Author Verification Using Syntactic N-grams

Notebook for PAN at CLEF 2015

Juan-Pablo Posadas-Durán, Grigori Sidorov, Ildar Batyrshin, and
Elibeth Mirasol-Meléndez

Center for Computing Research (CIC),
Instituto Politécnico Nacional (IPN),
Mexico City, Mexico
http://www.cic.ipn.mx/~sidorov
http://sites.google.com/site/batyr1/

Abstract This paper describes our approach to tackle the Author Verification task at PAN 2015. Our method builds a representation of an author's style by using the information contained in dependency trees. This information is represented as syntactic n-grams and used to conform a vector space. Using unsupervised machine learning approach, each instance is associated to the correponding author using the Jaccard distance. In this paper, we describe the features that were used and the employed unsupervised machine learning algorithm.

1 Introduction

The Author Verification task consists in determine if a given text was written by a person, given a small set (no more than 5, possibly as few as one) of document examples of its authorship. This task has many applications in different fields such as journalism, forensics, security and intellectual property among others. Authorship Verification differs from Authorship Attribution in the fact that in the former the number of examples is more reduced and the information is more limited.

For the task of Author Verification, the competitors were provided with training corpus in English, Spanish, Greek and Dutch. Unlike previous edition PAN 2014, this time the main difference is the variety in genres and topics included in the corpus. To perform the task we mainly used syntactic n-grams obtained from dependency trees as features to model an author's style. The concept of syntactic n-grams is described in the works [11,7,9]. This concept exploits the information about how an author form sentences at syntactic level, so in this manner syntactic n-grams can overcome the topic dependency that traditional n-grams suffer. The syntactic n-grams were used in other related tasks such as automatic English as second language grammar correction [8] and authorship attribution [11,6], but the main contribution of this work is to show that syntactic n-grams can be used to tackle the Author Verification task.

The paper is structured as follows: Section 2 introduces the proposed approach, Section 3 presents the results, Section 4 draws the conclusions and points the future work.

2 Methodology

Our proposal uses an unsupervised machine learning approach to decide if a given text was written by an author or not. First, each unknown and know text is represented as a vector in a space formed by syntactic n-gram, then we use a simple clustering algorithm that associates the unknown texts to the authors by measuring the similarity between unknown and known texts.

As the first step, we perform a standard preprocessing over each dataset before it is parsed. For obtaining syntacitc n-grmas, we use the following syntactic analyzers: Stanford CoreNLP [3] for the English dataset, FreeLing [5,4,1] for the Spanish dataset, and Alpino [1] for the Dutch one. In case of the Greek language we didn't submit results, since we were not able to find a syntactic parser publicly available for this language.

After each dataset is analyzed, we get the syntactic n-grams from the output of the analyzers. Different types of syntactic n-grams were proposed depending on the information used for their construction (lemmas, words, dependency relations, and POS tags). In our case we use the different types proposed in [6]. A simple feature selection of syntactic n-grams based on their frequency is implemented in order to eliminate those syntactic n-grams that rarely appear in the texts and therefore reduce the noise impact in our data representation[2].

The decision on the correspondence between the unknown text and the author is based on the similarity between their vector representations. We measure the similarity using the Jaccard distance defined in equation 1.

$$sim\,(v_i, v_j) = \frac{NNEQ}{NNZ} \tag{1}$$

where $NNEQ$ means the number of non-equal dimensions (the number of dimensions in which the first value is True, second is False and the number of dimensions in which the first value is False, second is True) and NNZ means the number of nonzero dimensions ($NNEQ$ and the number of dimensions in which both values are True).

The output for each unknown text is greater than 0.5 if the similarity measure is above the threshold θ, it is smaller than 0.5 if the similarity measure is below the threshold θ and it is equal to 0.5 if the similarity is zero.

3 Results

The results we obtained in the competition are presented in Table 1. We find our final scores for Spanish and English dataset around the middle of the preliminar results. This indicates that the use of syntactic n-grams for modelling the style of the authors is possible, but probably they should be complemented with other features. The low results that we got for Dutch dataset can be explained due to problems with the parser, because it showed parsing dialogues with errors, so most of the information were discarded.

[1] See `http://www.let.rug.nl/vannoord/alp/Alpino/`

Table 1. Results of our approach at PAN 15 competence

Language	AUC	C1	finalScore
Dutch	0.38165	0.34590	0.13201
Spanish	0.68000	0.68000	0.46240
English	0.68025	0.58800	0.39999

4 Conclusions and Future Work

In this paper, we presented our approach for the Author Verification task at PAN 2015. The main contribution of the approach is the use of syntactic n-grams as features to model an author's style.

We propose as future work the following ideas: (1) add new heuristics to handle bad constructed sentences in tweets instead of ignore them, (2) combine the proposed features with others of distinct nature (semantic features, lexical features, among others), and (3) to use the soft cosine measure [10] in order to take into account the similarity between the pairs of syntactic n-grams so the performance could increase.

Acknowledgments. This work was supported by project Conacyt 240844 and projects SIP-IPN 20151406, 20144274.

References

1. Carrera, J., Castellón, I., Lloberes, M., Padró, L., Tinkova, N.: Dependency grammars in Freeling. Procesamiento del Lenguaje Natural 41, 21–28 (September 2008)
2. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge University Press (2008)
3. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60 (2014), http://www.aclweb.org/anthology/P/P14/P14-5010
4. Padró, L.: Analizadores multilingües en freeling. Linguamatica 3(2), 13–20 (December 2011)
5. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, Istanbul, Turkey (May 2012)
6. Posadas-Duran, J.P., Sidorov, G., Batyrshin, I.: Complete syntactic n-grams as style markers for authorship attribution. In: LNAI, vol. 8856, pp. 9–17. Springer (2014)
7. Sidorov, G.: Non-continuous syntactic n-grams. Polibits 48(1), 67–75 (2013)
8. Sidorov, G.: Syntactic dependency based n-grams in rule based automatic english as second language grammar correction. International Journal of Computational Linguistics and Applications 4(2), 169–188 (2013)
9. Sidorov, G.: Should syntactic n-grams contain names of syntactic relations. International Journal of Computational Linguistics and Applications 5(1), 139–158 (2014)
10. Sidorov, G., Gelbukh, A., Gómez-Adorno, H., Pinto, D.: Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación y Sistemas 18(3), 491–504 (2014)

11. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L.: Syntactic n-grams as machine learning features for natural language processing. Expert Systems with Applications 41(3), 853–860 (2014)