Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1700–1711
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Universal Proposition Bank 2.0

Ishan Jindal[1], Alexandre Rademaker[1], Michał Ulewicz[2*], Linh Ha[4], Huyen Nguyen[3], Khoi-Nguyen Tran[1], Huaiyu Zhu[1], Yunyao Li[4*]
[1]IBM Research, [2]Kyndryl, [3]Vietnam National University, [4]Apple
ishan.jindal@ibm.com, alexrad@br.ibm.com, michal.ulewicz@kyndryl.com
{huyenntm, hamylinh}@hus.edu.vn, kndtran@ibm.com,
huaiyu@us.ibm.com, yunyaoli@apple.com

Abstract

Semantic role labeling (SRL) represents the meaning of a sentence in the form of predicate-argument structures. Such shallow semantic analysis is helpful in a wide range of downstream NLP tasks and real-world applications. As treebanks enabled the development of powerful syntactic parsers, high-quality training data in the form of propbanks is crucial to build models for accurate predicate-argument analysis. Unfortunately, most languages simply do not have corresponding propbanks due to the high cost required to construct such resources. To overcome such challenges, we released Universal Proposition Bank 1.0 (UP1.0) in 2017, with high-quality propbank data generated via a two-stage method exploiting monolingual SRL and multilingual parallel data. In this paper, we introduce Universal Proposition Bank 2.0 (UP2.0), with significant enhancements over UP1.0, including: (1) **propbanks with higher quality** by using a state-of-the-art monolingual SRL and improved auto-generation of annotations; (2) **expanded language coverage** (from 7 to 23 languages); (3) **span annotation** for the decoupling of syntactic analysis; and (4) **gold data** for a subset of the languages. We also share our experimental results that confirm the significant quality improvements of the generated propbanks. In addition, we present a comprehensive experimental evaluation on how different implementation choices impact the quality of the resulting data. We release these resources to the research community and hope to encourage more research on cross-lingual SRL.

Keywords: Annotation projection, Semantic role labeling, Multilingual, Span-based SRL, Dependency-based SRL

1. Introduction

Semantic role labeling (SRL) is a shallow semantic parsing task that identifies "*who did what to whom when*, *where* etc" for each predicate in a sentence. It provides an intermediate (shallow) level of a semantic representation that helps the map from syntactic parse structures to more fully-specified representations of meaning (Jurafsky and Martin, 2021). SRL has been shown to help a wide range of NLP applications such as natural language inference (Zhang et al., 2020b), question answering (Zhang et al., 2020b; Maqsud et al., 2014; Yih et al., 2016), machine translation (Shi et al., 2016) and information extraction (Niklaus et al., 2018; Zhang et al., 2020a).

The increasing availability of manually annotated meaning representation datasets such as FrameNet (Fillmore et al., 2004), NomBank (Meyers et al., 2004), Proposition Bank (PropBank) (Palmer et al., 2005)) as well as significant advances in modeling techniques such as deep learning techniques have led to increased interest and progress in computational models for English SRL. Parallelly, several attempts have been made to generate such resources for other languages such as German (Erk et al., 2003), Arabic (Zaghouani et al., 2010), Portuguese (Duran and Aluísio, 2011), Hindi (Vaidya et al., 2012), Finnish (Haverinen et al., 2015) and others. These propbanks are manually labeled by the corresponding language experts following different strategies. In some cases, language-specific rolesets are defined which are different from English PropBank rolesets. Therefore, despite the availability of such SRL resources in different languages, it is almost impractical to build a single multilingual SRL labeler because of the differences in semantic labels and rolesets. Furthermore, due to the high cost of manual annotation, such SRL resources are not available for most languages.

Our earlier work, Universal PropBank v1.0 (UP1.0) [1] provides a solution to these issues by annotating the text in different languages with a layer of **universal** semantic role labeling annotation. UP1.0 aims to automatically label texts in target languages with English PropBank rolesets (verb frames and semantic roles). A two-stage annotation projection approach was applied to cross-lingual transfer of semantic roles from English [Source Language (SL)] to low resource language [Target Language (TL)] (Akbik et al., 2015): (1) filtered annotation projection based on parallel corpora, focusing on high precision potentially at the expense of recall; (2) bootstrapped training of SRL that iteratively improves recall without reducing precision. UP1.0 includes a universal semantic layer for 7 treebanks from UD release 1.4.

UP1.0 has been regarded as an important resource for crosslingual SRL research since its release (Ak and Yildiz, 2019; Cai and Lapata, 2020; Fei et al., 2020; Gunasekara et al., 2020). However, it has several lim-

*Work done while at IBM

[1]https://github.com/System-T/UniversalPropositions

itations. First, given that the work was done over five years ago, the quality of propbanks can be improved by using the high quality state-of-the-art English SRL deep neural network model which seems to have substantial impact on annotation projection quality (Akbik et al., 2015). Similarly, state-of-the-art syntactic parser and word aligner can also be used to improve the quality. In addition, UP1.0 provides dependency-based SRL only, based on the assumption that argument spans can be obtained deterministically from the dependency tree following the dependents from the head node of the argument. However, annotating only the heads is insufficient to predict the spans of the argument.[2] Furthermore, high-quality syntactic parsers are often unavailable for most languages, and recent neural SRL models with SoTA performance are often syntax agnostic (Ouchi et al., 2018; Guan et al., 2019; Jindal et al., 2020a; Conia et al., 2021). Therefore it becomes necessary to provide span-based SRL for target languages so that syntax-agnostic SRL models can be trained.

2. Universal PropBank 2.0

Universal PropBank v2.0 (UP2.0) consists of automatically generated propbanks for 23 languages from 8 language families from UD release 2.9. For each language, UP2.0 provides broadly applicable SRL annotations with both span- and dependency-based semantic roles. See Table 1 for the statistics of the generated propbanks. It also includes a small set of manually annotated sentences for Polish, Portuguese and English as summarized in Table 2. We are annotating more gold data for additional languages and plan to include them as a part of future releases. The UP2.0 release is in the form of a sentence-wise semantic layer, representing predicates and their senses, and argument labels with head- and span-based annotations. A `python` script is included that combines the semantic layer annotations with UD syntactic layers. This arrangement decouples the SRL data from the evolving UD data. UP2.0 includes the following major enhancements over UP1.0:

Higher Quality We replaced the parsers, word aligners and underlying English SRL models in (Akbik et al., 2015) with recent state-of-the-art models, resulting in higher quality as measured on 3 TL language with gold SRL labels.

