Unsupervised Cross-Lingual Part-of-Speech Tagging for Truly Low-Resource Scenarios

Ramy Eskander
Department of Computer Science
Columbia University
rnd2110@columbia.edu

Smaranda Muresan
Data Science Institute
Columbia University
smara@columbia.edu

Michael Collins
Department of Computer Science
Columbia University
mc3354@columbia.edu

Abstract

We describe a fully unsupervised cross-lingual transfer approach for part-of-speech (POS) tagging under a truly low resource scenario. We assume access to parallel translations between the target language and one or more source languages for which POS taggers are available. We use the Bible as parallel data in our experiments: small size, out-of-domain and covering many diverse languages. Our approach innovates in three ways: 1) a robust approach of selecting training instances via cross-lingual annotation projection that exploits best practices of unsupervised type and token constraints, word-alignment confidence and density of projected POS, 2) a Bi-LSTM architecture that uses contextualized word embeddings, affix embeddings and hierarchical Brown clusters, and 3) an evaluation on 12 diverse languages in terms of language family and morphological typology. In spite of the use of limited and out-of-domain parallel data, our experiments demonstrate significant improvements in accuracy over previous work. In addition, we show that using multi-source information, either via projection or output combination, improves the performance for most target languages.

1 Introduction

Majority of world's languages do not have annotated datasets even for the most simple NLP tasks such as part-of-speech (POS) tagging. However, efforts in documenting low-resource languages often contain translations, usually of religious text, into other high-resource languages. One such parallel corpus is the Bible (Mayer and Cysouw, 2014): 484 languages have a complete Bible translation, while 2551 have a part of the Bible translated. Our goal is to learn POS taggers for a diverse set of target languages in a truly low-resource scenario, where only a limited and possibly out-of-domain set of translations into one or more high-resource languages is available (e.g., the Bible), together with supervised POS taggers for the high-resource source language(s).

Unsupervised cross-lingual POS tagging via annotation projection has a long research history (Yarowsky et al., 2001; Fossum and Abney, 2005; Das and Petrov, 2011; Duong et al., 2013; Agić et al., 2015, 2016; Buys and Botha, 2016). In contrast to our work, these approaches either use large and/or in-domain parallel data or rely on a large number of source languages for projection. However, since projection could suffer from bad translation, alignment mistakes or wrong assumptions, a key consideration for all these approaches is how to obtain high-quality training instances for the target language (i.e., sentences with accurate POS tags projected from the source-language(s)). Coupling token and type constraints (Das and Petrov, 2011; Täckström et al., 2013; Buys and Botha, 2016), word-alignment confidence (Duong et al., 2013), multi-source projection (Agić et al., 2016) and coverage (percentage of tokens covered by multi-source projection) (Plank and Agić, 2018) have shown to lead to training instances of better quality. However, only one or two of these have been usually employed.

Our first contribution is a robust approach for selecting training instances via cross-lingual annotation projection that exploits and expands all these best practices: coupling type and token constraints obtained in an unsupervised way, word-alignment confidence *together* with the density of the projected POS, and (optionally) multi-source projection (Sub-section 2.1).

Our second contribution is a BiLSTM (Hochreiter and Schmidhuber, 1997) neural architecture that uses pre-trained contextualized word embeddings, affix embeddings and hierarchical Brown clusters (Brown et al., 1992). As contextualized embeddings, we show gains by exploiting the mul-

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4820–4831,
November 16–20, 2020.

tilingual *XML-R* model (Conneau et al., 2019), while affix embeddings are particularly useful for morphologically-rich languages, and word clusters have been shown to be useful for non-neural POS tagging (Kupiec, 1992; Täckström et al., 2013; Owoputi et al., 2012). Moreover, in addition to the single-source setups, we propose an approach that utilizes multiple source languages by combining the outputs of single-source taggers via weighted voting at the token level (Sub-section 2.2).

Our third contribution is an extensive experimental setup, with 12 diverse target languages in terms of language family and morphological typology and six high-resource source languages (Section 3). While projecting from a single source language can be efficient, we show that using multiple sources, either via projection or output combination, further improves the tagging accuracy for most target languages. Our experiments, using limited and out-of-domain parallel data, demonstrate significant improvements over previous work (both unsupervised and semi-supervised), even when comparing our single-source setups to other multi-source ones. We also investigate how much gold data is needed to develop supervised taggers comparable to our best unsupervised models. In addition, we show that cross-lingual annotation projection generalizes across languages of different typologies better than the zero-shot model-transfer approach by Pires et al. (2019). Finally, our tagging scripts and models are made publicly available [1].

2 Approach

Our goal is to induce a neural POS tagger for a target language of interest without any direct supervision. Instead, we rely on parallel translations between the target and one or more source languages for which POS taggers are accessible. This section describes our approach: 1) cross-lingual annotation projection via word alignments to prepare the training instances of the target language, and 2) neural POS tagging for the target language.

