Bidirectional Echo State Network-based Reservoir Computing for Cross-domain Authorship Attribution

Notebook for PAN at CLEF 2018

Nils Schaetti

University of Neuchâtel
rue Emile Argand 11
2000 Neuchâtel, Switzerland
nils.schaetti@unine.ch

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.

1 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, has been discovered independently by researchers in the field of machine learning under the name Echo State Network (ESN) and by Neuroscientists as Liquid State Machine (LSM). 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. The training of an ESN is thus not only easier, since it is done only on the output layer, but also because it results in solving a system of linear equation. ESN has been applied to a large field of scientific domains, from astrophysics to robotic motor control and interaction, temporal series forecasting and classification in finance and weather forecasting, to image classification on the MNIST dataset [6, 7] and to gender profiling [5, 8].

As this year PAN18 challenge proposes a cross-domain authorship attribution task, we decided to evaluate an echo state network-based Resevoir Computing model on this task. This paper is organised as follow. Section 2 introduces the dataset used for development and testing, as well as the measures and methodology used to evaluate our approach. Section 3 explains the proposed Echo State Network-based Reservoir Computing model used to classify the texts. In section 4, we evaluate the strategy we created and compare results on the test collections. In the last section, we draw conclusions on the main findings and possible future improvements.

2 Corpus and Methodology

To compare different experimental results on cross-domain authorship attribution 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 cross-domain authorship attribution, the PAN CLEF evaluation campaign was launched [4]. Multiple research groups with different backgrounds from around the world have proposed a classification algorithm to be evaluated in the PAN CLEF 2018 campaign [9, 1] with the same methodology [3].

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

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

on a common evaluation dataset and with the same measures, but also on the base of the time need to produce the response. The access to this evaluation dataset is restricted so that there is no data leakage to the participants during a software run.

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. The number of known and unknown documents is given under the label ”Known” and ”Unknown” and the size of the author set for each problem under the label ”Authors”. 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. Fanfiction refers to a form of litterature produced by admirers (”fans”) of certain authors, novel or TV series, and is also known as transformative literature. The fandom refers the original work of art or genre.

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.

3 Echo State Network-based Reservoir Computing (ESN)

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

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

where $x_t \in R^{N_x}$, with N_x the number of neurons in the reservoir, is its activation vector at time t. The matrix $\overset{in}{W} \in R^{N_x \times N_u}$, with N_u the dimension of the input signal, represents the weights applied to the inputs u_t, and $W \in R^{N_x \times N_x}$ is the matrix of internal weights. Figure 1 shows the complete ESN architecture. We start usually by a null state $x_t = 0$ for the initial vector. 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,

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

where $\overset{out}{W} \in R^{N_y \times N_x}$, with N_y the number of outputs, and $\overset{out}{W}$ the output weights

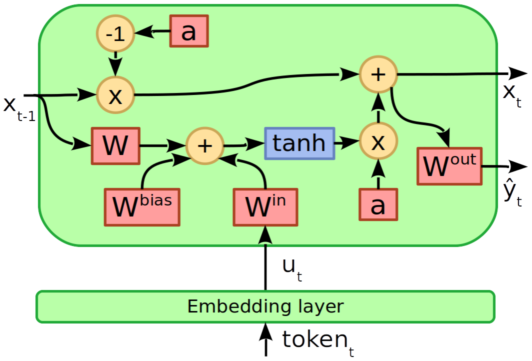

Fig. 1: Full reservoir architecture of an *Echo State Network* with an embedding layer.

matrix. g is a linear function, usually the identity function. 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, $\overset{out}{W} X = Y$, where $X \in R^{N_x \times T}$ is the matrix containing the reservoir states resulting of the training phase, and $Y \in R^{N_y \times T}$ is the matrix containing each target outputs, with T the length of the training set. To find $\overset{out}{W}$, it is possible to use the *Ridge Regression*, which minimise the magnitude of the output weights,

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

where λ is the regularisation factor which must be fined tuned for each specific task. The matrices of internal weights W and input-to-reservoir weights $\overset{in}{W}$ are generated randomly. The leaky rate α is tuned using grid search. Some parameters should be taken into account at the creation of the reservoir weights. More precisely, it is necessary to ensure the presence of the *Echo State Property* which guaranties that inputs will vanish with time and will not be amplified. The most commonly used method to ensure that a reservoir has the *Echo State Property* is to set its spectral radius below 1. The *spectral radius* of a matrix W, noted $\rho(W)$, is its highest absolute Eigen value.