[2]We have at least two reasons. First, the exact extent of the argument is not necessarily predictable from the head since some relations typically continue an entity (e.g., an `amod` is included in an entity span), and some are not (e.g., `nsubj` to a nominal predicate is not part of the predicate entity span). It is not 100% possible to formulate rules determining entity span from relation types since some are ambiguous (e.g., `dep`, which may or may not be part of the entity), and some are treebank or language-specific. Second, some tokens can be the head of two conflicting entities, especially coordination, where a token can be the head of two different entities with different spans.

Span-based SRL A deep learning technique is used to jointly train a single SRL model to jointly produce both span- and dependency-based SRL annotations for TL languages.

Gold SRL Annotations To enable the research community to perform fair evaluation of their multilingual and cross-lingual SRL systems we manually annotated English SRL labels for one language Polish and consolidated the propbanks of other two languages (Portuguese and English).

We release these resources containing generated propbanks and hand-annotated test sets to the research community through github project `https://github.com/UniversalPropositions`.
Following are more detailed description of the data generation process.

2.1. Automatic Data Generation

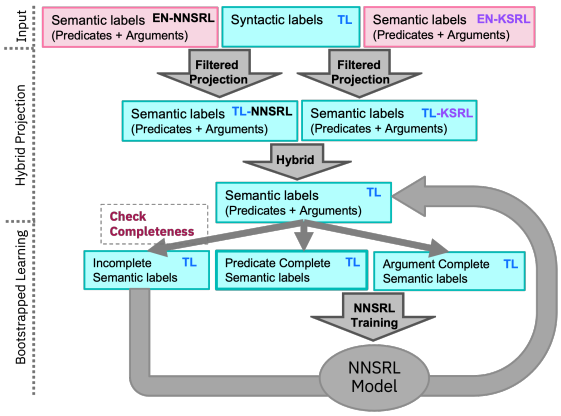

Figure 1: Overview of proposed approach projecting SRL labels from EN as source language onto TL as target language. `KSRL` means EN SRL labels using SRL model from (Akbik and Li, 2016).

We follow the same two-stage process in (Akbik et al., 2015) to generate high-quality proposition banks for multiple languages, as illustrated in Figure 1. First, we apply a filtered annotation projection to parallel corpora to achieve annotations with high precision for target-language. Then we bootstrap and retrain the TL SRL to iteratively improve recall of the generated propbank without reducing precision. This process assumes that the parallel corpus such as (Tiedemann, 2012) is available in EN-TL as well as the availability of the following components:

Syntactic parser We use Stanza parser (Qi et al., 2020) to obtain the dependency parse for sentences in EN and TL.

Word aligner We obtain the word alignment of the tokenized sentences from EN and TL using SimAlign (Sabet et al., 2020).

Semantic role labeler We developed a SoTA SRL model for EN to predict both span- and dependency-based semantic roles (Section 2.1.1).

Section 3.4 discusses more details on how the choices of the different syntactic components and how they impact the quality of resulting data.

2.1.1. Projection Quality Improvement

UP2.0 contains data with even higher quality than UP1.0 (Section 3.3), although we employ the same process flow as (Akbik et al., 2015). A few important enhancements made over (Akbik et al., 2015) lead to the quality improvements, as discussed next.

High Quality EN SRL As observed by (Akbik et al., 2015), the quality of the EN SRL annotations significantly impacts the quality of projected annotations on the target languages. The better the quality of EN SRL labels the better the quality of projected TL SRL labels. We observe the same in our experiments described in Section 3.3.

Inspired by the state-of-the-art neural SRL models such as (Shi and Lin, 2019; Jindal et al., 2020a), we develop a novel neural SRL (`NNSRL`) architecture to predict both span- and dependency-based SRL as depicted in Figure 2. The network consists of two branches where given the predicate-specific representation of each token, one branch predicts the head of the arguments, and the other predicts the span of the arguments. For EN SRL labeler, we train this network on OntoNotes data for which argument spans are available via LDC catalog (Weischedel et al., 2013) and argument heads are obtained via transforming the constituent analysis to dependency tree using CoreNLP Library (Schuster and Manning, 2016) with additional postprocessing to adapt the UD syntactic analysis to the most current UD guidelines.[3] The network is trained as an end-to-end system with the final loss function as a sum of all losses across the head and span SRL branches.

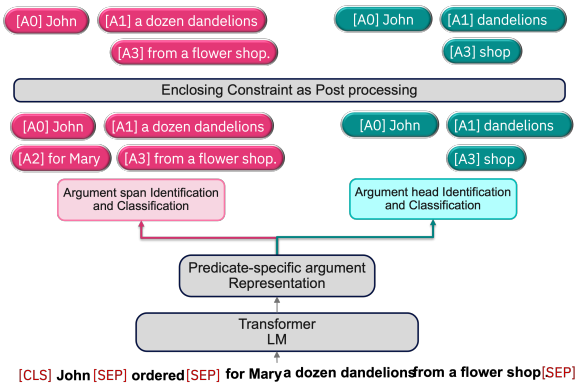

Figure 2: Proposed neural architecture (`NNSRL`) predicting both span- and dependency-based semantic roles simultaneously.

[3]The Ontonotes and EWT corpora with SRL on top of UD dependencies will be made available.

During inference, we apply the 'enclosing constraint' on model's prediction as a post-processing step. This constraint means that only those arguments of which head prediction is entirely enclosed by span prediction are retained.

Similar to (Shi and Lin, 2019; Jindal et al., 2020a), a practical end-to-end EN SRL model contains two neural models one for predicate identification and sense disambiguation, and another for argument identification and classification. We also train two different neural network models for predicates and arguments on predicate-complete and argument-complete subsets of the data, respectively.

Hybrid Projection Using a high-quality EN SRL model results in higher projection precision (See Section 3.2), but at the expense of low recall for predicate identification. One of the major advantages of syntax-based EN SRL model is its very high recall for predicate identification. To take advantage of both models, we adopt a hybrid approach that also utilizes projection of the EN SRL labels obtained from a syntax-based EN SRL model `KSRL` (Akbik and Li, 2016) to TL. As depicted in Figure 1 we first project the EN SRL labels predicted using two different SRL models, and then supplement the projected predicates in `TL-NNSRL` with the predicates from TL SRL projected labels in `TL-KSRL`. A summary of the number of predicates supplemented for each TL is provided in Table 1.