2.1 Cross-Lingual Projection via Word Alignments

Given sentence-aligned parallel data, we align the text of the source and target sides at the word level using GIZA++ (Och and Ney, 2003), while sentences of more than 80 tokens are eliminated. We construct bidirectional word alignments, by only considering the intersecting source-to-target and target-to-source alignments, and exclude the alignment points where the average of the alignment probabilities in the two directions is below some threshold α.

Tagging of Source Languages. Since cross-lingual projection requires a common POS tagset for all languages, we use the universal POS tagset of the Universal Dependencies (UD) project [2], which consists of 17 universal POS tags. We rely on off-the-shelf taggers to tag the source text prior to projecting the annotations as described next.

POS Projection using Token and Type Constraints. To project the POS tags from the source to the target language, we use token and type constraints based on the mapping induced by the word-level alignments. The idea of using both token and type constraints was first introduced by Täckström et al. (2013). Type constraints define the set of POS tags a word type can receive. In a semi-supervised leaning setup, type constraints can be obtained from an annotated corpus (Banko and Moore, 2004) or from a resource that serves as a POS lookup such as the Wiktionary [3] (Li et al., 2012; Täckström et al., 2013). For the extraction of type constraints in an unsupervised fashion, we follow the approach of (Buys and Botha, 2016), where we define a tag distribution for each word type on the target side by accumulating the counts of the different POS tags of the source-side tokens that align with the target-side tokens of that word type. The POS tags whose probability is equal to or greater than some threshold β constitute the type constraints of the underlying word type. As token constraint, every aligned token on the target side gets assigned the POS tag of its corresponding source-side token.

We combine both token and type constraints in a slightly different way than Täckström et al. (2013) and Buys and Botha (2016). If a token is not aligned, or its token constraint does not exist in the underlying type constraints, the token becomes unconstrained (i.e., receives a NULL tag). Otherwise, the token constraint is applied. Those applied token constraints represent the projected tags.

In contrast to the previous work, we do not use the type constraints to impose restrictions when training the model as they restrict the performance of our neural architecture.

[1] https://github.com/rnd2110/unsupervised-cross-lingual-POS-tagging

[2] https://universaldependencies.org/

[3] https://wiktionary.org/

Multilingual Projection. In addition to projecting the POS tags from one language to another, we experiment with a multilingual setup in which we follow Agić et al. (2016) by projecting the tags from multiple source languages prior to training the model ($Multi_{proj}$). The intuition is that the projection from a single source might suffer from inaccurate translation or wrong induced alignments. Moreover, the POS tags of two correctly aligned sentences might differ because of language-dependent specifications. Such problems can be resolved by inducing the tags from multiple sources.

For each target token T, we assign the projected tag that receives the maximum voting, weighted by the alignment confidence for each source.

$$tag(T) = \arg\max_{tag} \sum_{i,s} p(l_s|T) \times P(tag_{i,s}|T)$$

where $p(l_s|T)$ is in $\{0, 1\}$ to represent whether target token T is assigned a tag under the projection from language ls, while $P(tag_{i,s}|T)$ is the probability of the alignment resulting in the assignment of tag_i to target token T when projecting from language ls.

Selection of Training Instances. Prior to training a POS tagger using the projected tags as labels, we score the target sentences based on their "annotation" quality and exclude the ones whose scores are below a threshold γ. We define sentence score as the harmonic mean of density d_S and alignment confidence a_S, where d_S is the percentage of tokens with projected tags, and a_S is the average alignment probability of those tokens.

$$Score(S) = \frac{2\times(d_S \times a_S)}{(d_S + a_S)}$$

Filtering out sentences of low density and alignment confidence is crucial for training the model. While choosing the sentences with top alignment scores has proved successful in previous research (Duong et al., 2013), we add the density factor as our Bi-LSTM model benefits from longer contiguous labeled sequences.

2.2 Neural POS Tagging

The architecture of our POS tagger is a bidirectional long short-term memory (BiLSTM) neural-network model (Hochreiter and Schmidhuber, 1997). BiLSTMs have been widely used for POS tagging (Huang et al., 2015; Wang et al., 2015; Plank et al., 2016; Ma and Hovy, 2016; Cotterell and Heigold, 2017) and other sequence-labeling tasks. The input to our BiLSTM model is a labeled sentence where the word representation is the concatenation of word and sub-word information, namely pre-trained and randomly initialized word embeddings, affix embeddings and word clusters. Figure 1 shows the complete structure of our neural architecture. [4].

