If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages

Željko Agić, Dirk Hovy, and Anders Søgaard
Center for Language Technology
University of Copenhagen, Denmark
Njalsgade 140
{zeljko.agic|dirk.hovy|soegaard}@hum.ku.dk

Abstract

We present a simple method for learning part-of-speech taggers for languages like Akawaio, Aukan, or Cakchiquel – languages for which nothing but a translation of parts of the Bible exists. By aggregating over the tags from a few annotated languages and spreading them via word-alignment on the verses, we learn POS taggers for 100 languages, using the languages to bootstrap each other. We evaluate our cross-lingual models on the 25 languages where test sets exist, as well as on another 10 for which we have tag dictionaries. Our approach performs much better (20-30%) than state-of-the-art unsupervised POS taggers induced from Bible translations, and is often competitive with weakly supervised approaches that assume high-quality parallel corpora, representative monolingual corpora with perfect tokenization, and/or tag dictionaries. We make models for all 100 languages available.

1 Introduction

Most previous work in cross-lingual NLP has been limited to training and evaluating on no more than a dozen languages, typically all from the major Indo-European languages. While it has been observed repeatedly that using multiple source languages improves performance (Yarowsky et al., 2001; Yarowsky and Ngai, 2001; Fossum and Abney, 2005; McDonald et al., 2011), most available techniques work best for closely related languages.

In contrast, this paper presents an effort to learn POS taggers for truly low-resource languages, with minimum assumptions about the available language resources. Most low-resource languages are non-Indo-European, and typically, their typological and geographic neighbors have sparse resources as well. However, for a surprisingly large number of languages, translations of the Bible (or parts of it) exist. Due to the canonical nature and the verse format, these translations are viable parallel data, albeit lacking annotation. In our experiments, we use word alignments across all pairs of 100 parallel Bible translations to bootstrap annotation projections for those languages without any (even just weakly) supervised taggers. The projections provide both pseudo-annotated data as well as tag dictionaries for all languages. We use both resources to train semi-supervised POS taggers following Garrette and Baldridge (2013).

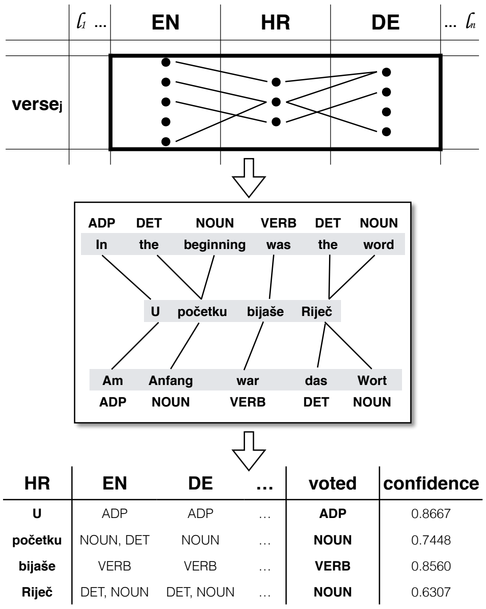
Our contribution We present a novel approach to learning POS taggers for truly low-resource languages, where only a translation of (parts of) the Bible is available. We obtain results competitive with approaches that assume the availability of larger volumes of more representative parallel corpora, perfectly tokenized monolingual corpora, and/or tag dictionaries for the target languages. Additionally, we make the POS tagging models for 100 languages publicly available and extend the mappings in Petrov et al. (2011) for six new languages (Hindi, Croatian, Icelandic, Norwegian, Persian, and Serbian). The models, mappings, as well as a complete list of all the resources used in these experiments, are available at https://bitbucket.org/lowlands/.

