i^l}$, by applying an FNN to the concatenation of node representation $\overrightarrow{h_{\alpha,i}^l}$ and $\overrightarrow{h_{\beta,i}^l}$. We concatenate the node representation from each attention head, k (Equation 7), except in the final layer where we take the average of node representations (Equation 8). L indicates the number of layers, while $\mathbf{W}$ and $\mathbf{b}$ indicate weight matrix and bias in the FNN. We optionally apply an activation function, σ.

$$\overrightarrow{h_i^l} = \sigma(||_{k=1}^K(\mathbf{W}^k[\overrightarrow{h_{\alpha,i}^l}||\overrightarrow{h_{\beta,i}^l}] + \mathbf{b}^k)), l < L \quad (7)$$

$$\overrightarrow{h_i^l} = \sigma(\frac{1}{K}\sum_{k=1}^{K}(\mathbf{W}^k[\overrightarrow{h_{\alpha,i}^l}||\overrightarrow{h_{\beta,i}^l}] + \mathbf{b}^k)), l = L \quad (8)$$

3.2.3 Gated Graph Convolutional Networks

Marcheggiani and Titov (2017) propose GGCNs to encode syntactic features from dependency trees in monolingual SRL models. Equation 9 shows how to calculate node representation, $\overrightarrow{h_i^l}$, in layer l. GGCNs separate the weight matrices, $\mathbf{W}_{dir}$, according to the direction of dependency arcs, i.e., original direction, opposite direction, and self-direction. Meanwhile, $\mathbf{b}_{rel}$ represents the dependency relation. Unlike regular GCNs (Kipf and Welling, 2017), GGCNs employ a scalar gate, g_{ij}, to measure the importance of neighbor node representations when updating the corresponding node representation.

$$\overrightarrow{h_i^l} = \mathrm{ReLU}(\sum_{j\in\mathcal{N}_i} g_{ij}(\mathbf{W}_{dir_{ij}}\overrightarrow{h_i^{l-1}} + \mathbf{b}_{rel_{ij}})) \quad (9)$$

3.2.4 Modified Self-attention Networks

Self-attention with relative position representations (Shaw et al., 2018) is an extension of the self-attention in Transformer (Vaswani et al., 2017) to include edges that represent relative positions between words. The self-attention can be categorized as GNNs since they receive input as a graph consisting of nodes and edges. Referring to the paper (Shaw et al., 2018), we modify edge representations from representing relative positions to representing dependency types. Instead of a fully connected graph as input, SATs take a graph formed over the sentence's dependency tree, as shown in Figure 3.

3.3 Decoder

We apply a linear scorer as the decoder. We concatenate node representation for each word, $\overrightarrow{h_i}$, with the predicate node representation, $\overrightarrow{h_{p,i}}$. Predicate node representation is taken from the node representation, $\overrightarrow{h_i}$, of the sentence's predicate. After that, we apply an FNN to produce the final node representation with an embedding size equal to the number of arguments, n. We then apply a softmax function, SM, to produce the probability for each label, z, as shown in Equation 10. We train the model to minimize the cross-entropy loss.

$$P(z) = \mathrm{SM}(\mathbf{W}[\overrightarrow{h_i}||\overrightarrow{h_{p,i}}] + \mathbf{b})_z, z \in [1, n] \quad (10)$$

4 Experiment Results

4.1 Datasets

We conduct experiments using datasets from UPB v1.0 and UPB v2.0. The dataset distribution can be found in Appendix A.1. In UPB v1.0, `UP_English-EWT` is annotated based on UD v2, while the other languages are annotated based on UD v1.4. Since `UP_English-EWT` in UPB v1.0 is annotated based on UD v2, we use its SRL annotations to construct a cross-lingual SRL model for target languages in UPB v2.0. To construct a cross-lingual SRL model for UPB v1.0, we convert syntactic annotations of `UP_English-EWT` (UPB v1.0) to UD v1.4 using the script available at Zhang et al. (2021)[2]. We merge annotations from English Web Treebank (EWT) (Bies et al., 2012), PropBank v3 (Kingsbury and Palmer, 2002; Palmer et al., 2005; Gildea and Palmer, 2002), and UD v1.4 (Nivre et al., 2016a).

4.2 Settings

We focus on conducting experiments in a zero-shot setting to examine the model's transferability across languages. We train the model in English and evaluate the model in seven languages and 23 languages from UPB v1.0 and UPB v2.0, respectively, as shown in Table 1. Furthermore, we compare the performance of the cross-lingual model against the monolingual model to illustrate the generalization achieved by each model.

We use gold and predicted dependency trees for model evaluation. We train dependency parsers from scratch using Stanza (Qi et al., 2020) to produce predicted dependency trees for languages in UPB v1.0. For languages in UPB v2.0, we use pre-trained models provided by Stanza[3]. Appendix B.1 shows each dependency parser's unlabeled attachment score (UAS) and labeled attachment score (LAS).

Following Conia et al. (2021), we train the model for 30 epochs by increasing the learning rate linearly for 1 epoch and decreasing linearly for 15 epochs. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer. We run the experiments five times and report the average F1 scores and standard errors. We choose the best epoch from each experiment based on the F1 score of the English validation set. Appendix B.2 explains the implementation details and hyperparameters we use throughout the experiments. We choose most hyperparameter values based on previous studies.

4.3 Comparison Among GNN-based SRL Models

We compare the performance of various GNN-based SRL models, i.e., GGCNs, SATs, GATs, and 2ATT-GATs, by substituting the encoder part in the architecture (Figure 2) with these GNNs. We stack three layers of GNNs and freeze mBERT as our contextualized word embedding to solely observe the influence of GNNs.

The left side of Table 2 shows the F1 score of each model evaluated using the test sets in UPB v1.0 and UPB v2.0 with predicted dependency trees. The results with gold dependency trees are available in Appendix C. In the monolingual setting, as indicated by EN, SATs achieve the highest F1 score for both datasets. Meanwhile, in the cross-lingual setting, as shown in AVG, GGCNs perform the best in UPB v1.0, while 2ATT-GATs perform the best in UPB v2.0.

