Unsupervised Stem-based Cross-lingual Part-of-Speech Tagging for Morphologically Rich Low-Resource Languages

Ramy Eskander[1] and **Cass Lowry**[2] and **Sujay Khandagale**[1],
Judith Klavans[3] and **Maria Polinsky**[3] and **Smaranda Muresan**[1]
[1]Columbia University, {rnd2110,sk4746,smara}@columbia.edu
[2]The Graduate Center, City University of New York, clowry@gradcenter.cuny.edu
[3]University of Maryland, {jklavans,polinsky}@umd.edu

Abstract

Unsupervised cross-lingual projection for part-of-speech (POS) tagging relies on the use of parallel data to project POS tags from a source language for which a POS tagger is available onto a target language across word-level alignments. The projected tags then form the basis for learning a POS model for the target language. However, languages with rich morphology often yield sparse word alignments because words corresponding to the same citation form do not align well. We hypothesize that for morphologically complex languages, it is more efficient to use the stem rather than the word as the core unit of abstraction. Our contributions are: 1) we propose an unsupervised stem-based cross-lingual approach for POS tagging for low-resource languages of rich morphology; 2) we further investigate morpheme-level alignment and projection; and 3) we examine whether the use of linguistic priors for morphological segmentation improves POS tagging. We conduct experiments using six source languages and eight morphologically complex target languages of diverse typologies. Our results show that the stem-based approach improves the POS models for all the target languages, with an average relative error reduction of 10.3% in accuracy per target language, and outperforms the word-based approach that operates on three-times more data for about two thirds of the language pairs we consider. Moreover, we show that morpheme-level alignment and projection and the use of linguistic priors for morphological segmentation further improve POS tagging.

1 Introduction

Low-resource languages lack annotated data even for basic syntactic information such as parts of speech (POS). To address this problem, two main unsupervised approaches have been adopted: zero-shot model transfer (Pires et al., 2019) and cross-lingual POS tagging via alignment and projection (Yarowsky et al., 2001; Fossum and Abney, 2005; Das and Petrov, 2011; Duong et al., 2013; Täckström et al., 2013; Agić et al., 2015, 2016; Buys and Botha, 2016; Eskander et al., 2020b). Eskander et al. (2020b) show that the alignment and projection approach is less sensitive to the morphological dissimilarities between the source and target languages than zero-shot model transfer.

In annotation projection, the word structure in the source and target languages impacts the quality of the alignment and projection phases, and hence affects the overall performance of the ultimate POS model. This becomes a concern for languages with rich word structure where affixation is common as they usually suffer from sparse alignment models that often fail to align words corresponding to the same citation form due to the extensive paradigms and translation inconsistencies. Sparse alignment hinders the ability of a system to project the tags properly and results in null tags on the target side. These null tags then reduce the number of qualifying training examples and impact the POS model by introducing non-continuous labeled sequences. Adding to these practical issues, the concept of word as a unit of structure has long been questioned in language sciences (Marantz, 2001).

We hypothesize that using the stem as the core unit of abstraction results in better POS models for low-resource languages of rich morphology. Our contribution is three-fold.

Unsupervised stem-based cross-lingual approach for POS tagging for morphologically complex low-resource languages, where we use the stem as the core unit of abstraction. In order to adapt a fully-unsupervised approach, we use a state-of-the-art unsupervised morphological segmenter, MorphAGram (Eskander et al., 2016, 2020a), to derive the stems and morphemes. We follow the setup of Eskander et al. (2020b) using the Bible as the only source of parallel data in order to emulate a low-resource scenario. We experiment with

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4061 - 4072

the same six source languages, namely English, Spanish, French, German, Russian and Arabic, but choose six morphologically complex target languages, namely Amharic, Basque, Finnish, Indonesian, Telugu and Turkish and add two new target languages, namely Georgian and Kazakh, where we contribute a small POS-annotated dataset for the former. We show that the stem-based approach outperforms the word-based one in 43 language pairs out of 48, with an average relative error reduction of 10.3% in accuracy per target language, up to 21.0% in the case of Kazakh. We also show that the stem-based approach outperforms the word-based approach which operates on three-times more data for about two thirds of the experimental pairs.

Morpheme-level alignment and projection, which allows for abstracting away from how the morphemes are combined in the source and target languages. We test the setup with Arabic as the source language and show improvements for seven out of the eight target languages.

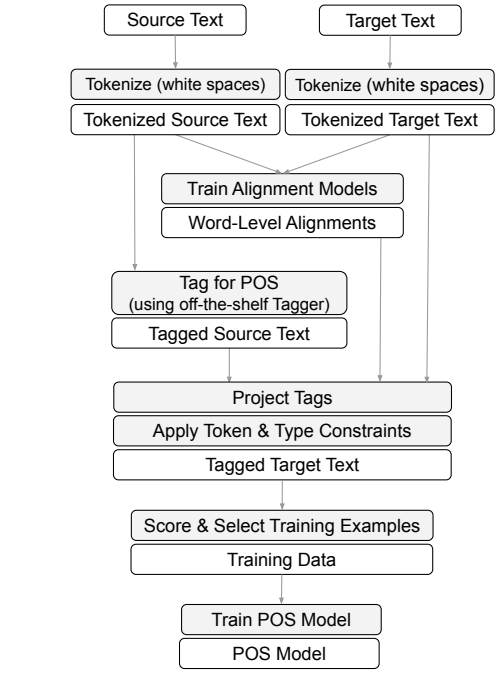
Using linguistic priors in morphological segmentation, which results in better segmentation models towards better alignment and projection. Using Georgian as a case study, we show that the use of linguistic priors, in the form of a set of affixes provided by an expert in the target language, improves the ultimate POS models.