Lang.	UP1.0	UP2.0		#Unique
	#Comp.	#Arg comp.	#Pred comp.	Frames
cs	251K	257K	71K	2991
de	438K	453K	262K	2977
el	262K	282K	80K	5044
es	579K	613K	139K	2833
fi	488K	512K	181K	1848
fr	676K	698K	180K	2517
hi	106K	109K	150K	413
hu	152K	162K	47K	2713
id	888K	920K	717K	4972
it	606K	606K	256K	2771
ja	120K	127K	100K	2942
ko	37K	42K	18K	1718
mr	11K	5K	6K	167
nl	442K	457K	136K	2656
pl	213K	223K	40K	2354
pt	775K	788K	152K	2978
ro	150K	147K	55K	1495
ru	622K	641K	417K	4683
ta	28K	22K	24K	458
te	16K	16K	14K	678
uk	123K	128K	81K	2396
vi	339K	359K	420K	1261
zh	366K	389K	314K	4408

Table 1: Characteristics of generated propbanks. `Complete` means both predicate and argument complete sentences. Hybrid projection introduces new predicates and arguments to UP2.0 argument complete sentences.

Definition 1 (Predicate Completeness)**.** *A sentence in TL is deemed predicate-complete if it has the same number of predicates as its corresponding EN sentence, where the predicates of the EN sentence are those identified using a trained* NNSRL *model and the predicates of the TL sentence are obtained via annotation projection.*

Definition 2 (Argument Completeness)**.** *(Equivalent to k-complete in (Akbik et al., 2015).) A direct component of a labeled sentence in TL is either a verb in TL or a syntactic dependent of a verb. Then a sentence in TL is k-complete if it contains equal to or fewer than k unlabeled direct components.* 0-complete *is abbreviated as argument-complete.*

2.1.2. Bootstrapped Training Enhancement

Similar to (Akbik et al., 2015), we employ bootstrapped training after label projection to address low recall issue with generated propbanks. The TL SRL model is trained iteratively over predicate-complete (Definition 1) and argument-complete (Definition 2) subsets of the data, supplemented by high precision labels produced from previous iteration. The quality of TL SRL model can be further improved by allowing the supervision from high quality EN training examples using polyglot training.

Polyglot Training The idea of training one model on multiple languages with multilingual word embeddings has previously been shown to outperform monolingual baselines, especially for low resource languages (Jindal et al., 2020b; Mulcaire et al., 2019). We expect models trained jointly on multiple languages with homogeneous annotations will be able to generalize better across languages. Therefore, we train NNSRL (described in Section 2.1.1) on both EN and TL simultaneously for both span- and dependency-based SRL.

Span-based SRL for TL We jointly train the NNSRL model on EN and the projected TL SRL labels simultaneously to not only improve the quality of the projected dependency-based SRL but to obtain span-based SRL for TL. In this process, instead of projecting the argument spans, we train a common encoder that uses the multilingual features to predict argument spans for TL. As the projected TL data contains only the dependency-based SRL, during the training phase of the NNSRL model, we update the parameters of span-branch only for the EN sentences. On the other hand, we update the parameters of the head-branch for both the EN and the target languages.

2.2. Gold Data

To assess the quality of the automatically generated propbanks, we curated human-annotated test sets for two TLs, in addition to one existing set. Table 2 provides the statistics of ground truth instances for each language.

Lang.	#Sentences	#Predicates	#Arguments
EN_{GOLD}	16622	50258	101603
FR^{*}_{GOLD}	1001	1979	5393
PL_{GOLD}	100	223	495
PT_{GOLD}	3779	6173	15097

Table 2: Characteristics of gold data for each language. $*$ means the ground truth are from (Van der Plas et al., 2011).

Polish (PL_{GOLD}) We select 100 English sentences from OntoNotes (Weischedel et al., 2013) and translate them into Polish. Then we manually label all the predicates and arguments according to English PropBank.

Portuguese (PT_{GOLD}) The Propbank.Br project annotated 3779 sentences (Duran and Aluísio, 2011), the Brazilian portion of Bosque, the manually revised subcorpus of Floresta Sintá(c)tica (Afonso et al., 2002) with manually verified constituent analysis. (Rademaker et al., 2017) converted the Bosque corpus to dependencies and incorporated it into the UD collection. We merge the two resources, projecting the SRL annotation from Propbank.Br on top of the dependencies from UD Bosque (UD 2.9) solving inconsistencies and fixing annotation errors [4] in the original Propbank.Br.

French (FR_{GOLD}) French ground truth data is obtained from (Van der Plas et al., 2011). It consists of 1001 manually labeled French sentences of Europarl corpus. All the predicates and arguments are labeled according to English PropBank. We noticed label noise in this dataset as observed in (Akbik et al., 2015). We therefore expect the true performance is somewhat higher than the performance in Table 2 Row 1.
We are also working on manually annotating EN PropBank labels for other languages and plan to release to the research community in the future.

3. Experiments and Evaluations

One of the main objectives of UP2.0 is to automatically generate high-quality propbanks for multiple languages sharing the homogeneous semantic role labels. In this section, we seek to answer the following questions: 1) What is the quality of the generated propbanks for dependency-based SRL and estimated quality for span-based SRL on a manually annotated gold set? 2) What effect does each component of the approach have on the quality of generated propbanks?

3.1. Data Preparation

Data Sources In our experiments, we examine 23 target languages Czech, German, Greek, Spanish,

[4] Inconsistencies are mainly related to 1) some Portuguese senses are mapped to more than one English sense; 2) some English senses are invalid (not PropBank senses); 3) some cases where more than one English sense was associated with the same token; 4) some English senses are annotated with different predicates than ones in the PropBank project.