Although GGCNs outperform 2ATT-GATs in UPB v1.0, if we extract the F1 scores of languages available in UPB v1.0 from UPB v2.0, i.e., FI, IT, ES, FR, DE, PT, and ZH, 2ATT-GATs still outperform GGCNs in all these languages. As we know, UPB v1.0 and UPB v2.0 are annotated semi-automatically (Akbik et al., 2015). UPB v2.0 has significantly improved over UPB v1.0 regarding SRL annotations quality (Jindal et al., 2022). Therefore, we believe the evaluation against UPB v2.0 is more crucial.

[2]https://github.com/zzsfornlp/zmsp

[3]Except for `UP_Japanese-GSDLUW` and `UP_French-Rhapsodie`, where we train from scratch since the models are unavailable.

Lang	GGCNs	SATs	GATs	2ATT-GATs	BiLSTM+ 2ATT-GATs	BiLSTM	3BiLSTMs
UPB v1.0							
EN	80.50±0.09	**80.89±0.04**	79.77±0.05	80.04±0.05	82.19±0.10	79.72±0.05	80.23±0.08
FI	**54.10±0.09**	53.95±0.17	48.95±0.32	53.17±0.10	52.83±0.15	39.85±0.19	40.18±0.22
IT	**57.57±0.13**	57.33±0.14	54.91±0.26	57.12±0.08	56.45±0.10	47.20±0.19	46.74±0.30
ES	54.32±0.11	**54.88±0.17**	50.44±0.15	54.69±0.12	52.41±0.16	41.77±0.15	41.52±0.05
FR	**46.92±0.12**	46.18±0.12	44.60±0.28	46.28±0.20	45.58±0.23	40.64±0.32	40.70±0.31
DE	58.75±0.21	58.65±0.05	57.06±0.10	**58.80±0.07**	56.18±0.13	41.45±0.24	41.99±0.12
PT	**53.60±0.04**	53.12±0.13	52.50±0.13	53.33±0.07	52.86±0.16	44.38±0.24	44.10±0.12
ZH	37.83±0.14	37.54±0.26	37.07±0.11	**38.42±0.16**	40.01±0.24	32.87±0.37	32.53±0.23
AVG	**51.87±0.05**	51.67±0.09	49.36±0.11	51.69±0.07	50.90±0.07	41.17±0.13	41.11±0.11
UPB v2.0							
EN	80.20±0.08	**80.71±0.06**	79.63±0.08	79.65±0.02	81.99±0.10	79.51±0.08	80.08±0.08
CS	56.60±0.20	56.86±0.12	55.93±0.06	**57.79±0.11**	55.19±0.23	47.13±0.06	47.50±0.13
EL	58.39±0.30	58.56±0.52	59.68±0.27	**60.69±0.36**	59.48±0.21	56.05±0.16	55.53±0.27
KO	39.31±0.24	38.26±0.77	41.25±0.25	**42.20±0.45**	35.01±0.54	27.75±0.29	27.84±0.99
RO	52.43±0.07	52.65±0.19	**53.51±0.20**	53.24±0.17	54.41±0.09	48.71±0.30	47.83±0.31
HI	47.26±0.07	47.93±0.08	44.80±0.22	**48.40±0.14**	42.69±0.44	27.49±0.84	27.30±0.44
MR	35.43±0.59	37.86±0.86	**39.53±1.70**	38.94±1.01	36.17±0.96	27.22±1.89	26.31±1.60
TA	31.64±0.27	33.74±0.54	**34.58±0.66**	34.25±0.40	30.70±0.75	23.64±0.84	22.20±0.88
HU	50.08±0.27	50.90±0.26	48.27±0.14	**51.00±0.12**	46.65±0.21	39.05±0.29	39.52±0.39
PL	57.73±0.12	57.54±0.08	57.57±0.14	**59.85±0.16**	57.01±0.33	51.65±0.11	51.19±0.33
TE	46.27±0.64	45.09±0.55	42.66±0.72	**47.01±0.82**	40.12±0.99	35.97±1.30	32.84±0.55
NL	62.94±0.20	62.81±0.21	63.00±0.19	**63.39±0.15**	58.39±0.28	51.95±0.36	52.42±0.10
ID	53.35±0.34	53.43±0.23	**58.82±0.25**	54.24±0.10	56.80±0.32	56.55±0.34	55.00±0.35
JA	36.07±0.15	36.12±0.39	36.30±0.51	**36.82±0.11**	33.39±1.08	23.76±1.07	21.40±1.07
RU	57.63±0.22	58.69±0.35	**59.43±0.27**	58.88±0.23	58.95±0.29	55.49±0.24	55.48±0.23
UK	57.39±0.04	58.45±0.16	57.18±0.26	**58.62±0.12**	57.59±0.27	52.80±0.39	52.98±0.17
ZH	42.88±0.33	44.13±0.31	**45.55±0.33**	44.77±0.16	47.73±0.19	48.54±0.46	47.95±0.34
VI	25.78±0.05	26.11±0.14	**27.48±0.03**	26.51±0.14	27.32±0.08	29.39±0.29	27.34±0.36
FI	53.54±0.11	54.09±0.13	52.92±0.13	**54.46±0.04**	54.37±0.17	51.54±0.13	51.83±0.23
IT	57.31±0.10	57.81±0.15	58.00±0.12	**58.52±0.07**	58.14±0.19	55.23±0.16	54.97±0.12
ES	54.10±0.06	54.85±0.10	**55.82±0.06**	55.68±0.09	55.02±0.21	51.02±0.13	50.51±0.06
FR	61.55±0.14	61.82±0.10	62.05±0.19	**62.43±0.04**	60.60±0.25	59.81±0.15	59.88±0.26
DE	58.34±0.06	58.80±0.14	59.34±0.08	**59.39±0.07**	53.92±0.31	41.40±0.38	43.10±0.23
PT	65.55±0.08	65.82±0.10	**66.37±0.07**	66.09±0.07	65.85±0.11	64.05±0.17	63.71±0.10
AVG	50.50±0.04	50.97±0.13	51.31±0.08	**51.88±0.09**	49.81±0.16	44.62±0.11	44.11±0.23

Table 2: F1 scores (%) in UPB v1.0 and UPB v2.0 test sets with predicted dependency trees (frozen mBERT). The bold score in each language indicates the highest F1 score among GNN-based models, i.e., GGCNs, SATs, GATs, and 2ATT-GATs. The underlined score in each language indicates the highest F1 score among 2ATT-GATs, BiLSTM+2ATT-GATs, BiLSTM, and 3BiLSTMs. AVG indicates the average F1 score of each model for all languages except English.

