Increasing the quality and quantity of source language data for unsupervised cross-lingual POS tagging

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

Abstract

Bilingual corpora offer a promising bridge between resource-rich and resource-poor languages, enabling the development of natural language processing systems for the latter. English is often selected as the resource-rich language, but another choice might give better performance. In this paper, we consider the task of unsupervised cross-lingual POS tagging, and construct a model that predicts the best source language for a given target language. In experiments on 9 languages, this model improves on using a single fixed source language. We then show that further improvements can be made by combining information from multiple source languages.

1 Introduction

Supervised part-of-speech (POS) taggers perform very well in cases where substantial manually-annotated data is available, as is the case for languages such as English, Portuguese, German, French and Arabic. For example, Petrov et al. (2012) built supervised POS taggers for 22 European languages using the TNT tagger (Brants, 2000), with an average accuracy of 95.2%. However, creating annotated linguistic resources is expensive and time-consuming. Many widely-spoken languages, such as Vietnamese, Javanese, and Lahnda have little or no manually annotated data, making a supervised approach impossible.

However, parallel texts are becoming increasingly available through sources such as multilingual websites and documents, and large archives of translation memory from books, news, etc. Moreover, the number of languages with parallel data is increasing. The era of English dominating one side of parallel texts is shifting to a far wider range of languages. Parallel data can be exploited to bridge languages, and to transfer annotated information from a highly-resourced *source* language to a lesser-resourced *target* language, to build unsupervised POS taggers (e.g., Das and Petrov, 2011; Duong et al., 2013).

One issue in building such a tagger is choosing the source language. English is commonly used, because parallel data which has English on one side is often most readily available. However, the appropriate source language might depend on the target language. For example, Snyder et al. (2008) found that a better tagger for Slovene could be built by using data from Serbian – a closely related language – than from English. Moreover, if parallel data for a target language with more than one source language is available, it might be possible to exploit this additional information; however, this issue has not been explored to date.

In this paper we build unsupervised POS taggers for 72 language pairs. We identify features based on monolingual and parallel corpora that we use to predict the best source language to build a tagger for a given target language. We show that choosing an appropriate source language can improve the accuracy of a state-of-the-art unsupervised POS tagging methodology, compared to using a single fixed source language. This prediction can be done based on features of the source and target language derived from monolingual corpora – important if parallel data is not available for our target language, and we need to choose which data to collect – although further improvements can be obtained using features based on parallel corpora. We then show that if multiple source languages are available, even better accuracy can be obtained by combining information from them.

2 Related work

One approach to build an unsupervised POS tagger is to *project* tag information from a resource-rich source language to a resource-poor target lan-

International Joint Conference on Natural Language Processing, pages 1243–1249,
Nagoya, Japan, 14-18 October 2013.

guage. Das and Petrov (2011) and Duong et al. (2013) both achieve state-of-the-art performance on eight European languages using this cross-lingual approach. The two approaches are similar in the following respects. First, both project tag information from source to target language, applying some kind of noise reduction along the way: Das and Petrov use high confidence alignments, while Duong et al. use high confidence sentences. Second, both use a semi-supervised method to obtain more labeled data: Das and Petrov use graph based label propagation, while Duong et al. use self-training. Finally, both apply noise reduction/filtering on the (automatically) labeled data: Das and Petrov only extract the tag dictionary from labeled data, while Duong et al. heuristically revise tags after each self-training step. Crucially, in both of these approaches, once a tagger is built from parallel data, it can be used to tag monolingual text. The method of Duong et al. is less computationally intensive than that of Das and Petrov, as the graph-based propagation algorithm used by the latter requires convex optimisation. Because of its relative simplicity, yet comparable accuracy, in this paper we extend the method of Duong et al.

Both Das and Petrov and Duong et al. exploit the Europarl Corpus with English as the source language (Koehn, 2005).[1] However, as recent work has shown, it is worth considering other choices of source language. For example, Snyder et al. (2008) found that the accuracy of a Slovene tagger improved by 7.7% when paired with Serbian, a closely related language, but only 1.3 percentage points when paired with English. Reddy and Sharoff (2011) and Hana et al. (2004) showed that for closely related languages, transition probabilities for an HMM tagger can be used interchangeably. This suggests that the source language might have a drastic effect on tagger performance. In this paper we investigate the problem of making a good choice of source language(s).

