Authorship identification using correlations of frequent features
Notebook for PAN at CLEF 2013

Timo Petmanson

Institute of Computer Science, University of Tartu
timo_p@ut.ee

Abstract In this work, we explore how well can we perform the PAN‘13 authorship identification task by using correlation of frequent significant features found in documents written by a known author. We extract features from the context of four types of words: first words of sentences or lines, nouns, verbs, punctuation. We compute the Matthews Correlation Coefficient for all pairs of extracted features and by using principal component analysis, transform them into a form suitable for using simple Euclidian distance as a metric. By hypothesizing that the distances of different documents written by a same author belong to same distribution, we can provide educated guesses for the authorship identification. Our system achieves F_1-score of 66.7% for English, 56.7% for Greek and 80% for Spanish texts with an overall F_1-score of 67.1% on the PAN‘13 dataset.

Introduction

A single subtask of the PAN‘13 authorship identification task contains up to ten documents written by a known author and a single document, that needs to be classified as either written by the same author or not. The dataset contains documents in English, Greek and Spanish languages. Similar task was also presented at PAN‘11 [2], but with a fixed set of authors.

Our approach is to first extract all frequent and statistically significant features from all given documents in a single subtask. Then, we use the correlation of the features to compute the distances between all given documents. We use Student‘s t-test to determine if the distances come from the same distribution, hence same author. On the preliminary data, our approach achieved F_1-score of 66.7% for English, 56.7% for Greek and 80.0% for Spanish texts with an overall F_1-score of 67.1%

1 Feature extraction

Given a single subtask, we started by extracting lexical and morphological features. Lexical features are the original word, does the word start with an uppercase character, are all characters uppercase, does contain punctuation, does contain digits, is first or last word in a sentence or a line. Morphological features are the word lemma, part-of-speech, other language specific features such as the case, mood etc. Similar features have traditionally been used in authorship identification tasks [8,2].

For the feature extraction, we used NLTK toolkit for POS-tagging English [3], Tree-Tagger for Spanish [7] and AUEB Tagger for Greek [4]. We encoded the features for each word as lists of tuples

(feature, offset, value) ,

where *offset* determines the feature/value combination of the word relative to the position of current word. For instance, a feature *(case, -1, nominative)* would say that the case of the previous word was nominative. We enumerated all possible features with the offset ranging from -2 to 2. Additionally, we were interested in extracting composite features that could express the co-occurrence of two or more simple features. For example, a composite feature could be *(pos-tag, 1, noun) & (case, 0, partitive)*.

Frequent composite features. We a say a feature is *frequent*, if its support (number of occurrences) is greater or equal to a predefined threshold. We set the treshold so that every feature matching at least 5% of tokens would be frequent. Particularly, we are interested in extracting frequent features for four different types of tokens: first words of sentences or lines, nouns and adjectives, verbs, punctuation. The tresholds were set respectively to the number of tokens in each particular group.

Counting the support of simple features is straightforward. To obtain frequent composite features, we can use the *monotonicity* property of support. Given two features A and B, we can assure that

$$\mathrm{support}(A\&B) \leq \min(\mathrm{support}(A), \mathrm{support}(B)) \enspace , \tag{1}$$

which means that if either of the features A and B are not frequent, then also their conjunction $A\&B$ is not frequent. The monotonicity property makes it possible to extract all frequent composite features efficiently using the Apriori algorithm used in frequent itemset mining [1]. Basic and composite features can be encoded as *itemsets* and tokens as *transactions*.

Statistical significance. Not all frequent features are relevant to authorship identification due to common frequent words and language-specific grammatical and stylistic patterns. For this purpose, we compiled a random subsets of documents from Brown corpus [5], Spanish and Greek Wikipedia. Each subset contained 100 documents.

For each frequent feature, we first compute the percentage of matched tokens on every document in the reference corpus of the subtask language. Then, we estimate the p-value as the fraction of documents with higher or equal percentage of matches than was obtained on the documents in the subtask. Next, we filter out the features that are not statistically significant (p-value 0.025). This will help to reduce the number of irrelevant features.

2 Feature correlation as a distance metric

We use *Matthews correlation coefficient* (MCC) [6] to measure the similarity between two different features in a single document as we represent each feature as a bitvector of matched positions and MCC has proved to be more stable with binary vectors. We compute MCC between all pairs of features of document and compile a vector length $m = k(k-1)/2$, where k is the number of features. Given n documents, we compile a $n \times m$ matrix to store the correlation vectors. Each row represents the feature correlations of a particular document.

