Author Profiling using Word Embeddings with Subword Information

Notebook for PAN at CLEF 2018

Rafael Felipe Sandroni Dias and Ivandré Paraboni

School of Arts, Sciences and Humanities (EACH)
University of São Paulo (USP)
São Paulo, Brazil
{rafaelsandroni,ivandre}@usp.br

Abstract. We present a simple experiment on multilingual author profiling as proposed by the PAN-CLEF 2018 shared task, focusing on the issue of gender identification from Twitter text in English, Spanish and Arabic. Our proposal makes use of word embeddings enriched with char n-gram information, and outperforms a majority class baseline.

1 Introduction

Author profiling (AP) is the computational task of determining author's demographics from the text they produce. Systems of this kind make use of document classification methods to predict a wide range of traits, including author's gender [8], age [2], personality [7,16], religiosity [5], and many others. AP is a popular research topic in NLP, and has been the focus of a number of shared tasks in the PAN-CLEF series [10,17].

At PAN-CLEF 2018 [14], a gender identification task from a combination of text and/or images has been proposed. The languages addressed in the task are English, Arabic and Spanish, all of which in the Twitter domain.

This paper describes our own entry to the AP gender identification task. This consists of a simple experiment involving word embedding models enriched with subword information as proposed in [4] to predict gender from Twitter text, hence disregarding the image information also made available for the task. Preliminary results suggest that the model outperforms a majority class baseline in the three target languages.

The rest of this paper is structured as follows. Section 2 discusses related work on AP. Section 3 describes our main AP approach, and Section 4 describes its evaluation over the PAN-CLEF 2018 AP dataset. Finally, section 5 suggests future work.

2 Related work

2.1 Author profiling

Gender detection is an increasingly popular research topic, with a wide range of computational approaches achieving high accuracy for different languages and domains.

Some of these studies, including the top-performing participants of the previous PAN-CLEF AP gender detection task [10] and other recent initiatives, are briefly discussed as follows.

The work in [2] presents a model called *N-GrAM* for predicting gender in English, Spanish, Portuguese and Arabic. The model makes use of word and character n-grams evaluated using decision tree, MLP, Naive Bayes and SVM classifiers. The SVM-based model was the overall best performing system among participants in the PAN-CLEF 2017 AP gender prediction task.

Also in the context of PAN-CLEF 2017, the work in [8] presents a method for preprocessing Twitter publications, feature extraction, weighted features and a number of classifier models for gender prediction and other tasks. The method obtained the second best result for gender prediction in that shared task.

The work in [15] proposes a document representation model for gender prediction in English using document and term weighting with a combination of POS n-grams and term frequencies (TF-IDF). The model outperforms a BoW baseline on a corpus of hotel reviews.

The growing interest in neural methods for NLP is also evident in the case of gender recognition from text. In [1], a recurrent convolutional neural network model with contextual window (WRCNN) is applied to the task of gender prediction from blogs and from Project Gutenberg's books using extensions of previous RCNN models. Reported results are, on average, 4% higher than those obtained by a baseline system on both domains.

The work in [6] classifies graph vertices using recursive networks to identify gender, age and Twitter user type in the English language. The method combines network, text and label information and converts a graph to a series of tree structure, and then uses individual RNNs on each tree. The proposed approach is shown to outperform four robust baseline systems, namely, logistic regression, label propagation, text-associated DeepWalk and Tri-Party Deep Network Representations.

Finally, the work in [3] compares three neural network architectures to address the problem of predicting gender in Twitter texts: character-level models with convolution layers and bidirectional LSTM, word-level models with bidirectional LSTM using GloVe [12] representation, and document-level models with feed-forward using Bag-of-Words features. The study also explores an ensemble method that combines the three architectures by majority vote. The combination of character-level and word-level information outperforms the individual strategies, whereas the use of document-level information actually reduces overall accuracy.

2.2 Subword models

Popular word vectors representations such as the Skip-gram model [9] use feed-forward networks to predict a word based on the words on its left and right context. This method represents each word of the vocabulary by a unique vector, without shared parameters. In particular, the internal structure of the word is disregarded, which is a major limitation for morphologically rich languages. One possible way of overcoming these difficulties is by adding character-level information as in the case of the subword models proposed in [4].

A subword model is a an extension of the standard Skip-gram model that takes subword information into account. In this model, each word is represented by the sum of their character n-grams. The symbols $<$ and $>$ are added at the beginning and end of each word to distinguish prefixes and suffixes from other sequences, and the word itself is added to the set of its n-grams in order to learn word representations as well. For instance, given the word *author* and n = 3, the corresponding character n-grams will be represented as follows.

$$< au, aut, uth, tho, hor, or >$$

The experiments in [4] make use of all n-grams for n between 3 to 6, although other strategies are possible (e.g., extracting all prefixes and suffixes etc.) Given a n-gram dictionary of size G and a word w, the set of n-grams in w is denoted as $G_w \subset \{1, ..., G\}$. Each n-gram g is associated to a vector representation $\mathbf{z}_g$, and a word is represented by the sum of the vector representations of its n-grams, as illustrated in Equation 1, proposed in [4].

$$s(w, c) = \sum_{g \in G_w} z_g^\top v_c \tag{1}$$

Models of this kind allow representations to be shared across words, which will arguably improve the learning of rare forms. Crucially to our own work, this may be helpful in author profiling tasks such as gender detection, which may often rely on prefix or affix information (and particularly so in the case of morphologically rich languages.) The use of subword models as proposed in [4] makes the core of the author profiling experiment described in the next sections.