3 Parallel data

We would like to conduct experiments on a resource-poor target language, however, it would be much harder to evaluate. We instead experiment with nine languages: English, Danish, Dutch, Portuguese, Swedish, Greek, Italian, Spanish, and German. We use the JRC-Acquis corpus which provides parallel data for every pair of 22 European languages (Steinberger et al., 2006). We thus, extract a subset of 72 language pairs. It's worth nothing that we consider (x–y) and (y–x) to be distinct language pairs. To the best of our knowledge, JRC-Acquis is the biggest corpus providing parallel data for all language pairs we consider. Table 1 shows some statistics about the data.

[1] Das and Petrov also use the ODS United Nations dataset.

Language	No. of Texts	No. of Words ($\times 10^6$)
en	23545	55.5
da	23624	50.9
nl	23564	56.8
pt	23505	59.6
sv	20243	47.0
el	23184	55.9
it	23472	57.2
es	23573	62.1
de	23541	50.9

Table 1: The number of texts and words for each language considered in the JRC-Acquis corpus.

Language	Corpus Size		Voc. Size
	JRC-Acquis	Europarl	
en	-	-	14810
da	1000785	1968800	29867
nl	1132352	1997775	21316
pt	1121460	1960407	19333
sv	1061156	1862234	29403
el	792732	1235976	34992
it	1122016	1909115	19310
es	1117322	1965734	18496
de	1136452	1920209	29860

Table 2: Corpus size (number of tokens) for each language, with English as the source language. The vocabulary size for a 1M word sample from JRC-Acquis for each language is also shown.

4 Features

In this section, we consider factors that influence the choice of source language. We divide the features into two categories: *monolingual features* which exploit only monolingual data, and *bilingual features* which exploit parallel data.

4.1 Monolingual features

Morphological complexity. Morphologically rich languages introduce complexity when aligning parallel data because there is much greater ambiguity in alignment. Given the reliance of our approach on alignments, morphological complexity is an important factor to consider. We can estimate morphological complexity by counting the number of types, i.e. the vocabulary size, in a fixed amount of text. Table 2 shows the vocabulary size for each language, based on a one million word sample from JRC-Acquis (although any monolingual corpus could be used).

Language relatedness. Our nine languages belong to three language families: Germanic (English, Danish, Dutch, Swedish, German); Romance (Portuguese, Italian, Spanish), and Baltic (Greek). Duong et al. (2013) note that their tagger performs better on Germanic languages than that of Das and Petrov (2011), which might be because this is the same family as the source language used (English). Thus, language relatedness is an important factor to consider.

We quantify language relatedness using lexicostatistics on the Swadesh 200 Wordlist (Dyen et al., 1992). Lexicostatistics involves the judgment of a linguist about whether a given pair of words are cognates. The relatedness of two languages is just the percentage of cognates in the wordlist. Dyen et al. provides a table showing this number for all 84 Indo-European languages. We thus, extract a subset of 36 language pairs from this list.[2] Note that this measure is symmetric.

4.2 Bilingual features

Corpus size. The most obvious factor is corpus size. The more data we have, the better. We count the number of parallel sentences in the corpus. Table 2 shows the corpus size for each language pair with English as the source side.

One-to-One alignment proportion. We believe that one-to-one mappings are more meaningful for this task than many-to-one mappings. The intuition is that, if there is only one possible way to copy a tag from the source language to the target language, we can be more confident about the mapping. The proportion of 1–1 mappings is calculated using a fixed number of parallel sentences (800k sentences) for all language pairs.

Sentence alignment score. Sentence alignment scores are provided by the aligner for IBM Model 3. Duong et al. (2013) used these scores to rank sentences in building their tagger, showing this to be effective in choosing high quality sentences. Higher alignment scores might therefore correspond to a more accurate tagger. We use the average sentence alignment score for each language pair as a feature.

Lexical translation entropy. We adopt the idea of translation model entropy from Koehn et al. (2009). However, instead of scanning all possible sentence segmentations and calculating the phrase-based entropy, we use a simpler method based on the lexical translation table. That is, the entropy for each lexical entry is calculated as

$$H(s) = -\sum_{t \in T} p(t|s) \times log_2 p(t|s)$$

where T is the set of possible translations of word s, and t is a translation. For each language, we pick a fixed amount of text (1 million words) and calculate the average entropy for all words.

5 Build taggers

In this section we construct 72 taggers, using parallel data for 72 language pairs, and then evaluate the performance of each pair. We use an open source unsupervised cross-lingual POS tagger (UMPOS) from Duong et al. (2013), a state-of-the-art system. UMPOS employs the consensus 12 Universal Tagset (Petrov et al., 2012),[3] to avoid the problem of transliterating between different tagsets for different languages, and to enable comparison across languages.