2 Experiments

Our approach is a combination of simple techniques. Part of the process is depicted in Figure 1, and the algorithm is presented in Algorithm 1. Assume we have n languages for which we assume the availability of m verses of the Bible. We run IBM-2[1] on all $n(n-1)$ pairs of languages. Assume also manually POS-annotated training data

[1] github.com/clab/fast_align

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 268–272,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics



HR	EN	DE	...	voted	confidence
U	ADP	ADP	...	ADP	0.8667
početku	NOUN, DET	NOUN	...	NOUN	0.7448
bijaše	VERB	VERB	...	VERB	0.8560
Riječ	DET, NOUN	DET, NOUN	...	NOUN	0.6307

Figure 1: An illustration of our approach.

is available for the first k of these languages. We then run taggers for these languages on the corresponding translations of the Bible to predict tags for all tokens in these translations.

We can think of this partially annotated multi-parallel corpus as a tensor object. Each column is a language l_i, and each row a verse v_j (trivially sentence-aligned to the corresponding verses in the other columns). In each cell of this matrix $M(i, j, \cdot)$, we have a sequence of word tokens. For two languages, l_1 and l_2, the word tokens in $M(1, j, \cdot)$ can be aligned (by IBM-2) to multiple word tokens in $M(2, j, \cdot)$, but not all words need to be aligned.

After running supervised POS taggers on the k languages for which we have training data, we have POS-annotated the word tokens in k columns of our tensor object. We then project the POS tag of each word token w to all other word tokens aligned to w. In our experiments, $k = 17$ or 18 (if the target language is not one of the languages for which we have training data), which means each word token will potentially have many POS tags projected onto it. Note that the number of tags can exceed 18, since many-to-many word alignments are allowed.

We now use these projections to train POS taggers for the remaining $n - k$ languages. We use aggregated projected annotations as token-level supervision. We aggregate from the incoming projected POS tags by majority voting. We also use the complete set of projections onto each word type in the target language as a type-level tag dictionary. We combine the tag dictionary and the token-level projections to train discriminative, type-constrained POS taggers (Collins, 2002; Täckström et al., 2013). Below we refer to these POS taggers as using k sources (k-SRC).

These n many POS taggers can now also be used to obtain predictions for all word tokens in our tensor object. This corresponds to doing the second loop over lines 8–17 in Algorithm 1. For each of our n languages, we thus complete the tensor by projecting tags into word tokens from the $n - 1$ remaining source languages. For the k supervised languages, we project the tags produced by the supervised POS taggers rather than the tags obtained by projection. We can then train our final POS taggers for all n languages – 100, in our case – using projections from 99 languages (n-1-SRC). Note that we also train projected taggers for those languages for which we have annotated data. This is to enable us to evaluate our methodology on more languages.

Algorithm 1 Train n taggers with supervision for k

```
 1: Let $M$ be a tensor with $M(i, j, \cdot)$ the word-aligned token
    sequence in the $j$th verse of the Bible in language $i$
 2: for $i \leq k$ do
 3:     Train TNT tagger for $l_i$ using manually annotated data
 4:     for $j \leq m$ do
 5:         Obtain POS predictions for $M(i, j, \cdot)$
 6:     end for
 7: end for
 8: for $I \in \{0, 1\}$ do
 9:     if $i > k, I = 1$ then
10:         Train TNT tagger for $l_i$ using projected annota-
            tions in $M(i, \cdot, \cdot)$
11:     end if
12:     Populate $M(i, \cdot, \cdot)$ by propagating tags across align-
        ments
13:     for $i \leq n$ do
14:         Use majority voting to obtain one tag per word
15:         Obtain type-level tag dictionary from all the data
16:         Train TNT/GAR tagger for $l_i$ using projected an-
            notations in $M(i, \cdot, \cdot)$ and tag dictionary
17:     end for
18: end for
```

DataWe use the 100 translations of (parts of) the Bible available as part of the Edinburgh Multilingual Parallel Bible Corpus (Christodouloupoulos and Steedman, 2014).[2] This dataset includes translations into languages such as Akawaio, Aukan or Cakchiquel. The majority of these languages are non-Indoeuropean, and 39 of them have less than one million speakers. For 54 of