Finally, we make our code publicly available to encourage further research [1].

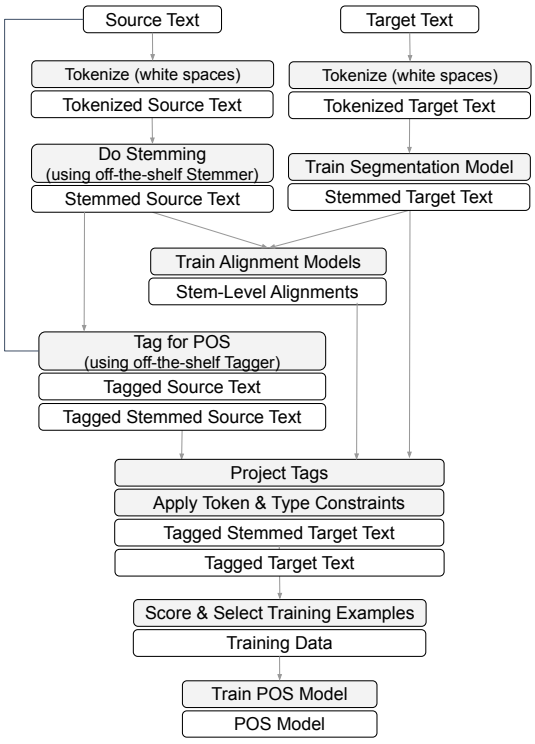
2 Approach

We perform fully unsupervised cross-lingual POS tagging via alignment and projection. We follow the main architecture presented by Eskander et al. (2020b) (Section 2.1). A primary difference is that we harness unsupervised morphological segmentation to use the stems as the core unit of abstraction for both alignment and projection (Section 2.2). In addition, we experiment with morpheme-level alignment and projection (Section2.3) and examine the use of linguistic priors towards better morphological segmentation and POS tagging (Section 2.4). This allows for less sparse alignment models and denser projections, which in turn produces larger POS training data of a better quality.

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



(a) Word-based approach (Eskander et al., 2020b)

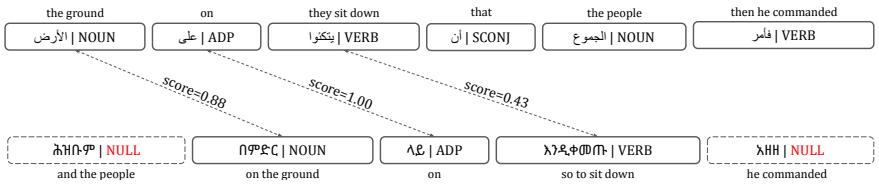


(b) Our proposed stem-based approach

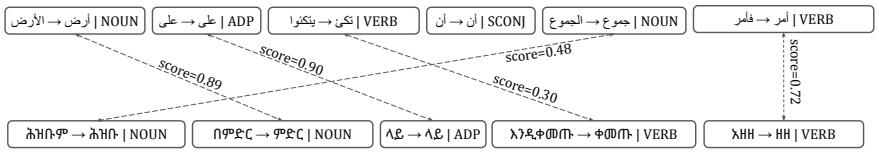
Figure 1: Unsupervised cross-lingual POS tagging via alignment and projection

2.1 Word-based Alignment and Projection

Figure 1a shows the pipeline presented by Eskander et al. (2020b). The only input to the process is a parallel text between the target language and a source one for which a POS tagger is accessible. First,



(a) Word-based alignment and projection



(b) Stem-based alignment and projection

Figure 2: An example of alignment and projection from Arabic onto Amharic. The alignment models are trained on the New Testament. Arabic reads right to left.

the parallel text gets white-space tokenized and is used to train two word-alignment models (source-to-target and target-to-source) using GIZA++ (Och and Ney, 2003). The alignment models are then applied to align the source and target sides on the word level. Next, the source side is tagged for POS using an off-the-shelf tagger (e.g., Stanza (Qi et al., 2020)). The source tags are then projected onto the target across the word-based alignments, where only those bidirectional alignments whose confidence is above a particular threshold are considered in order to eliminate one-to-many and many-to-one alignments and those alignments of low confidence.

The projected tags for each token represent what are called token constraints, while the tag distribution of each word type across the whole target side forms type constraints. The token and type constraints are then coupled by nullifying those tag assignments whose type-level probabilities are below some threshold. The target text, along with its projected tags, then constitutes the training data for a neural POS tagger, where only the top scoring sentences, in terms of tag-assignment density and alignment confidence, are considered.

The neural tagger is a bidirectional long short-term memory (BiLSTM) model (Hochreiter and Schmidhuber, 1997) that uses a custom softmax activation to handle the null tags. It uses word embeddings, both randomly initialized and the contextual multilingual embeddings *XLM-R* (Conneau et al., 2019); prefix and suffix n-gram character embeddings, where n is in $\{1, 2, 3, 4\}$; and hierarchical Brown-cluster (Brown et al., 1992) embeddings.

