Experiments in Authorship-Link Ranking and Complete Author Clustering

Notebook for PAN at CLEF 2016

Valentin Zmiycharov[1], Dimitar Alexandrov[1], Hristo Georgiev[1], Yasen Kiprov[1], Georgi Georgiev[1], Ivan Koychev[1], and Preslav Nakov[2]

[1] FMI, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
{valentin.zmiycharov, dimityr.alexandrov, hristo.i.georgiev}@gmail.com,
{yasen.kiprov, g.d.georgiev}@gmail.com, koychev@fmi.uni-sofia.bg

[2] Qatar Computing Research Institute, HBKU, Doha, Qatar
pnakov@qf.org.qa

Abstract The paper presents the approach we developed for the Authorship-Link Ranking and Complete Author Clustering task at the PAN 2016 competition. Given a document collection, the task is to group documents written by the same author, so that each cluster corresponds to a different author. This task can also be viewed as one of establishing authorship links between documents. We use a combination of classification and agglomerative clustering with a rich set of features such as average sentence length, function words ratio, type-token ratio and part of speech tags.

1 Introduction

For this task, we are given a collection of up to 100 documents. All of them are single-authored, in the same language, and belong to the same genre: the language and the genre are given. The topic and the length of the documents vary, and the number of distinct authors whose documents are included in the collection is unknown.

The participating systems have to provide two outputs for each instance:

– **Complete author clustering result**: Each cluster should contain all documents found in the collection by a specific author. The clusters should be non-overlapping, i.e., each document should belong to exactly one cluster (Figure 1).

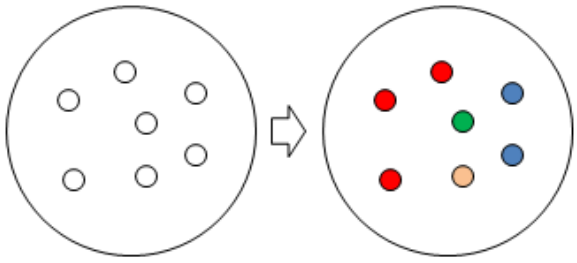

Figure 1. Complete author clustering explained. (Taken from [6])

– **Authorship-link ranking result**: A list of document pairs ranked according to a real-valued score in [0,1], where higher values denote higher confidence that the pair of documents are written by the same author (Figure 2).

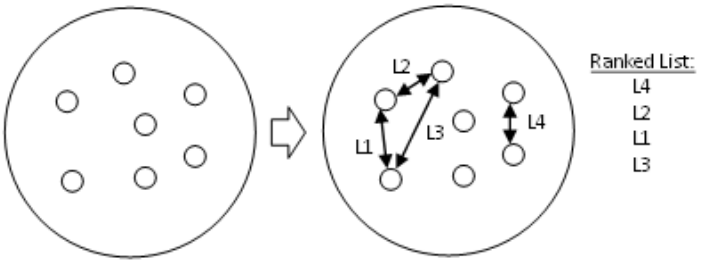

Figure 2. Authorship-link ranking. (Taken from [6])

2 Method

After analyzing the training documents, we concluded that we cannot use typical document similarity as a feature, e.g., based on TF-IDF, which is good for classification into topics, but here different authors may write on the same topic. Instead, we focused on features that model author style and are orthogonal to topic-related features [3]. Using these features, we first perform classification for each pair of documents about whether they are written by the same author. Then, we use agglomerative clustering using the classifier's confidence scores for each pair of documents.

2.1 Training Set Analysis

The training set contains 18 folders for three languages and two genres, i.e., three folders per language/genre pair:

- **English / Articles**
- **English / Reviews**
- **Dutch / Articles**
- **Dutch / Reviews**
- **Greek / Articles**
- **Greek / Reviews**

Each folder contains between 50 and 100 documents. The clusters are represented by thresholded authorship-link values of 0 and 1, where 1 means that the two documents are by the same author, and thus should be in the same cluster (only those with a value of 1 are presented). Therefore, the task may be seen as a classification task asking whether two documents are by the same author or not. In order to distinguish one author from another one, author style features have to be implemented, as content-based similarity is not very helpful in this case.

2.2 Features

We used the following features:

- **Average sentence length:** When writing a document, an author could sometimes unconsciously use conjunctions or commas instead of ending the sentence with a period. Thus, this feature could be very indicative for authorship attribution.
- **Function words ratio:** Different authors have different biases and preferences with respect to the use of function words, which makes these words some of the most popular features for stylometry and authorship attribution. We use three separate lists of function words[3] for English (173 words), Greek (250 words), and Dutch (104 words).
- **Type-token ratio:** The richness of the vocabulary used by an author is another indicator of style. We use the number of unique word types in a document divided by the total number of tokens in the document. We consider two documents to be written by the same author if they have the same (or similar) type-token ratios. This feature also reflects the author's tendency to repeat words.
- **Features, based on part of speech:** described below in detail.