3 Method

We developed a simple experiment to assess the use of word embedding models enriched with char n-gram information [4] for the gender identification task from text. The underlying assumption for our experiment is that these subword models may help capture morphological clues (including prefixes, suffixes etc.) that represent gender information in certain languages, and which are otherwise unavailable in standard word embedding models.

To investigate this, we used the Twitter data provided for the PAN-CLEF 2018 Author Profiling task [14] and created a group of documents for each of the three target languages (English and Spanish, with 3000 authors each, and Arabic, with 1500 authors.) The groups were evenly balanced for gender (feminine / masculine). Each author was represented by 100 tweets, which were grouped together as separate documents for each language and author.

The models made use of pre-trained size 300 word vectors for each target language in the Wikipedia domain, and the subword skip-gram extension in [4] with default parameters. In these models, each word is represented by the sum of its character n-gram vectors, where n ranges from 3 to 6, and every tweet is represented as the average

sum of its individual word vectors. Words that do not appear in the vector vocabulary are represented by zero vectors of size 300. No text preprocessing was performed. All models were trained using the scikit-learn liblinear logistic regression implementation [11] with its default parameters and L2 regularization.

4 Results

Table 1 presents mean accuracy results for the PAN-CLEF 2018 author profiling gender prediction task [14] using 10-fold cross validation over the training and (undisclosed) test datasets as provided by TIRA [13]. Given that classes are balanced, we notice that a hypothetical majority-class baseline would obtain 0.50 mean accuracy.

Table 1. 10-fold cross validation mean accuracy results

Language	N	Training data	Test data
English	3000	0.67	0.66
Spanish	3000	0.67	0.67
Arabic	1500	0.68	0.68

After submission, we decided to further investigate whether learning the embedding model directly from the training data (and not from Wikipedia, as in the actual submission) would improve our test results. In this post hoc analysis, we found that mean accuracy scores for the Spanish data remained essentially the same, whereas a small increase was observed in the case of English (from 0.66 to 0.70) and Arabic (from 0.68 to 0.71).

5 Final remarks

This paper presented a simple experiment on multilingual author profiling, focusing on the issue of gender identification based on text only. Our proposal makes use of word embeddings enriched with char n-gram information, and outperforms a majority class baseline. As future work, we intend to evaluate a wide range of alternative models and make use of more robust baseline systems.

Acknowledgements. The second author received financial support from FAPESP grant nro. 2016/14223-0.

References

1. Bartle, A., Zheng, J.: Gender classification with deep learning. Tech. rep., Stanford Technical Report (2015)
2. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-GrAM: New groningen author-profiling model. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum. Dublin (2017)
3. Berg, P., Gopinathan, M.: A deep learning ensemble approach to gender identification of tweet authors. Master's thesis, Norwegian University of Science and Technology (2017)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the ACL 5, 135–146 (2017)
5. Hsieh, F.C., Dias, R.F.S., Paraboni, I.: Author profiling from facebook corpora. In: 11th International Conference on Language Resources and Evaluation (LREC-2018). pp. 2566–2570. ELRA, Miyazaki, Japan (2018)
6. Kim, S.M., Xu, Q., Qu, L., Wan, S., Paris, C.: Demographic inference on Twitter using recursive neural networks. In: Proceedings of ACL-2017. pp. 471–477. Vancouver (2017)
7. Mairesse, F., Walker, M., Mehl, M., Moore, R.: Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research (JAIR) 30, 457–500 (2007)
8. Martinc, M., Skrjanec, I., Zupan, K., Pollak, S.: PAN 2017: Author profiling - gender and language variety prediction. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum. Dublin (2017)
9. Mikolov, T., Wen-tau, S., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proc. of NAACL-HLT-2013. pp. 746–751. Atlanta, USA (2013)
10. Pardo, F.M.R., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum. Dublin (2017)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research 12(Oct), 2825–2830 (2011)
12. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. In: Proceedings of EMNLP-2014. pp. 1532–1543 (2014)
13. 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: Evangelos Kanoulas et. al. (ed.) 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 (2014)
14. Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. 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 (2018)
15. Reddy, T.R., Vardhan, B.V., Reddy, P.V.: N-Gram approach for gender prediction. In: Advance Computing Conference (IACC). pp. 860–865 (2017)
16. dos Santos, V.G., Paraboni, I., Silva, B.B.C.: Big five personality recognition from multiple text genres. In: Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415. pp. 29–37. Springer-Verlag, Prague, Czech Republic (2017)
17. 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: Patrice Bellot et. al. (ed.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Initiative (CLEF 18). Springer, Berlin Heidelberg New York (2018)