Word and Affix Embeddings We use two types of word-embedding features: pre-trained contextualized embeddings (*PT*) and randomly initialized embeddings (*RI*). For the pre-trained contextualized embeddings, we use the final layer of the multilingual *XLM-RoBERTa* model, *XLM-R* (Conneau et al., 2019) [5] *XLM-R* is a transformer-based multilingual masked language model that is pre-trained on texts of 100 languages, and its performance is competitive with strong monolingual models when tested on a variety of NLP tasks. It also shows better performance than multilingual *BERT*, *mBERT* (Devlin et al., 2019), particularly for low-resource languages. We use the average of the embedding vectors of the first and last sub-tokens of each word to represent its pre-trained embeddings.

It is worth noting that when using our architecture for a target language that is not present in the *XLM-R* model, one can consider training a custom *XLM* transformer-based model [6] given the availability of monolingual data and suitable computational resources, and thus our architecture is not limited to the languages available in the *XLM-R* model.

The randomly initialized embeddings are learned as part of training the model. Coupling both the randomly initialized embeddings and the pre-trained ones is essential when the domain of the training data is different from the one of the pre-trained embeddings, which is the case in our learning setup, where we use the Bible data for training, while the *XLM-R* model is trained on text from Wikipedia [7] and a CommonCrawl corpus (See Conneau et al. (2019) for more details).

In addition to word embeddings, we use randomly initialized prefix and suffix n-gram character embeddings, where n is in $\{1, 2, 3, 4\}$, as the use of affix information has proved effective in POS tagging (Ratnaparkhi, 1996; Martins and Kreutzer,

[4] We also experimented with BiLSTM+CRF, but the CRF layer did not improve the model, which is in line with previous research (Yang et al., 2018; Plank and Agić, 2018).

[5] We get better results when using the XLM-R embeddings as features as opposed to performing fine tuning, where the latter is more suitable to sentence-level predictions.

[6] https://github.com/facebookresearch/XLM

[7] https://wikipedia.org

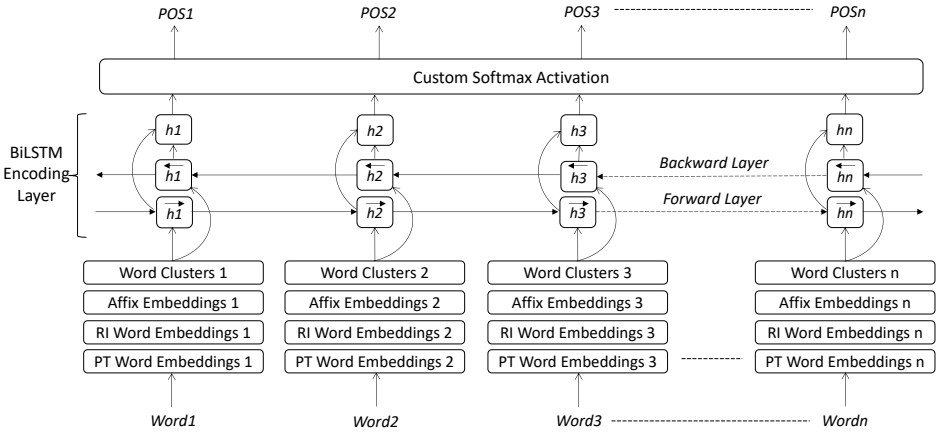

Figure 1: The architecture of our BiLSTM neural-network model. *PT* = pre-trained, *RI* = randomly initialized.

2017).

Word Clusters. The use of word clusters for POS tagging was first proposed by Kupiec (1992) in a supervised tagging setup, and has then proved efficient for unsupervised learning (Täckström et al., 2013; Buys and Botha, 2016). In this work, we follow Owoputi et al. (2012) by utilizing hierarchical Brown clustering (Brown et al., 1992), which is an HMM-based clustering of a binary merging criterion based on the logarithmic probability of a context under a class-based language model, where the objectives is to reduce the loss in adjusted mutual information (AMI).

The output of hierarchical Brown clustering is a binary tree of n leaf nodes that represent n word clusters, where each word in the vocabulary belongs to a single leaf cluster. Leaf clusters are recursively grouped into parent ones (interior nodes) until a super cluster of the entire vocabulary is reached (the root).

We produce hierarchical brown clusters for each target language by applying Percy Liang's implementation of Brown clustering [8] (Liang, 2005) on monolingual text that is a combination of the Wikipedia and Bible texts of the target language.

For each word, we use the main cluster (the binary representation of the corresponding leaf node) and all of its ancestors (the prefixes of the binary representation) as features. This allows us to use the hierarchical clustering information and thus avoid the commitment to a specific granularity level, where high-level clusters may be insufficient, while the lower ones may represent over-clustering.

Custom Softmax Activation. We use softmax activation on top of the BiLSTM encoding layer for the computation of the final output. However, since some words have NULL tags as a result of missing alignments or non-intersecting token and type constraints (Sub-section 2.1), we set the value of the output neuron corresponding to the NULL tag to $-\infty$ so that it does not contribute to the calculation of the softmax probabilities and thus prohibit the model from decoding NULL. Moreover, we mask the words with NULL tags when computing the cross-entropy network loss.

