UniNE at CLEF 2018: Bidirectional Echo State Network-based Reservoir Computing for Cross-domain Authorship Attribution

Nils Schaetti[1]

[1]Computer Science Institute, University of Neuchâtel

UNIVERSITÉ DE NEUCHÂTEL

Abstract

This paper describes and evaluates a model for **cross-domain authorship attribution** using **Bidirectional Echo State Network-based (ESN) Reservoir Computing**. We applied this model to the cross-domain authorship attribution task of the PAN18 challenge and show that it can be applied to this task. This BD-ESN based on a word embedding layer of dimension 300 reaches an averaged F-1 score of 0.408 on the development corpus and 0.3870 on the test corpus. The evaluation is based on a collections of Fanfiction gathered online, covering different original work of art [3].

Introduction

In natural language processing, one might ask : **who is the author of a given text or document?**, based on training corpus and a set of corresponding authors. This task is known as authorship attribution and the motives behind this are multiple. For example, authorship attribution can be applied to forensic linguistic investigation for cases of phishing, spam, threats or cyber-bullying, or for any investigation requiring to identify the true author of a threatening letter or email based on some text written by potential suspects.

A common variety of authorship attribution task is **cross-domain authorship attribution** where given sample of documents coming from a finite and restricted set of candidate authors are used to determine the most likely author of a previously unseen document with unknown authorship. This tasks is made harder when documents of known and unknown authorship are from different genre and thematic area.

The classical line of research on authorship attribution is based on statistical methods and researchers applied neural network methods with good results but with long training time and high complexity. Neural models like Deep-learning are known to be very efficient on image or video classification tasks. However, **Deep-learning** have face more troubles on NLP tasks and **recurrent neural networks** (RNNs) have been applied successfully to tasks like authorship attribution.

However, RNNs are known to be difficult to train and suffer from the problem of vanishing gradient, and use **back-propagation through time** (BPTT) which unfolds a network in time. It's in this context of slow and painful progress that a new approach named **Reservoir Computing** was introduced [1]. The key concept is to separate the part where the computing is done and the output layer where the training is done. The reservoir part is randomly constructed and training only the output layer is often enough to have good performances in practice.

Methodology

Language	Known	Unknown	Authors
English	35	106	5
	140	22	20
French	35	50	5
	140	22	20
Italian	35	81	5
	140	47	20
Spanish	35	104	5
	140	16	20
Polish	35	118	5
	140	65	20

Table 1: The training collection

To compare different experimental results on cross-domain authorship attribution task with different models, we need a common ground composed of the same data sets and evaluation measures. In order to create this common ground, and to allow the large study in the domain of cross-domain authorship attribution, the PAN CLEF evaluation campaign was launched [4].

All teams have used the *TIRA* platform to evaluate their strategy. This platform can be used to automatically deploy and evaluate a software [2]. The algorithms are evaluated on a common evaluation dataset and with the same measures, but also on the base of the time need to produce the response.

To create our algorithm for the PAN CLEF 2018 evaluation campaign, a development corpus was created with highly similar characteristics to the evaluation corpus comprising a set of cross-domain authorship attribution problems for 5 languages, English, French, Italian, Spanish and Polish. The term "training corpus" is not used because the sets of possible authors of the development and evaluation corpora is not overlapping. Based on these collection, the problem to address was to identify the authors of a set of unknown documents given another set of documents (known fanfics) written by a small set (5 to 20) of candidate authors.

The development corpus is composed of 10 problems, 2 per language with various number of known and unknown documents. An overview of this corpus is depicted in table 1. Each author has written at least one of the unknown document and all documents belong to the same fandom. However, known document belong to several fandoms excluding target fandom and is not necessarily the same for all candidate authors.

A corpus with similar characteristics will be used to compare the participants' software of the PAN CLEF 2018 campaign, and we don't have information about its size due to the *TIRA* system. The response of the software is the name of the predicted author for each unknown document belong to each language. The overall performance of the system is the macro-averaged F-1 score.