[2] homepages.inf.ed.ac.uk/s0787820/bible/

			UNSUPERVISED						UPPER BOUNDS			
			BASELINES		OUR SYSTEMS				WEAKLY SUP		SUPERVISED	
		OOV	BROWN	2HMM	TNT-k-SRC	TNT-n-1-SRC	GAR-k-SRC	GAR-n-1-SRC	DAS	LI	GAR	TNT
bul	YT	31.8	54.5	71.8	**78.0**	77.7	75.7	75.7	-	-	83.1	96.9
ces	YT	44.3	51.9	66.3	71.7	**73.3**	70.9	71.4	-	-	-	98.7
dan	YT	28.6	58.6	69.6	78.6	**79.0**	73.7	73.3	83.2	83.3	78.8	96.7
deu	YT	36.8	45.3	70.0	**80.5**	80.2	77.6	77.6	82.8	85.8	87.1	98.1
eng	YT	38.0	58.2	62.6	72.4	**73.0**	72.2	72.6	-	87.1	80.8	96.7
eus	NT	64.6	46.0	41.6	**63.4**	62.8	57.3	56.9	-	-	66.9	93.7
fra	YT	26.1	42.0	76.5	76.1	76.6	78.6	**80.2**	-	-	85.5	95.1
ell	YT	63.7	43.0	49.8	51.9	52.3	57.9	**59.0**	82.5	79.2	64.4	-
hin	Y	36.1	59.5	69.2	**70.9**	67.6	70.8	71.5	-	-	-	-
hrv	Y	34.7	52.8	65.6	**67.8**	67.1	67.2	66.7	-	-	-	-
hun	YT	41.2	45.9	57.4	70.0	70.4	71.3	**72.0**	-	-	77.9	95.6
isl	Y	19.7	42.6	65.9	**70.6**	69.0	68.7	68.3	-	-	-	-
ind	YT	29.4	52.6	73.1	76.6	**76.8**	74.9	76.0	-	-	87.1	95.1
ita	YT	24.0	45.1	78.3	76.5	76.9	78.5	**79.2**	86.8	86.5	83.5	95.8
plt	Y	35.0	48.9	44.3	56.4	56.6	62.0	**64.6**	-	-	-	-
mar	Y	33.0	**55.8**	45.8	52.0	52.9	52.8	52.3	-	-	-	-
nor	YT	27.5	56.1	73.0	**77.0**	76.7	75.4	76.0	-	-	84.3	97.7
pes	Y	33.6	57.9	**61.5**	59.3	59.6	59.1	60.8	-	-	-	-
pol	YT	36.4	52.2	68.7	**75.6**	75.1	70.8	74.0	-	-	-	95.7
por	YT	27.9	54.5	74.3	82.9	**83.8**	81.1	82.0	87.9	84.5	87.3	96.8
slv	Y	15.8	42.1	78.1	79.5	**80.5**	68.7	70.1	-	-	-	-
spa	YT	21.9	52.6	47.3	81.1	81.4	**82.6**	**82.6**	84.2	86.4	88.7	96.2
srp	Y	41.7	59.3	47.3	**69.6**	69.2	67.9	67.2	-	-	-	94.7
swe	YT	31.5	58.5	68.4	74.7	**75.2**	71.4	71.9	80.5	86.1	76.1	94.7
tur	YT	41.6	53.7	46.8	60.5	**61.3**	56.5	57.9	-	-	72.2	89.1
average		≤ 50	52.2	64.4	72.1	72.2	70.8	71.5				

Table 1: Results on 25 test languages. Y=entire Bible available. N=only New Testament available. T=manually annotated data available for training (but not used to obtain results for the language itself). Unsupervised baselines are evaluated using optimal 1:1 mappings.

these languages, we have a translation of the entire Bible. For 42, we only have the New Testament, and for the remaining four we only have parts of the New Testament. We note that Bible translations typically have fewer POS-unambiguous words than newswire (Christodouloupoulos and Steedman, 2014). We also note that in rare cases sentences span multiple verses, which means, we sometimes train POS taggers on partial sentences. See Christodouloupoulos and Steedman (2014) for further discussion of the resource. Most of the manually annotated resources were obtained from the CoNLL 2006-2007 releases of various treebanks, the NLTK corpora, the HamleDT resources, and the Universal Dependencies project. We provide a complete overview of the resources at `https://bitbucket.org/lowlands/`