Multilingual Decoding. In addition to the Mul_{proj} setup presented in Sub-section 2.1, we conduct another multilingual setup where we combine the outputs of the single-source taggers through weighted maximum voting at the token level (Mul_{out}). The weight of a language pair, $w(l_s, l_t)$, is measured as a softmax function whose input vector is the average sentence-level alignment probabilities when aligning the source language l_s to the underlying target language l_t.

$$tag(T) = \arg\max_{tag} \sum_{i,s} w(l_s, l_t) \times P(tag_{i,s}|T)$$

Where $P(tag_{i,s}|T)$ is in $\{0, 1\}$ to represent whether target token T is assigned tag_i by the model trained on the projection from language l_s.

3 Experiments and Evaluation

3.1 Languages and Data

We run our experiments on six source languages and 12 target ones [9] for a total of 72 languages pairs.

[8] https://github.com/percyliang/brown-cluster

[9] Although the majority of our target languages are high-resource, we use them in a simulated low-resource scenario.

We choose six widely-spoken source languages as the assumption is that for a low-resource language, a parallel text is highly likely to involve one of them. These languages are English (Indo-European (IE), Germanic), Spanish (IE, Romance), French (IE, Romance), German (IE, Germanic), Russian (IE, Slavic) and Arabic (Afro-Asiatic, Semitic). On the other side, we choose 12 diverse target languages in terms of language family and morphological typology: Afrikaans (IE, Germanic), Amharic (Afro-Asiatic, Semitic), Basque (language isolate), Bulgarian (IE, Slavic), Finnish (Uralic, Finnic), Hindi (IE, Hindi), Indonesian (Austronesian, Malayo-Sumbawan), Lithuanian (IE, Baltic), Persian (IE, Iranian), Portuguese (IE, Romance), Telugu (Dravidian, South Central) and Turkish (Turkic, Southwestern).

We use the multilingual parallel Bible corpus [10] (Christodouloupoulos and Steedman, 2015) as the source of our parallel data, where we perform the alignment on the verse and word levels. The Bible text is available in full for our source and target languages except Basque, where only the new testament is available.

We use Stanza [11] (Qi et al., 2020) to tag the source-side text of the source languages except for Arabic, for which we apply MADAMIRA (Pasha et al., 2014) for performance gain. However, since MADAMIRA was trained on PTB tags and was not designed to follow the UD guidelines, we mapped the Arabic PTB tags into their UD cognates and manually corrected the analyses of the most frequent 2,500 Arabic POS and lemma pairs by selecting the most likely analysis for each.

We evaluate our models in terms of POS accuracy on the test sets of the Universal Dependencies, UD v2.5 (Zeman et al., 2019) [12]. We also report our results on older versions in order to compare to the state-of-the-art systems, whenever needed.

3.2 Experimental Settings

The alignment and projection thresholds as well as the hyperparameters of the model are manually tuned on Bulgarian, Basque, Finnish and Indonesian when projecting from English using the UD development sets. We set the alignment threshold α to 0.1 and the threshold γ for the selection of training instances to 0.5. The POS type distribution threshold β is set to 0.3 as this has proved effective by Banko and Moore (2004) and Buys and Botha (2016). Table 1 lists the number of training sentences per target language based on the tuned thresholds.

Language	Number of Training Sentences	
	One-Source Average	Mul_{proj}
Afrikaans	23,800	30,900
Amharic	10,000	26,700
Basque	7,200	7,900
Bulgarian	21,600	30,400
Finnish	24,000	30,900
Hindi	16,100	30,900
Indonesian	9,600	28,900
Lithuanian	25,700	31,100
Persian	17,500	30,900
Portuguese	26,800	31,100
Telugu	10,100	30,000
Turkish	16,000	30,100

Table 1: Number of training sentences per language (rounded to the nearest 100)

Our BiLSTM networks are one layer deep with 128 nodes, while the size of all the randomly initialized word and affix embeddings is 64, and the number of Brown clusters is set to 128. We use Adam for optimization (Kingma and Ba, 2014) with a learning rate of 0.0001 and a learning decay rate of 0.1 at each epoch for a total of 12 epochs. To avoid overfitting, we apply L2 regularization and two dropout layers, before and after the BiLSTM encoder, with a dropout rate of 0.7. The training rate is approximately 2,500 sentences per hour when utilizing a single 2.00 GHz CPU.

3.3 Results

Table 2 reports the accuracy of our POS taggers for all 72 language pairs, in addition to the two multi-source setups Mul_{out} and Mul_{proj}, based on the average of three runs. As upper bound, we report the state-of-the-art supervised results when training on the UD training sets [13] using Stanza [14] (Qi et al., 2020).