In UPB v2.0, 2ATT-GATs outperform other GNNs, indicating that 2ATT-GATs can generalize better across languages. However, GATs sometimes perform better in languages that have relatively low LAS, i.e., MR, TA, ID, VI, and RU[4]. We conjecture that GATs are more robust to inaccurate dependency relation labels because GATs treat dependency trees as unlabeled. Clearer evidence can be shown in ID, where the F1 scores decrease significantly when we replace gold with predicted dependency trees except for GATs[5] because the UAS of predicted dependency trees in ID is relatively high compared to LAS, i.e., 87.31 and 77.33.

[4]UAS/LAS of each language: 79.85/70.63 (MR), 80.89/72.30 (TA), 87.31/77.33 (ID), 77.58/74.16 (VI), 84.42/81.41 (RU Taiga), and 90.44/87.2 (RU GSD).

[5]F1 scores of ID using gold trees: 57.18±0.26 (GGCNs), 57.05±0.20 (SATs), 59.05±0.21 (GATs), and 58.06±0.13 (2ATT-GATs).

4.4 Comparison of 2ATT-GATs Against BiLSTM-based SRL Models

We compare 2ATT-GATs with the widely used network in cross-lingual SRL, i.e., BiLSTMs. We build the first BiLSTM-based SRL model by stacking one BiLSTM layer with two layers of 2ATT-GATs (BiLSTM+2ATT-GATs). We also compare 2ATT-GATs with syntax-agnostic models, i.e., a model with one layer of BiLSTM as the encoder (BiLSTM) and a model with three layers of BiLSTMs as the encoder (3BiLSTMs).

The right side of Table 2 shows the F1 score of each model. The results with gold dependency

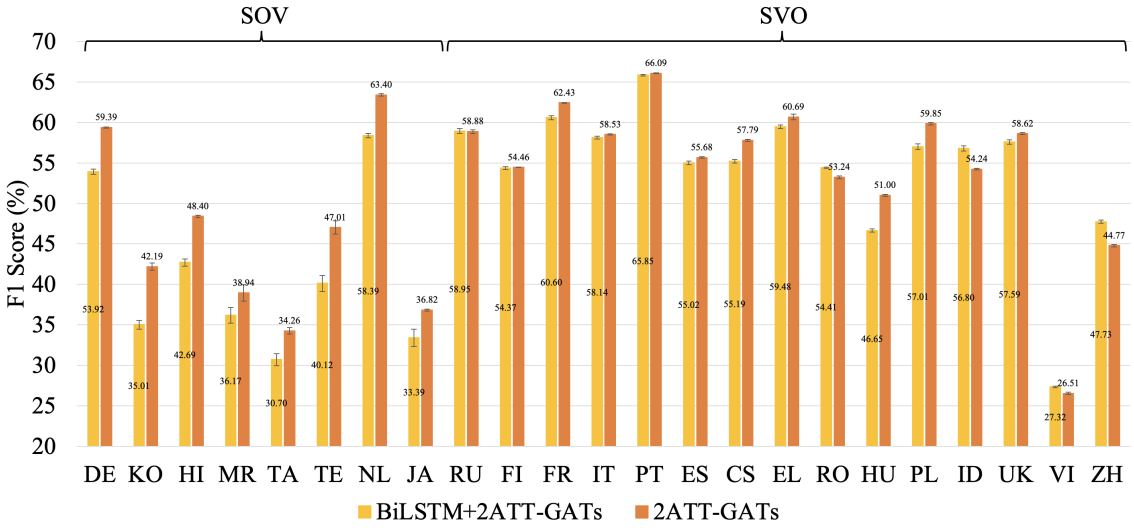

Figure 4: F1 scores (%) in UPB v2.0 test set using predicted dependency trees. Each language has two bars, i.e., the left bar indicates the F1 score for BiLSTM+2ATT-GATs, and the right bar indicates the F1 score for 2ATT-GATs.

trees are available in Appendix C. From EN, we can see that BiLSTM+2ATT-GATs perform best in the monolingual setting. BiLSTM+2ATT-GATs also perform better than syntax-agnostic models, i.e., BiLSTM and 3BiLSTMs, in the cross-lingual setting. This indicates that the help of syntax improves performance in both settings even though the syntax provided is not accurate.

Among the syntax-aware models, 2ATT-GATs perform the best in the cross-lingual setting. The context for each word learned through BiLSTMs, i.e., what words precede and follow each word in a sentence, might be too specific to the language it is trained with. By encoding the sentence over its dependency tree, we let information flow based on the sentence's grammatical structure, which is more universal across languages.

Figure 4 compares the F1 scores of BiLSTM+2ATT-GATs and 2ATT-GATs for each language in UPB v2.0. We know that the models are trained in English that has a subject-verb-object (SVO) word order. If we look deeper, for certain languages with subject-object-verb (SOV) word order, i.e., DE, KO, HI, MR, TA, TE, NL, and JA, 2ATT-GATs show significant improvements over BiLSTM+2ATT-GATs. This proves that 2ATT-GATs have better transferability to languages with diverse word orders than the BiLSTM-based models.

BiLSTM+2ATT-GATs sometimes perform better than 2ATT-GATs in languages with the same word order as English, i.e., SVO word order, including RU, RO, ID, VI, and ZH. We confirm that in RU, the predicted dependency trees decrease the F1 score a little more significantly for 2ATT-GATs[6]. For the other languages, the quality of the dependency trees (especially the dependency relations), which is relatively low in the original dataset, might be why 2ATT-GATs perform worst.

