Multilingual NLP

Homework 3: Cross-lingual POS tagging

Ronald Cardenas Acosta

April 12, 2018

1 Introduction

In this report we present experiments on cross-lingual POS tagging for Shipibo-Konibo, a Peruvian native language that belongs to the Panoan language family. In terms of syntactic profile, languages of this family are (mainly) post-positional and agglutinating languages with highly synthetic verbal morphology, and a basic but quite flexible agent-object-verb (AOV) word order in transitive constructions and subject-verb (SV) order in intransitive ones.

We project POS tags through aligned Spanish-Shipibo parallel data. We find that the tagger trained on gold data alone outperforms the model trained on data extended by the projected corpora. We hypothesize that the reason of this is that Spanish and Shipibo are too dissimilar languages.

2 Corpora

The corpora used was kindly provided by the Artificial Intelligence Research Group at PUCP University in Lima, Peru. The parallel Spanish-Shipibo corpora consists of 29,755 aligned sentences covering domains like the bible (Old testament), legal scripts, educational material, folk tales, and manually gathered common phrases. Table 1 presents details for each domain.

In addition, we use a corpus annotated with POS tags and lemmas, consisting of 1,478 sentences, 13,046 tokens, and 1,593 lemmas. The details per POS tag are shown in Table 2. It is worth noting that the provided corpus presented incompletely and in some cases incorrect annotations. For this reason, the author manually fixed and expanded the tagged corpus to the size now reported.

Table 1: Statistics of the parallel corpora used.

Domain	#sentences	#tokens(SPA)	#tokens(SHK)
Bible (Old Testament)	13,482	210,990	212,539
Common phrases	6,281	14,814	17,060
Educational material	5,979	53,708	49,133
Folk tales	1,302	13,541	10,934
Legal scripts	875	14,756	12,319

Table 2: Statistics of tagger corpus per POS tag.

VERB	PUNCT	NOUN	PRON	AUX	ADJ	ADV	DET
3,064	2,932	2,779	851	671	570	476	462
CONJ	PROPN	INTW	POSTP	NUM	INTJ	SYM	ONOM
316	293	242	194	128	28	21	19

3 Experiments

The steps followed in the experiments were:

- Preprocessing of parallel corpus: Tokenization, truecasing, cleaning and filtering of sentences longer than 80 tokens.
- Tagging of Spanish side of corpus with the pre-trained UDPipe Ancora-UD-2.0 model. Command:

  udpipe –tag –input=horizontal ~/udpipe-ud-2.0-170801/spanish-ancora-ud-2.0-170801.udpipe < data/all.clean.spa > data/all.clean.spa.conllu.tagged
- Formatting of Shipibo side of corpus in CONLL-U format. Command:

  udpipe –tokenize –tokenizer=presegmented –input=horizontal –output=conllu ~/udpipe-ud-2.0-170801/spanish-ancora-ud-2.0-170801.udpipe < data/all.clean.shp > data/all.clean.shp.conllu
- Training of alignment models with FastAlign tool. Training of SPA− >SHK:
  fast_align -i data/all.clean.fa -d -o -v > alignment/spa-shp.clean.fa-ali

  Training of SHK− >SPA:

  fast_align -i data/all.clean.fa -d -o -v -r > alignment/spa-shp.clean.fa-ali.rev

  Obtaining the intersection simmetrization:
  atools -i alignment/spa-shp.clean.fa-ali -j alignment/spa-shp.clean.fa-ali.rev -c intersect > alignment/spa-shp.clean.fa-ali.intersect
- Projection of POS tags SPA− >SHK. Command:

  python project_align.py –ts data/all.clean.spa.conllu.tagged –ut data/all.clean.shp.conllu –a alignment/spa-shp.clean.fa-ali.intersect
- Train SHK tagger. We propose as baseline a tagger trained only on manually annotated data (Gold). This is compared with a model trained on the manually annotated data + the data tagged by projection (Gold+Projected).

In the last step, a UDPipe tagger model was trained following a 10-fold cross-validation strategy. At each step, the gold corpus is divided in training, test and validation. In the case of Gold+Projected, training folds are aggregated with the whole projected data. We train the

Table 3: Accuracy for POS tagging and Lemmatization.

Model	UPOS(%)
GOLD	87.42
GOLD+PROJECTED	63.38

model with 10 different hyper parameter configurations, choose the best performing one in the validation fold, and report its performance on the test fold for the final cross-validation score.

The trained UDPipe model binaries presented along with this report were trained using the last division obtained from the cross-validation procedure described above.

The complete list of commands run for each step of the experiments can be found in the file *README.md*.

4 Discussion and Conclusion

It can be observed from Table 3 that a tagger is further confused when training it with projected tags. We hypothesize that both syntactic and morphological differences between Spanish and Shipibo-Konibo are too great to allow a tagger to benefit from cross-lingual projection.

We plan to experiment with the projection from languages syntactically more similar, such as Basque and Quechua, using bible parallel corpus. Such experiments were not presented in this report due to the fact that the previously provided Bible corpus did not have (book, versicle) annotation, it was only aligned at the sentence level. However, we recently received a Shipibo corpus of the New Testament with this kind of annotation.