Bidirectional Echo State Network-base Reservoir Computing

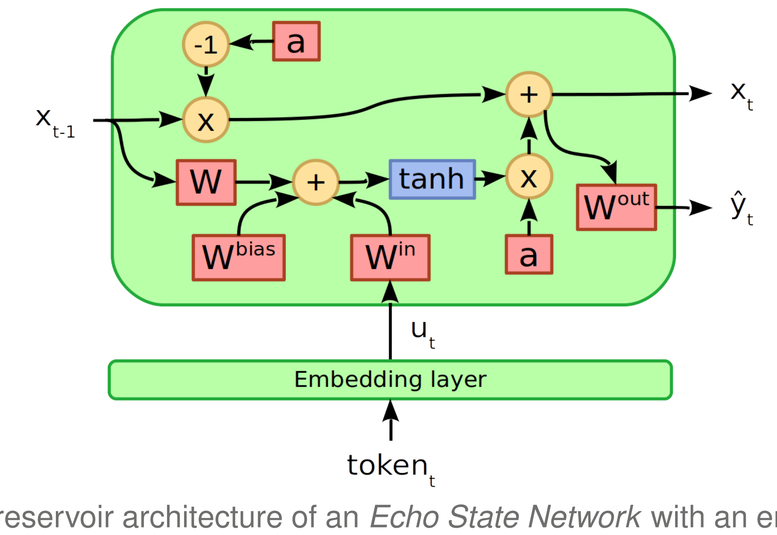

Figure 1: Full reservoir architecture of an *Echo State Network* with an embedding layer.

The main kind of network used in the paper comes directly from equation 1, the highly non-linear dimensional vector at time t, x_t, is defined by

$$x_{t+1} = (1-a)x_t + af(\overset{in}{W} u_{t+1} + Wx_t + \overset{bias}{W}) \quad (1)$$

Figure 1 shows the complete ESN architecture. a is the *leaky rate* which allows to adapt the network's dynamic to the one of the task to learn and $\overset{bias}{W}$ is bias to the reservoir's units. The function f is a nonlinear function, usually the sigmoid function. The network's outputs $\hat{y}$ is then defined by,

$$\hat{y}_t = g(\overset{out}{W} x_t) \quad (2)$$

The learning phase consists to solve a system of linear equations to minimise the error $E(Y, \overset{out}{W} X)$ between the target to be learned and the network's output. The matrix $\overset{out}{W}$ can be computer by linear regression. To find $\overset{out}{W}$, it is possible to use the *Ridge Regression*, which minimise the magnitude of the output weights,

$$\overset{out}{W} = YX^T(XX^T + \lambda I)^{-1} \quad (3)$$

where λ is the regularisation factor which must be fined tuned for each specific task.

To transform input text to input signal, we used word embedding pre-trained with *Glove*. The text is fed into the reservoir word after word which result in an input time series of dimension 300.

We used a specific variety of ESN named Bidirectional-ESN (BDESN). With this model, the inputs are fed into the reservoir in normal and reverse order. The resulting joined states, $x_t^{(lr)} \cup x_{t-T}^{(rl)}$, are used to compute the output $\hat{y}_t$, where $x^{(lr)}$ and $x^{(rl)}$ are respectively the states resulting from inputs in normal (left-to-right) and reverse order (right-to-left).

Results

Our approach was evaluated with macro-averaged F1 on the TIRA platform. The table 2 shows that this approach arrives 9 out of 12 with 0.282, 0.352, 0.378 and 0.539 respectively for 20, 15, 10 and 5 authors.

	Team	20 Authors	15 Authors	10 Authors	5 Authors
1	Custódio and Paraboni	0.648	0.676	0.739	0.677
2	Murauer et al.	0.609	0.642	0.680	0.642
3	Halvani and Graner	0.609	0.605	0.665	0.636
4	Mosavat	0.569	0.575	0.653	0.656
5	Yigal et al.	0.570	0.566	0.649	0.607
6	Martín dCR et al	0.556	0.556	0.660	0.582
7	PAN18-BASELINE	0.546	0.532	0.595	0.663
8	Miller et al.	0.556	0.550	0.671	0.552
9	**Schaetti**	**0.282**	**0.352**	**0.378**	**0.538**
10	Gagala	0.204	0.240	0.285	0.339
11	López-Anguita et al.	0.064	0.065	0.195	0.233
12	Tabealhoje	0.012	0.015	0.030	0.056

Table 2: Positioning in the PAN18 challenge

References

[1] M. Lukoševičius and H. Jaeger.
Reservoir computing approaches to recurrent neural network training.
Computer Science Review, 3(3):127–149, 2009.

[2] 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.

[3] 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.

[4] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein.
Overview of the 4th author profiling task at pan 2016: cross-genre evaluations.
Working Notes Papers of the CLEF, 2016.