4.5 Obtaining the Best Model

We conduct thorough experiments with the number of layers and contextualized multilingual word embedding that we use. Using frozen mBERT as contextualized multilingual word embedding, we compare the result of stacking one layer, two layers, and three layers of 2ATT-GATs as the encoder. We find that two layers of 2ATT-GATs and three layers of 2ATT-GATs give the best F1 score in UPB v1.0 and UPB v2.0, respectively.

We further experiment with two and three layers of 2ATT-GATs using mBERT and XLM-R as the contextualized multilingual word embeddings and fine-tune them. We provide the evaluation results in Appendix C. In UPB v1.0, a combination of frozen mBERT and two layers of 2ATT-GATs gives the best F1 score. Meanwhile, in UPB v2.0, fine-tuned XLM-R with two layers of 2ATT-GATs gives the best F1 score.

We further analyze why two layers of 2ATT-GATs perform better than three layers of 2ATT-GATs. In Figure 5, we group the F1 score difference between two 2ATT-GATs and three 2ATT-GATs for each dependency range, i.e., the number

[6]F1 scores of RU using gold trees: 59.93±0.18 (2ATT-GATs) and 59.59±0.36 (BiLSTM+2ATT-GATs).

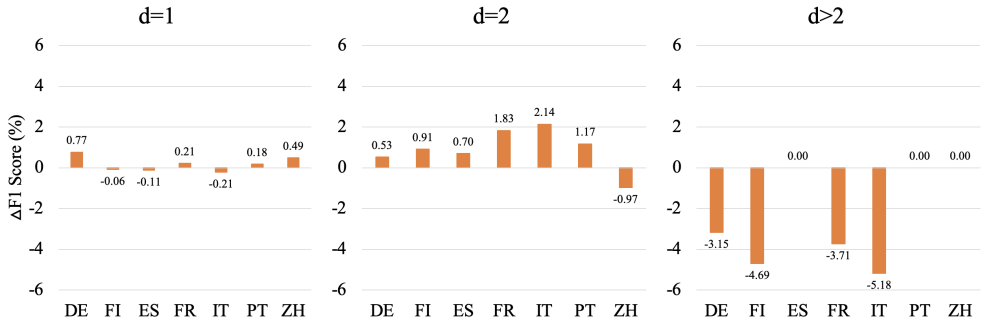

Figure 5: F1 score differences (%) between 2layers+mBERT and 3layers+mBERT in UPB v1.0 dev set (predicted).

of edges lies between the predicate and its argument. The evaluation shows that 2ATT-GATs perform better than three 2ATT-GATs because they perform better in $d = 1$ and $d = 2$, and the dependency range (d) in languages mostly lies in 1-2.

4.6 Comparison of the Best Model Against Existing Works

We compare the best model on UPB v1.0 with existing works, i.e., Fei et al. (2020) and Zhang et al. (2021), as shown in Table 3. Fei et al. (2020) translate and project the SRL annotations from the source language to the target language to be included as part of the training set. Zhang et al. (2021) employ multi-task learning consisting of dependency parsing task and SRL task and include the dependency parsing task of the source and target language as part of the training process. Since both previous works have not been evaluated using UPB v2.0, we compare the F1 score in UPB v1.0. The evaluation results of our best model in UPB v2.0 can be seen in Appendix C.

It can be seen in Table 3 that 2ATT-GATs underperform previous models. Since our goal is to find the most transferable networks to build the cross-lingual model, unlike the other models, we do not include any knowledge of sentences from the target language in the training process. Our work can complement these previous works and provide important insights regarding the architecture design for building future cross-lingual SRL models.

5 Conclusions and Future Works

Through a simple encoder-decoder architecture, we show that GNNs are better than BiLSTMs for building cross-lingual SRL models, especially in distant languages. Encoding sentences based on their dependency trees helps create a more generalized cross-lingual SRL model rather than using word

Lang	2layers+ mBERT (gold)	2layers+ mBERT (predicted)	Fei	Zhang
FI	57.47±0.24	53.82±0.21	54.5	**59.9**
FR	48.80±0.28	46.27±0.25	**64.8**	56.6
DE	62.50±0.24	58.78±0.13	**65.0**	60.2
IT	**60.96±0.10**	57.73±0.13	58.7	60.6
PT	57.21±0.04	53.41±0.12	56.0	**59.5**
ES	57.74±0.10	54.47±0.07	**62.5**	57.3
ZH	43.03±0.10	38.87±0.11	-	-

Table 3: F1 scores (%) in UPB v1.0 test set.

sequences. Furthermore, through empirical experiments comparing four types of GNNs using 23 languages available in UPB, we conclude that two-attention relational GATs are the most effective GNNs.

In the future, we can extend our model to incorporate language-specific information to distinguish characteristics between languages. This can be useful for training the model in a few-shot setting where we include the target sentences in the training set. Furthermore, in the experiments, we compare two-attention relational GATs with modified self-attention networks from Transformer. We can further analyze the effect of incorporating two-attention relational GATs to replace self-attention in Transformer in the cross-lingual SRL domain.

Limitations

The limitation of this work is that we focus on argument detection and argument labeling in cross-lingual SRL, assuming that the sentences' gold predicates are easy to obtain. Furthermore, we focus on conducting experiments in a zero-shot setting. The availability of target sentences in the training set might affect the models' behavior, which should be investigated further.

Ethics Statement

We believe there is no ethical issue raised in this work. SRL is a low-level task to support other advanced NLP applications. Therefore, increasing the coverage of SRL models in various languages is beneficial for developing NLP tools that can help solve the problems in this diverse society.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition Banks for multilingual semantic role labeling. 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)*, pages 397–407, Beijing, China. Association for Computational Linguistics.

Alan Akbik, Vishwajeet Kumar, and Yunyao Li. 2016. Towards semi-automatic generation of proposition Banks for low-resource languages. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 993–998, Austin, Texas. Association for Computational Linguistics.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank.