The input for UMPOS is a tagger for the source language, $Tagger(s)$, along with parallel data (s–t). The source language s is tagged using $Tagger(s)$, and then the tagged labels are projected to the target language t. Sentences are then ranked, and a seed model tagger T_0 is built on just the high scoring sentences. By applying self-training with revision, a series of new models $T_1, T_2, \ldots, T_m$ is constructed where T_i is the tagger after i iterations. The target language tagger, $Tagger(t)$, is then the last model, T_m.

$Tagger(s)$ is trained from manually annotated data $Data(s)$ which is mainly derived from the CoNLL 2006 and CoNLL 2007 Shared Tasks. Using the matching provided by Petrov et al., we map the individual tagsets to the Universal Tagset. We train a supervised POS tagger $Tagger(s)$ on the annotated data using the TNT tagger (Brants, 2000). Table 3 shows the source and size of annotated data, and the 5 fold cross-validation accuracy of $Tagger(s)$, for each language.

We evaluate each $Tagger(t)$ using $Data(t)$; results are shown in Table 4. The average tagger per-

[2]This estimate of language relatedness is not based on parallel text, and is therefore considered a monolingual feature.

[3]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).

		Target language									
		en	da	nl	pt	sv	el	it	es	de	Average
Source language	en	-	76.17	72.97	79.57	**73.83**	50.38	72.20	75.37	73.95	71.81
	da	55.73	-	53.28	50.53	66.08	34.13	46.03	50.34	53.90	51.25
	nl	**75.70**	**76.31**	-	78.92	70.24	54.22	70.49	76.90	**79.47**	72.78
	pt	72.40	69.49	63.07	-	66.67	61.82	**74.23**	80.50	64.70	69.11
	sv	66.56	75.82	61.20	65.51	-	52.74	58.93	63.88	64.48	63.64
	el	47.67	49.50	49.75	57.11	46.64	-	47.33	62.29	55.16	51.93
	it	74.50	71.60	68.19	**84.50**	67.92	47.33	-	**81.80**	68.28	70.52
	es	68.76	68.83	66.34	80.72	68.83	**62.29**	74.07	-	70.36	70.03
	de	72.24	74.48	**76.54**	70.87	66.56	55.16	56.98	70.84	-	67.96
	Baseline	30.28	23.27	24.28	24.53	26.35	24.00	25.09	21.98	26.50	25.14

Table 4: Percentage accuracy for the tagger for each source–target language pair. The best tagger for each target language is shown in bold.

Language	Source	No. of Words	% accuracy
en	WSJ/PennTB	1289k	96.74
da	DDT/CoNLL06	94k	96.20
nl	Alpino/CoNLL06	203k	96.42
pt	Floresta/CoNLL06	206k	96.38
sv	Talbanken/CoNLL06	191k	93.95
el	GDT/CoNLL07	65k	97.68
it	ISST/CoNLL07	76k	94.48
es	Cast3LB/CoNLL06	89k	95.36
de	Tiger/CoNLL06	712k	97.79

Table 3: Source and size of annotated data for each language. The accuracy of each source language tagger is also shown.

Language	JRC-Acquis	Europarl
da	76.2	85.6
nl	73.0	84.0
pt	79.6	86.3
sv	73.8	81.0
el	50.4	80.0
it	72.2	81.4
es	75.4	83.3
de	74.0	85.4
Average	71.8	83.4

Table 5: Accuracy on JRC-Acquis and Europarl using English as the source language.

formance for each source language is also given. It turns out that choosing Dutch instead of English as the source language gives the best average accuracy. The tagger performance on each target language is much better than the baseline that always picks the most frequent tag for each word.

The Greek tagger performs poorly. From Table 2, Greek is the most morphologically complex language in this set, and has the smallest corpus size, two factors which partially explain why tagger performance for Greek is low whether Greek occupies either the source or target language role.

From Table 4, it seems that taggers perform better if the source and target language are in the same language family. For example, the top four source languages for Danish are Dutch, English, Swedish, and German, and the top two source languages for Portuguese are Italian and Spanish. This confirms the intuition in adding language relatedness features in section 4.