The architecture can benefit from parallel data of multiple source languages, where either the projections from multiple source languages (Mul_proj) or the decoded outputs that are based on multiple single-source models (Mul_out) can be combined through maximum-voting mechanisms.

2.2 Stem-Based Alignment and Projection

While the architecture by Eskander et al. (2020b) yields the state-of-the-art results for unsupervised POS tagging when evaluated on 12 languages of diverse typologies, the complexity of word structure in the source and target languages has a direct impact on the quality of both alignment and projection. Rich word structure where affixation is common increases the ratio of word types to word tokens, which in turn results in sparse alignment models and incomplete projections that form null tags on the target side. Null tags result in a score that is too low for the underlying sentence to qualify as a training example and introduce missing information for the training of the POS model, which negatively impacts the overall quality of POS tagging.

An example is shown in Figure 2a, where Arabic and Amharic are the source and target languages, respectively. The example corresponds to verse *MAT 15:35*, *"He commanded the multitude to sit down on the ground"*, where the word-alignment models are trained on the New Testament. As shown, the two Arabic-Amharic pairs {ሕዝቡም, الجموع} (*and the people*, *the people*) and {አዘዘ, فأمر} (*he commanded*, *then he commanded*) are not aligned, resulting in null tags. The sparse word-

Verse	Amharic Word		Arabic Word	
MAT 15:35	ሕዝቡም	(*and the people*)	الجموع	(*the people*)
MAT 26:55	ለሕዝቡ	(*to the people*)	للجموع	(*to the people*)
LUK 9:11	ሕዝቡም	(*and the people*)	فالجموع	(*and the people*
LUK 23:4	ለሕዝቡ	(*to the people*)	والجموع	(*and the people*)
MAT 1:24	እንዳዘዘው	(*as he commanded*)	أمره	(*he commanded him*)
MAT 15:35	አዘዘ	(*he commanded*)	فأمر	(*then he commanded*)
b.LUK 5:14	እንዳዘዘ	(*as he commanded*)	أمر	(*he commanded*)
b.ACT 21:34	አዘዘ	(*he commanded*)	أمر	(*he commanded*)

Table 1: Paired inflected forms that correspond to the same citation form across Arabic and Amharic parallel verses in the New Testament

alignment models are simply unable to properly align words that correspond to the same citation form because of the extensive paradigms, which, along with translation inconsistencies, leads to the loss of the one-to-one correspondence between word structures across parallel texts (examples are shown in Table 1). Using the stem instead of the word as the core unit of abstraction is more productive; the stem is usually shared by all the members of a paradigm, which reduces misalignment.

Figure 2b shows that stemming the Arabic and Amharic texts yields complete one-to-one alignments and projections, which in turn eliminates the word-based null assignments and assigns each word on the Amharic side a valid POS tag.

Figure 1b illustrates our overall stem-based approach. We first stem the source and target sides and train two stem-level alignment models, one in each direction. Next, we assign the stems of the source side the POS tags of their corresponding words, which are then projected onto the target stems through the stem-level alignments. We then apply the token and type constraints on the labeled stems on the target side. However, since we train the ultimate POS model on the word level, we replace each target stem by its corresponding word and assign that word the stem-based projected POS tag. The rest of the pipeline for sentence selection and training the POS model are the same as in the word-based architecture described in Section 2.1.

We assume that the source language is a high-resource one for which an off-the-shelf stemmer is accessible. On the other hand, for the target languages, we use MorphAGram[2] (Eskander et al., 2020a) to train an unsupervised morphological segmentation model using the target side of the parallel text. MorphAGram is a state-of-the-art framework for unsupervised morphological segmentation based on Adaptor Grammars (AGs) (Johnson et al., 2007), nonparametric Bayesian models that generalize Probabilistic Context Free Grammars (PCFGs). We run MorphAGram in a cascaded setup of two learning rounds. In the first round, we train a segmentation model using a language-independent high-precision grammar (PrStSu2a+SM [3]) to obtain a list of morphemes. We then seed these morphemes into the best performing language-independent grammar (PrStSu+SM) for the second round of learning as described by Eskander et al. (2016, 2020a). Both PrStSu2a+SM and PrStSu+SM grammars model the word as a sequence of prefixes, a stem and suffixes, where the affixes are recursively defined in order to model multiple consecutive items.

2.3 Morpheme-Based Alignment and Projection

Next, we perform morpheme-based alignment and projection in a similar fashion as in the stem-based approach (Section 2.2). This approach abstracts away from whether the morphemes in the source and target languages are free-standing or not.

On the source side, each morpheme receives a separate POS tag using an off-the-shelf POS tagger. These tags are then projected onto the target morphemes through bidirectional morpheme-level alignments. We obtain the target morphemes using MorphAGram, where the output of the PrStSu+SM grammar yields prefixes, a stem, and suffixes for each word. However, since we train the POS model on the word level, we replace each sequence of morphemes on the target side by its corresponding word and assign that word the POS tag of the representative morpheme. We define the representative morpheme either as the morpheme whose POS tag ranks the highest among those of the other morphemes[4] (RANK) or as the stem morpheme (STEM).

