Şeyda Portillo Palma

Exercise 2B: Evaluation of spaCy's dependency parser

Introduction

The process of evaluating the relationships between the phrases in a sentence to ascertain its grammatical structure is known as dependency parsing (DP). Based mainly on this, a sentence is broken into numerous components. The method is predicated on the notion that each linguistic component of a phrase has a direct link with the others. Dependencies are the names given to these relationships.

Material and methods

spaCy's dependency parser is evaluated in English and Spanish pre-trained modules. Since it offers a set of treebanks for multiple languages including English and Spanish, Universal Dependency Treebank 2.x is used. For each language, two sentences are taken from the data corpus in the relevant links:

https://github.com/UniversalDependencies/UD_English-LinES/blob/master/en_lines-ud-dev.conllu

https://github.com/UniversalDependencies/UD_Spanish-AnCora/blob/master/es_ancora-ud-dev.conllu

The available pre-labeled data that correlate to reality is also taken from the aforementioned link.

Having the necessary pack imported, each sentence is parsed into their dependencies by the pre-trained language models. Model performance is assessed in comparison to human-annotated data. Using confusion matrix, the outcome, is compared to the gold annotation or ground truth. If the dataset has more than two classes or if each class has an unequal amount of observations, classification accuracy alone may be deceiving. By calculating a confusion matrix, a better understanding of the categorization model's successes and failures can be acquired.

The following two things are required as input parameters for the confusion matrix:

1. Predicted result set: We have a forecast of tags for the input words once the model has ran on the input. This parameter was essentially created by the tagger using the corpus. This is the prediction set.

2. Actual result set: This group accurately depicts what each word's intended tag should be. This is the ground truth.

Results

Based on the confusion matrix, the pre-trained model gives roughly the same tags with the ground truth. Despite slight deviations, if every row and every column is thought as one of the dependencies and that overall visual is in the diagonal the pre-trained model works close to 100% correct. However, given that the data used is

very limited and consists of only two sentences, it is imperative that this result be presented in a modest manner. With a larger corpus and a proper statistical analysis, the possibility that different results may be obtained is reserved.
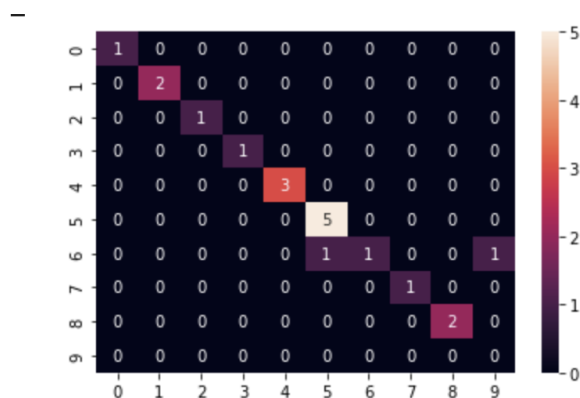
Code

https://github.com/seydaportillo/nlp/blob/3c8007b5fbd7d81b5900e6bb1fa3f674aa68b606/nlp_exercise2.py

—---

Ex: spaCy's dependency parser in Spanish

```
0 Desde NOUN nsubj entonces
1 entonces VERB ROOT entonces
2 entró ADJ acomp entonces
3 en ADP prep entró
4 silencio PROPN compound absoluto
5 absoluto NOUN pobj en
6 . PUNCT punct entonces
```
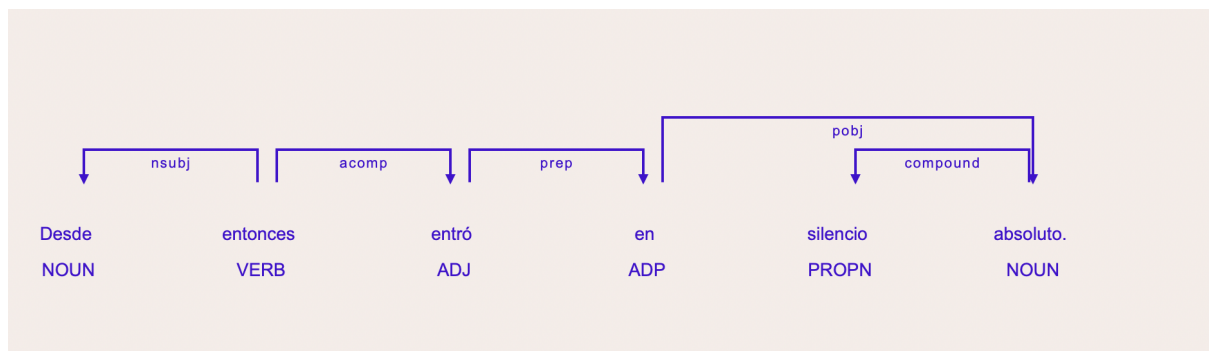

1. Matrices for the samples in English   2.Matrices for the samples in Spanish