TNA ACTIVITY REPORT

"Improving Part of Speech Tagging for Latin-American Spanish Corpora"

Author: Riva Quiroga

Current position: 1. PhD Student in Linguistics / 2. Editor

Affiliation: 1. Pontifical Catholic University of Chile / 2. Programming Historian

Host institution: Charles University

Mentor: Silvie Cinkova

Period of stay: April 13 - July 8 2022

Project Overview

Part of Speech (POS) tagging is a key step when preprocessing a corpus, as the correct assignment of parts of speech to words will impact the results of lemmatization and dependency parsing. This is particularly important in a field such as Computational Literary Studies, where identifying actions, characters, and instances of particular linguistic constructions can help researchers to gain insights about a corpus.

Although POS taggers are currently available for Spanish, most of them have been trained exclusively with texts written in European Spanish, a variety that is used only by around ten percent of all native speakers of this language (Instituto Cervantes, 2020). For example, AnCora (Taulé, Martí, & Recasens, 2008) –the biggest Spanish corpus available at Universal Dependencies, and the one used by tools such as UDPipe an the Python library spaCy– is composed only of newspaper and newswire articles written by Spanish media. As a consequence, POS taggers have lower accuracy with texts written in other varieties of Spanish, and with genres that do not rely exclusively in formal speech. This presents a serious limitation to explore corpora of Latin-American Literature using computational methods. To fill this gap, this project aims to build a corpus that can be used to train models for POS tagging texts written in Latin-American Spanish.

Activities during the research stay

During the research stay at the Institute of Formal and Applied Linguistics at Charles University three phases of the project were conducted:

1. Exploring the corpora used to train the available Spanish language models
2. Building a corpus of Latin-American Literature Dialogues
3. Annotating the corpus

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

Exploring the corpora used to train the available Spanish language models

In order to understand why the current models behave the way they do, an exploration of the corpora used for their training was conducted. The idea was to quantify the presence of grammatical phenomena such as the use of second person singular and *voseo* to determine if their low frequency might be the reason for the low accuracy of these models labeling this phenomena. To that end, the following treebanks were explored:

- AnCora
- GSD
- DEFT Spanish Treebank, which is composed by:
    - International Spanish Newswire Treebank
    - Latin American Spanish Discussion Forum Treebank

These treebanks are the ones used in the main existing tools for tokenization, tagging, lemmatization and dependency parsing texts written in Spanish. In UDPipe it is possible to choose between AnCora and GSD, SpaCy uses AnCora, and CoreNLP combines AnCora with both of the DEFT Spanish Treebanks.

As shown in Table 1, the second person singular was scarcely used in all of the four corpora.

Table 1: Use of second person singular

Treebank	Number of cases	Percentage of use in the corpus
International Spanish Newswire	5	0.10%
GSD	4	0.64%
AnCora	193	0.79%
Discussion Forum	674	9.87%

These results showed the need of building a corpus with a higher presence of this phenomena.

Building de corpus

The second phase of the project was to build a corpus of Latin-American Literature that could be used for training language models in the future. Considering the low presence of second person, the corpus was compiled focusing on dialogues from Latin-American novels and short stories. Fifty ~100-tokens fragments from different Latin-American Spanish varieties were chosen to reach the 5000 token threshold that was suggested.

Annotating the corpus

The corpus was annotated using UDPipe (Straka, 2018) and then manually checked and corrected using INCEpTION (Klie et al., 2018) and ConlluEditor (Heinecke, 2019). To avoid parsing errors, some preprocessing was done to the text:

- The use of dashes was standardized (some texts were using em dashes to introduce a dialogue and others used en dashes).
- An extra line break was added after every sentence to ensure that UDPipe would not join different sentences of a dialogue (it does it sometimes when they start with an en dash and there is only one line break).
- The format of quotation marks was standardized.

UDPipe

The annotation process was performed using the AnCora Spanish Model through the UDPipe API. A bash script was created to automate the whole process. Each annotated text was saved in CoNLL-U format.

INCEpTION

A first round of annotation error fixing was done using INCEpTION. Some of the advantages of this tool are that it allows:

- to perform batch editions to the files using CQL
- to use version control with git
- to calculate inter annotator agreement (in case there are two or more annotators, although it was not the our case).

One of the limitations of INCEpTION for the kind of task we were undertaking, is that it does not allow editing the original text, so fixing problems with word segmentation was not possible. This kind of fix was needed to correct some contractions that were not recognized by UDPipe. To solve this issue, other tools were evaluated, and among them ConlluEditor was chosen.

ConlluEditor

As mentioned before, the editing of the original text was not possible using ICEpTION, so ConlluEditor was used to perform this task. Another advantage of ConlluEditor is that it makes it very easy to edit the dependency trees, as it allows to toggle between a tree graph layout, a flat graph, and a table view. Its main limitation is that it only allows to work with one file at a time, so it is not possible to do batch editing.

Future work

Currently, the corpus built is under a second round of revision to ensure the consistency of the annotation. After this revision is done (hopefully, by the end of September), we will start

the next step, that is, training the model using UDPipe. In parallel, we will work on preparing the annotated corpus for its publication on Universal Dependencies.

We are also working on developing a lesson for Programming Historian that shows other researchers how to use UDPipe for tokenization, tagging, lemmatization and dependency parsing of their corpora, and how to edit their annotations using tools such as ICEpTION and ConlluEditor. The documentation of these tools is not always targeted to people from the Humanities, so creating tutorials that allow researchers from Computational Literary Studies and related fields to use them with ease will be a great way to share the knowledge gained in this research stay.

References

Heinecke, J. (2019). ConlluEditor: A fully graphical editor for Universal Dependencies treebank files. *Universal Dependencies Workshop 2019*. https://syntaxfest.github.io/syntaxfest19/proceedings/papers/paper_55.pdf

Klie, J.-C., Bugert, M., Boullosa, B., Castilho, R. E. de, & Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 197–207. https://doi.org/10.18653/v1/K18-2020