2.4 Using Linguistic Priors for Segmentation

We hypothesize that better detection of stems yields more robust alignment and projection towards improved POS tagging. Accordingly, instead of conducting morphological segmentation on the target side in a fully unsupervised manner, we follow Eskander et al. (2021) by seeding affix morphemes into the grammar tree prior to training the segmentation model textcolorbluein a minimally super-

[2] https://github.com/rnd2110/MorphAGram

[3] See Eskander et al. (2020a) for grammar definitions.

[4] We use the default POS ranking at https://github.com/coastalcph/ud-conversion-tools

vised fashion; these affixes are generated manually by an expert in the target language. With Georgian as a case study, we examine this setup in the stem-based approach using the PrStSu+SM grammar.

2.5 Featurizing Segmented Data

In this setup, we utilize the unsupervised morphological-segmentation model that is trained on the target side of the parallel text to produce stem, complex-prefix and complex-suffix features, and leverage these features as part of the neural POS model. For training, we use these features as randomly initialized embeddings that we concatenate with the existing word, affix and words-cluster embeddings prior to applying the BiLSTM encoding layer.

3 Experiments and Evaluation

3.1 Languages and Data

We conduct our experiments on six source languages and eight target ones [5], for a total of 48 language pairs. We use the same source languages used by Eskander et al. (2020b), namely English, Spanish, French, German, Russian, and Arabic [6], and experiment with eight typologically diverse target languages: six morphologically rich languages that are largely agglutinative, namely Basque, Finnish, Georgian, Kazakh, Telugu, and Turkish; morphologically rich Amharic, where many morphological alterations rely on consonantal roots; and less morphologically rich Indonesian.

We conduct the experiments in a truly low-resource scenario, where we use the New Testament as the source of our parallel data (unless noted otherwise): limited in size and out-of-domain with respect to the evaluation sets. We use the Multilingual Parallel Bible Corpus [7] (Christodouloupoulos and Steedman, 2015) as the source of data for all the languages, except for Georgian and Kazakh [8].

3.2 Experimental Settings

For the tagging of the source languages, we use the same off-the-shelf taggers as in Eskander et al. (2020b): Stanza [9] (Qi et al., 2020) for English, Spanish, French, German and Russian and MADAMIRA (Pasha et al., 2014) for Arabic. We also use MADAMIRA for Arabic morphological segmentation. For the stemming of the other source languages, we use the *Snowball* Stemmer (Porter, 2001) as part of *NLTK* [10] (Bird and Loper, 2004). On the other hand, we use MorphAGram [11] (Eskander et al., 2020a) to train and apply the morphological-segmentation models for the target languages as described in Section 2.2.

We follow Eskander et al. (2020b) by using the same thresholds for alignment and projection, along with the same neural hyperparameters of the POS tagger. We also evaluate our models in terms of POS accuracy on the same Universal Dependencies (UD) v2.5 (Zeman et al., 2019) test datasets. However, since the UD project does not currently contain Georgian datasets, we developed a small POS dataset for Georgian (100 sentences) following the UD-tagging scheme. The sentences are taken from the Modern Georgian and Political texts sub-corpora of the Georgian National Corpus [12], and they are hand-tagged and carefully revised by a linguist who specializes in and speaks Georgian as a second language [13]. Finally, all the results are averaged over three runs.

3.3 System Performance

Table 2 reports the POS accuracy for the baseline word-based approach and our stem-based approach for all the 48 target-source language pairs using the New Testament as the source of parallel data. In addition, we report the results for the two multilingual setups Mul_out and Mul_proj per target language. The stem-based approach outperforms the word-based one in 43 language pairs and all the multilingual setups except Mul_proj in the case of Indonesian, which stands out in our language sample as the least complex in terms of morphology. The biggest improvement in the stem-based approach is achieved in the cases of Russian $\rightarrow$ Turkish, Russian $\rightarrow$ Kazakh, Spanish $\rightarrow$ Kazakh and Arabic $\rightarrow$ Georgian, with relative error reductions of 33.8%, 30.2%, 28.6% and 27.2%, respectively. When averaging across the sources (includ-

[5] Although most of our target languages are high-resource ones, we use them in a simulated low-resource setup.

[6] The source languages are commonly spoken ones as the assumption is that translations that involve those languages are easily accessible.

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

[8] We collected the New-Testament texts for Georgian and Kazakh from https://github.com/cysouw/MissingBibleVerses.

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

[10] https://www.nltk.org

[11] https://github.com/rnd2110/MorphAGram

[12] http://gnc.gov.ge

[13] https://github.com/rnd2110/unsupervised-cross-lingual-POS-tagging/blob/main/data/KAT-POSUD.txt

