Simpler unsupervised POS tagging with bilingual projections

Long Duong, [12] **Paul Cook,** [1] **Steven Bird,** [1] and **Pavel Pecina**[2]
1 Department of Computing and Information Systems, The University of Melbourne
2 Charles University in Prague, Czech Republic
lduong@student.unimelb.edu.au, paulcook@unimelb.edu.au,
sbird@unimelb.edu.au, pecina@ufal.mff.cuni.cz

Abstract

We present an unsupervised approach to part-of-speech tagging based on projections of tags in a word-aligned bilingual parallel corpus. In contrast to the existing state-of-the-art approach of Das and Petrov, we have developed a substantially simpler method by automatically identifying "good" training sentences from the parallel corpus and applying self-training. In experimental results on eight languages, our method achieves state-of-the-art results.

1 Unsupervised part-of-speech tagging

Currently, part-of-speech (POS) taggers are available for many highly spoken and well-resourced languages such as English, French, German, Italian, and Arabic. For example, Petrov et al. (2012) build supervised POS taggers for 22 languages using the TNT tagger (Brants, 2000), with an average accuracy of 95.2%. However, many widely-spoken languages — including Bengali, Javanese, and Lahnda — have little data manually labelled for POS, limiting supervised approaches to POS tagging for these languages.

However, with the growing quantity of text available online, and in particular, multilingual parallel texts from sources such as multilingual websites, government documents and large archives of human translations of books, news, and so forth, *unannotated parallel data* is becoming more widely available. This parallel data can be exploited to bridge languages, and in particular, transfer information from a highly-resourced language to a lesser-resourced language, to build unsupervised POS taggers.

In this paper, we propose an unsupervised approach to POS tagging in a similar vein to the work of Das and Petrov (2011). In this approach, a parallel corpus for a more-resourced language having a POS tagger, and a lesser-resourced language, is word-aligned. These alignments are exploited to infer an unsupervised tagger for the target language (i.e., a tagger not requiring manually-labelled data in the target language). Our approach is substantially simpler than that of Das and Petrov, the current state-of-the art, yet performs comparably well.

2 Related work

There is a wealth of prior research on building unsupervised POS taggers. Some approaches have exploited similarities between typologically similar languages (e.g., Czech and Russian, or Telugu and Kannada) to estimate the transition probabilities for an HMM tagger for one language based on a corpus for another language (e.g., Hana et al., 2004; Feldman et al., 2006; Reddy and Sharoff, 2011). Other approaches have simultaneously tagged two languages based on alignments in a parallel corpus (e.g., Snyder et al., 2008).

A number of studies have used *tag projection* to copy tag information from a resource-rich to a resource-poor language, based on word alignments in a parallel corpus. After alignment, the resource-rich language is tagged, and tags are projected from the source language to the target language based on the alignment (e.g., Yarowsky and Ngai, 2001; Das and Petrov, 2011). Das and Petrov (2011) achieved the current state-of-the-art for unsupervised tagging by exploiting high confidence alignments to copy tags from the source language to the target language. Graph-based label propagation was used to automatically produce more labelled training data. First, a graph was constructed in which each vertex corresponds to a unique trigram, and edge weights represent the syntactic similarity between vertices. Labels were then propagated by optimizing a convex function to favor the same tags for closely related nodes

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 634–639,
Sofia, Bulgaria, August 4-9 2013.

Model	Coverage	Accuracy
Many-to-1 alignments	88%	68%
1-to-1 alignments	68%	78%
1-to-1 alignments: Top 60k sents	91%	80%

Table 1: Token coverage and accuracy of many-to-one and 1-to-1 alignments, as well as the top 60k sentences based on alignment score for 1-to-1 alignments, using directly-projected labels only.

while keeping a uniform tag distribution for unrelated nodes. A tag dictionary was then extracted from the automatically labelled data, and this was used to constrain a feature-based HMM tagger.

The method we propose here is simpler to that of Das and Petrov in that it does not require convex optimization for label propagation or a feature based HMM, yet it achieves comparable results.

3 Tagset

Our tagger exploits the idea of projecting tag information from a resource-rich to resource-poor language. To facilitate this mapping, we adopt Petrov et al.'s (2012) twelve universal tags: NOUN, VERB, ADJ, ADV, PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), "." (punctuation), and X (all other categories, e.g., foreign words, abbreviations). These twelve basic tags are common across taggers for most languages.

Adopting a universal tagset avoids the need to map between a variety of different, language-specific tagsets. Furthermore, it makes it possible to apply unsupervised tagging methods to languages for which no tagset is available, such as Telugu and Vietnamese.