Duong et al. (2013) used English as the source language to build taggers for the same eight other languages. The only difference between these two experiments is that Duong et al. used Europarl (Koehn, 2005) data instead of JRC-Acquis. Table 2 also compares the size of parallel data with English as the source language for JRC-Acquis and Europarl. Given that Europarl is larger, higher performance is expected. Table 5 compares the tagger accuracy for each target language using English as the source language, for the two datasets. As expected, the accuracies are higher for Europarl. However, there is a strong correlation between the results for the two experiments (Pearson's r = 0.7). This suggests that, if we had as much data as Europarl for every language pair (not just English), we would expect all numbers in Table 4 to improve substantially (not only the first row where English is the source language).

6 Source language selection

In this section, using features defined in section 4 and tagger performance in Table 4, we build a model that can predict the performance of the target language tagger given a source language.

6.1 Individual feature correlation

Table 6 shows the Pearson's correlation (r) and coefficient of determination (r^2) of each feature with tagger accuracy.

Surprisingly, the one-to-one alignment proportion is very strongly correlated with tagger performance ($r = 0.745$). Lexical translation entropy

Features	r	r^2
Source vocabulary size	-0.613	0.376
Target vocabulary size	-0.202	0.041
Language relatedness	0.497	0.247
Corpus size	0.620	0.385
One-to-one alignment proportion	0.745	0.556
Sentence alignment score	0.492	0.242
Lexical translation entropy	-0.590	0.348

Table 6: Pearson's r and r^2 for each feature.

has a negative correlation, as expected, because lower entropy leads to a better alignment and therefore better tagger performance. The source language vocabulary size is highly negatively correlated, but that strong relationship is not found for the target language. This suggests that the model is not affected much by the target language, but prefers a morphologically simple source language.

Corpus size also has a high positive correlation, confirming the intuition that more data is better. This strong relationship, together with the negative correlation for morphological complexity, consolidates the explanation above about the poor performance of the tagger for Greek, where the availability of data is very limited, and where Greek has the richest morphology of any language considered.

6.2 Building a predictive model

In this experiment we build a model to predict the performance of a target language tagger given a source language. We fit all features into a multiple linear regression model. The r^2 value improved greatly to 0.74, compared to 0.556 for one-to-one alignment proportion, the best individual feature.

We evaluate our model in a leave-one-out cross validation experiment. To build a predictive model for language t, we remove data in Table 4 associated with t and train the multiple linear regression model $model(t)$ on the remaining data. So, given source language s and (s–t) parallel data, $model(t)$ outputs the predicted performance of the tagger trained on (s–t) parallel data. The correlation of the predicted value with the original value (Table 4) is very high ($r = 0.81$).

We also build another predictive model based solely on monolingual features (morphology complexity and language relatedness). The intuition here is that, if we want to build a tagger for a target language, but only have monolingual data for that language, what parallel data would we want to collect first? This monolingual model also shows a high correlation with the original table ($r = 0.74$). If we only use language relatedness, the correlation is very weak ($r = 0.13$), showing that language relatedness on its own is not effective at predicting the best source language.

The predicted best source language for each target language is the language predicted to produce the highest accuracy tagger. Table 7 shows the source language prediction from models exploiting all features, and only monolingual features. The Fixed model always chooses Dutch (nl) as the source language, because Dutch gives the highest average accuracy (Table 4). The Oracle model always picks the best language, and gives the upperbound for the predictive model as a point of comparison. As expected, the model exploiting all features achieves a higher average accuracy than the monolingual model, which nevertheless still outperforms Fixed (although there is some variation for individual languages). With respect to the Oracle upperbound, and Fixed baseline, the error rate reduction for the monolingual and all features models is 10.9% and 52.3%, respectively, showing the effectiveness of using a predictive model.

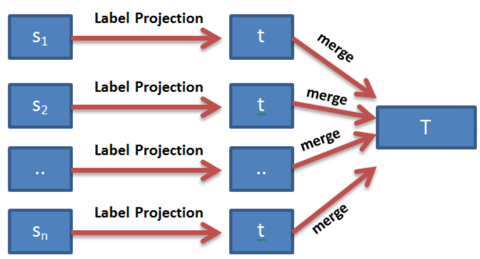

Figure 1: Combining multiple source languages to produce a single file.

7 Multiple Source Languages

In this section, we combine information from multiple source languages to build a single target language tagger. We take a simple approach to doing so, as shown in Figure 1. Each s_i is a tagged corpus for source language i. POS tags are then projected to the target language side t for each corpus. We merge all of these partially-tagged target language corpora (in which unaligned words are untagged) to form T.[4] We build the target lan-

[4]Because the JRC-Acquis corpus consists of translations of documents into multiple languages, in some cases the same target language sentence occurs in the parallel corpus for multiple source languages. In this preliminary approach to combining information from multiple source languages, we simply treat these as different target language sentences. Because the sentences are aligned with different source languages, they might contain different partial tag information.