Target	Approach	Source for Unsupervised Learning								Ave. Error Reduction
		English	Spanish	French	German	Russian	Arabic	*Mul_out*	*Mul_proj*	
Amharic	Word-based	75.9	74.9	75.5	76.4	72.1	72.6	76.6	78.0	
	Stem-based	**79.6***	**77.5**	**77.7**	**77.8**	**76.2**	**74.5**	**78.6**	**79.6**	9.9
Basque	Word-based	67.3	64.6	65.8	66.7	61.7	55.6	66.4	67.1	
	Stem-based	**69.1**	**70.4***	**70.5**	**69.6**	**65.2**	**60.8**	**71.0**	**71.4**	11.5
Finnish	Word-based	81.0	78.8	77.4	79.8	77.8	66.1	81.0	81.7	
	Stem-based	**81.9**	**80.1**	**80.9***	**82.3**	**79.0**	**70.3**	**82.4**	**82.9**	8.8
Georgian	Word-based	**82.8**	80.1	80.2	**82.5**	83.1	71.2	83.6	84.3	
	Stem-based	82.0	(**80.4**)	**81.0**	82.2	**83.4**	**79.0***	**84.3**	**84.7**	4.4
Indonesian	Word-based	82.3	**81.6**	**81.0**	77.1	76.8	69.8	80.9	**81.7**	
	Stem-based	(**82.5**)	81.0	80.1	(**77.3**)	**81.2***	**72.3**	**81.4**	81.0	2.5
Kazakh	Word-based	73.6	64.7	67.3	68.9	62.1	63.6	69.7	70.3	
	Stem-based	**76.4**	**74.8**	**75.5**	**73.2**	**73.6***	**70.8**	**75.3**	**76.7**	21.0
Telugu	Word-based	76.7	68.4	67.9	70.4	63.5	59.5	68.6	71.3	
	Stem-based	**78.6**	**72.7**	**72.2**	**71.9**	**69.6**	**66.8**	**72.9***	**73.8**	12.1
Turkish	Word-based	**73.9**	70.1	70.5	69.2	66.2	64.7	71.0	73.3	
	Stem-based	73.7	**73.1**	**73.0**	**71.9**	**77.6***	**71.9**	**75.4**	**73.6**	12.1
Ave. Error Reduction		5.0	10.4	10.5	6.8	16.3	15.6	10.6	7.1	

Table 2: POS-tagging performance (accuracy) of the word-based and stem-based approaches when using the New Testament as the source of parallel data. The best result per target-source pair is in **bold**. The highest relative error reduction in the stem-based approach per target language is marked by *. The stem-based improvements that are not statistically significant for $p\text{-}value < 0.01$ are between parentheses.

Target	Approach			
	Word-Based	Stem-Based	Morpheme-Based (RANK)	Morpheme Based (STEM)
Amharic	72.6	**74.5**	72.5	73.6
Basque	55.6	60.8	61.9	**62.2**
Finnish	66.1	70.3	73.8	**74.2**
Georgian	71.2	79.0	**80.5**	80.0
Indonesian	69.8	72.3	75.5	**75.6**
Kazakh	63.6	70.8	71.8	**71.9**
Telugu	59.5	66.8	**74.7**	71.8
Turkish	64.7	71.9	73.2	**73.4**

Table 3: POS-tagging performance (accuracy) of the word-based, stem-based and morpheme-based approaches when projecting from Arabic using the New Testament as the source of parallel data. The best result per target language is in **bold**. All the morpheme-based improvements are statistically significant for $p\text{-}value < 0.01$.

ing the multilingual ones), Kazakh, Telugu and Turkish experience the highest relative error reductions of 21.0%, 12.1% and 12.1%, respectively. On the other hand, Russian and Arabic yield the highest relative error reductions of 16.3% and 15.6%, respectively, when averaging across the target languages, which is in line with the morphological complexity of the two languages.

Eskander et al. (2020b) show that related languages transfer best across each other. This results in efficient word-based baselines for related language pairs, which in turn limits the corresponding gains in the stem-based approach. On the other hand, a low word-based baseline makes room for improvement when operating on the stem level. For instance, both Georgian and Telugu witness the highest stem-based gains when transferring from Arabic, their lowest performing source language in the word-based approach. For more information about the correlation between language relatedness and cross-lingual learning, see Eskander (2021).

Next, we evaluate the morpheme-based approach (Section 2.3). Table 3 reports the performance of the word-based, stem-based and morpheme-based approaches using Arabic as the source language since it is more morphologically complex than the other sources. The morpheme-based approach results in dense training instances as both alignment and projection are performed in a more fine-grained level compared to the word-based and stem-based approaches. It therefore improves POS tagging for all the target languages except Amharic, where Telugu benefits the most with relative error reductions of 23.9% and 15.3% over the stem-based approach using the RANK and STEM mechanisms, respectively. The difference in the performance of the RANK and STEM mechanisms is only statisti-

Target	Approach	Source for Unsupervised Learning								Ave. Error Reduction
		English	Spanish	French	German	Russian	Arabic	Mul_out	Mul_proj	
Georgian	Word-based	82.8	80.1	80.2	**82.5**	83.1	71.2	83.6	84.3	
	Stem-based	82.0	80.4	81.0	82.2	83.4	**79.0**	84.3	84.7	4.6
	LP Stem-based	**82.9**	**(80.8)**	**82.2**	(82.4)	**83.9**	77.4	**85.3**	**(85.1)**	6.7

Table 4: POS-tagging performance (accuracy) of the word-based and stem-based (with and without linguistic priors (LP)) approaches when using the New Testament as the source of parallel data. The best result per source language is in **bold**. The improvements in the LP stem-based approach that are not statistically significant for $p\text{-}value < 0.01$ are between parentheses.

cally significant for $p\text{-}value < 0.01$ in the cases of Amharic and Basque, where the STEM mechanism yields better performance, and in the case of Telugu, where the RANK mechanism is superior. Finally, we hypothesize that the quality of morphological segmentation highly affects the efficiency of the morpheme-based setups, which explains the variation in the performance across the different target languages and mechanisms.