Finnish, French, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Marathi, Dutch, Polish, Portuguese, Romanian, Russian, Tamil, Telugu, Ukrainian, Vietnamese and Chinese. Further statistics on all the languages are made available in Appendix A. These experiments mainly use languages from the Indo-European, Austroasiatic, Uralic, Austronesian, Japonic, Korenaic, Balto-Slavic, Dravidian and Sino-Tibetan language families because 1) these language families are among the top 10 Language Families by Number of Speakers in the world[5]; 2) We could easily find the language experts to label few gold sentences for evaluation. We use English as SL in all our experiments.
Parallel corpora for these languages were downloaded from OPUS web page[6]. Our experiments use three parallel corpora:

Europarl a parallel corpus extracted from the European Parliament[7] (Koehn and others, 2005). Abbreviated as **EP**.

Tatoeba a database of translated sentences[8]. Abbreviated as **TB**. This dataset contains real-world examples which are collaboratively manually translated to sentences in different languages by a large number of volunteers. We include this corpus because of its known good quality as it contains manually translated sentences.

Open Subtitles a collection of translated movie subtitles[9] (Lison and Tiedemann, 2016). Abbreviated as **OS**.

Model	Training	Predicate	Argument	
			Head	Span
SoTA				
JIN*	Span only	-	-	**86.60**
SHI*	Span only	-	-	86.50
JIN†	Head only	89.0	85.29	-
KSRL	Head only	84.8	81.70	-
UP1.0-SRL	Head only	84.0	80.00	-
NNSRL	Polyglot	**93.4**	**87.29**	83.25
NNSRL	Head only	93.4	87.82	-
NNSRL	Span only	93.4	-	83.14

Table 3: Performance comparison of NNSRL on OntoNotes test-set both for span- and dependency-based SRL. – means the model does not have predictions.* means reported number directly from research paper. † means model retrained for OntoNotes. JIN=(Jindal et al., 2020a), SHI=(Shi and Lin, 2019)

[5]https://www.vistawide.com/languages/language_families_statistics1.htm
[6]https://opus.nlpl.eu
[7]https://statmt.org/europarl/
[8]https://tatoeba.org/en
[9]http://www.opensubtitles.org

For all the languages, we use different sources of Bitext to improve the domain adaptability of the Target languages SRL model, as evident from Table 4.

Data Pre-processing Once multiple parallel corpora are combined into one parallel corpus for a language pair, we apply pre-processing and remove sentences with certain properties: (1) duplicate sentences, (2) sentences having character encoding problems, (3) sentences with less than 5 tokens, as these sentences do not generally contain a predicate and unnecessarily blow up the data size, (4) sentences with more than 80 tokens. In addition to this, we replace multiple spaces with one space. In Table 7 we show the number of sentences removed after pre-processing. Around 15% of the sentences for each language pair are removed.

3.2. Model Preparation

EN SRL Model Since for EN we have span- and dependency-based SRL datasets available (OntoNotes), we train the NNSRL model described in Section 2.1.1. We use *BERT-base-multilingual-cased* transformer model as text encoder in all the experiments. For BERT fine-tuning, we used the Transformer (Vaswani et al., 2017) implementation by Wolf et al. (2019).[10] We convert the SRL as token classification task both for predicate and arguments. We use the same architecture for dependency-based SRL datasets except that we do not compute loss for the span-branch and do not apply enclosing constraint at inference. We compare our model with the current state-of-the-art SRL models in Table 3. To the best of our knowledge, our proposed NNSRL model is the first ever model that can predict both the head and the span of each argument.
Existing SRL model are either span-based or dependency-based, making it hard to compare the performance of proposed NNSRL with existing SoTA SRL models. Therefore, to facilitate a fair comparison with SoTA, we also train NNSRL individually for each SRL type. For example, for training NNSRL only for dependency-based SRL we freeze the parameters of span-branch and vice versa. From Table 3, NNSRL provides the best performance for predicates and dependency-based arguments as compared to existing approaches. However, better performance on dependency-based SRL comes at the expense of a little performance drop on span-based SRL.
In all our experiments we use Stanza as syntactic parser and SimAlign as word aligner. We refer readers to Appendix B and Appendix C where we provide the comparisons of Stanza and SimAling with existing approaches, respectively.

3.3. Results: Quality Evaluation

We measure the quality of generated propbanks in standard measures of precision, recall, and F1 for both

[10]https://github.com/huggingface/transformers

Lang	Bitext	Projection	Model	Data	Predicate			Argument		
					P	R	F1	P	R	F1
FR_{GOLD}	EP	Hybrid	NNSRL	EN+TL	79.92	**59.76**	**68.38**	**58.47**	**43.94**	**50.17**
		Hybrid	NNSRL-Head	TL only	**80.20**	56.93	66.59	55.87	43.12	48.68
		UP1.0	NNSRL-Head	TL only	76.72	40.40	53.01	45.54	42.23	43.82
		UP1.0	SRL-Head	TL only	33.20	52.50	40.70	44.30	12.30	19.30
	TB	Hybrid	NNSRL	EN+TL	77.08	68.5	72.54	**56.54**	44.39	**49.73**
		Hybrid	NNSRL-Head	TL only	76.85	**69.16**	**72.80**	50.60	**44.91**	47.58
		UP1.0	NNSRL-Head	TL only	**79.44**	47.67	59.59	49.82	40.71	44.80
		UP1.0	SRL-Head	TL only	31.00	43.80	36.30	19.70	05.60	08.70
	EP+TB	Hybrid	NNSRL	EN+TL	73.59	**76.34**	**74.94**	56.27	**46.11**	**50.69**
		Hybrid	NNSRL-Head	TL only	**77.81**	66.84	71.91	50.64	42.46	46.19
		UP1.0	NNSRL-Head	TL only	72.24	46.71	56.74	51.71	39.88	45.03
		UP1.0	SRL-Head	TL only	36.90	55.40	44.30	**57.60**	26.40	36.20
PL_{GOLD}	EP	Hybrid	NNSRL	EN+TL	**76.64**	44.87	56.60	**55.68**	**39.60**	**46.28**
		Hybrid	NNSRL-Head	TL only	74.83	**48.29**	**58.70**	47.04	38.59	42.40
		UP1.0	NNSRL-Head	TL only	75.47	34.19	47.06	40.20	40.61	40.40
		UP1.0	SRL-Head	TL only	29.30	29.30	29.30	31.50	13.40	18.80
	EP+TB	Hybrid	NNSRL	EN+TL	**78.06**	**51.71**	**62.21**	**61.28**	**44.44**	**51.52**
		Hybrid	NNSRL-Head	TL only	73.08	48.72	58.46	58.33	43.84	50.06
		UP1.0	NNSRL-Head	TL only	70.48	31.62	43.66	41.08	**40.00**	40.53
		UP1.0	SRL-Head	TL only	29.80	30.60	30.20	36.20	18.40	24.40
PT_{GOLD}	EP	Hybrid	NNSRL	EN+TL	**66.85**	**65.56**	**66.20**	**60.69**	46.12	**52.41**
		Hybrid	NNSRL-Head	TL only	66.33	55.79	60.61	57.24	**46.76**	51.47
		UP1.0	NNSRL-Head	TL only	62.03	37.75	46.93	45.27	42.94	44.07
		UP1.0	SRL-Head	TL only	37.10	42.30	39.50	46.60	33.80	39.20
	TB	Hybrid	NNSRL	EN+TL	66.49	**71.13**	**68.73**	**55.51**	**51.28**	**53.31**
		Hybrid	NNSRL-Head	TL only	64.51	70.79	67.51	51.04	44.97	47.81
		UP1.0	NNSRL-Head	TL only	**69.70**	49.34	57.78	50.76	46.25	48.40
		UP1.0	SRL-Head	TL only	33.90	41.90	37.40	45.50	31.60	37.30
	EP+TB	Hybrid	NNSRL	EN+TL	66.73	**71.47**	**69.02**	**59.17**	**54.60**	**56.79**
		Hybrid	NNSRL-Head	TL only	**68.36**	65.04	66.66	50.75	44.13	47.21
		UP1.0	NNSRL-Head	TL only	56.87	49.20	52.76	49.7	45.85	47.70
		UP1.0	SRL-Head	TL only	37.20	44.00	40.30	48.10	37.10	41.90