Target language	All features	Monolingual features	Fixed	Oracle
en	pt (72.40)	**nl (75.70)**	**nl (75.70)**	nl (75.70)
da	sv (75.82)	en (76.17)	**nl (76.31)**	nl (76.31)
nl	**en (72.97)**	**en (72.97)**	-	de (76.54)
pt	**it (84.50)**	es (80.72)	nl (78.92)	it (84.50)
sv	**en (73.83)**	**en (73.83)**	nl (70.24)	en (73.83)
el	**es (62.29)**	en (50.38)	nl (54.22)	es (62.29)
it	**pt (74.23)**	es (74.07)	nl (70.49)	pt (74.23)
es	**pt (80.50)**	**pt (80.50)**	nl (76.90)	it (81.80)
de	en (73.95)	en (73.95)	**nl (79.47)**	nl (79.47)
Average	**74.50**	73.14	72.78	76.07

Table 7: Best source language prediction (and % accuracy of the corresponding tagger) for models exploiting all features, only monolingual features, and a fixed source language, as well as an oracle model that always picks the best language. The best (non-oracle) source language and accuracy for each target language is shown in bold.

Language	1-best	3-best	5-best	7-best
en	75.70	76.66	76.36	**78.16**
da	76.31	78.40	**82.45**	82.43
nl	76.54	76.17	80.00	**81.45**
pt	84.50	84.91	**85.00**	84.24
sv	73.83	74.65	74.10	**76.66**
el	62.29	**70.23**	67.22	67.69
it	74.23	**78.71**	78.47	76.05
es	81.80	82.53	82.13	**82.64**
de	**79.47**	79.28	77.92	77.35
Average	76.07	77.95	78.18	**78.52**

Table 8: Tagger accuracy when combining the 1-, 3-, 5-, and 7-best source languages. The best system for each target language is shown in bold.

guage tagger from T by adapting the method from Section 5. The typical steps for this method are (1) tag the source language, (2) project labels from the source to target language, (3) build the seed model, and (4) apply self-training with revision to produce the final model. Here we simply start from step (3) and build the seed model from T.

In these experiments we assume that when building a tagger for a target language we have access to all other source languages. Table 8 shows accuracy when combining information from the 1-, 3-, 5-, and 7-best source languages, as determined by an oracle. As more source languages are added, average accuracy increases, demonstrating that the method of Duong et al. (2013) can be substantially improved by combining information from multiple source languages. Having established this, in future work we will consider using the best languages as identified by the various feature sets. Moreover, for individual target languages, the best accuracy is not always achieved using the most source languages, suggesting that further work could be done to identify the best set of source languages. There is also a trade-off between accuracy and efficiency; taggers built from more source languages are generally slower.

8 Conclusions

In this paper, we have investigated the problem of choosing the best source language(s) to use in unsupervised cross-lingual POS tagging based on tag projection in parallel corpora. We have shown that our predictive model can select a source language – based on only monolingual features of the source and target languages – that improves tagger accuracy compared to choosing the single, best-overall source language. However, if parallel data is available, our predictive model is able to leverage this to select a more appropriate source language and obtain further improvements in accuracy. Finally, we showed that if multiple source languages are available, even better accuracy can be obtained by combining information from them.

Based on these findings, a synopsis for building a tagger for a resource-poor target language t is as follows: (1) if parallel data for t is unavailable, use monolingual features to predict the best source language s and collect (s–t) parallel data; (2) if there are multiple parallel corpora for t, and there is sufficient time, combine all the corpora to produce a tagger with the best expected accuracy; (3) if time is limited, use all features to identify the n-best source languages.

In future work, we would like to apply the methods described in this paper for identifying "good" source languages for other cross-lingual NLP tasks which exploit parallel data to transfer annotations between languages, including grammar induction, parsing, and morphological analysis. We further intend to expand our experiments to consider more source and target languages.

9 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).

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*, pages 600–609. Portland, Oregon, USA.

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. Sofia, Bulgaria.

Isidore Dyen, Joseph B. Kruskal, and Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5):iii–iv+1–132.

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. Phuket, Thailand.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for Europe. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. Ottawa, Canada.

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

Siva Reddy and Serge Sharoff. 2011. Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. In *Proceedings of the Fifth International Workshop on Cross Lingual Information Access (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 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, pages 1041–1050. Honolulu, Hawaii.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2142–2147. Genoa, Italy.