As mentioned earlier, we use Georgian as a case study to examine the impact of using linguistic priors for morphological segmentation on the quality of POS tagging (Section 2.4). The results are listed in Table 4. The use of linguistic priors improves the stem-based approach except when projecting from Arabic. The lack of improvement in the case of Arabic can be explained by over-segmentation that produces incorrect POS tags for the common conjunction ಮತ್ತು (*and*). The characters ಮತ್ತು also correspond to a verbal prefix that is manually seeded as a prior. This seeding causes erroneous projections labeling ಮತ್ತು as a verb or an adverb when projecting from Arabic.

Finally, we experiment with the use of the stem and affix information as training features in the POS neural model (Section 2.5). However, most of the improvements due to the use of these features are not statistically significant (See Appendix A for full results) since such features are surpassed by the prefix and suffix n-gram character embeddings.

3.4 Analysis of the Stem-Based Approach

Upon alignment and projection, the highest scoring target sentences are selected as training examples, where sentence score is defined as the harmonic mean of the percentage of tokens with projected tags and the average alignment probability of those tokens. The fine-grained stem-level alignments allow for better alignment confidence and more dense sentences, which in turn increases sentence scores and the number of training examples, and hence reduces the number of out-of-vocabulary words (OOVs). Table 5 lists the average number of training examples, average relative increase in the number of training examples, average relative increase in sentence scores and average relative decrease in the number of OOVs for each target language in the stem-based approach with respect to the word-based one. We witness improvements in the examined aspects for each target language, which explains the considerable improvements in the stem-based approach.

Next, we examine the average relative error reduction in the detection of open-class tags (nouns, verbs and adjectives) in the stem-based approach as compared to the word-based one per target language (Table 6) and per source language (Table 7). Kazakh benefits the most from the stem-based approach at the detection of nouns and adjectives, while Amharic receives the highest gains for verbs. On the other hand, projecting from Russian in the stem space achieves the highest gains for nouns, while the stem-based projection from Arabic yields the highest gains for both verbs and adjectives.

Finally, Figure 3 illustrates the absolute improvements in POS tagging when applying the stem-based approach using the New-Testament as the source of parallel data compared to the word-based approach using the entire Bible as the source of parallel data (three-times more data). As illustrated, the stem-based approach achieves better performance in about two thirds of the language pairs with an average absolute gain of 1.7%. This means using the stem as the core unit of abstraction compensates for the lack of adequate parallel data.

4 Related Work

The line of work most closely related to ours is unsupervised cross-lingual POS tagging via alignment and projection, which was first introduced

Target	Ave. No. of Training Examples	Ave. Relative Increase in Training Examples %	Ave. Relative Increase in Sentence Scores %	Ave. Relative Decrease in OOVs %
Amharic	2,605	15.0	132.6	9.2
Basque	7,225	15.8	3.9	0.1
Finnish	7,125	6.9	5.5	0.7
Georgian	7,794	12.0	1.9	1.6
Indonesian	5,286	5.5	11.9	0.3
Kazakh	4,330	7.8	21.6	3.5
Telugu	4,719	2.8	14.1	0.7
Turkish	6,280	12.9	14.6	2.7

Table 5: Average number of training examples, average relative increase in training examples, average relative increase in sentence scores and average relative decrease in OOVs per target language in the stem-based approach w.r.t. to the word-based one

Target	Ave. Relative Error Reduction %		
	Noun	Verb	Adjective
Amharic	16.2	13.2	26.2
Basque	9.0	23.1	14.6
Finnish	13.0	16.6	15.8
Georgian	4.7	1.8	17.8
Indonesian	8.3	25.1	11.7
Kazakh	29.8	32.1	23.1
Telugu	17.4	23.7	-1.1
Turkish	26.8	23.0	20.9

Table 6: Average relative error reductions for the detection open-class tags per target language

Target	Ave. Relative Error Reduction %		
	Noun	Verb	Adjective
English	7.1	13.4	5.1
Spanish	16.5	25.2	15.2
French	12.7	25.5	16.9
German	9.5	11.0	22.8
Russian	21.3	21.2	25.7
Arabic	16.0	25.4	25.6

Table 7: Average relative error reductions for the detection of open-class tags per source language

by Yarowsky et al. (2001). They applied noise-reduction techniques to improve the alignments and used the resulting transition and emission probabilities to define an HMM POS tagger.

Exploiting multiple source languages via maximum voting was then explored by Fossum and Abney (2005), by voting among the outputs of different single-source models, and by Agić et al. (2015), by projecting the annotations from multiple languages before training the POS tagger.

In order to increase the size of the training data, Das and Petrov (2011) proposed graph-based label propagation, while Duong et al. (2013); Agić et al. (2015) applied self-training and revision. On another hand, Täckström et al. (2013) and Buys and Botha (2016) investigated the use of token and type constraints to reject projections of low confidence.

Eskander et al. (2020b) derived a cross-lingual POS-tagging pipeline that utilizes the best practices in alignment and projection. In addition, they examined the use of pretrained multilingual contextual embeddings, along with affix embeddings and Brown clusters, within a rich neural architecture, which achieves the state-of-the-art results for unsupervised POS tagging. We follow their approach by presenting stem-based alignment and projection for morphologically complex low-resource languages.