As we extracted features for four different groups of words: first words of a sentence or a line, verbs, nouns, punctuation, we compile a total of four such matrices for each subtask and by row-wise concatenation put them in a single matrix M containing all the pair-wise correlations for features in every particular group of words.

Next, we find $n-1$ principal components from matrix M and project the data as a new matrix P. As a result, we can view each document as a point in $(n-1)$-dimensional Euclidian space and use Euclidian distance to compute the distance between any two documents. As an alternative, we also measured the distance between the documents by a computing *mean-squared errors* (MSE) between respective correlations, but MSE did not prove to be as discriminating as the approach using PCA.

We hypothesize that if the unknown document is not written by the same author, its mean distance to known documents should be significantly larger than the mean similarity between known documents. If average mean distance of known vs unknown document was less or equal to mean distance between known and known, we automatically report the unknown document to written by the same author. Otherwise, we use one-tailed Student's t-test to obtain the p-value, which describes the statistical significance between the means. We use rather large threshold 0.5 to determine the authorship:

$$result = \begin{cases} pvalue \geq 0.5 & \text{Y (same author)} \\ pvalue < 0.5 & \text{N (different author)} \end{cases} .$$

Some subtasks of PAN‘13 dataset only contained one known document from the author. In such cases, we split the known document into two and compared the first and second part of the document as our approach requires at least two known documents from an author.

3 Evaluation

In Figure 1, we have depicted the distances our method for all three languages on training subset of the data that was available during writing this notebook. We see that our distance metric is quite discriminating for English and Spanish languages as documents written by same authors are on average more similar than the documents written by someone else. Our approach seems to have most difficulty with Greek language as the distances do not reflect very well the authorship of the documents.

On the full evaluation dataset, our approach achieved following F_1-scores: English - 66.7%, Greek - 56.7%, Spanish - 80.0% with overall F_1-score of 67.1%. On Spanish subset, our approach was ranked second. Overall, our solution shared fourth and fifth ranking from a total of eighteen participants.

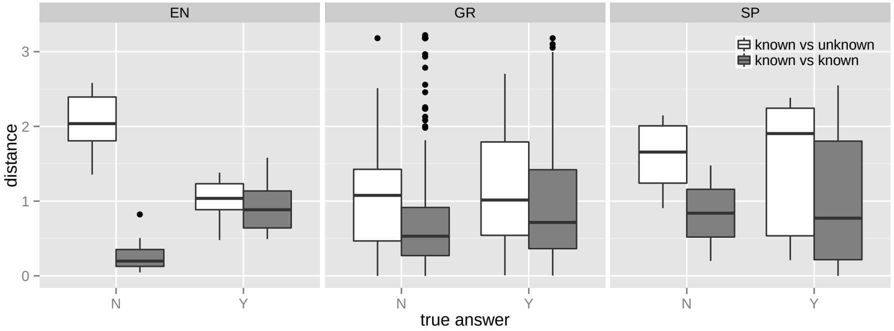

Figure 1. Feature correlation based distances between known and unknown documents for English, Greek and Spanish languages in training subset of PAN'13 authorship identification dataset.

Summary

In this work, we have shown that the features we extract contain rather strong signal relevant for authorship identification. At least our t-test based approach handles certain cases correctly. However, we might improve the results by using more sophisticated *novelty detection* algorithms instead. Also, we might get stronger features by performing PCA separately on each of the four token groups. The next important step in our future work is evaluating these options.

References

1. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB. vol. 1215, pp. 487–499 (1994)
2. Argamon, S., Juola, P.: Overview of the international authorship identification competition at pan-2011. In: CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers), Amsterdam, The Netherlands (2011)
3. Bird, S.: Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions. pp. 69–72. Association for Computational Linguistics (2006)
4. Koleli, E.: A new Greek part-of-speech tagger, based on a maximum entropy classifier. Master's thesis, Department of Informatics, Athens University of Economics and Business (2011)
5. Kučera, H., Francis, W.N.: Computational analysis of present-day American English. Dartmouth Publishing Group (1967)
6. Matthews, B.W.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 405(2), 442–451 (1975)
7. Schmid, H.: Treetagger. TC project at the Institute for Computational Linguistics of the University of Stuttgart (1994)
8. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3), 538–556 (2009)