Final project
Out-of-Domain PoS tagging

Caio Corro, originally by Guillaume Wisniewski
caio.corro@limsi.fr

2019-2020

Abstract
The goal of this project is to design a PoS tagger. This is a 'research project': the description of the work to do is voluntarily fuzzy and you will have to use your creativity to solve the problem completely while showing us that you have understood the main concepts of statistical learning.

The work has to be done in pairs. The project will be evaluated by a (written) report and an oral presentation.

Please visit the course webpage for information about deadline:
http://teaching.caio-corro.fr/2019-2020/introml/docs/project/

1 Overview

This project aims at developing a system able to predict, the Part-of-Speech of each word of a sentence.

The approach we will consider was originally proposed in [1]

Part-of-Speech tagging is a relatively easy task: as long as there are engouh training data of the same domain, a simple multi-class classifier with very simple features achieve human-comparable performance [2]. However, PoS taggers performance degrades significantly when they are applied to sentences that depart from training data. The goal of this project is twofolds:

- evaluate and characterize the impact of changes in domain;
- develop features that are robust to changes in domain.

2 Data

2.1 Treebanks

Your PoS tagger can be trained (and evaluated) on the corpora of the *Universal Dependencies* project [3] that aims at developing cross-linguistically consistent treebank annotations for a wide array of languages. We will only consider the French treebanks.[1]

There are 6 French treebanks in the UD project. Most of them contain both a train and a test set. We will also consider, in our experiments, two extra treebanks that contain tweets annotated with PoS labels.

All these treebanks can be downloaded from the lecture website.

2.2 Comparing datasets

This first set of experiments aim at measuring the similarity between the different data sets we will consider in our experiments.

Question 1.

Describing shortly the different data sets we will consider : where do the data come from? what is the size (number of words or sentences) of the different of the train and test sets? ...

To characterize the differences between UGC corpora (denoted as 'in-domain' in the following) and 'canonical' corpora (denoted as 'out-domain' in the following) we consider three measures of the noisiness of a corpus:

- The percentage of Out-of-Vocabulary (OOV) words, i.e. words appearing in the test set that are not contained on the train set;

[1] If you prefer, you can use the English treebanks, also avalaible on the lecture website.

- The KL divergence of 3-grams characters distributions estimated on the train and test sets [4]. The divergence is defined as:

$$\mathrm{KL}\left(c_{\text{test}}||c_{\text{train}}\right) = \sum_{n\in\mathcal{N}} p_{\text{test}}(n) \cdot \log \frac{p_{\text{test}}(n)}{p_{\text{train}}(n)}$$

  where the sum runs over $\mathcal{N}$ the set of all the 3-gram of characters in the train and test sets, and $p_d(c_{i-2}, c_{i-1}, c_i) = \frac{\#\{c_{i-2},c_{i-1},c_i\}+1}{\#\mathcal{N}+\#\mathcal{V}\cdot(\#d-2)}$ is the probability to observe the 3-gram $c_{i-2}c_{i-1}c_i$ in data set d with Laplace-smoothing; $\#\mathcal{V}$ is the number of distinct 3-grams of characters in the train and in the test sets and $\#d$ is the number of characters in the corpus d (and, consequently $d-2$ is the total number of 3-grams in the corpus).

- perplexity on the test set of a (word level) Language Model estimated on the test set. The language model can be estimated by KenLM (this tools can also be used to compute the perplexity).

Question 2.

Give an intuitive explanation of each of this metric: what do we try to measure? why? what does it mean when a given metric has a low (or high) value.

compute the value of the different metric for the different combination of train and test sets. What can you conclude?

3 Model

A PoS tagger is simply a mutli-class classifier. You can use the multi-class classifier of your choice, considering the four following set of features:

- the word;
- a window of 5 words around the word of interest (i.e. the word we want to predict a label for, the two previous words and the two following words)
- the features described in [1];

Question 3.

Explain (intuitively) how is each features relevant to predict the PoS of word.

Implement and evaluate the different PoS taggers on the different combination of train and test sets. A PoS tagger can be evaluated by considering:

- its precision over the whole test set (i.e. the percentage of labels that have been correctly predicted)
- its precision over the ambiguous words (i.e. the precision computed only on words that appear with more than one label in the train set;

- its precision over OOV.

What can you conclude? What is the impact of the similarity between the train and test on prediction performance.

4 Work to do

The project has to be done in pairs. You have to:

- answer the various questions on the subject
- implement and evaluate the different classifiers;
- analyze the results obtained.

References

[1] Tobias Schnabel and Hinrich Schütze. Flors: Fast and simple domain adaptation for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 2:15–26, 2014.

[2] Christopher D. Manning. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Proceedings of the Computational Linguistics and Intelligent Text Processing, 12th International Conference*, CICLing 2011, pages 171–189. Springer, 2011.

[3] Joakim Nivre, Željko Agić, Lars Ahrenberg, and other. Universal dependencies 2.3, November 2018. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[4] Héctor Martínez Alonso, Djamé Seddah, and Benoît Sagot. From noisy questions to minecraft texts: Annotation challenges in extreme syntax scenario. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 13–23. The COLING 2016 Organizing Committee, 2016.