Regarding unsupervised morphological segmentation, several generative and discriminative frameworks have been developed over the last two decades. The two most notable frameworks are: 1) Morfessor (Creutz and Lagus, 2007; Grönroos et al., 2014), a commonly-used HMM framework that utilizes the MDL principle to segment into morphemes of a hierarchical structure; and 2) MorphAGram (Eskander et al., 2020a), a segmentation framework that is based on Adaptor Grammars (AGs) (Johnson et al., 2007), Bayesian models that utilize Probabilistic Context Free Grammars (PCFGs). We use MorphAGram to train morphological-segmentation models as it achieves the state-of-the-art performance and allows for deriving affix and stem information (as opposed to a sequence of unlabeled morphemes).

5 Conclusion and Future Work

We presented a fully unsupervised stem-based approach for cross-lingual POS tagging via alignment and projection, where we use the stem as the core unit of abstraction to abstract away from complex affixation. Our experiments using six source languages and eight morphologically rich target lan-

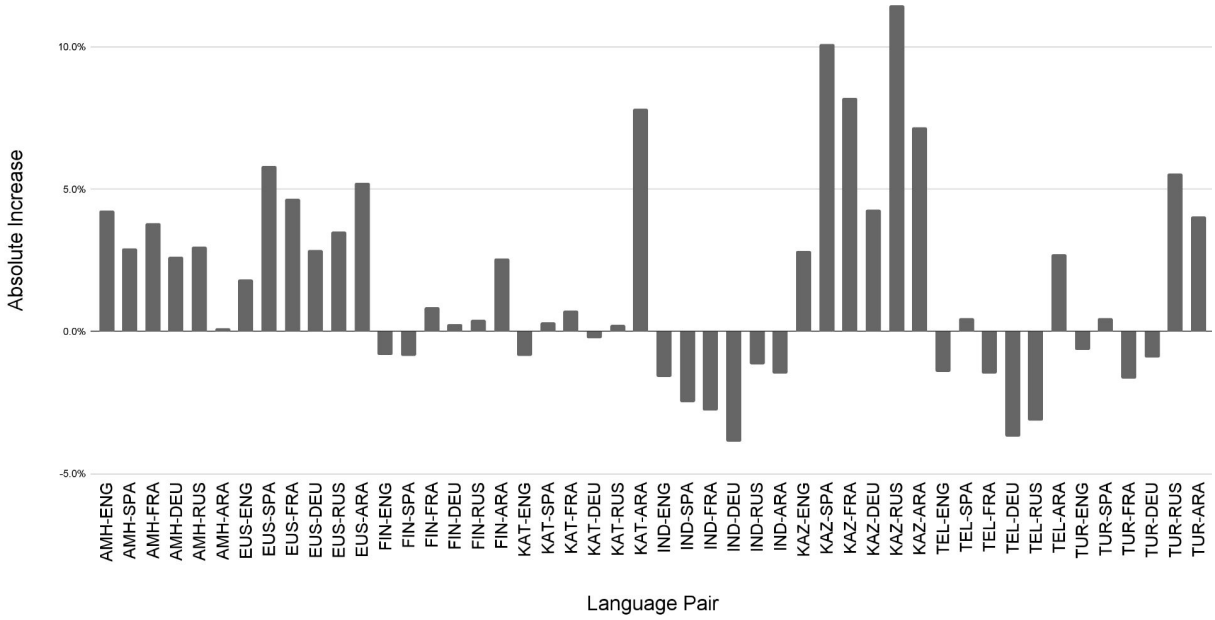


Figure 3: Absolute performance increases (accuracy) when applying the stem-based approach using the New Testament as the source of parallel data as compared to the word-based approach using the entire Bible as the source of parallel data

guages in low-resource setups show improvements over the word-based approach in 43 language pairs out of 48, with an average relative error reduction of 10.3% in accuracy per target language. In addition, we examined morpheme-based alignment and projection and the use of linguistic priors in morphological segmentation, which further improve POS tagging.

In the future, we plan to study the role of morphological typology in cross-lingual learning. This allows for deriving disciplined guidelines for the selection of an appropriate source language that transfers well to the target language of interest.

Acknowledgements

This research is based upon work supported by the Intelligence Advanced Research Projects Activity (IARPA), (contract #FA8650-17-C-9117) and the National Science Foundation (awards #1941742 and #1941733). The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing official policies, expressed or implied, of ODNI, IARPA, NSF or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

6 Ethical Considerations

The Georgian annotations were done by a linguist with appropriate compensation after educating them about the research purpose and the annotation process. We claim ownership of the Georgian dataset for open distribution as the text is taken from the Modern Georgian and Political texts subcorpora of the Georgian National Corpus, which is an open-sourced work that allows for modifications, derived work and redistribution. The quality of the annotations was examined manually and empirically. The source code and the data will be released open-source. Finally, the limitations of the work lay within the reported performance. There should be no potential risks given these stated limitations.

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.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. 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*.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.

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.

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.

Ramy Eskander. 2021. *Unsupervised Morphological Segmentation and Part-of-Speech Tagging for Low-Resource Scenarios*. Columbia University.

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith L Klavans, and Smaranda Muresan. 2020a. Morphagram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7112–7122.

