A Machine Learning-based Intrinsic Method for Cross-topic and Cross-genre Authorship Verification

Notebook for PAN at CLEF 2015

Yunita Sari and Mark Stevenson

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello
Sheffield S1 4DP, United Kingdom
E-mail:{y.sari, mark.stevenson}@sheffield.ac.uk

Abstract This paper presents our approach for the Author Identification task in the PAN CLEF Challenge 2015. We identified the challenges of this year's are the limited amount of training data and the problems in the sub-corpora are independent in terms of topic and genre. We adopted a machine learning based intrinsic method to verify whether a pair of documents have been written by same or different authors. Several content-independent features, such as function words and stylometric features, were used to capture the difference between documents. Evaluation results on the test corpora show our approach works best on the Spanish data set with 0.7238 and 0.67 for the AUC and C@1 scores respectively.

Keywords: Machine Learning, Intrinsic Method, Authorship Verification

1 Cross-topic and Cross-genre Authorship Verification

Given a pair of documents (X,Y), the task of author verification is to identify whether the documents have been written by same or different authors. Compared to authorship attribution, the authorship verification task is significantly more difficult. Verification does not learn about the specific character of each author, but rather about the differences between a pair of documents. The problem is complicated by the fact that an author may consciously or unconsciously vary his/her writing style from text to text [5].

This year's PAN lab Author Identification task focuses on cross-genre and cross-topic authorship verification. In this case, the genre and/or topic may differ significantly between the known and unknown documents. This task is more representative of real world applications where we could not control the genre/topic of the documents.

The PAN Author idenfitication task is defined as follows: *"Given a small set (no more than 5, possibly as few as one) of* ***known*** *documents by a single person and a* ***questioned*** *document, the task is to determine whether the questioned document was written by the same person who wrote the known document set. The genre and/or topic may differ significantly between the known and unknown docu-ments"* [1]

1.1 Data set

The data set consists of author verification problems in four different languages. In each problem, there are some known documents written by single person and only one un-

known document. The genre and/or topic between documents may differ significantly. The document length varies from a few hundred to a few thousand words. Table 1 shows the sub-corpora including their language and type (cross-genre or cross-topic)

Table 1: The authorship verification problems training data set

Language	Type	Total_problems
Dutch	cross-genre	100
English	cross-topic	100
Greek	cross-topic	100
Spanish	cross-genre	100

1.2 Performance measure

The author verification systems are tested on a set of problems. The system must provide a probability score for each unknown document. The performance of the system will be evaluated using area under the ROC curve (AUC). In addition, the output will also be measured based on c@1 score [7]. Probability score which is greater than 0.5 is considered as positive answer, while a score lower than 0.5 is considered as negative. If the score is 0.5, then it will be considered as an *i don't know answer*. The c@1 measure can be define as follows:

$$c@1 = \left(\frac{1}{n}\right) * \left(n_c + \left(n_u * \frac{n_c}{n}\right)\right) \quad (1)$$

where:

- n = number of problems
- n_c = number of correct answer
- n_u = number of unanswered problems

The overall performance will be evaluated on the product of AUC and c@1

2 Methodological Approach

We adopted a machine learning-based intrinsic method to address this verification problem. Intrinsic methods use only the provided documents (in this case known and unknown documents) to determine whether they are written by same author or not. A machine learning algorithm then will be trained on the labeled document pairs to construct a model which can be used to classify the unlabeled pairs. Note that in the verification problems, the machine learning does not learn about the specific character of each author, but rather about the differences between a pair of documents [6]. Texts are represented by various types of features such as function word, character n-grams, word n-grams and several stylometric features.

2.1 Textual Representation

As the genre and/or topic may differ significantly between the known and unknown documents, we can not rely on the content based features to capture the differences between documents. We therefore focused more on content-independent features such as function words and stylometric features. In addition, those features can be applied to any of the language used in the task. We used six types of features in total including: stylometric features (10), function words, character 8-grams, character 3-grams, word bigrams, and word unigrams.

Given collection of problems $P = \{P_i : \forall_i \in I\}$ where $I = \{1, 2, 3, .., n\}$ is the index of P. P_i contains exactly one unknown document U and a set of known documents $K = \{K_j : \forall_j \in J\}$ where J is the index of K and $1 \leq J \leq 5$. Our approach represented each problem P_i as vector $P_i = \{R_1, R_2, .., R_n\}$ where n is the maximum number of feature types (in our case are six). R_i is the distance of two similar feature vector representation of a set of known documents K and unknown document U. If K contains more than one document, then the generated feature vector is an average vector of J documents. Table 2 shows details of the features vector representation and comparison measures used.