4 A Simpler Unsupervised POS Tagger

Here we describe our proposed tagger. The key idea is to maximize the amount of information gleaned from the source language, while limiting the amount of noise. We describe the seed model and then explain how it is successively refined through self-training and revision.

4.1 Seed Model

The first step is to construct a seed tagger from directly-projected labels. Given a parallel corpus for a source and target language, Algorithm 1 provides a method for building an unsupervised tagger for the target language. In typical applications, the source language would be a better-resourced language having a tagger, while the target language would be lesser-resourced, lacking a tagger and large amounts of manually POS-labelled data.

Algorithm 1 Build seed model

1: Tag source side.
2: Word align the corpus with Giza++ and remove the many-to-one mappings.
3: Project tags from source to target using the remaining 1-to-1 alignments.
4: Select the top n sentences based on sentence alignment score.
5: Estimate emission and transition probabilities.
6: Build seed tagger T.

We eliminate many-to-one alignments (Step 2). Keeping these would give more POS-tagged tokens for the target side, but also introduce noise. For example, suppose English and French were the source and target language, respectively. In this case alignments such as English *laws (NNS)* to French *les (DT) lois (NNS)* would be expected (Yarowsky and Ngai, 2001). However, in Step 3, where tags are projected from the source to target language, this would incorrectly tag French *les* as NN. We build a French tagger based on English–French data from the Europarl Corpus (Koehn, 2005). We also compare the accuracy and coverage of the tags obtained through direct projection using the French Melt POS tagger (Denis and Sagot, 2009). Table 1 confirms that the one-to-one alignments indeed give higher accuracy but lower coverage than the many-to-one alignments. At this stage of the model we hypothesize that high-confidence tags are important, and hence eliminate the many-to-one alignments.

In Step 4, in an effort to again obtain higher quality target language tags from direct projection, we eliminate all but the top n sentences based on their alignment scores, as provided by the aligner via IBM model 3. We heuristically set this cutoff to 60k to balance the accuracy and size of the seed model.[1] Returning to our preliminary English–French experiments in Table 1, this process gives improvements in both accuracy and coverage.[2]

[1] We considered values in the range 60–90k, but this choice had little impact on the accuracy of the model.

[2] We also considered using all projected labels for the top 60k sentences, not just 1-to-1 alignments, but in preliminary experiments this did not perform as well, possibly due to the previously-observed problems with many-to-one alignments.

The number of parameters for the emission probability is $|V| \times |T|$ where V is the vocabulary and T is the tag set. The transition probability, on the other hand, has only $|T|^3$ parameters for the trigram model we use. Because of this difference in number of parameters, in step 5, we use different strategies to estimate the emission and transition probabilities. The emission probability is estimated from all 60k selected sentences. However, for the transition probability, which has less parameters, we again focus on "better" sentences, by estimating this probability from only those sentences that have (1) token coverage $> 90\%$ (based on direct projection of tags from the source language), and (2) length > 4 tokens. These criteria aim to identify longer, mostly-tagged sentences, which we hypothesize are particularly useful as training data. In the case of our preliminary English–French experiments, roughly 62% of the 60k selected sentences meet these criteria and are used to estimate the transition probability. For unaligned words, we simply assign a random POS and very low probability, which does not substantially affect transition probability estimates.

In Step 6 we build a tagger by feeding the estimated emission and transition probabilities into the TNT tagger (Brants, 2000), an implementation of a trigram HMM tagger.

4.2 Self training and revision

For self training and revision, we use the seed model, along with the large number of target language sentences available that have been partially tagged through direct projection, in order to build a more accurate tagger. Algorithm 2 describes this process of self training and revision, and assumes that the parallel source–target corpus has been word aligned, with many-to-one alignments removed, and that the sentences are sorted by alignment score. In contrast to Algorithm 1, all sentences are used, not just the 60k sentences with the highest alignment scores.

We believe that sentence alignment score might correspond to difficulty to tag. By sorting the sentences by alignment score, sentences which are more difficult to tag are tagged using a more mature model. Following Algorithm 1, we divide sentences into blocks of 60k.

In step 3 the tagged block is revised by comparing the tags from the tagger with those obtained through direct projection. Suppose source language word w_i^s is aligned with target language word w_j^t with probability $p(w_j^t|w_i^s)$, T_i^s is the tag for w_i^s using the tagger available for the source language, and T_j^t is the tag for w_j^t using the tagger learned for the target language. If $p(w_j^t|w_i^s) > S$, where S is a threshold which we heuristically set to 0.7, we replace T_j^t by T_i^s.

