Syntactic N-grams as Features for the Author Profiling Task

Notebook for PAN at CLEF 2015

Juan-Pablo Posadas-Durán, Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, Alexander Gelbukh, and Obdulia Pichardo-Lagunas

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

Abstract This paper describes our approach to tackle the Author Profiling task at PAN 2015. Our method relies on syntactic features, such as syntactic based n-grams of various types in order to predict the age, gender and personality traits that has the author of a given text. In this paper, we describe the used features, the employed classification algorithm, and other general ideas concerning the experiments we conducted.

1 Introduction

The Author Profiling task consists in identifying author's personality features based on a sample of the author writing. This challenging task has a growing importance in several applications related to forensics, security, and terrorism prevention: identifying the author of a suspicious text. Also for marketing purposes, the identification of author's profile proved to be useful for better market segmentation.

This year, in PAN 2015, the task consisted in prediction of the age, gender, and personality traits of authors based on their published tweets. Thus, the participants were provided with tweets in English, Spanish, Italian, and Dutch in order to extract the information concerning author's personality out of them. To perform the task we used syntactic n-grams (the concept is introduced in detail in [12,9,10]) of various types (words, POS tags, syntactic relations, etc.) along with other features such as frequencies of emoticons, hashtags, retweets and others. Syntactic n-grams differ from traditional ones in the way that the neighbors are taken by following syntactic relations in syntactic trees, while in traditional n-grams, the words are taken from the surface strings, as they appear in a text [12,9,10]. The application of syntactic n-grams gives better results than using traditional ones for the task authorship attribution [12,7]. This makes it important to study its impact in the author profiling task.

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

2 Methodology

Presented approach considers the task of Author Profiling as a multilabel classification problem, where an instance is associated with seven labels. The set of labels was defined by the committee of PAN 2015; it consists of features related to the personal traits of an author. Two of these labels, gender and age, were used at PAN 2014 while the rest of the labels (open, agreeable, conscientious, extroverted, and stable) were added in this new edition of the competence and measure some aspects of author's behavior assigning a value that varies from -0.5 to $+0.5$, where the positive extreme means a very strong presence in the author's behavior, while the negative extreme implies the opposite.

Our method uses a supervised machine learning approach, where a classifier is trained independently for each label. In this way, the prediction for an instance is the union of the outputs of each classifier. The vector space model was used to represent the tweets of an author and introduce the use of syntactic n-grams as markers of personality along with the use of traditional SVM classifiers.

Data representation and feature selection details are presented in the following subsections.

2.1 Syntactic N-grams

The main motivation behind our approach is the use of syntactic n-grams as markers to model author's features. There are different types of syntactic n-grams depending on the information used for their construction (lemmas, words, relations, and POS tags); all of them are related through a dependency tree but explore different linguistic aspects of a sentence. We use ten different types proposed in [7], so that the most information from the dependency tree is used.

A syntactic parser is required for our approach, since it allows constructing syntactic n-grams from dependency trees. Different syntactic analyzers where used: Stanford CoreNLP [5] for the English dataset, FreeLing [6,1] for the Spanish dataset, and Alpino[1] for the Dutch one. We do not present the results for the Italian dataset, since we were not able to find a syntactic parser publicly available for this language.

The size of n-grams is another important aspect to be considered. In this proposal, we use the sizes in range between 3 and 5, because several studies related to the use of general n-grams in authorship attribution demonstrated that particularly these sizes correspond to the most representative features [13,2,3].

We perform a standard preprocessing over each dataset before it is parsed by a respective parser. In the preprocessing phase, we also extract several specific characteristics of tweets such as number of retweets, frequency of hashtags, frequency of emoticons, usage of referencing urls and treat them as features.

In the preprocessing phase, the sentences to be parsed are selected depending on their size, so the criteria concerning the limitations on the size of syntactic n-grams are satisfied. We also treat in a specific way the sentences whose size is less than 5, since they provide only a few syntactic n-grams and are generally related to expressions that parsers do not process well.

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

2.2 Data Representation

While using a vector space model approach, an instance is represented as a vector space, in which each dimension corresponds to a specific syntactic n-gram and the value is its frequency. Let's suppose that $\{d_1, \ldots, d_n\}$ are the instances in the training corpus and $\{s_1, \ldots, s_m\}$ are different syntactic n-grams. We build the vectors $v_j = \langle f_{1j}, \ldots, f_{ij} \rangle$, where f_{ij} represents the frequency of the syntactic n-gram s_i inside the instance d_j.