Rui Cai and Mirella Lapata. 2020. Alignment-free cross-lingual semantic role labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3883–3894, Online. Association for Computational Linguistics.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 52–60, Los Angeles, California. Association for Computational Linguistics.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations*.

Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021. Unifying cross-lingual semantic role labeling with heterogeneous linguistic resources. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–351, Online. Association for Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.

Daniel Gildea and Martha Palmer. 2002. The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 239–246, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Aleksa Gordić. 2020. pytorch-gat. `https://github.com/gordicaleksa/pytorch-GAT`.

Junfeng Jiang, An Wang, and Akiko Aizawa. 2021. Attention-based relational graph convolutional network for target-oriented opinion words extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1986–1997, Online. Association for Computational Linguistics.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Ha Linh, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal Proposition Bank 2.0. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1700–1711, Marseille, France. European Language Resources Association.

Atif Khan, Naomie Salim, and Yogan Jaya Kumar. 2015. A framework for multi-document abstractive summa-

rization based on semantic role labelling. *Applied Soft Computing*, 30:737–747.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Eckhard Bick, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Claudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lucia Lam, Phuong Lê Hồng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Robert Östling, Lilja Øvrelid, Valeria Paiva, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Mats Wirén, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2016a. Universal dependencies 1.4. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016b. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. 2018. Searching for activation functions.

Reinhard Rapp. 2022. Using semantic role labeling to improve neural machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3079–3083, Marseille, France. European Language Resources Association.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238, Online. Association for Computational Linguistics.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Aibek

Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, Lorena Martín-Rodríguez, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2021. Universal dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2021. On the benefit of syntactic supervision for cross-lingual transfer in semantic role labeling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6229–6246, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lang	Train	Dev	Test
UPB v1.0			
English (EN)	12,542	1,974	2,060
Finnish (FI)	12,217	716	648
Italian (IT)	12,837	489	489
Spanish (ES)	28,492	3,206	1,995
French (FR)	14,553	1,596	298
German (DE)	14,118	799	977
Portuguese (PT)	7,494	938	936
Chinese (ZH)	3,997	500	500
UPB v2.0			
English (EN)	12,542	1,974	2,062
Czech (CS)	102,993	11,311	12,203
Greek (EL)	1,662	403	456
Korean (KO)	27,410	3,016	3,276
Romanian (RO)	35,911	2,247	2,272
Hindi (HI)	13,304	1,659	1,684
Marathi (MR)	373	46	47
Tamil (TA)	400	80	120
Hungarian (HU)	910	441	449
Polish (PL)	31,496	3,960	3,942
Telugu (TE)	1,051	131	146
Dutch (NL)	18,078	1,394	1,472
Indonesian (ID)	4,482	559	557
Japanese (JA)	14,100	1,014	1,086
Russian (RU)	19,894	1,525	1,482
Ukrainian (UK)	5,496	672	892
Chinese (ZH)	3,997	500	500
Vietnamese (VI)	1,400	800	800
Finnish (FI)	27,198	3,239	3,422
Italian (IT)	29,685	2,277	2,518
Spanish (ES)	28,474	3,054	2,147
French (FR)	17,968	2,970	1,712
German (DE)	166,849	19,233	19,436
Portuguese (PT)	16,633	2,376	2,367

Table 4: Number of sentences available for each language in UPB v1.0 and UPB v2.0.

Treebank	License
UD v1.4	
UD_English	CC BY-SA 4.0
UD_Chinese	CC BY-NC-SA 4.0
UD_Finnish	CC BY-SA 4.0
UD_Spanish	CC BY-NC-SA 3.0 US
UD_Spanish-AnCora	GNU GPL 3.0
UD_French	CC BY-NC-SA 3.0 US
UD_German	CC BY-NC-SA 3.0 US
UD_Italian	CC BY-NC-SA 3.0
UD_Portuguese-Bosque	CC BY-SA 4.0
UD v2.9	
UD_English-EWT	CC BY-SA 4.0
UD_Chinese-GSD	CC BY-SA 4.0
UD_Czech-CAC	CC BY-SA 4.0
UD_Czech-CLTT	CC BY-SA 4.0
UD_Czech-FicTree	CC BY-NC-SA 4.0
UD_Czech-PDT	CC BY-NC-SA 3.0
UD_Dutch-Alpino	CC BY-SA 4.0
UD_Dutch-LassySmall	CC BY-SA 4.0
UD_Finnish-FTB	CC BY 4.0
UD_Finnish-TDT	CC BY-SA 4.0
UD_French-GSD	CC BY-SA 4.0
UD_French-Rhapsodie	CC BY-SA 4.0
UD_French-Sequoia	LGPL-LR
UD_German-GSD	CC BY-SA 4.0
UD_German-HDT	CC BY-SA 4.0
UD_Greek-GDT	CC BY-NC-SA 3.0
UD_Hindi-HDTB	CC BY-NC-SA 4.0
UD_Hungarian-Szeged	CC BY-NC-SA 3.0
UD_Indonesian-GSD	CC BY-SA 4.0
UD_Italian-ISDT	CC BY-NC-SA 3.0
UD_Italian-ParTUT	CC BY-NC-SA 4.0
UD_Italian-PoSTWITA	CC BY-NC-SA 4.0
UD_Italian-TWITTIRO	CC BY-SA 4.0
UD_Italian-VIT	CC BY-NC-SA 3.0
UD_Japanese-GSD	CC BY-SA 4.0
UD_Japanese-GSDLUW	CC BY-SA 4.0
UD_Korean-GSD	CC BY-SA 4.0
UD_Korean-Kaist	CC BY-SA 4.0
UD_Marathi-UFAL	CC BY-SA 4.0
UD_Polish-LFG	GNU GPL 3.0
UD_Polish-PDB	CC BY-NC-SA 4.0
UD_Portuguese-Bosque	CC BY-SA 4.0
UD_Portuguese-GSD	CC BY-SA 4.0
UD_Romanian-Nonstandard	CC BY-SA 4.0
UD_Romanian-RRT	CC BY-SA 4.0
UD_Romanian-SiMoNERo	CC BY-SA 4.0
UD_Russian-GSD	CC BY-SA 4.0
UD_Russian-Taiga	CC BY-SA 4.0
UD_Spanish-AnCora	CC BY 4.0
UD_Spanish-GSD	CC BY-SA 4.0
UD_Tamil-TTB	CC BY-NC-SA 3.0
UD_Telugu-MTG	CC BY-SA 4.0
UD_Ukrainian-IU	CC BY-NC-SA 4.0
UD_Vietnamese-VTB	CC BY-SA 4.0

Table 5: License for each treebank in UD v1.4 and UD2.9 that we use in the experiments.