Algorithm 2 Self training and revision

1: Divide target language sentences into blocks of n sentences.
2: Tag the first block with the seed tagger.
3: Revise the tagged block.
4: Train a new tagger on the tagged block.
5: Add the previous tagger's lexicon to the new tagger.
6: Use the new tagger to tag the next block.
7: Goto 3 and repeat until all blocks are tagged.

Self-training can suffer from over-fitting, in which errors in the original model are repeated and amplified in the new model (McClosky et al., 2006). To avoid this, we remove the tag of any token that the model is uncertain of, i.e., if $p(w_j^t|w_i^s) < S$ and $T_j^t \neq T_i^s$ then $T_j^t = \text{Null}$. So, on the target side, aligned words have a tag from direct projection or no tag, and unaligned words have a tag assigned by our model.

Step 4 estimates the emission and transition probabilities as in Algorithm 1. In Step 5, emission probabilities for lexical items in the previous model, but missing from the current model, are added to the current model. Later models therefore take advantage of information from earlier models, and have wider coverage.

5 Experimental Results

Using parallel data from Europarl (Koehn, 2005) we apply our method to build taggers for the same eight target languages as Das and Petrov (2011) — Danish, Dutch, German, Greek, Italian, Portuguese, Spanish and Swedish — with English as the source language. Our training data (Europarl) is a subset of the training data of Das and Petrov (who also used the ODS United Nations dataset which we were unable to obtain). The evaluation metric and test data are the same as that used by Das and Petrov. Our results are comparable to theirs, although our system is penalized by having less training data. We tag the source language with the Stanford POS tagger (Toutanova et al., 2003).

	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
Seed model	83.7	81.1	83.6	77.8	78.6	84.9	81.4	78.9	81.3
Self training + revision	**85.6**	**84.0**	**85.4**	80.4	81.4	86.3	83.3	**81.0**	83.4
Das and Petrov (2011)	83.2	79.5	82.8	**82.5**	**86.8**	**87.9**	**84.2**	80.5	**83.4**

Table 2: Token-level POS tagging accuracy for our seed model, self training and revision, and the method of Das and Petrov (2011). The best results on each language, and on average, are shown in bold.

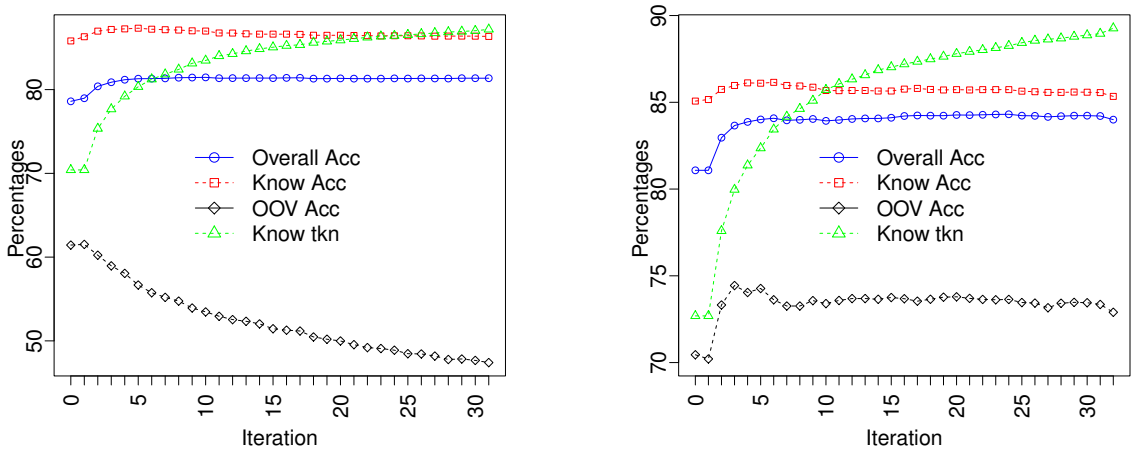

Figure 1: Overall accuracy, accuracy on known tokens, accuracy on unknown tokens, and proportion of known tokens for Italian (left) and Dutch (right).

Table 2 shows results for our seed model, self training and revision, and the results reported by Das and Petrov. Self training and revision improve the accuracy for every language over the seed model, and gives an average improvement of roughly two percentage points. The average accuracy of self training and revision is on par with that reported by Das and Petrov. On individual languages, self training and revision and the method of Das and Petrov are split — each performs better on half of the cases. Interestingly, our method achieves higher accuracies on Germanic languages — the family of our source language, English — while Das and Petrov perform better on Romance languages. This might be because our model relies on alignments, which might be more accurate for more-related languages, whereas Das and Petrov additionally rely on label propagation.