Table 4: Performance comparison on gold data with respect to the different projection methods and different SRL model. EN is source language in all these experiments. First row of each block is **UP2.0**. **Bold** is best in each block. **Underline** is best for that PropBank.

predicate identification and sense disambiguation (as Predicate), and argument identification and classification (as Argument) for three languages (FR, PL, PT) that have hand-annotated dependency-based SRL labels as described in Section 2.2. We also provide an estimate on span-base SRL quality for these propbanks as no such gold span-based SRL test sets are available.

Dependency-based TL SRL Quality We empirically demonstrate the effectiveness of the proposed approach via extensive experiments in Table 4 for dependency-based SRL quality of 3 generated propbanks on gold SRL labels. All experiments have consistently shown that generated propbanks in UP2.0 have the best quality (Row 1 in each block in Table 4 measures the UP2.0 generated propbank quality). Precision for predicate labels is over 30 points and recall is over 20 points better than UP1.0, and for arguments labels, precision is over 10 points and recall is over 20 points better than UP1.0, demonstrating the effectiveness of the proposed projection approach. We further measure whether these performance differences are statistically significant in Table 5. We find that p-value is less than the significance level alpha (e.g., 0.05) for all the propbanks. As a result, generated propbank quality in UP2.0 is substantially better than UP1.0 as measured by these encouraging results.

Span-based TL SRL Quality While gold annotation for dependency-based SRL for a few of the languages

			FR_{GOLD}	PL_{GOLD}	PT_{GOLD}
UP2.0	Predicate	Mean	66.28	54.45	61.80
		SD	07.94	07.35	07.88
	Argument	Mean	47.41	45.2	49.91
		SD	02.56	04.85	03.88
UP1.0	Predicate	Mean	40.43	32.43	39.10
		SD	04.01	04.67	01.44
	Argument	Mean	23.73	20.43	39.47
		SD	07.55	03.45	02.31
p-value	Predicate		1.1e-4	1.4e-3	1.4e-4
	Argument		0.024	1.6e-4	1.2e-3

Table 5: Two sample T-test showing the UP2.0 performance is statistically significant both for predicate and arguments F1. Rejecting the NULL hypothesis that UP2.0 quality is same as UP1.0 quality.

is available, no such resources are available for span-based SRL for target languages. Consequently, we decided to estimate the quality of span-based SRL as span label quality.

Span Label Quality Estimating the argument span label quality is straightforward and is close to the F1 score of argument head classification. As we know from Section 2.1.1 we apply enclosing constraints on the trained `NNSRL` at the inference time, this means that if the semantic labels for argument head are correct it is correct for the argument spans. On FR, PL, and PT we find that 97.2%, 95.14%, and 98.24% of the instances for which the predicted head is enclosed by the predicted span have the same semantic label, respectively. We estimate a lower bound on the span label quality as the product of dependency-based SRL quality and the fraction of instances for which the predicted head is enclosed by the predicted span have the same semantic label (Table 6). From Table 6 it can observed that the span label quality is very close to the dependency-based SRL quality, therefore, better the dependency-based SRL quality better will be the span-based SRL quality.

Lang	Model	pred head	gold head	Est. F1
EN	Gold	99.54	99.90	-
FR_{GOLD}	NNSRL	97.20	72.04	50.69
PL_{GOLD}	NNSRL	95.14	55.36	51.52
PT_{GOLD}	NNSRL	98.24	70.62	56.79

Table 6: % of arguments for which head is inside the predicted span having same semantic label. Performance is computed for corresponding best NNSRL-Polyglot dependency-based SRL performance.

3.4. Results: Effect of Individual Components

In this section, we present a comprehensive experimental evaluation on how different implementation choices impact the quality of the resulting data.

Bitext Selection is the first step in our projection pipeline. We observe that the selection of Bitext impacts the quality of generated propbanks. We choose the Bitext based on the criteria that promote generalizability to target domains (EP and OS) and contain sentences designed for foreign language learners (TB). We also experimented with the combinations of these corpora in Table 4. For FR_{GOLD} we know from the outset that gold data is extracted from the Europarl corpus; therefore, an FR PropBank generated from EP alone has better quality than an FR PropBank generated from TB alone. While EP by itself produces the best F1 score, the combination with TB improves the predicate recall by 16.58 points and argument recall by 2.17 points over EP alone, but this comes at the cost of a little loss of precision. We observe a similar trend that propbanks generated with TB corpus are not of high quality by themselves but jointly with EP corpus results in high-quality propbanks. Therefore, generating propbanks with sentences from diverse domains improves the generalizability of the SRL models trained on these propbanks.