A Artifacts

A.1 Dataset Distribution

Table 4 shows the dataset distribution in Universal Proposition Bank (UPB). Since we run our experiments in a zero-shot setting, we only use the dev set and test set for languages other than English.

A.2 Licenses

Complete UPB v1.0 and UPB v2.0 contain annotations from Universal Dependencies (UD) v1.4 and UD v2.9. Therefore, we provide the license for each UD treebank in Table 5. Despite UD's inherited license, UPB v1.0 and UPB v2.0 also have a CDLA-Sharing-1.0 license. We also retrieve annotations from PropBank v3, English Web Treebank, and UD v1.4 to form English SRL annotations based on UD v1.4. Therefore, we provide the license for PropBank v3, i.e., CC BY-SA 4.0, and English Web Treebank, i.e., LDC User Agreement for Non-Members.

We refer to publicly available codes to help preprocess the data and build the model. We provide the list of repositories with their corresponding licenses, i.e., Zhang et al. (2021)[7] (GPL-3.0), https://github.com/UniversalPropositions/tools (Apache-2.0), Gordić (2020)[8] (MIT), and Marcheggiani and Titov (2017)[9] (Apache-2.0).

We access all the resources we mentioned above solely for academic research. We make sure that we obey the intended use for each artifact.

B Training

B.1 Dependency Parsers

We train the dependency parsers using Stanza with a 0.0005 learning rate, $70,000$ max steps, and $10,000$ max steps before stopping. Table 6 shows

[7] https://github.com/zzsfornlp/zmsp
[8] https://github.com/gordicaleksa/pytorch-GAT
[9] https://github.com/diegma/neural-dep-srl

Treebank	Dev		Test	
	UAS	LAS	UAS	LAS
UPB v1.0				
UP_English-EWT	92.18	90.27	91.39	89.48
UP_Chinese	87.28	85.11	88.31	86.13
UP_Finnish	92.15	90.82	90.28	89.04
UP_Spanish	91.68	89.75	91.15	88.95
UP_Spanish-AnCora	94.06	92.81	93.80	92.39
UP_French	92.50	91.08	89.89	87.61
UP_German	92.70	91.13	89.45	87.31
UP_Italian	94.12	92.75	93.70	92.36
UP_Portuguese-Bosque	93.20	92.02	92.45	91.05
UPB v2.0				
UP_English-EWT	92.46	90.86	91.42	89.82
UP_Chinese-GSD	85.11	83.19	87.06	85.13
UP_Czech-CAC	92.97	91.62	93.43	91.68
UP_Czech-CLTT	89.13	86.98	88.32	86.09
UP_Czech-FicTree	94.68	93.11	94.61	92.76
UP_Czech-PDT	93.74	92.24	93.50	91.87
UP_Dutch-Alpino	94.53	92.24	92.87	90.42
UP_Dutch-LassySmall	90.77	87.62	92.12	89.11
UP_Finnish-FTB	93.77	92.29	94.03	92.41
UP_Finnish-TDT	91.97	90.41	92.24	90.74
UP_French-GSD	95.66	94.45	93.47	91.87
UP_French-Rhapsodie	87.75	83.25	86.42	81.88
UP_French-Sequoia	93.54	92.23	93.10	91.70
UP_German-GSD	91.78	88.61	89.65	85.62
UP_German-HDT	95.18	93.64	95.30	93.72
UP_Greek-GDT	91.77	90.43	92.93	91.19
UP_Hindi-HDTB	96.62	94.49	96.68	94.43
UP_Hungarian-Szeged	87.64	84.10	86.72	83.25
UP_Indonesian-GSD	86.49	76.25	87.31	77.33
UP_Italian-ISDT	94.41	92.84	94.37	93.16
UP_Italian-ParTUT	92.76	90.52	93.10	91.40
UP_Italian-PoSTWITA	87.21	83.20	88.33	84.41
UP_Italian-TWITTIRO	87.25	81.64	84.85	79.77
UP_Italian-VIT	90.63	88.82	91.54	89.05
UP_Japanese-GSD	96.09	95.47	95.11	94.21
UP_Japanese-GSDLUW	96.12	95.82	95.35	95.12
UP_Korean-GSD	88.22	85.41	89.65	87.07
UP_Korean-Kaist	91.35	90.39	90.41	89.45
UP_Marathi-UFAL	74.55	64.32	79.85	70.63
UP_Polish-LFG	97.56	96.73	97.80	96.92
UP_Polish-PDB	94.17	92.69	94.58	93.16
UP_Portuguese-Bosque	94.25	92.51	94.85	93.54
UP_Portuguese-GSD	94.44	93.34	94.21	93.23
UP_Romanian-Nonstandard	93.18	90.04	91.43	87.75
UP_Romanian-RRT	91.96	88.60	91.93	88.45
UP_Romanian-SiMoNERo	93.38	91.21	93.78	91.86
UP_Russian-GSD	90.55	87.80	90.44	87.21
UP_Russian-Taiga	83.94	79.32	84.42	81.41
UP_Spanish-AnCora	93.83	92.16	93.82	92.00
UP_Spanish-GSD	91.91	89.79	91.93	89.58
UP_Tamil-TTB	81.24	73.48	80.89	72.30
UP_Telugu-MTG	92.90	86.25	93.07	85.58
UP_Ukrainian-IU	91.14	89.34	90.10	88.24
UP_Vietnamese-VTB	78.92	74.99	77.58	74.16

Table 6: UAS and LAS of each treebank's dependency parser.