Compared to Das and Petrov, our model performs poorest on Italian, in terms of percentage point difference in accuracy. Figure 1 (left panel) shows accuracy, accuracy on known words, accuracy on unknown words, and proportion of known tokens for each iteration of our model for Italian; iteration 0 is the seed model, and iteration 31 is the final model. Our model performs poorly on unknown words as indicated by the low accuracy on unknown words, and high accuracy on known words compared to the overall accuracy. The poor performance on unknown words is expected because we do not use any language-specific rules to handle this case. Moreover, on average for the final model, approximately 10% of the test data tokens are unknown. One way to improve the performance of our tagger might be to reduce the proportion of unknown words by using a larger training corpus, as Das and Petrov did.

We examine the impact of self-training and revision over training iterations. We find that for all languages, accuracy rises quickly in the first 5–6 iterations, and then subsequently improves only slightly. We exemplify this in Figure 1 (right panel) for Dutch. (Findings are similar for other languages.) Although accuracy does not increase much in later iterations, they may still have some benefit as the vocabulary size continues to grow.

6 Conclusion

We have proposed a method for unsupervised POS tagging that performs on par with the current state-of-the-art (Das and Petrov, 2011), but is substantially less-sophisticated (specifically not requiring convex optimization or a feature-based HMM). The complexity of our algorithm is $O(nlogn)$ compared to $O(n^2)$ for that of Das and Petrov

(2011) where n is the size of training data.[3] We made our code are available for download.[4]

In future work we intend to consider using a larger training corpus to reduce the proportion of unknown tokens and improve accuracy. Given the improvements of our model over that of Das and Petrov on languages from the same family as our source language, and the observation of Snyder et al. (2008) that a better tagger can be learned from a more-closely related language, we also plan to consider strategies for selecting an appropriate source language for a given target language. Using our final model with unsupervised HMM methods might improve the final performance too, i.e. use our final model as the initial state for HMM, then experiment with different inference algorithms such as Expectation Maximization (EM), Variational Bayers (VB) or Gibbs sampling (GS).[5] Gao and Johnson (2008) compare EM, VB and GS for unsupervised English POS tagging. In many cases, GS outperformed other methods, thus we would like to try GS first for our model.

7 Acknowledgements

This work is funded by Erasmus Mundus European Masters Program in Language and Communication Technologies (EM-LCT) and by the Czech Science Foundation (grant no. P103/12/G084). We would like to thank Prokopis Prokopidis for providing us the Greek Treebank and Antonia Marti for the Spanish CoNLL 06 dataset. Finally, we thank Siva Reddy and Spandana Gella for many discussions and suggestions.

References

Thorsten Brants. 2000. TnT: A statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing (ANLP '00)*, pages 224–231. Seattle, Washington, USA.

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 (ACL 2011)*, pages 600–609. Portland, Oregon, USA.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, pages 721–736. Hong Kong, China.

Anna Feldman, Jirka Hana, and Chris Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'06)*, pages 549–554. Genoa, Italy.

Jianfeng Gao and Mark Johnson. 2008. A comparison of bayesian estimators for unsupervised hidden markov model pos taggers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 344–352. Association for Computational Linguistics, Stroudsburg, PA, USA.

Jiri Hana, Anna Feldman, and Chris Brew. 2004. A resource-light approach to Russian morphology: Tagging Russian using Czech resources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP '04)*, pages 222–229. Barcelona, Spain.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 79–86. AAMT, Phuket, Thailand.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL '06)*, pages 152–159. New York, USA.

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

Siva Reddy and Serge Sharoff. 2011. Cross language POS Taggers (and other tools) for Indian

[3] We re-implemented label propagation from Das and Petrov (2011). It took over a day to complete this step on an eight core Intel Xeon 3.16GHz CPU with 32 Gb Ram, but only 15 minutes for our model.

[4] https://code.google.com/p/universal-tagger/

[5] We in fact have tried EM, but it did not help. The overall performance dropped slightly. This might be because self-training with revision already found the local maximal point.

languages: An experiment with Kannada using Telugu resources. In *Proceedings of the IJCNLP 2011 workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies (CLIA 2011)*. Chiang Mai, Thailand.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, pages 1041–1050. Honolulu, Hawaii.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03)*, pages 173–180. Edmonton, Canada.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL '01)*, pages 1–8. Pittsburgh, Pennsylvania, USA.