UniNE at CLEF 2018: Character-based Convolutional Neural Network for Style Change Detection

Nils Schaetti[1]

[1]Computer Science Institute, University of Neuchâtel

UNIVERSITÉ DE NEUCHÂTEL

Abstract

This paper describes and evaluates a model for style change detection using **character-based Convolutional Neural Networks** (CNN). We applied this model to the style change detection task of the PAN18 challenge and show that its architecture allows this model to be applied to any language. This CNN based on a character-embedding layer, 25 filters and a temporal max-pooling layer reaches a classification accuracy of 62.13%. The evaluation is based on a collections of text gathered from various sites of the StackExchange network, covering different topics (PAN STYLE CHANGE DETECTION task at CLEF 2018 [2]).

Introduction

Today, the use of the work of an author without its authorisation, known as textural plagiarism, is a major problem in fields such as education and research. The field of **automatic plagiarism detection** raises new questions : how to find if a text has been written by one or more author? The increasing access to the Word Wide Web make millions of textual resources easily accessible and providing and enormous amount of sources for potential plagiarism. Therefore, technology and methods to automatically detect plagiarism has received increasing intention in the software in industry and in the academia.

There is two kinds of tasks in plagiarism analysis, **external plagiarism detection** and **intrinsic plagiarism detection**. The first refers to the use of a given reference corpus to identify pairs of very similar passages in a suspicious document. In the second, no reference corpus is given and we must rely on the detection of irregularities, inconsistencies or anomalies within a document. The second is more ambitious since no reference corpus is given, but the first one is the target of most studies.

To face this challenge, the main line of research known as 'stylometry' attempted to quantify the writing style using a variety of measures, representing kind of stylistic information, such as lexical features (word frequencies, word n-grams) or syntactic features (part-of-speech) and some studies have demonstrated the effectiveness of character n-grams.

As this year PAN18 challenge propose a style change detection task, we decided to evaluate a character-based CNN (Deep-Learning) model on this task.

Methodology

Corpus	Document	Changes	No changes
Training	2980	1490	1490
Validation	1492	746	746
Extended training	18913	13007	5906

Table 1: Training, validation and extended training collections

To compare different experimental results on the style change detection task with different models, we need a common ground composed of the same datasets and evaluation measures. In order to create this common ground, and to allow the large study in the domain of intrinsic plagiarism detection, the PAN CLEF evaluation campaign was launched ([2]). Multiple research groups with different backgrounds from around the world have proposed a detection algorithm to be evaluated in the PAN CLEF 2018 campaign with the same methodology.

All teams have used the *TIRA* platform to evaluate their strategy. This platform can be used to automatically deploy and evaluate a software [1]. The algorithms are evaluated on a common test dataset and with the same measures, but also on the base of the time need to produce the response. The access to this test dataset is restricted so that there is no data leakage to the participants during a software run. For the PAN CLEF 2018 evaluation campaign, a collection of texts was created. Based on this collection, the problem to address was to predict if the text is the work of one or more author.

The training and validation data were collected from various site of the StackExchange network. The texts come from the same language. For each text, there is a two-class label we can predict which can take the value *True* (stylistic change(s) in the text) or *False* (no stylistic change(s)). The test sets are also texts collected from the StackExchange network and the task is therefore to predict the *changes* label for each text in the test data.

The training collection is composed of 2'980 text, 1'490 for each class. To allow our classification model to reach higher accuracy, we extended the training collections by switching the parts written by different authors to create new examples. This result in a final training set of 18'913 texts, 13'007 for the class of multi-authored document and 5'906 for the class single authored documents. An overview of these collections is depicted in table 1. The number of documents from each collection is given under the label "Documents" and the total number of document per class in the collection are indicated respectively under the labels "Changes" and "No changes". The training and validation data set are well balanced as for each collection, there is the same number of documents for each class.

Character 2-grams convolutional neural network

In our system, we applied a character based CNN to each text in a collection. A text is fed into the model as an array of character with a fixed size of 12'000. If the document is shorter than 12'000, the additional space is felt with zeros as each character is represented by an index. For each document, we passed it to lower cases and transformed it into a list of character. Each character are transformed to indexes with a vocabulary V constructed during the training phase.

The **first layer is a embedding layer** with a size equal to the vocabulary size $|V|$ and a dimension of 50 for each character. This layer has two purposes, first to reduce the dimensionality of the inputs to 50, compared to $|V|$ for one-hot encoded vectors, and secondly, to encode similarities between character into a multi-dimensional space where two character appearing in similar context are near each others.