Further, for some of the target languages Bitext corpus size is quite limited thus limiting the number of unique frames for that target language. This can be seen from Table 1, where HI, MR, TA and TE have limited frame coverage as compared to other languages. Improving the frame coverage for these languages will be addressed in later release.

Why Better EN SRL Model? Following Bitext selection, the EN SRL model is used to predict EN SRL labels for the EN subset of Bitext, and errors made by EN SRL are often propagated to the TL SRL via projection. We observe a substantial impact of the EN SRL model on the quality of generated propbanks. The last two rows in each block in Table 4 show the impact of the EN SRL model. On all the propbanks with gold standards, F1 scores of both predicates and arguments are $\sim$15 and $\sim$20 points better as compared to the SRL model used in UP1.0 (Akbik et al., 2015) except for PT_{GOLD} where the argument performance is only $\leq$10 points better. Hence, the quality EN SRL model has substantial impact on the quality of generated propbanks.

Why Hybrid Projection? No doubt the *NNSRL-Head* provides better quality propbanks when labels are projected according to UP1.0 (Last two rows in each block in Table 4), one important observation is that for predicates the projection recall with the SRL model in UP 1.0 is consistently better than *NNSRL-Head* model. Therefore, we introduce hybrid projection to get the best of both worlds that is supplementing the predicates and arguments from `AK` projection to `NNSRL` projection (Figure 1). We further observe that hybrid projection decreases the size of generated propbanks (Table 1 in terms of the number of `Complete` sentences

as compared to Complete sentence from UP1.0, however, the quality of generated propbank using hybrid projections is substantially better as shown in the last three rows in each block in Table 4). Therefore, hybrid projections guarantees better precision and recall of the TL SRL model when trained on these generated propbanks.

Why Polyglot Training? In addition to quality improvement of the generated propbanks, we also aim to generate Span-based SRL for TL such that a syntax-agnostic span-based SRL model can easily be learned for TL. Therefore, we train NNSRL model both on EN and TL simultaneously and only updating the parameters of head branch for TL sentences. Polyglot training not only generate spans-based SRL for TL it also consistently improves the generated propbank quality for dependency-based SRL (Top row in each block in Table 4) by allowing parameter sharing across multiple languages. Thus allowing the cross-lingual transfer between EN and TL further enhance the generated TL PropBank quality.

4. Related Work

Given a sentence, the SRL task is generally divided into four sub-tasks: predicate identification, predicate sense disambiguation, arguments identification and the assignment of semantic role labels to each argument. Error introduced at any of these stages drifts to the entire pipeline and results in incorrect annotations. Given the complexity of the task it is inevitable to make mistakes during highly time-consuming manual semantic annotation process. Therefore, several alternatives to manual annotation has been studied in the past such as approaches with (semi-)automatic semantic annotations (Padó, 2007; Van der Plas et al., 2011; Akbik et al., 2015; Exner et al., 2016; Mille et al., 2018; Gotham and Haug, 2019).

Annotation Projection Approaches Main objective of annotation projection approaches is to *project* semantic labels from resource rich language (English) [Source Language(SL)] to low resource language [Target Language(TL)] assuming the parallel corpora SL-TL exists. Projecting SRL labels can be traced back to (Padó, 2007; Padó and Lapata, 2009) where semantic labels of FrameNet and PropBank was projected from English to resource-poorer languages (German). Van der Plas et al. (2011) improves the projection quality by jointly learning syntactic-semantic features and showed results on French. They also hand labeled one thousand French sentence with English PropBank semantic labels to facilitate evaluation of projection techniques. We also evaluate our approach on this set. Akbik et al. (2015) further improves the annotation projection approach by applying bootstrapped learning on top of filtered projection to improve the recall without reducing precision. Alternatively, Exner et al. (2016) proposed an approach to transfer labels to the aligned entities in SL-TL pairs thus projecting labels for French and Swedish. Most recently, Mille et al. (2018) proposed deep datasets, where Deep Track dataset consists of trees that contain only content words linked by predicate-argument edges in the PropBank fashion (available for French and Spanish).

Universal PropBank Resources Large scale annotation projection was first introduced in (Akbik et al., 2015) as Universal PropBank v1.0, where two step approach was used to project SL SRL labels to TL providing universal semantic layer for 7 treebanks from UD release 1.4. Recently, (Droganova and Zeman, 2019) proposed Deep Universal Dependencies, based on UD release 2.4[11] where deep annotations are derived semi-automatically from surface trees with acceptable quality. Though the authors show that the approach can be extended to any language, this approach does not transfer predicates senses and contextual arguments.
Additionally, none of the these approaches provide span-based annotation of SRL for TL.

5. Conclusion and Future Work

In this work, we introduces Universal PropBank 2.0 (UP2.0)—a high quality automatically generated propbanks for 23 languages from 8 language families and manually annotated instances for 3 languages. Through comprehensive experiments we show that the generated propbanks data quality in UP2.0 is significantly better than UP1.0. We also analyze the impact of essential design decisions and implementation details on the quality of generated propbanks.
We plan to perform an extensive performance analysis of generated argument spans in future by evaluating against the hand-annotated gold spans. Further, we plan to include hand-annotated SRL datasets for multiple other languages from different languages families to facilitate proper benchmarking. It would be interesting to perform an qualitative comparison against existing propbanks such as Chinese PropBank, Hindi PropBank, Finnish PropBank or Arabic PropBank.

6. Acknowledgements

Linh Ha was funded by Vingroup Joint Stock Company and supported by the Domestic Master/ PhD Scholarship Program of Vingroup Innovation Foundation (VINIF), Vingroup Big Data Institute (VINBIGDATA), code VINIF.2021.TS.028. We would like to thank Dao Do, Trang Tran, Hue Phan, and Cuong Le for manually annotating sentences from VI Tatoeba corpus.

7. Bibliographical References

Afonso, S., Bick, E., Haber, R., and Santos, D. (2002). Floresta sintá (c) tica: a treebank for portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. ELRA.

[11] https://ufal.mff.cuni.cz/deep-universal-dependencies

Ak, K. and Yildiz, O. T. (2019). Automatic propbank generation for turkish. In *RANLP*.

Akbik, A. and Li, Y. (2016). K-srl: instance-based learning for semantic role labeling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 599–608.