As in the case with traditional n-grams, syntactic n-grams also suffer from noise, since many of them appear only once, and therefore these rare features may not be useful to build author's profiles. In order to reduce the noise in the training dataset, we perform chi-square test as a feature selection strategy, which proved to give good results for the Information Retrieval task [14,8]. The chi-square test measures the importance of a feature for a specific class. Let's suppose that $s = \{s_1, \ldots, s_m\}$ are different syntactic n-grams, and $c = \{c_1, \ldots, c_k\}$ are all possible classes for a specific label. The chi-square with one degree of freedom assigns a score to the syntactic n-gram according to the following equation 1[4]:

$$\chi^2(s_i, c_j) = \frac{(N_{11} + N_{10} + N01 + N_{00}) * (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) * (N_{11} + N_{10}) * (N_{10} + N_{00}) * (N_{01} + N_{00})}, \quad (1)$$

where N_{11} means the number of instances, in which s_i occurs in class c_j; N_{01} means the number of instances, in which s_i does not occur in class c_j; N_{10} means the number of instances, in which s_i occurs out of the class c_j; and N_{00} means the number of instances, in which neither s_i nor c_j occur.

The chi-square test with one degree of freedom transforms the space into a binary class space. Thus, for this task, where the number of classes is greater than two, we take $max\left(\chi^2\right)$ over the different classes and select those whose score is greater than a certain threshold θ.

The final set of features for a specific label is the union of all the selected features via the chi-square test. Based on this, we train the SVM classifier using rbf kernel and typical normalization of vectors. The procedure is repeated for each label, and then each classifier is trained for each label.

3 Results

Our approach greatly depends on the use of syntactic parsers that construct dependency trees. While implementing our proposal, we could only find syntactic parsers for English, Spanish and Dutch. Therefore, the results were obtained only for these three languages (table 1). Our approach showed a relatively good performance for the Dutch language; however, the results for English and Spanish are not that high.

Our global results are not as high as the of the other systems. Analyzing the reasons for this performance, we saw that the main problem is in predicting the age and gender, while for the personal traits (RMSE) the results are comparable with the rest of competitors.

Table 1. Results of our approach at PAN 15 competence

Language	GLOBAL	age	gender	RMSE
English	0.5890	0.5845	0.5915	0.1882
Spanish	0.5874	0.5114	0.6591	0.2116
Dutch	**0.6798**	–	**0.5313**	**0.1716**

4 Conclusion

In this paper, we presented our approach for the Author Profiling task at PAN 2015. The main contribution of the approach is that it shows that syntactic n-grams can be used as features to model author's aspects such as gender, age and personal traits. Considering syntactic n-grams as dimensions in a vector space model and using a supervised machine learning approach, it is possible to tackle the problem of Author Profiling.

The preliminary results show that the use of syntactic n-grams along with other specific tweet features (such as number of retweets, frequency of hashtags, frequency of emoticons, and usage of referencing urls) gives good results when predicting personal traits; however, their usage is not that successful when predicting the age and gender.

As our approach exploits information contained in the dependency trees, its performance is influenced by the use of external syntactic parsers. Although most of the syntactic parsers have recently undergone important improvements, they still have several problems concerning the noise data analysis. The use of the external tools adds noise to the data, and this is one of the reasons why our approach did not show very good results when processing tweets.

In order to improve the approach, we propose the following steps: (1) to add new heuristics to handle grammatical mistakes in tweets instead of ignoring them, (2) to use a weight scheme that will help the approach to handle imbalance training data, (3) to combine the proposed features with other features of distinct nature (semantic features, lexical features, among others), and (4) to use the soft cosine measure [11] in order to consider the similarity between the pairs of syntactic n-grams.

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. Escalante, H.J., Solorio, T., Montes-y Gómez, M.: Local histograms of character n-grams for authorship attribution. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 288–298. Association for Computational Linguistics (2011)
3. Kešelj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the conference pacific association for computational linguistics, PACLING. vol. 3, pp. 255–264 (2003)

4. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
5. 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
6. 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)
7. 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)
8. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
9. Sidorov, G.: Non-continuous syntactic n-grams. Polibits 48(1), 67–75 (2013)
10. Sidorov, G.: Should syntactic n-grams contain names of syntactic relations. International Journal of Computational Linguistics and Applications 5(1), 139–158 (2014)
11. 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)
12. 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)
13. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3), 538–556 (2009)
14. Zheng, Z., Wu, X., Srihari, R.: Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter 6(1), 80–89 (2004)