There is a noticeable variance in the performance of the different taggers. However, languages of the same families transfer best across each other. For instance, English and German transfer best to Afrikaans (IE, Germanic), while Spanish yields the best results for Portuguese (IE, Romance), and Russian is the best source for Bulgarian (IE, Slavic).

[10] http://christos-c.com/bible/

[11] https://github.com/stanfordnlp/stanza

[12] Evaluation Corpora: Afrikaans-AfriBooms, Amharic-ATT, Basque-BDT, Bulgarian-BTB, Finnish-TDT, Hindi-HDTB, Indonesian-GSD, Lithuanian-ALKSNIS, Persian-Seraji, Portuguese-Bosque, Telugu-MTG and Turkish-IMST

[13] One exception is Amharic; only a test set is available.

[14] https://stanfordnlp.github.io/stanza/performance.html

Target	Source for Unsupervised Learning								Upper-Bound Supervised
	English	Spanish	French	German	Russian	Arabic	Mul_{out}	Mul_{proj}	
Afrikaans	86.9	83.1	83.9	84.1	76.4	66.1	83.3	**89.3***	97.6
Amharic	75.3	74.6	73.9	75.2	73.3	73.7	77.7	**79.3***	–
Basque	**67.3**	64.6	65.8	66.7	61.7	55.7	66.6	67.1	96.2
Bulgarian	85.6	83.2	83.7	80.7	87.2	72.5	86.9	**88.2***	98.7
Finnish	82.8	80.9	80.0	82.0	78.6	67.2	82.1	**83.4***	97.0
Hindi	73.9	72.3	72.6	60.9	66.9	61.5	**74.1**	72.8	97.6
Indonesian	84.1	83.5	82.9	81.2	82.4	71.3	**84.4**	83.0	93.7
Lithuanian	80.9	78.2	79.0	78.7	**83.3**	70.5	81.5	82.5	93.4
Persian	77.2	78.1	76.1	76.5	78.1	70.6	**79.0***	77.3	97.3
Portuguese	86.1	**88.7**	86.6	81.2	79.5	69.5	88.6	87.8	97.0
Telugu	**80.0**	72.3	73.7	75.6	72.7	65.1	75.6	77.1	92.9
Turkish	74.3	72.7	74.7	72.8	72.0	67.6	**74.9**	74.6	94.2
Average	79.5	77.7	77.7	76.3	76.0	67.6	79.5	**80.2***	96.0

Table 2: POS tagging results (accuracy) when evaluating on the test sets of UD v2.5. The best unsupervised result for each target language is in **bold**, while statistically significant multilingual improvements are marked by *. The last column reports the supervised performance by Stanza

One exception is the case of transferring from Arabic to Amharic (Afro-Asiatic, Semitic). One possible reason is that the Arabic analyzer does not follow the UD guidelines (Sub-section 2.1), which also affects the performance of all the taggers that use Arabic as the source.

Since English is the most vital language, where its morphological-annotation guidelines were the basis for those of other languages, transferring from English yields the best performance for seven target languages. On the target side, the Basque taggers suffer from the lowest performance since the parallel data is only available for the New Testament of the Bible, along with the fact that Basque is a language isolate, which is challenging for cross-lingual transfer learning.

Multi-Source Performance. As expected, the multi-source setups achieve the best on-average results and the best tagging performance for eight target languages. In addition, Mul_{proj} outperforms Mul_{output} in seven occasions, which highlights the importance of producing projected tags of high quality prior to training the taggers. As shown in Table 1, Mul_{proj} results in a significant increase in the number of training sentences, which, along with the quality of the projected tags, gives the best overall performance.

Per-Tag accuracy. Table 3 reports the accuracy of nouns, verbs and adjectives for each target language in the Mul_{proj} setup. The accuracy of adjectives is the lowest across all target languages. The only exception is Persian, where the performance of verbs is lower than that of nouns and adjectives, and it is the lowest among all target languages. In contrast, the accuracy on nouns is the highest on average and across nine languages, where it exceeds 90% in Afrikaans, Bulgarian and Portuguese, while verbs achieve the highest accuracy in Amharic, Indonesian and Telugu. Each of the three tags is ranked second to the lowest in Basque, an isolate with the least available data.

Languages	NOUN	VERB	ADJ
Afrikaans	91.5	85.0	83.6
Amharic	78.1	81.1	43.6
Basque	70.1	61.0	24.9
Bulgarian	92.5	87.8	71.0
Finnish	84.7	79.7	63.1
Hindi	75.2	63.6	53.5
Indonesian	80.2	84.3	56.0
Lithuanian	88.7	86.0	56.2
Persian	85.4	49.0	54.9
Portuguese	92.2	89.0	76.6
Telugu	68.0	80.5	15.0
Turkish	77.2	80.5	44.2
Average	82.0	77.3	53.6

Table 3: Accuracies of nouns, verbs and adjectives for each target language in the Mul_{proj} setup

Ablation Experiments. We conduct two ablation experiments: 1) no XLM-R embeddings (i.e., the target language is not present in the XLM-R model, and no computational resources are available to train one), denoted by *No_XLM*, and 2) no XML-R embeddings and no word clusters (i.e., no monolingual data is available for the target language), denoted by *No_Mono*.

