NLP2 - Assignment 2 Report
Multilingual part-of-speech tagging

Minh Ngo 10897402[1], Peter Dekker 10820973[1]

June 19, 2015

[1] University of Amsterdam
{minh.ngole, peter.dekker}@student.uva.nl

Abstract

A part-of-speech (POS) tagger for a low-resource language has been created by running a POS tagger on a known rich-resource language and transferring the tags to the low-resource language using a parallel corpus. This approach yields promising results, especially when combining taggers, trained on multiple source languages. Several ideas have been introduced to improve accuracy. Results of these experiments are described in this report.

1 Introduction

Part-of-speech (POS) taggers are not available for all languages. When parallel corpora are available for a pair of a low-resource language and a high-resource language with a POS tagger, the knowledge about POS tags can be transferred from one language to the other. To do this, machine translation alignments are used. By applying smoothing techniques and bagging POS tagger results competitive results can be obtained for the target language.

The research of POS tags projection from multiple source languages across aligned corpora has been done by Fossum and Abney (2005) based on the work of Yarowsky and Ngai (2001). They used different taggers to tag different languages and mapped this information later to the target language using alignments provided by the GIZA++ SMT tool and then built HMM (Hidden Markov Model) bigram-based taggers. Results of individual taggers can be combined later to improve an accuracy of the model.

The difference between Fossum and Abney (2005) and the current work is that the word aligner *fast_align* (Dyer et al., 2013) has been used instead of GIZA++, The Stanford POS tagger (Toutanova et al., 2003) has been used instead of Treetag, Brill, TNLT, SVMTool taggers and the final result has been evaluated on the ready annotated data instead of the state-of-art POS taggers of the target language. In addition, several improvements in the HMM model and linear combination of individual POS taggers have been considered and evaluated. Models have been tested on the same subset of input languages (English, Spanish, French, German), but on another parallel corpus (Europal) and for more target languages (Hungarian and Czech) that are supposed to be typologically different.

2 Method

The method proposed by Fossum and Abney (2005) has been implemented with several modifications and improvements. In this project the statistical machine translation tool GIZA++

has been replaced by the alternative fast unsupervised aligner *fast_align* proposed by Dyer et al. (2013).

2.1 Source language POS tagging

The source language sentences were tagged using the Stanford POS tagger (Toutanova et al., 2003). Pre-trained models were used to tag the different source languages: English, French, German and Spanish. The local tagsets of the source languages were mapped to the universal tagset with mappings provided by Petrov et al. (2011).

2.2 Alignment

Word alignment between the two sides of the parallel corpus has been performed using the *fast_align* aligner (Dyer et al., 2013), based on IBM model 2. The algorithm is run for 20 iterations, at this point the perplexity stabilizes in our test setting.

2.3 Handling unknown words

For estimating the probability of an unknown word given the POS tag, probabilities of all words occurring once in the corpus have been counted. All these words have been replaced by the "UNK" token to form the common probability of "UNK" given some POS tag.

2.4 POS tag projection

A target language word receives the POS tag of the source word to which it has been aligned. Now, we want to build a tagger for the target language from our tagged target corpus. An HMM tagger wants to find the best tag sequence t':

$$t' = argmax_{t_1^n} \quad = \quad \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}) \tag{1}$$

We need emission probabilities $P(w_i|t_i)$, the probability of observing a word given a POS tag. The noisy channel model has been used to determine a probability of the word given the POS-tag:

$$P(w_i|t_i) \quad = \quad \frac{P(t_i|w_i)P(w_i)}{P(t_i)} \tag{2}$$

where $\mathbf{P(t_i|w_i)}$ is a probability of the POS tag $\mathbf{t_i}$ given the word $\mathbf{w_i}$ in the i-th position of the sequence, $\mathbf{P(w_i)}$ and $\mathbf{P(t_i)}$ are probabilities of the word $\mathbf{w_i}$ and the tag $\mathbf{t_i}$ respectively. Counts of given POS tags and observed words are kept, and thus these probabilities can be calculated. Two estimates of tag probabilities for 1-to-1 alignments and 1-to-n alignments are calculated separately and combined linearly. This is done to handle many-to-one alignments.