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.

In this paper, 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). To implement our model, we used EchoTorch [1] and TorchLanguage [2], two packages based on pyTorch designed respectively for Reservoir Computing and Natural Language Processing.

[1] https://github.com/nschaetti/EchoTorch

[2] https://github.com/nschaetti/TorchLanguage

Language	Problem	Authors	Macro-averaged F1
English	Problem0001	5	0.153
	Problem0002	20	0.595
			0.374
French	Problem0003	5	0.353
	Problem0004	20	0.501
			0.427
Italian	Problem0005	5	0.286
	Problem0006	20	0.529
			0.408
Polish	Problem0007	5	0.289
	Problem0008	20	0.533
			0.411
Spanish	Problem0009	5	0.327
	Problem0010	20	0.512
			0.420
Overall			0.408

Table 2: Evaluation on the development corpus

4 Evaluation

To evaluate our model we tested their macro-averaged F1 on the development and test corpora. The table 2 shows macro-averaged F1 for each 10 problems in the development set. For English, our model attains respectively an macro-averaged F1 score of 0.153 and 0.595 for problem 1 and 2, and an average of 0.374. For French, the model got 0.353 and 0.501 respectively for problem 3 and 4, and an average of 0.427. For Italian, the model got 0.286 and 0.529 respectively for problem 5 and 6, and an average of 0.408. For Polish, the model got 0.289 and 0.533 respectively for problem 7 and 8, and an average of 0.411. Finally, for Spanish, the model got 0.327 and 0.512 respectively for problem 9 and 10, and an average of 0.420. Our model got an average of 0.408 on the development corpus.

The table 3 shows the results on the test corpus obtained on the *TIRA* platform. Our model got an average of 0.387 on the test corpus, not far from the result obtained on the development corpus.

5 Conclusion

This paper evaluated an Echo State Network-based Reservoir Computing model for *cross-domain authorship attribution* based on fan-fiction gathered on the internet. Based on the hypothesis that textual documents of known authors can be used to identify the author of unknown documents, we introduced an ESN classifier for document classification that can predict this characteristics. However, this model shows low performance compared to results obtained on other datasets with the same model. We think there is two reasons for our model's low performance.

First, even if ESN is known to be less greedy in data, the training set is may be too

Corpus	Macro-averaged F1
Development	0.408
Test	0.3870

Table 3: Evaluation on the two corpora

small for this neural model. Secondly, our model is based on a word embedding layer and therefore on words meaning, this is probably not appropriate for *cross-domain authorship attribution* and we obtained very good results on other datasets with character embedding and we plan therefore to test this solution in the future.

References

1. Kestemont, M., Tschugnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
2. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
3. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, Septembre 11-14, 2017, Proceedings. Springer, Berlin Heidelberg New York (Sep 2017)
4. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. Working Notes Papers of the CLEF (2016)
5. Schaetti, N.: Unine at clef 2017: Tf-idf and deep-learning for author profiling. PAN CLEF 2017 (2017)
6. Schaetti, N., Couturier, R., Salomon, M.: Reservoir computing: Étude théorique et pratique en reconnaissance de chiffres manuscrits, mémoire de master (2015)
7. Schaetti, N., Salomon, M., Couturier, R.: Echo state networks-based reservoir computing for mnist handwritten digits recognition. 19th IEEE International Conference on Computational Science and Engineering (CSE 2016) (2016)
8. Schaetti, N., Savoy, J.: Comparison of Neural Models for Gender Profiling. In: Domenica Fioredistella Iezzi, Livia Celardo, M.M. (ed.) Proceedings of the 14th international conference on statistical analysis of textual data (Jun 2018)
9. Stamatatos, E., Rangel, F., Tschuggnall, M., Kestemont, M., Rosso, P., Stein, B., Potthast, M.: Overview of PAN-2018: Author Identification, Author Profiling, and Author Obfuscation. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Initiative (CLEF 18). Springer, Berlin Heidelberg New York (Sep 2018)