ModelsWe train TNT POS taggers (Brants, 2000) using only token-level projections. We also train semi-supervised POS taggers using the approach in Garrette and Baldridge (2013) (GAR), using both projections and dictionaries, as well as the unlabelled Bible translations.[3] We use the English data as development data. We train TNT and GAR using k or $n-1$ source languages, leading to four taggers in total.

Baselines Our baselines are two standard unsupervised POS induction algorithms: Brown clustering using the implementation by Percy Liang[4] and second-order unsupervised HMMs using logistic regression for emission probabilities (Berg-Kirkpatrick et al., 2010; Li et al., 2012), with and without our Bible tag dictionaries.[5]

Upper bounds The weakly supervised system in Das and Petrov (2011) (DAS) relies on larger volumes of more representative and perfectly tokenized parallel data than we assume, as well as a representative sample of unlabeled data. Such data is simply not available for many of the languages considered here. The weakly supervised system in Li et al. (2012) (LI) also relies on crowd-sourced type-level tag dictionaries, not available for most of the languages of concern to us. We present their reported results. Finally, we train the two base POS taggers (GAR and TNT) on the manually annotated data available for 17 of our languages, to be able to compare against state-of-the-art performance of supervised POS taggers.

[3] `github.com/dhgarrette/low-resource-pos-tagging-2014/`

[4] `github.com/percyliang/brown-cluster`

[5] `code.google.com/p/wikily-supervised-pos-tagger/`

Results Our results on the 25 test languages are consistently better than the unsupervised baselines, with the exceptions of Marathi and Persian, and by a very large margin. Our average performance across the languages with OOV rates smaller than 50% is above 70%. While previous papers on weakly supervised POS tagging (e.g., Das and Li) have presented slightly better results for the small set of Indo-European languages in the CoNLL 2006–7 shared tasks, we emphasize again that our set-up requires fewer resources and does *not* rely on perfectly tokenized training data. Our parallel data also suffers from a severe, but more realistic domain bias. Note that doing the second round of projections (n-1-Src) often improves performance by about a percentage point, but this improvement is not consistent across languages. We observe that most errors are due to our systems predicting too many nouns. Note that for the two languages with underlined OOV rates ($\geq$ 50), performance is very low. This is due to differences in orthography and tokenization. We leave out those results in the averages, but leave them in the results table.

To evaluate on more low-resource languages, we also extracted tag dictionaries from Wiktionary for another 10 languages, from Afrikaans to Swahili. Figure 2 presents the type-level in-vocabulary tag errors of the projected tags in the Bible. This figure is similar to the ones used in Li et al. (2012). We also computed token-level accuracies, where every tag assignment licensed by Wiktionary counts as correct. For three languages, results were 80-90%: Afrikaans, Lithuanian, and Russian. For another three languages, results were 50-70%: Hebrew, Romanian, and Swahili. Results were 35-50% for the remaining four languages: Latin, Maori, Albanian, and Ewe.

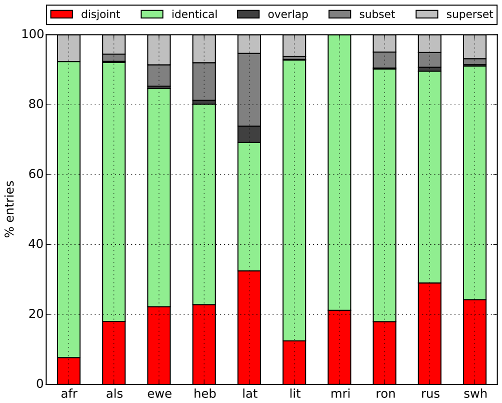


Figure 2: Type-level in-vocabulary tag errors as the percentage of word types assigned a set of tags that is disjoint, identical to, overlaps, is a subset, or is a superset of the Wiktionary tags.