Akbik, A., Chiticariu, L., Danilevsky, M., Li, Y., Vaithyanathan, S., and Zhu, H. (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.

Cai, R. and Lapata, M. (2020). Alignment-free cross-lingual semantic role labeling. In *EMNLP*.

Conia, S., Bacciu, A., and Navigli, R. (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, June. Association for Computational Linguistics.

Dou, Z.-Y. and Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora.

Droganova, K. and Zeman, D. (2019). Towards deep universal dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 144–152.

Exner, P., Klang, M., and Nugues, P. (2016). Multilingual supervision of semantic annotation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1007–1017.

Fei, H., Zhang, M., Li, F., and Ji, D. (2020). Cross-lingual semantic role labeling with model transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2427–2437.

Gotham, M. and Haug, D. T. T. (2019). Glue semantics for universal dependencies. In *Proceedings of the LFG'18 Conference*. CSLI Publications.

Guan, C., Cheng, Y., and Zhao, H. (2019). Semantic role labeling with associated memory network. 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 3361–3371, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Gunasekara, S., Chathura, D., Jeewantha, C., and Dias, G. (2020). Using annotation projection for semantic role labeling of low-resourced language: Sinhala. *2020 International Conference on Asian Language Processing (IALP)*, pages 98–103.

Jindal, I., Aharonov, R., Brahma, S., Zhu, H., and Li, Y. (2020a). Improved semantic role labeling using parameterized neighborhood memory adaptation. *arXiv preprint arXiv:2011.14459*.

Jindal, I., Li, Y., Brahma, S., and Zhu, H. (2020b). Clar: A cross-lingual argument regularizer for semantic role labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3113–3125.

Jurafsky, D. and Martin, J. H. (2021). Speech and language processing. Draft of December 29.

Maqsud, U., Arnold, S., Hülfenhaus, M., and Akbik, A. (2014). Nerdle: Topic-specific question answering using wikia seeds. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 81–85.

Mille, S., Belz, A., Bohnet, B., and Wanner, L. (2018). Underspecified Universal Dependency structures as inputs for multilingual surface realisation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 199–209, Tilburg University, The Netherlands, November. Association for Computational Linguistics.

Mulcaire, P., Kasai, J., and Smith, N. A. (2019). Polyglot contextual representations improve crosslingual transfer. *arXiv preprint arXiv:1902.09697*.

Niklaus, C., Cetto, M., Freitas, A., and Handschuh, S. (2018). A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Ouchi, H., Shindo, H., and Matsumoto, Y. (2018). A span selection model for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1630–1642, Brussels, Belgium, Oct. Association for Computational Linguistics.

Padó, S. and Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Padó, S. (2007). *Cross-lingual annotation projection models for role-semantic information*. Saarland University.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Sabet, M. J., Dufter, P., Yvon, F., and Schütze, H. (2020). Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*.

Sabet, M. J., Dufter, P., Yvon, F., and Schütze, H. (2021). Simalign: High quality word alignments without parallel training data using static and contextualized embeddings.

Schuster, S. and Manning, C. D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Shi, P. and Lin, J. (2019). Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Shi, C., Liu, S., Ren, S., Feng, S., Li, M., Zhou, M., Sun, X., and Wang, H. (2016). Knowledge-based semantic embedding for machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2245–2254.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October. Association for Computational Linguistics.

Van der Plas, L., Merlo, P., and Henderson, J. (2011). Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 299–304.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Yih, W.-t., Richardson, M., Meek, C., Chang, M.-W., and Suh, J. (2016). The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206.

Zhang, H., Wang, H., and Roth, D. (2020a). Unsupervised label-aware event trigger and argument classification. *CoRR*, abs/2012.15243.

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., and Zhou, X. (2020b). Semantics-aware bert for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9628–9635.

8. Language Resource References

Duran, M. S. and Aluísio, S. (2011). Propbank-br: a brazilian portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Erk, K., Kowalski, A., Padó, S., and Pinkal, M. (2003). Towards a resource for lexical semantics: A large german corpus with extensive semantic annotation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 537–544.

Fillmore, C. J., Ruppenhofer, J., and Baker, C. F. (2004). Framenet and representing the link between semantic and syntactic relations. *Frontiers in linguistics*, 1:19–59.

Haverinen, K., Kanerva, J., Kohonen, S., Missilä, A., Ojala, S., Viljanen, T., Laippala, V., and Ginter, F. (2015). The finnish proposition bank. *Language Resources and Evaluation*, 49(4):907–926.

Koehn, P. et al. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The nombank project: An interim report. In *Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004*, pages 24–31.

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

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy, September.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Vaidya, A., Choi, J. D., Palmer, M., and Narasimhan, B. (2012). Empty argument insertion in the hindi propbank. In *LREC*, pages 1522–1526.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., et al. (2013). Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., and Palmer, M. (2010). The revised arabic propbank. In *Proceedings of the fourth linguistic annotation workshop*, pages 222–226.

A. Languages

In Table7, we present different statistics on the use of Bitext for each language pair and the number of sentences that were eliminated after preprocessing.

Lang.	Bitext Source	#Sentences	#After processing
cs	EP+TB	0.67M	0.57M
de	EP+TB	2.26M	1.7M
el	EP+TB	1.31M	1.05M
es	EP+TB	2.22M	1.85M
fi	EP+TB	2.05M	1.67M
fr	EP+TB	2.31M	1.90M
hi	OS+TB	2.04M	1.43M
hu	EP+TB	0.64M	0.58M
id	OS+TB	3.69M	3.16M
it	EP+TB	2.00M	1.77M
ja	OS+TB	1.33M	1.10M
ko	OS+TB	0.45M	0.32M
mr	CC+TB	0.42M	0.23M
nl	EP+TB	1.97M	1.57M
pl	EP+TB	0.62M	0.55M
pt	EP+TB	2.00M	1.85M
ro	EP+TB	0.38M	0.36M
ru	OS+TB	2.27M	1.79M
ta	CC+TB	0.55M	0.36M
te	CC+TB	0.38M	0.17M
uk	OS+TB	0.43M	0.37M
vi	OS+TB	1.72M	1.42M
zh	OS+TB	2.00M	1.57M

Table 7: Statistics on bitext source for each language. Languages are encoded in ISO 639-1 Codes.