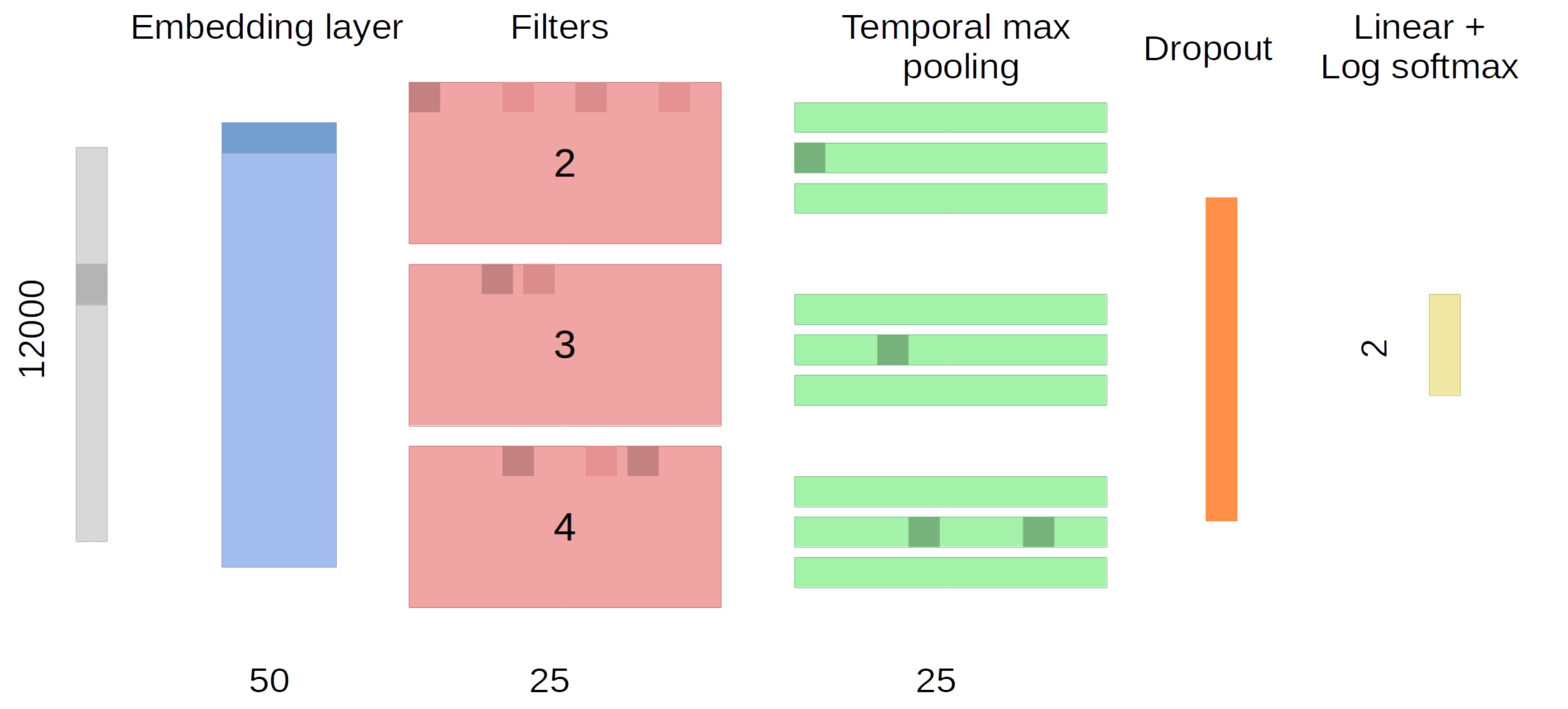

Figure 1: Structure of the character 2-grams based *Convolutional Neural Network* with the following layers : embedding layer (dim=300), three convolutional layers (kernel size 2, 3 and 4), three temporal max pooling layers, a final linear layer of size 2 with log softmax outputs.

The **second layer is composed of three different convolutional layers** with kernel sizes of 2, 3 an 4. Theses layers encode patterns of 2, 3 or 4 consecutive character and each layer has 25 filters and 75 patterns can thus be represented. During the training phase, our model will then find the 75 most effective patterns of character to encode the irregularities.

The **third layer is composed of three max pooling layers** with size 700 and stride 350, one for each preceding convolutional layers. These layers encode the pattern matching for each part of size 700 of the texts. We pass the output through a *ReLU* non-linearity and a **fourth dropout layer**.

The **last layer is a linear one** with a **log-softmax**, with two outputs for each class. We trained our model with the **gradient descent algorithm** with **cross-entropy** as loss function.

Results

To evaluate our model we tested its accuracy on the extended training corpus. The table 2 shows the results of accuracy on the validation set. Our model attained an accuracy of 63.40% compared to 50% for a random classifier.

Corpus	10-Fold CV	**Random**
Validation	0.6340	0.5000
Test	0.6213	0.5000

Table 2: Evaluation for the three collections

The table 3 shows the ranking evaluation for the style change detection task. Our model arrives last with 62.10%, but second in terms of runtime.

	Team	Accuracy	Accuracy solved	Runtime
1	zlatkova18	89.34%	89.34%	01:35:25
2	hosseinia18	82.47%	82.47%	10:12:28
3	ogaltsov18	80.32%	80.32%	00:05:15
4	khan18	64.27%	64.27%	00:01:10
5	schaetti18	62.13%	62.13%	00:03:36

Table 3: Positioning in the PAN18 challenge

References

[1] M. Potthast, T. Gollub, F. Rangel, P. Rosso, E. Stamatatos, and B. Stein. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, and E. Toms, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Berlin Heidelberg New York, Sept. 2014. Springer.

[2] F. Rangel, P. Rosso, M. Montes-y Gómez, M. Potthast, and B. Stein. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working Notes Papers of the CLEF*, 2018.

www.unine.ch

September 10, 2018

nils.schaetti@unine.ch