When testing the *No_XLM* and *No_Mono* setups on all 72 language pairs [15], the average accuracy

[15] We double the learning rate as the complexity decreases.

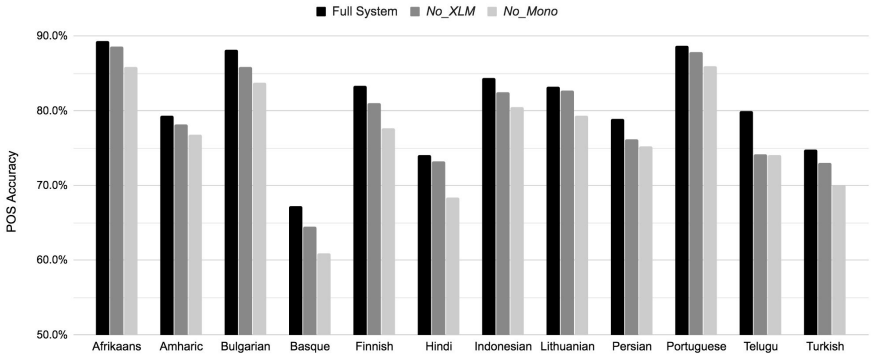

Figure 2: The best performance of each target language in three settings: full system (black), *No_XLM* (dark gray) and *No_Mono* (light gray)

decreases by absolute 2.2% and 5.1%, respectively. However, when projecting from multiple sources in the Mul_{proj} setup, this is reduced to only 1.8% and 4.1%, respectively.

Figure 2 reports the best performance for each target language in three setups: no ablation (full system), *No_XLM* and *No_Mono*. The impact of eliminating the XLM embeddings is most noticeable in Telugu, while it is negligible in Lithuanian, with absolute reduction of 5.8% and 0.6% in POS accuracy, respectively. On the other hand, Hindi benefits most from word clustering, where the *No_Mono* performance is 4.9% below that of *No_XLM*.

The performance drop in the *No_Mono* setup highlights the importance of monolingual data, which is key to the competitive performance of our taggers, especially when compared to other systems that utilize linguistic resources. However, the performance of the system in the absence of only the *XLM-R* embeddings decreases by a small percent, which provides a relatively good compromise when one lacks adequate computational resources.

3.4 Comparison w.r.t. State-of-the-Art

Next, we show that our system outperforms the state-of-the-art unsupervised and semi-supervised cross-lingual POS taggers, where the robust selection of training instances and the rich word representation in the neural architecture are more efficient than using larger and/or domain-appropriate parallel data, some labeled data or off-the-shelf resources encapsulating linguistic knowledge.

We first compare our models to two state-of-the-art systems that perform fully unsupervised cross-lingual POS tagging via annotation projection: AGIC (Agić et al., 2016) and BUYS (Buys and Botha, 2016). AGIC is a multilingual annotation-projection system that is the basis of our Mul_{proj} setup and uses a TnT POS tagger (Brants, 2000) for training. BUYS is a neural model that is based on the Wsabie algorithm (Weston et al., 2011) and utilizes morphological tags projected via coupling token and type constraints.

We report the performance of our system versus AGIC and BUYS on the test sets of UD v1.2 in Table 4. Our taggers outperform both AGIC and BUYS on all the common language pairs with error reduction of 49.1% and 9.0%, respectively, despite the use of smaller and out-of-domain parallel data and only six source languages in the multi-source setup. In contrast, AGIC has the advantage of utilizing 21 source languages for projection, while BUYS uses large-size parallel data, taken from Europarl [16], that is up to 2M tokens whose domain is similar to the one of the UD test sets.

Target	Source	AGIC	Our System
Bulgarian	Multilingual	70.0	**85.6**
Finnish	Multilingual	69.6	**81.2**
Hindi	Multilingual	50.5	**72.9**
Indonesian	Multilingual	75.5	**84.8**
Persian	Multilingual	33.7	**76.7**
Portuguese	Multilingual	84.2	**88.7**
Target	**Source**	**BUYS**	**Our System**
Bulgarian	English	81.8	**83.3**
Finnish	English	77.1	**80.4**
Portuguese	English	84.3	**84.6**
Portuguese	Spanish	88.0	**89.1**

Table 4: Comparison to AGIC and BUYS

Next, we compare our system to two semi-supervised cross-lingual POS tagging systems: CTRL (Cotterell and Heigold, 2017) and DsDs

[16] http://www.statmt.org/europarl/

(Plank and Agić, 2018). CTRL is a character-level RNN tagger that jointly learns the morphological tags of a high-resource language and the target one, where it has two experimental setups that utilize 100 and 1000 manually annotated target tokens, denoted by D100 and D1000, respectively. DsDs is a BiLSTM tagger that follows the annotation-projection approach by Agić et al. (2016) and utilizes the Polyglot embeddings (Al-Rfou' et al., 2013) and lexical information from the Wiktionary.