B. Selection of Syntactic Parser

Since the release of UP1.0 several high quality deep learning based parsers have been proposed. In UP2.0 we choose the parser with SoTA performance. We evaluate Stanza (Qi et al., 2020), UDPipe (Straka, 2018) and spaCy parsers using pre-trained models on UniversalDependencies 2.5[12] test datasets for 8 languages: German (DE), English (EN), Spanish (ES), French (FR), Italian (IT), Polish (PL), Portuguese (PT) and Chinese (ZH) in Table 8. We use part of speech (POS) accuracy, unlabeled attachment score (UAS) and labeled attachment score (LAS) to measure the performance of each parser. Table 8 summarizes the performance of each parser. Except ES and ZH Stanza provides the best parsing results, however Stanza has comparable performance for these languages. Therefore, we choose Stanza as syntactic parser in all our experiments. During pre-processing step we remove all the sentences from consideration for which Stanza returns multiple sentences regardless of which subset of Bitext it belongs to. We also remove all Multi-word tokens[13] from Stanza results to avoid any word alignments issues.

C. Selection of Word Aligner

The quality of word alignments also impacts the quality of projected labels. Table 9 presents the results of evaluation on the three word aligners: one statistical - BerkeleyAligner[14] and other two are deep learning based models: Awesome-align (Dou and Neubig, 2021) and SimAlign (Sabet et al., 2021). We evaluate the performance of word aligners on EN-DE, EN-FR and EN-ZH language pairs[15] in terms of alignment error rate (AER), precision (P), recall (R) and F1 measures (Och and Ney, 2003). Possible alignments are ignored in the EN-FR evaluations. For the neural aligners we use the text features at the 8th layer of the multilingual BERT model bert-base-multilingual-cased[16]

Lang.	Parser	POS	UAS	LAS
DE	spaCy/de_core_news_lg	93.79	47.78	3.85
	spaCy/de_core_news_sm	93.58	47.65	3.77
	spaCy/de_dep_news_trf	92.55	48.08	3.89
	Stanza	**94.95**	**88.38**	**82.74**
	UDPipe	92.47	80.91	74.06
EN	spaCy/en_core_web_lg	91.40	59.04	42.89
	spaCy/en_core_web_sm	90.82	58.54	42.47
	spaCy/en_core_web_trf	93.12	61.50	44.71
	Stanza	**96.30**	**89.84**	**87.03**
	UDPipe	93.63	84.85	80.62
ES	spaCy/es_core_news_lg	91.39	82.73	70.61
	spaCy/es_core_news_sm	90.95	81.72	69.34
	spaCy/es_dep_news_trf	**91.82**	**85.80**	**73.43**
	Stanza	91.68	84.90	72.55
	UDPipe	91.22	80.97	68.53
FR	spaCy/fr_core_news_lg	88.33	78.07	69.11
	spaCy/fr_core_news_sm	85.87	73.11	63.90
	spaCy/fr_dep_news_trf	96.62	91.50	85.16
	Stanza	**97.99**	**93.15**	**88.68**
	UDPipe	94.05	87.42	81.03
IT	spaCy/it_core_news_lg	95.75	90.97	87.44
	spaCy/it_core_news_sm	94.02	88.61	83.64
	Stanza	**97.93**	**93.49**	**90.02**
	UDPipe	97.33	90.39	86.94
PL	spaCy/pl_core_news_lg	97.24	90.01	82.24
	spaCy/pl_core_news_sm	96.37	87.49	78.71
	Stanza	**98.40**	**93.81**	**87.88**
	UDPipe	97.12	86.93	78.64
PT	spaCy/pt_core_news_lg	97.56	91.02	87.32
	spaCy/pt_core_news_sm	97.51	91.08	87.62
	Stanza	**98.10**	**95.33**	**93.29**
	UDPipe	91.30	80.74	67.74
ZH	spaCy/zh_core_web_lg	69.55	47.24	14.26
	spaCy/zh_core_web_sm	61.09	37.45	9.43
	spaCy / trf	74.02	54.21	19.21
	Stanza	83.31	69.03	59.73
	UDPipe	**88.58**	**72.23**	**59.98**

Table 8: Comparison of different syntactic parsers on UD release 2.5. The best performance for each language is in **bold**.

[12] https://github.com/UniversalDependencies

[13] https://stanfordnlp.github.io/stanza/mwt.html

[14] https://code.google.com/archive/p/berkeleyaligner/

[15] https://github.com/neulab/awesome-align

[16] https://github.com/google-research/bert/blob/master/multilingual.md

Lang.	Method	AER	P	R	F1
EN-DE	Awesome-align	**20.40**	89.19	71.88	**79.60**
	BerkeleyAligner	39.40	67.38	55.07	60.60
	SimAlign/argmax	24.16	**92.15**	64.44	75.84
	SimAlign/itermax	21.73	84.92	**72.58**	78.27
	SimAlign/match	26.03	78.34	70.06	73.97
EN-FR	Awesome-align	24.54	63.05	93.96	75.46
	BerkeleyAligner	34.71	52.99	85.04	65.29
	SimAlign/argmax	**21.85**	**68.28**	91.36	**78.15**
	SimAlign/itermax	27.43	58.67	**95.10**	72.57
	SimAlign/match	30.48	55.54	92.92	69.52
EN-ZH	Awesome-align	**18.43**	81.82	**81.32**	**81.57**
	BerkeleyAligner	37.89	62.11	62.12	62.11
	SimAlign/argmax	22.96	**88.39**	68.28	77.04
	SimAlign/itermax	22.08	78.00	77.83	77.92
	SimAlign/match	26.15	72.65	75.08	73.85

Table 9: Comparison of different word aligners. The best performance for each language pair is in **bold**.

for alignments extraction. For Awesome-align we use the default parameters as defined in Dou and Neubig (2021). For SimAlign we run word tokens evaluations for three alignments extraction methods: argmax, itermax, match.

From Table 9, all neural aligners are better than the statistical aligner. However, it is not immediately clear which neural aligner is the best as both have the comparable performance. Therefore, we analyse the impact of AER, precision and recall measures evaluated for neural word aligners on annotation projection performance. We observe that aligner with high recall has larger impact on the quality of generated propbanks. Based on these observations we use SimAlign with itermax word alignments extraction method *SimAlign/itermax* as word aligner for all our experiments.