3 Related work

The Bible has been used as a resource for machine translation and multi-lingual information retrieval before, e.g., (Chew et al., 2006). It has also been used in cross-lingual POS tagging (Yarowsky et al., 2001; Fossum and Abney, 2005), NP-chunking (Yarowsky et al., 2001; Yarowsky and Ngai, 2001) and cross-lingual dependency parsing (Sukhareva and Chiarcos, 2014) before. Yarowsky et al. (2001) and Fossum and Abney (2005) use word-aligned parallel translations of the Bible to project the predictions of POS taggers for several language pairs, including English, German, and Spanish to Czech and French. The resulting annotated target language corpora enable them to train POS taggers for these languages. Yarowsky and Ngai (2001) showed similar results using just the Hansards corpus on English to French and Chinese. Our work is inspired by these approaches, yet broader in scope on both the source and target side.

Das and Petrov (2011) use word-aligned text to automatically create type-level tag dictionaries. Earlier work on building tag dictionaries from word-aligned text includes Probst (2003). Their tag dictionaries contain target language trigrams to be able to disambiguate ambiguous target language words. To handle the noise in the automatically obtained dictionaries, they use label propagation on a similarity graph to smooth and expand the label distributions. Our approach is similar to theirs in using projections to obtain type-level tag dictionaries, but we keep the token supervision and type supervision apart and end up with a model more similar to that of Täckström et al. (2013), who combine word-aligned text with crowdsourced type-level tag dictionaries. Täckström et al. (2013) constrain Viterbi search via type-level tag dictionaries, pruning all tags not licensed by the dictionary. For the remaining tags, they use high-confidence word alignments to further prune the Viterbi search. We follow Täckström et al. (2013) in using our automatically created, *not* crowdsourced, tag dictionaries to prune tags during search, but we use word alignments to obtain token-level annotations that we use as annotated training data, similar to Fossum and Abney (2005), Yarowsky et al. (2001), and Yarowsky and Ngai (2001).

Duong et al. (2013) use word-alignment probabilities to select training data for their cross-lingual POS models. They consider a simple single-source training set-up. We also tried ranking projected training data by confidence, using an ensemble of projections from 17–99 source languages and majority voting to obtain probabilities for the token-level target-language projections, but this did not lead to improvements on the English development data.

4 Conclusions

We present a novel approach to learning POS taggers, assuming only that parts of the Bible are available for the target language. Our approach combines annotation projection, bootstrapping, and label propagation to learn POS taggers that perform significantly better than unsupervised baselines, and often competitive to state-of-the-art weakly supervised POS taggers that assume more and better resources are available.

Acknowledgements

This research is funded by the ERC Starting Grant LOWLANDS No. 313695.

References

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Cote, John DeNero, , and Dan Klein. 2010. Painless unsupervised learning with features. In *Proceedings of NAACL*.

Thorsten Brants. 2000. Tnt: a statistical part-of-speech tagger. In *ANLP*.

Peter Chew, Steve Verzi, Travis Bauer, and Jonathan McClain. 2006. Evaluation of the bible as a resource for cross-language information retrieval. In *ACL Workshop on n Multilingual Language Resources and Interoperability*.

Cristos Christodouloupoulos and Mark Steedman. 2014. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*.

Michael Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *EMNLP*.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. 2013. Simpler unsupervised pos tagging with bilingual projections. In *Proceedings of ACL*.

Victoria Fossum and Steven Abney. 2005. Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora. In *IJCNLP*.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *NAACL*.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *EMNLP*.

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

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. CoRR abs/1104.2086.

Katharina Probst. 2003. Using 'smart' bilingual projection to feature-tag a monolingual dictionary. In *CoNLL*.

Maria Sukhareva and Christian Chiarcos. 2014. Diachronic proximity vs. data sparsity in cross-lingual parser projection. In *COLING Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*.

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

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np brackets via robust projection across aligned corpora. In *Proceedings of NAACL*.

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*.