Table 2: List of features and comparison measures

Feature	**Model**	**Comparison method**
(R1) Stylometric features	average feature's presence	min-max similarity
(R2) Function words	ratio function word to total number of words in the document	Manhattan distance
(R3) Character 8-grams	tf-idf	cosine similarity
(R4) Character 3-grams	tf-idf	cosine similarity
(R5) Word bigrams	tf-idf	cosine similarity
(R6) Word unigrams	tf-idf	cosine similarity

Stylometric Features There are ten sytlometric features used in our experiment. Some features were adapted from Guthrie's work [3] on anomalous text detection and were among the most effective features to separate anomalous segments from normal segments of the text. The complete list of stylometric features are:

1. Average number of non standard word[1]
2. Average number of words per sentence
3. Percentage of short sentences (less than 8 words)
4. Percentage of long sentences (greater than 15 words)
5. Percentage word with three syllables
6. Lexical diversity (ratio of total number of unique words to total number of words in a document)

[1] Enchant spell checking library (http://www.abisource.com/projects/enchant/) was used to identify non-standard English words

7. Total number of punctuations

We also implemented three readibility measures:

1. Flesch-Kincaid Reading Ease [4]

$$ReadingEase = 206.835 - 1.015\Big(\frac{total\_words}{total\_sentences}\Big) - 84.6\Big(\frac{total\_syllables}{total\_words}\Big) \quad (2)$$

2. Flesch-Kincaid Grade Level [4]

$$GradeLevel = 11.8\Big(\frac{total\_syllables}{total\_words}\Big) + 0.39\Big(\frac{total\_words}{total\_sentences}\Big) - 15.59 \quad (3)$$

3. Gunning-Fog Index [2]

$$FogIndex = \Big(\Big(\frac{total\_words}{total\_sentences}\Big) + \Big(\frac{words\_with\_3\_or\_more\_syllables}{total\_words}\Big) \times 100\Big) \quad (4)$$

2.2 Distance Measures

We experimented with several different comparison measures for computing similarity between a pair of vectors. We noticed particular comparison metric performs better in certain type of features, thus we applied different measure for each features type. Three different distance measures were used:

Cosine similarity measure

$$d(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}.\boldsymbol{y}}{||\boldsymbol{x}||||\boldsymbol{y}||} = \frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} {x_i}^2}\sqrt{\sum_{i=1}^{p} {y_i}^2}} \quad (5)$$

Minimum maximum similarity measure

$$minmax(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{p} min(x_i, y_i)}{\sum_{i=1}^{p} max(x_i, y_i)} \quad (6)$$

City block distance (also called Manhattan distance or L_1 distance)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} |x_i - y_i| \quad (7)$$

2.3 Feature selection and classifier

Our authorship identification software was written in Python. We applied feature selection using Extratreeclassifier and the SVM classifier. The classifier hyperparameters were optimized using the GridSearchCV. Scikit-learn library[2] was used for both feature selection and classification.

3 Evaluation and Result

3.1 Training corpora

We evaluated the approach on the training data using 10-fold cross validation. Since there are some incompatibility issues, we did not perform the verification task on Greek data. Table 3 shows the result of our approach on three of the sub-language corpora. The best result was achieved on the Spanish data set with 0.846 and 0.807 for the AUC and C@1 scores respectively. Compared to other sub-language corpora, the Spanish data set contains more known documents; which may explain why the results on this data set are better than the results on the other sub-languages data. We also observed that the use of some NLP libraries which are mainly trained on English data did not perform well on the non-English language data sets. Thus, feature selection was applied to remove unhelpful features.

Table 3: 10-fold cross validation on the training corpora

Data set	AUC	C@1	finalScore
English	0.662	0.606	0.401
Dutch	0.618	0.553	0.342
Spanish	0.846	0.807	0.683

3.2 Testing corpora

Table 4 shows the official result of our approach on test data released by PAN 15 organizer. As predicted, our approach performed well on the data set which has more known documents. The best results were achieved on the Spanish data with final score of 0.48495. Our approach applied supervised learning where the performance depends strongly on the amount of training data. Thus, as can be seen in Table 4, our verification task did not obtain good results on English data which has only one known document per problem. However, in term of runtime, our approach generally more efficient since all necessary processing were performed in the training phase.

[2] http://scikit-learn.org

Table 4: Result on test data set

Data set	AUC	C@1	finalScore	Runtime
English	0.4011	0.5	0.20055	00:05:46
Dutch	0.61306	0.62075	0.38056	00:02:03
Spanish	0.7238	0.67	0.48495	00:03:47

4 Conclusion

This year's author verification problem is considerably harder than last year's since the number of known documents is very limited and the genre/topic between known and unknown documents differ significantly. In addition, for English, the data set was derived from Project Gutenberg's opera play scripts which are an unusual type of text.

We identified that the most challenging part of this task was to find suitable features which could capture the differences between documents. In addition, for certain data set, not all features were helpful. Thus applying feature selection were beneficial and greatly improved the accuracy of the classifier.

References

1. PAN Authorship Identification Task 2015, https://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/author-identification.html
2. Gunning, R.: The Technique of Clear Writing. McGraw-Hill (1952)
3. Guthrie, D.: Unsupervised Detection of Anomalous Text. Ph.D. thesis (2008)
4. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Tech. Rep. February (1975)
5. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. Twenty-first international conference on Machine learning - ICML '04 p. 62 (2004)
6. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology 65(1), 178–187 (Jan 2014), http://doi.wiley.com/10.1002/asi.22954
7. Peñas, A., Rodrigo, A.: A simple measure to assess non-response. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 1415–1424 (2011)