Table 5 reports the performance of our system versus CTRL on the test sets of UD v2, and versus DsDs on the development sets of UD v2.1 using the 12 universal tags of Petrov et al. (2012) (only Basque is evaluated on the test set). Our system outperforms CTRL except in the D1000 setup of Portuguese, where our results are still comparable. Our system also outperforms DsDs when evaluated on four language pairs out of six, with an overall error reduction of 43.7%.

Target	Source	CTRL	Our System
Bulgarian	Russian-D100	68.8	**87.2**
Bulgarian	Russian-D1000	83.1	**87.2**
Portuguese	Spanish-D100	81.8	**88.7**
Portuguese	Spanish-D1000	**88.9**	88.7

Target	Source	DsDs	Our System
Basque	Multilingual	62.7	**76.5**
Bulgarian	Multilingual	**89.7**	89.3
Finnish	Multilingual	82.4	**85.6**
Hindi	Multilingual	66.2	**84.0**
Persian	Multilingual	43.8	**80.6**
Portuguese	Multilingual	**92.2**	**92.2**

Table 5: Comparison to CTRL and DsDs

3.5 Annotation Projection vs. Supervision

The comparison to the upper-bound supervised results in Table 2 shows that the unsupervised Afrikaans, Indonesian and Portuguese taggers successfully predict at least 90% of the correct decisions made by their corresponding supervised ones. The impact of such small gaps could be tolerable when utilizing the taggers as part of downstream tasks, and thus the trade-off between developing an unsupervised tagger versus an expensive supervised one (if possible) should be considered.

Next, for each target language, except Amharic, we estimate the amount of manual annotations needed to develop a supervised tagger that approximates the performance of the unsupervised one. We do so by iteratively training [17] and evaluating POS taggers in increments of 100 words until the target performance is reached. We list the results in Table 6 with respect to the best unsupervised results in Table 2. On average, it is required to annotate 3,773 words to develop an equivalent supervised tagger, where the training sizes range from 1,200 words, for Basque and Telugu, to 9,000 words, for Lithuanian.

Languages	Annotation Size	POS Acc.
Afrikaans	5,700	89.3
Basque	1,200	67.3
Bulgarian	2,400	88.2
Finnish	5,600	83.4
Hindi	1,800	74.1
Indonesian	2,900	84.4
Lithuanian	9,000	83.3
Persian	2,200	79.0
Portuguese	6,900	88.7
Telugu	1,200	80.0
Turkish	2,600	74.9
Average	3,773	81.2

Table 6: Training sizes of equivalent supervised taggers

3.6 Annotation Projection vs. Model Transfer

One approach of zero-shot cross-lingual POS tagging is to apply a tagging model trained for a related language. Pires et al. (2019) investigate zero-shot model transfer by fine-tuning the multilingual *BERT* language model, *mBERT* (Devlin et al., 2019), for the POS tagging of some language and applying the fine-tuned model to another. While the approach does not require any translation or annotations on the source side, the pre-trained models do not generalize well across languages of different typologies.

We compare our approach versus zero-shot model transfer when transferring from English to Japanese (different language families and morphological typologies). We utilize the Bible translation, where we use *mBERT* instead of *XLM-R* and train our model for only three epochs in order to replicate the experimental settings by Pires et al. (2019). As shown in Table 7, our approach achieves relative error reduction of 27.6% when evaluated on the Japanese test set from the CoNLL 2017 shared task (Zeman et al., 2017), This result suggests that annotation projection is less sensitive to the relatedness between the source and target languages (which is in line with the results in Table 2), and thus can better generalize across languages of different typologies.

[17] We use the UD training data and the same parameters of the unsupervised setting but for 100 epochs instead of 12.

Target	Source	PIRES	Our System
Japanese	English	49.4	**65.4**

Table 7: Comparison to Pires et al. (2019)

Table 8 reports the macro-average POS accuracies when transferring between languages depending on their typological features: Subject/Object/Verb order (SVO and SOV) and Adjective/Noun order (AN and NA) [18]. In the work of Pires et al. (2019), the best performance is achieved when transferring from a language with similar typological features. In contrast, our system is less sensitive to typological similarities, where the performance of transferring from SVO languages is comparable to that of SOV sources, while both AN and NA targets equally benefit from NA sources. This could be explained since the typological features of the source only contribute to the alignment and projection phases, while training the POS model is fully conducted in the target space after eliminating erroneous annotations.