UAS and LAS for each treebank in UPB v1.0 and UPB v2.0.

B.2 SRL Models

We build the code for training and evaluation using PyTorch library[10]. We use the transformers library provided by Hugging Face[11] to produce contextualized multilingual word embedding from mBERT (i.e., `bert-base-multilingual-cased`) and XLM-R (i.e., `xlm-roberta-base`).

We refer to Gordić (2020) to implement GATs. Following the implementation, we use ELU (Clevert et al., 2016) as the activation function in each layer, σ, except in the final layer where we use $Swish$, the same activation function that we use in the input layer. We tried to use the same activation function for every layer, but this setting works best in this task. We use the same skeleton as GATs for implementing 2ATT-GATs and SATs, where node and edge dropouts are placed at the beginning of every layer.

In SATs, we experiment with two settings for the edge representations, i.e., a_{ij}. First, we used the same edge representation for each attention head, k. Then, we used different edge representations for each attention head. The results show that using the same edge representation in each attention head works best for this task. Therefore, we use this setting when comparing SATs with other GNNs.

For the implementation of GGCNs, we refer to the original implementation, i.e., Marcheggiani and Titov (2017)[12], and its reimplementation in PyTorch[13]. Following the reimplementation in PyTorch, we use LeakyReLU as the activation function for GGCNs in each layer, except in the final layer where we use $Swish$. We place the dropout at the end of every layer.

[10]https://pytorch.org/

[11]https://huggingface.co/

Hyperparameter	Value
learning rate	10^{-3}
minimum learning rate	10^{-5}
weight decay	10^{-4}
word embedding learning rate	10^{-5}
word embedding minimum learning rate	10^{-6}
word embedding weight decay	10^{-4}
batch size	32
epochs	30
warmup epochs	1
cooldown epochs	15
number of attention heads, K	8
predicate indicator embedding size, $\overrightarrow{p_i}$	64
multilingual word embedding size, $\overrightarrow{c_i}$	512
edge embedding size, $\overrightarrow{r_{ij}}$	64
LSTM hidden size	512
GATs, SATs, and 2ATT-GATs hidden size	64
GATs, SATs, and 2ATT-GATs output size	same as input (576)
GGCNs hidden size	same as input (576)
GGCNs output size	same as input (576)
dropout in input layer	0.2
dropout before decoder	0.3
GATs best node and edge dropout	0.2
SATs best node and edge dropout	0.2
2ATT-GATs best node and edge dropout	0.3
GGCNs best dropout	0.5

Table 7: Hyperparameter values for cross-lingual and monolingual SRL models.

[12]https://github.com/diegma/neural-dep-srl

[13]https://github.com/kdrivas/Graph-convolutional

Lang	Gold	Predicted
CS	61.44±0.08	61.42±0.07
EL	68.80±0.09	67.90±0.16
KO	46.76±0.19	47.06±0.20
RO	55.55±0.05	55.36±0.04
HI	48.73±0.15	48.97±0.16
MR	41.78±0.41	42.19±0.80
TA	40.73±0.34	37.79±0.52
HU	54.86±0.07	53.94±0.10
PL	62.85±0.10	63.01±0.11
TE	56.51±0.47	56.19±0.46
NL	65.36±0.14	64.69±0.11
ID	63.83±0.13	63.22±0.07
JA	37.82±0.75	38.22±0.80
RU	62.40±0.19	62.02±0.16
UK	60.85±0.17	60.63±0.15
ZH	42.33±0.72	41.68±0.75
VI	28.80±0.14	29.30±0.15
FI	58.28±0.10	57.95±0.11
IT	60.82±0.10	60.53±0.07
ES	57.97±0.09	58.29±0.07
FR	64.38±0.10	63.65±0.11
DE	61.35±0.11	60.86±0.09
PT	67.75±0.07	67.31±0.07

Table 8: F1 scores (%) in UPB v2.0 test set with gold dependency trees (left) and predicted dependency trees (right) (2layers+XLM-R+fine-tuned).

We provide the hyperparameter values we use to train the SRL models in Table 7. For GNN-based models, we search for the best dropout in $0.1 - 0.5$ with a 0.1 increment.

We use Tesla P100 to train the models. Training time for GNN-based models ranges from 3-4 hours. Training time for BiLSTM+2ATT-GATs and BiLSTM ranges from 4-5 hours. Meanwhile, for 3BiLSTMs, it takes around 10 hours to train a model. We run the experiment 5 times and produce two models for each setting, i.e., a model trained on UPB v1.0 and a model trained on UPB v2.0. The estimation of total GPU hours is 2,000.

C Supporting Results

Table 9 shows the F1 scores for GNN-based and BiLSTM-based cross-lingual and monolingual SRL models using gold dependency trees.

Table 10 shows the average F1 scores of target languages in dev sets using different numbers of layers, i.e., two layers and three layers, and different contextualized multilingual word embeddings, i.e., mBERT and XLM-R.

Table 8 shows the F1 scores of the best model (2layers+XLM-R+fine-tuned) for each language in the UPB v2.0 test set.