Before calculating $\mathbf{P(w_i|t_i)}$, the smoothing approach proposed by Fossum and Abney (2005) has been performed. In this case for each unique word $\mathbf{w_i}$ in the vocabulary a set of possible POS tags $\{t_j\}$ with probabilities $\mathbf{P(t_j|w_i)}$ has been considered. For each word two most frequent core tags have been chosen. For calculating probabilities of core tags only two most frequent generic tags owned by the core tag have been considered. $\mathbf{P(t_i)}$ and $\mathbf{P(w_i)}$ have been calculated using frequency obtained from the corpus.

Furthermore, we need transition probabilities $P(t_i|t_{i-1})$, the probability of a POS tag given the previous POS tag. This is done by counting every occurring bigram of POS tags in the tagged target corpus. The counts are then smoothed using Witten-Bell smoothing, so unobserved bigrams also receive a probability.

Using the emission and transition probabilities, the best tag sequence is calculated using an approximation of Equation 1, the Viterbi algorithm.

2.5 Combination of POS taggers

To increase the accuracy, POS taggers trained on different source languages, mapping to the same target language, can be combined. The POS taggers are all run on the same evaluation corpus, their different tag decisions for every word then have to be combined. This can be done by *majority tag* or *linear combination*.

In the majority tag approach, the tag which has been chosen as the best tag by most POS taggers is picked. In the linear combination approach, the distribution of tags for every POS tagger is taken into account. The probability of a POS tag is a linear combination of the probabilities of this tag for all POS taggers. Subsequently, from this new distribution, the best tag is chosen. The final probability of the tag **T** given the word **w** can be calculated as following:

$$P_{linear}(T|w) \quad = \quad \sum_{i=1}^{n} k_i(P_i(T|w)) \tag{3}$$

where $\mathbf{k_i}$ is a weight of the i-th POS-tagger decision in the final result.

2.6 Possible improvements

Several potential improvements have been proposed and will be evaluated in the *Experiments* section.

- For the linear combination of POS-taggers, instead of defining the weight of each i-th POS-tagger $\mathbf{k_i}$ as $\frac{1}{n}$ (where n is a number of individual POS taggers), k_i can be defined depending on the accuracy of the tagger on a small annotated subset. More accurate taggers will receive more weight.

- Instead of defining the weight of each i-th POS-tagger for each POS tag equally, it can be determined in a more clever way, by assuming that some POS taggers are "experts" at tagging specific POS tags. For each POS tag and each tagger the weight in the 3 can be determined depending on the accuracy of their results for particular POS tags on the small annotated subset.

- POS taggers can be combined with an optimal weight by training a machine learning classifier

- A bidirectional tagger can be designed by not only taking into account the forward relation of words in the sentence (from the i-th word to (i+1)-th word), but also model the backward relation of words in the sentence(from the (i+1)-th word to i-th word) (Toutanova et al., 2003).

3 Experiments

We trained our approach on the parallel corpus of the European parliament proceedings (Europarl). We used a cleaned version of the corpus (Graën et al., 2014), to be able to parallelize the sentences between all involved languages. As source languages, English, French, German and Spanish were used. As target languages, Czech and Hungarian were used. Czech is from the same language family (Indo-European) as the source languages, but in contrast with the source languages, from the Slavic group. Hungarian even originates from a different language

Language	Accuracy (Ac.) (%)	Ac. majority tag (%)	Ac. linear combination (%)
English	57.02		
German	59.03		
French	60.47		
Spanish	53.95		
en-de		59.31	58.44
en-fr		59.95	61.36
en-es		57.13	59.58
de-fr		61.82	62.27
de-es		59.76	60.05
fr-es		59.22	58.87
en-de-fr		61.85	62.35
en-de-es		60.77	61.09
en-fr-es		61.28	61.67
de-fr-es		61.79	**62.87**
en-de-fr-es		62.73	62.79

Table 1: Accuracy for our Czech POS tagger, trained on 10,000 lines of the Europarl corpus in English/French/German/Spanish and evaluated on a 10,000 annotated lines of the Czech corpus. The bold value shows the best accuracy that has been obtained.

family than the source languages (Uralic). We used these target languages to ensure the general applicability of our approach.