PIRES	SVO	SOV
SVO	**81.6**	66.5
SOV	64.0	**64.2**

Our System	SVO	SOV
SVO	**82.5**	72.4
SOV	**81.3**	71.3

PIRES	AN	NA
AN	**73.3**	70.9
NA	75.1	**79.6**

Our System	AN	NA
AN	**77.5**	76.8
NA	74.3	**74.4**

Table 8: Macro-average POS accuracies when transferring between SVO/SOV languages and AN/NA languages. Rows = sources, columns = targets

4 Related Work

Unsupervised POS tagging through annotation projection was first proposed by Yarowsky et al. (2001), where they transferred POS tags from English to French and Chinese. The work was then extended by Fossum and Abney (2005), where they combined the outputs of single-source taggers based on different source languages. The multilingual setups were then further explored by Agić et al. (2015) and Agić et al. (2016).

In efforts to increase the coverage of the projected data, Das and Petrov (2011) proposed graph-based label propagation to expand the projected tags on the target side, while Duong et al. (2013) and Agić et al. (2015) applied self-training and revision, where they performed the projection and training in iterations. On another side, Täckström et al. (2013) and Buys and Botha (2016) organized the projection process through the use of token and type constraints, which we adapt in our approach.

Semi-supervised setups have been explored by either restricting the type constraints through the use of a POS dictionary (Täckström et al., 2013) or by adding additional signals in training, either by using a POS dictionary (Kirov et al., 2018; Plank and Agić, 2018) or by combining manual and projected annotations (Fang and Cohn, 2016). In contrast, our system is fully unsupervised, where we show that the robust construction of the training data can surpass the use of external resources.

While most prior work does tagging for several target languages, and so is our work, some research focuses on specific language pairs such as projecting from German to Hittite (Sukhareva et al., 2017) and from Russian to Ukrainian (Huck et al., 2019).

5 Conclusion and Future Work

We presented a fully unsupervised cross-lingual POS tagger that does annotation projection by utilizing translation from one or more source languages into the target one. We showed that despite the use of limited and out-of-domain parallel data, our models outperform the state-of-the-art systems. We also showed that the robust selection of training instances and the rich word representation in our neural architecture are more efficient than utilizing some labeled data or external linguistic resources.

In the future, we plan to enhance the system for handling morphologically complex languages trough unsupervised morphological segmentation. One approach is to perform the alignment and projection on the stem and morpheme levels. In addition, stem and morpheme information can be utilized as additional signals in training.

Acknowledgements

This research is based upon work supported by the Intelligence Advanced Research Projects Activity (IARPA), (contract FA8650-17-C-9117). The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing official policies, ex-pressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copy-right annotation therein.

[18] Strictly speaking, the numbers are not comparable as the languages are different. However, they provide insight into how the two approaches perform across languages of different typological features.

References

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Banko and Robert C Moore. 2004. Part of speech tagging in context. In *Proceedings of the 20th international conference on Computational Linguistics*, page 556. Association for Computational Linguistics.

Thorsten Brants. 2000. Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231. Association for Computational Linguistics.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Jan Buys and Jan A. Botha. 2016. Cross-lingual morphological tagging for low-resource languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1954–1964, Berlin, Germany.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 600–609. 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 NAACL 2019*.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. 2013. Simpler unsupervised pos tagging with bilingual projections. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 634–639.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource POS tagging using cross-lingual projection. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 178–186, Berlin, Germany. Association for Computational Linguistics.

Victoria Fossum and Steven Abney. 2005. Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora. In *International Conference on Natural Language Processing*, pages 862–873. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Matthias Huck, Diana Dutka, and Alexander Fraser. 2019. Cross-lingual annotation projection is effective for neural part-of-speech tagging. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 223–233.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Julian Kupiec. 1992. Robust part-of-speech tagging using a hidden markov model. *Computer speech & language*, 6(3):225–242.

Shen Li, Joao V Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1389–1398. Association for Computational Linguistics.

Percy Liang. 2005. *Semi-supervised learning for natural language*. Ph.D. thesis, Massachusetts Institute of Technology.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

André FT Martins and Julia Kreutzer. 2017. Learning what's easy: Fully differentiable neural easy-first taggers. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 349–362.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. 2012. Part-of-speech tagging for twitter: Word clusters and other advances. *School of Computer Science, Carnegie Mellon University, Tech. Rep*.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, volume 14, pages 1094–1101.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Barbara Plank and Željko Agić. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 614–620, Brussels, Belgium. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*.

Maria Sukhareva, Francesco Fuscagni, Johannes Daxenberger, Susanne Görke, Doris Prechel, and Iryna Gurevych. 2017. Distantly supervised pos tagging of low-resource languages under extreme data sparsity: The case of hittite. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 95–104.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2015. Part-of-speech tagging with bidirectional long short-term memory recurrent neural network. *arXiv preprint arXiv:1510.06168*.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, and et al. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Zeman, Martin Popel, Milan Straka, and et al. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.