Ramy Eskander, Cass Lowry, Sujay Khandagale, Francesca Callejas, Judith L Klavans, Maria Polinsky, and Smaranda Muresan. 2021. Minimally-supervised morphological segmentation using adaptor grammars with linguistic priors. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3969–3974.

Ramy Eskander, Smaranda Muresan, and Michael Collins. 2020b. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4820–4831.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2016. Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of he Twenty-Sixth International Conference on Computational Linguistics (COLING)*, Osaka, Japan.

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.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of the 2014 International Conference on Computational Linguistics (COLING)*, pages 1177–1185.

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

Mark Johnson, Thomas L Griffiths, Sharon Goldwater, et al. 2007. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems*, 19:641.

Alec Marantz. 2001. Words and things. *handout, MIT*.

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

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.

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.

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

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*.

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.

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.

A Appendix: Segmentation Information as Training Features

Table 2 shows the POS tagging results (accuracy) when using the stem and affix information as training features in the neural POS model, as described in Subsection 2.5.

		Source for Unsupervised Learning							
Target	Approach	English	Spanish	French	German	Russian	Arabic	Mul_out	Mul_proj
Amharic	Word-based	75.9	74.9	75.5	76.4	72.1	72.6	76.6	78.0
	Stem-based	79.6	77.5	77.7	**77.8**	76.2	74.5	78.6	79.6
	Stem-based + Stem Features	(**80.2**)	77.5	(**78.0**)	77.6	(**76.6**)	74.6	**78.7**	**79.7**
	Stem-based + Stem+Affix Features	79.8	**77.7**	77.8	**77.8**	76.5	**74.7**	**78.7**	79.4
Basque	Word-based	67.3	64.6	65.8	66.7	61.7	55.6	66.4	67.1
	Stem-based	**69.1**	70.4	70.5	**69.6**	65.2	60.8	**71.0**	71.4
	Stem-based + Stem Features	68.7	70.5	70.5	69.3	(**65.6**)	60.3	70.9	71.6
	Stem-based + Stem+Affix Features	69.0	**70.6**	**70.8**	69.1	(65.3)	(**62.0**)	70.9	(**71.8**)
Finnish	Word-based	81.0	78.8	77.4	79.8	77.8	66.1	81.0	81.7
	Stem-based	**81.9**	80.1	80.9	82.3	79.0	70.3	82.4	**82.9**
	Stem-based + Stem Features	**81.9**	(**80.4**)	80.9	**82.4**	**79.1**	(70.5)	(**82.7**)	82.7
	Stem-based + Stem+Affix Features	81.8	80.1	(**81.2**)	**82.4**	78.9	(70.6)	(**82.7**)	**82.9**
Georgian	Word-based	**82.8**	80.1	80.2	**82.5**	83.1	71.2	83.6	84.3
	Stem-based	82.0	80.4	81.0	82.2	**83.4**	**79.0**	84.3	84.7
	Stem-based + Stem Features	82.1	**80.5**	(**81.3**)	82.1	83.3	78.7	**84.4**	(**85.0**)
	Stem-based + Stem+Affix Features	81.5	80.3	80.9	81.7	83.1	78.8	83.7	84.3
Indonesian	Word-based	82.3	**81.6**	**81.0**	77.1	76.8	69.8	80.9	**81.7**
	Stem-based	**82.5**	81.0	80.1	77.3	81.2	**72.3**	**81.4**	81.0
	Stem-based + Stem Features	**82.5**	80.8	79.9	(**77.6**)	**81.3**	71.7	81.1	80.9
	Stem-based + Stem+Affix Features	**82.5**	80.9	80.0	77.3	81.0	72.0	81.0	80.6
Kazakh	Word-based	73.6	64.7	67.3	68.9	62.1	63.6	69.7	70.3
	Stem-based	76.4	74.8	75.5	**73.2**	**73.6**	**70.8**	75.3	76.7
	Stem-based + Stem Features	76.3	74.8	75.7	72.8	(**73.6**)	70.7	**75.4**	76.5
	Stem-based + Stem+Affix Features	(**76.6**)	(**75.2**)	(**75.8**)	73.1	**73.6**	**70.8**	75.3	(**76.8**)
Telugu	Word-based	76.7	68.4	67.9	70.4	63.5	59.5	68.6	71.3
	Stem-based	**78.6**	**72.7**	72.2	**71.9**	**69.6**	**66.8**	72.9	**73.8**
	Stem-based + Stem Features	77.9	71.5	**72.7**	**71.9**	**69.6**	66.7	73.1	73.1
	Stem-based + Stem+Affix Features	78.4	72.4	**72.7**	71.4	68.7	67.1	(**73.6**)	73.7
Turkish	Word-based	**73.9**	70.1	70.5	69.2	66.2	64.7	71.0	73.3
	Stem-based	73.7	**73.1**	73.0	**71.9**	**77.6**	**71.9**	**75.4**	73.6
	Stem-based + Stem Features	73.5	73.0	**73.1**	71.5	**77.6**	71.7	75.1	(73.7)
	Stem-based + Stem+Affix Features	73.6	73.0	**73.1**	71.8	**77.6**	71.7	75.3	(**73.9**)

Table 8: Performance with segmentation features

B Appendix: Hardware

We use a Google-Cloud virtual instance of 48 2.00GHz cores and 240GB of RAM to run all of our experiments. The training rate is nearly 2,500 sentences per hour.