Our HMM tagger is run on an evaluation corpus. The resulting tag sequence is then compared with the annotated tags. The accuracy, the number of correct POS tags divided by the total number of POS tag positions, is calculated. As an evaluation set, the POS tagged corpora from the *Universal dependency treebank* (Agić et al., 2015) were used. For Czech, the first 10,000 lines of the test section were used. For Hungarian, the train section was used.

4 Results

Table 1 and 2 show the result of training our tagger for Czech and Hungarian. It can be seen that the scores of the two languages do not differ much. In most cases the accuracy of combinations of taggers are higher than the highest accuracy of the constituent taggers. The results show that combining more languages tends to boost accuracy, but not every combination of more languages is better than a combination of fewer languages. For Hungarian, a linear combination of two languages, French and German, gave the highest accuracy. For Czech, a combination of three languages (English, French and Spanish) gave the highest accuracy.

From the results, it does not become very clear whether a combination of related or unrelated source languages is better. In both results, for Czech and Hungarian, the combination English-German and English-Spanish perform worst. English and German are from the same Germanic language group, whilst Spanish and English are from different groups (Spanish is Romance) (Fortson, 2004). Furthermore, English contains a large number of Romance words in its lexicon. It is hard to draw a conclusion about the benefits of source language relatedness here. More interesting to see is that the combination German-French performs well for both target languages. German is from the Germanic group and French from the Romance group, so this could point in the direction of a benefit for unrelated source languages.

Language	Accuracy (Ac.) (%)	Ac. majority tag (%)	Ac. linear combination (%)
English	49.71		
French	57.09		
Spanish	56.90		
German	58.57		
En-Fr		60.02	60.13
En-Es		50.13	50.31
En-De		51.45	50.65
Fr-Es		59.68	59.72
Fr-De		61.15	**61.67**
Es-De		59.35	59.41
En-Fr-Es		60.40	60.46
En-Fr-De		61.25	61.41
En-Es-De		60.35	59.75
Fr-Es-De		61.30	61.48
En-Fr-Es-De		61.3	61.00

Table 2: Accuracy for our Hungarian POS tagger, trained on 10,000 lines of the Europarl corpus in English/French/German/Spanish and evaluated on a 1032 annotated lines of the Hungarian corpus. The bold value shows the best accuracy that has been obtained.

4.1 Improvements

The accuracy can be slightly approved by tuning weights of individual POS taggers in the linear combination depending on their accuracy in different decision making tasks. Table 4.1 shows that the proposed approach based on the accuracy of a POS tagger on predicting a certain POS tag has outperformed the approach based on just the accuracy of individual POS taggers in general.

Bidirectional HMM tagging proposed by Toutanova et al. (2003) has significantly boosted the result (2.5% in the combined English-French-Spanish-German model) [Table 3].

Several machine learning classifiers (SVM, Random Forest, Gradient Boost, AdaBoost) have been trained on the results of POS taggers, to combine them in an optimal way. This however gave significantly worse results (22% of accuracy).

Experiment	HMM Forward	HMM Bidirectional	CW1	CW2
En-Fr-Es (Majority)	60.07	60.73	**61.25**	60.07
En-Fr-Es (Linear)	60.50	-	**61.30**	60.50
En-Fr-De (Majority)	61.20	**62.17**	61.86	61.20
En-Fr-De (Linear)	61.71	-	**63.20**	61.71
En-Es-De (Majority)	60.26	**61.19**	60.78	60.26
En-Es-De (Linear)	60.61	-	**61.00**	60.61
Fr-Es-De (Majority)	60.09	60.84	**61.79**	60.80
Fr-Es-De (Linear)	61.92	-	**62.55**	61.92
En-Fr-Es-De (Majority)	61.93	**63.73**	62.75	61.97
En-Fr-Es-De (Linear)	62.01	**64.50**	62.35	62.01

Table 3: Improvements comparison. Models have been evaluated on the Czech corpus. CW1 is the HMM forward bigram model with different weights for POS tags depending on the accuracy of individual POS taggers. CW2 is the HMM forward model with different weights for individual POS taggers depending on their accuracy in general. Bold values show experiments where a particular tagger outperforms other taggers.

