RESEARCH PROJECT

ECOLE CENTRALE LILLE

MASTER DATA SCIENCE

POS tagging with simple features

Authors:
Ayoub Youssoufi
Yassin El Hajj Chehade

professor:
Mathieu Dehouck

January 18, 2023

1 Introduction:

Part-of-Speech (POS) tagging is a process of classifying words in a sentence into their corresponding grammatical categories, such as noun, verb, adjective, etc. This is a fundamental task in Natural Language Processing (NLP) and is used in a wide range of applications such as text summarization, information extraction, and machine translation.

In this code, several simple features are used to perform POS tagging on French text using the Perceptron algorithm. The first feature used is the length of the word, the second feature is grammatical rules based on the word's suffix and prefix, and the third feature is the grammatical rules of the word's left and right neighbors. These features were then combined with the use of PCA (Principal Component Analysis) to improve the performance of the POS tagging. The results of the different approaches were compared, and it was found that using PCA on the concatenation of the word's length and grammatical rules had the best performance, followed by the combination between the PCA and the grammatical feature of the left and right neighbors.

2 Ideas & Results :

In the code provided, several approaches were taken to perform POS tagging on French text using the Perceptron algorithm. The first approach used the length of the word as the only feature for tagging. The second approach used grammatical rules based on the word's suffix and prefix as the only feature for tagging. The third approach used a combination of the word's length and grammatical rules as features for tagging. The fourth approach used PCA (Principal Component Analysis) on the concatenation of the word's length and grammatical rules as features for tagging. The fifth approach used a combination of the grammatical rules of the word's left and right neighbors and PCA on the concatenation of the word's length and grammatical rules as features for tagging. The results of the different approaches are summarized in the table bellow:

The method used in POS tagging	Accuracy
Accuracy using only the length feature	2.5%
Accuracy using only the grammatical rules feature	15.7%
Accuracy using the combination between the length and grammatical features	18.0%
Accuracy using PCA on the concatenation	34.0%
Accuracy using the combination between the PCA and the grammatical rules on the neighbor	20.3%
Accuracy using the combination between the PCA and the grammatical feature of the left and right neighbors	23.5%

It can be seen from the results that using only the length feature has the worst performance, while using PCA on the concatenation of the word's length and grammatical rules has the best performance. Using the combination between the PCA and the grammatical feature of the left and right neighbors is also a good approach with a good accuracy.

3 Conclusion

The code provided demonstrates that using a combination of different features and techniques can improve the performance of POS tagging. However, it is worth noting that the results obtained are not very high, and that there is still a lot of room for improvement. Overall, the best performance was achieved by using a combination of techniques such as PCA and grammatical feature of left and right neighbors. However, it is worth noting that these results are not very high and it is not the final solution and more research should be in this field.