Lang	GGCNs	SATs	GATs	2ATT-GATs	BiLSTM+ 2ATT-GATs	BiLSTM	3BiLSTMs
				UPB v1.0			
EN	83.82±0.06	**84.19±0.05**	82.77±0.04	83.21±0.07	85.35±0.14	79.72±0.05	80.23±0.08
FI	**58.02±0.09**	57.77±0.19	52.07±0.28	56.82±0.08	56.18±0.24	39.85±0.19	40.18±0.22
IT	**61.09±0.12**	60.77±0.23	57.18±0.29	60.52±0.09	59.00±0.10	47.20±0.19	46.74±0.30
ES	57.39±0.12	**58.12±0.17**	52.70±0.18	57.97±0.16	55.02±0.20	41.77±0.15	41.52±0.05
FR	**49.42±0.14**	48.99±0.20	46.04±0.19	48.68±0.17	47.87±0.15	40.64±0.32	40.70±0.31
DE	**62.78±0.27**	62.38±0.05	60.31±0.15	62.63±0.09	59.55±0.22	41.45±0.24	41.99±0.12
PT	**57.60±0.05**	56.90±0.15	55.26±0.17	57.18±0.08	56.35±0.24	44.38±0.24	44.10±0.12
ZH	42.12±0.20	42.00±0.20	39.63±0.14	**42.60±0.16**	43.51±0.14	32.87±0.37	32.53±0.23
AVG	**55.49±0.06**	55.28±0.13	51.88±0.13	55.20±0.08	53.93±0.14	41.17±0.13	41.11±0.11
				UPB v2.0			
EN	83.93±0.08	**84.39±0.06**	82.82±0.09	83.26±0.07	85.37±0.10	79.51±0.08	80.08±0.08
CS	56.42±0.19	56.75±0.11	55.83±0.07	**57.70±0.10**	55.16±0.24	47.13±0.06	47.50±0.13
EL	59.57±0.33	59.64±0.57	60.84±0.26	**61.80±0.35**	60.66±0.21	56.05±0.16	55.53±0.27
KO	39.07±0.22	38.20±0.80	40.83±0.26	**42.14±0.48**	34.94±0.59	27.75±0.29	27.84±0.99
RO	52.64±0.09	52.83±0.19	**53.82±0.20**	53.37±0.17	54.47±0.14	48.71±0.30	47.83±0.31
HI	46.77±0.08	47.41±0.08	44.63±0.23	**47.93±0.12**	42.45±0.46	27.49±0.84	27.30±0.44
MR	34.57±0.99	37.08±0.68	37.43±1.66	**39.16±1.02**	36.73±1.44	27.22±1.89	26.31±1.60
TA	34.83±0.34	36.51±0.36	**38.25±0.57**	37.74±0.63	33.76±0.79	23.64±0.84	22.20±0.88
HU	52.33±0.29	**53.14±0.17**	49.34±0.14	52.81±0.11	48.44±0.25	39.05±0.29	39.52±0.39
PL	57.56±0.17	57.26±0.08	57.50±0.13	**59.75±0.15**	56.80±0.34	51.65±0.11	51.19±0.33
TE	**44.50±0.72**	42.94±0.62	40.91±0.68	44.44±0.51	39.21±1.22	35.97±1.30	32.84±0.55
NL	63.81±0.26	63.95±0.21	63.38±0.14	**64.34±0.19**	59.15±0.23	51.95±0.36	52.42±0.10
ID	57.18±0.26	57.05±0.20	**59.05±0.21**	58.06±0.13	60.49±0.23	56.55±0.34	55.00±0.35
JA	35.70±0.13	35.87±0.43	35.96±0.54	**36.35±0.09**	33.15±1.13	23.76±1.07	21.40±1.07
RU	58.84±0.23	59.82±0.25	59.59±0.24	**59.93±0.18**	59.59±0.36	55.49±0.24	55.48±0.23
UK	57.89±0.08	58.88±0.18	57.53±0.21	**59.14±0.15**	58.18±0.25	52.80±0.39	52.98±0.17
ZH	44.22±0.27	45.51±0.20	45.89±0.31	**45.99±0.22**	48.68±0.20	48.54±0.46	47.95±0.34
VI	25.14±0.04	25.32±0.15	**27.95±0.04**	26.04±0.07	27.07±0.11	29.39±0.29	27.34±0.36
FI	53.89±0.11	54.37±0.14	53.12±0.11	**54.84±0.03**	54.74±0.19	51.54±0.13	51.83±0.23
IT	57.80±0.05	58.29±0.18	58.24±0.08	**58.91±0.09**	58.73±0.20	55.23±0.16	54.97±0.12
ES	53.59±0.07	54.37±0.08	**55.47±0.07**	55.27±0.11	54.75±0.20	51.02±0.13	50.51±0.06
FR	63.03±0.11	63.13±0.13	62.67±0.19	**63.79±0.06**	61.52±0.18	59.81±0.15	59.88±0.26
DE	58.89±0.05	59.31±0.15	59.80±0.06	**59.89±0.07**	54.44±0.31	41.40±0.38	43.10±0.23
PT	66.14±0.06	66.31±0.13	**66.74±0.09**	66.51±0.03	66.30±0.13	64.05±0.17	63.71±0.10
AVG	51.06±0.07	51.48±0.13	51.51±0.07	**52.43±0.07**	50.41±0.19	44.62±0.11	44.11±0.23

Table 9: F1 scores (%) on UPB v1.0 and UPB v2.0 test sets with gold dependency trees (frozen mBERT). The bold score in each language indicates the highest F1 score among GNN-based models, i.e., GGCNs, SATs, GATs, and 2ATT-GATs. The underlined score in each language indicates the highest F1 score among 2ATT-GATs, BiLSTM+2ATT-GATs, BiLSTM, and 3BiLSTMs. AVG indicates the average F1 score of each model for all languages except English.

	2layers+ mBERT	2layers+ mBERT+ fine-tuned	2layers+ XLM-R	2layers+ XLM-R+ fine-tuned	3layers+ mBERT	3layers+ mBERT+ fine-tuned	3layers+ XLM-R	3layers+ XLM-R+ fine-tuned
				UPB v1.0				
AVG (gold)	**56.45±0.05**	55.00±0.05	55.87±0.15	54.33±0.10	56.15±0.11	55.05±0.08	56.00±0.13	54.45±0.13
AVG (pred)	**53.34±0.05**	52.41±0.033	52.93±0.13	51.95±0.11	52.99±0.11	52.43±0.08	52.96±0.09	51.95±0.11
				UPB v2.0				
AVG (gold)	52.41±0.10	54.63±0.11	54.05±0.082	**55.34±0.06**	52.67±0.08	54.30±0.09	53.71±0.08	55.26±0.08
AVG (pred)	51.64±0.12	53.95±0.10	53.25±0.12	**54.93±0.09**	51.93±0.07	53.62±0.07	52.89±0.09	54.72±0.08

Table 10: Average F1 scores (%) of target languages in UPB v1.0 and UPB v2.0 dev sets using gold and predicted dependency trees.