Core POS	HMM Forward Tagger (%)				HMM Bidirectional Tagger (%)			
	$en \rightarrow cs$	$de \rightarrow cs$	$fr \rightarrow cs$	$es \rightarrow cs$	$en \rightarrow cs$	$de \rightarrow cs$	$fr \rightarrow cs$	$es \rightarrow cs$
ADJ	18.67	19.36	24.65	20.08	**25.62**	**28.80**	**28.62**	**24.51**
ADP	54.76	71.11	76.80	70.55	**78.26**	**74.61**	**80.97**	**75.44**
ADV	38.75	37.26	35.65	28.40	**48.10**	**42.53**	**41.42**	**33.93**
CONJ	69.35	**67.31**	67.94	71.15	**70.03**	64.02	**70.93**	**74.55**
DET	21.89	0.34	60.29	47.36	**24.76**	**0.46**	**61.60**	**55.35**
NOUN	83.00	90.72	62.90	55.04	**86.92**	**95.68**	**73.61**	**63.06**
PRON	19.10	54.29	43.48	42.10	**20.00**	**30.47**	**66.88**	**45.22**
VERB	**39.70**	27.03	70.92	66.11	39.43	**28.11**	**72.30**	**69.22**

Table 4: POS tag accuracy of different individual POS taggers. Models have been evaluated on the Czech corpus of 10000 lines. Bold values in the table show where one HMM tagger outperforms the other on a certain language pair.

5 Conclusion

In this report, we researched the possibilities of transferring POS tags from one language to the other using machine translation word alignments. The method shows good results, especially when taggers from different source languages are combined. It is promising to see that the method works well for typologically different target languages, this suggests that the method is applicable to a wide range of target languages.

The current accuracy (that has been obtained by training on the relatively small dataset of 10000 lines of parallel corpora) hovers around 60%. With an even higher accuracy, this approach could become a real alternative to human annotation for low-resource languages. With inexpensive POS tagging, other NLP tasks also become available for these languages. Therefore, we tried a number of improvements to boost accuracy. A number of these improvements have yielded increases in accuracy, but there is still room for more improvement.

References

Ž. Agić, M. J. Aranzabe, A. Atutxa, C. Bosco, J. Choi, M.-C. de Marneffe, T. Dozat, R. Farkas, J. Foster, F. Ginter, I. Goenaga, K. Gojenola, Y. Goldberg, J. Hajič, A. T. Johannsen, J. Kanerva, J. Kuokkala, V. Laippala, A. Lenci, K. Lindén, N. Ljubešić, T. Lynn, C. Manning, H. A. Martínez, R. McDonald, A. Missilä, S. Montemagni, J. Nivre, H. Nurmi, P. Osenova, S. Petrov, J. Piitulainen, B. Plank, P. Prokopidis, S. Pyysalo, W. Seeker, M. Seraji, N. Silveira, M. Simi, K. Simov, A. Smith, R. Tsarfaty, V. Vincze, and D. Zeman. Universal dependencies 1.1, 2015. URL http://hdl.handle.net/11234/LRT-1478.

C. Dyer, V. Chahuneau, and N. A. Smith. Simple, fast, and effective reparameterization of ibm model 2. 2013.

B. Fortson. *Indo-European language and culture: an introduction*. Cambridge Univ Press, 2004.

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

J. Graën, D. Batinic, M. Volk, and G. Faaß. *Cleaning the Europarl Corpus for Linguistic Applications*. Universitätsbibliothek Hildesheim, 2014.

S. Petrov, D. Das, and R. McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 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*, pages 173–180. Association for Computational Linguistics, 2003.

D. Yarowsky and G. Ngai. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. 2001.