For the part-of-speech (POS) features, we used the following taggers:

- **English** - Stanford Log-linear Part-Of-Speech tagger [7]
- **Dutch** - Stanford Log-linear Part-Of-Speech tagger [7]
- **Greek** - AUEB's Natural Language Processing Group Part-of-Speech tagger[4]

After tagging the documents, we extracted the following part-of-speech features: **Nouns ratio, Adjectives ratio, Verbs ratio, and Conjunctions ratio**. These features are based on the same principle, which we explain below. Before comparing two documents, we perform the following steps:

1. The number of occurrences of the part-of-speech tags we are interested in (nouns, adjectives, verbs and conjunctions) is to be extracted for each sentence in the document and stored (in-memory or persisted in the file system).
2. Order statistics: all sample values in the document have to be put in ascending order.
3. After extracting the statistical distribution, based on the occurrences of the required part-of-speech tags in the document, we are now able to get some statistical knowledge from this data. For that purpose, we calculate the minimum and the maximum values (first and last values from the statistical order), first and third quantiles, and the median value (which represents the second quantile)[5].
4. For each of the features in 3, we calculate the information gain, which can strongly distinguish authors. This complex feature provides information, which is not visible when reading a document and is immune to conscious control.
5. Create a vector with the following attributes:
   - Minimum value of the distribution for the POS tag (e.g., verb) in the document

[3] http://www.ranks.nl/stopwords
[4] http://nlp.cs.aueb.gr/software.html

- Maximum value
- First quantile
- Median (second quantile)
- Third quantile
- Maximum value

By using the constructed vector, we are now able to measure the distance between any two documents from the same language, regardless of their length. The idea of this feature was inspired from [2], but we extended it with more attributes, which we think would be able to give better results. The distance between any two vectors is calculated using Euclidean distance. After that, the measured distance is added as an attribute to the classifier (see 2.3).

2.3 Classification

A classification example represents a pair of documents and whether they are by the same or by different authors. This treats the task as a binary classification problem. For each of the described features, the absolute difference of its values between the two documents in the pair is calculated. We obtained better results without normalization, and thus we pass the vector containing the differences to the classifier without modification. Note that we trained six different classifiers: one for each language/genre pair.

An approach that did not work well was to use linear regression, which computed for each document pair a number between 0 and 1 (closer to 1 are documents that are more likely to be by the same author). After some experiments and unsatisfactory results, we decided to model the task as a classification problem: whether a document pair is by the same author or not.

2.4 Clustering

We decided to iteratively create clusters **based on authorship-link results** calculated by an SVM classifier using the above features. Our clustering algorithm consists of the following steps:

1. Start with a single document from the test set, then construct and classify pairs with all other documents. The ones that are classified as positive are added to the cluster, and the rest are ignored.
2. We select some of the remaining documents and loop through all existing clusters. We add the document to all clusters which have more than 50% documents that are close to it (pairs classified as positive). If there are no such clusters, the document forms a new cluster.
3. Repeat the previous step for all the remaining documents.
4. Find documents that are contained in more than one cluster and remove it from all but one clusters – the one with the highest total similarity to the documents in the cluster. As a result, all documents are in exactly one cluster.

3 Results and Analysis

Table 1 shows results on the training corpus. Here, a positive/negative document pair means two documents written by the same/different author(s). For each feature, **the absolute difference** between the calculated values for each two documents is measured, and we show the average value of the feature for positive and for negative pairs. The results for each feature are presented, followed by the results by the trained classifier.

Feature	Avg for positive	Avg for negative
Avg sentence length	14.67	22.55
Function words ratio	0.14	0.15
Type-token ratio	0.04	0.07
Nouns ratio	4.26	4.70
Adjectives ratio	2.21	2.50
Verbs ratio	4.57	4.62
Postag conjunction	2.06	2.47
CLASSIFIER	**0.096**	**0.95**

Table 1. Average results for each feature.

We created a double variable which indicates what is the minimum rank for which we consider a pair of documents to be written by the same author. After many experiments, we chose its value to be **0.8**.

The clustering output is evaluated according to BCubed F-score [1] The ranking of authorship links is evaluated according to Mean Average Precision [4]. The results of our submission are shown in Table 2.

4 Conclusion and Future Work

We have described our authorship-link clustering solution, where we recognized different authors, based on their style of writing. All baseline features that we use in our work (average metrics for word and sentence, function words usage, dictionary richness) are useful, but not precise enough to separate authors correctly. That is the main motivation for experimenting with POS features and statistical distributions. We see great potential in such features, and we believe the current implementation can be greatly improved.

In future work, we plan to experiment with different machine learning algorithms (Neural Networks, Bagging, Active Learning), and with more stylometry features.

Acknowledgments

This research was performed by a team of students from MSc programs in Computer Science in the Sofia University “St Kliment Ohridski”. We also thank the Sofia University “St Kliment Ohridski” for the support and guidance to our team participation at the CLEF 2016 Conference.

Language	Genre	F-Cubed	R-Cubed	P-Cubed	Av-Precision
english	articles	0.77326	0.71429	0.84286	0.0030303
english	articles	0.64987	0.50408	0.91429	0.0024442
english	articles	0.88479	0.91429	0.85714	0
english	reviews	0.80802	0.725	0.9125	0
english	reviews	0.91233	0.9	0.925	0
english	reviews	0.65973	0.525	0.8875	0
dutch	articles	0.77824	0.73684	0.82456	0
dutch	articles	0.86833	0.87719	0.85965	0
dutch	articles	0.6529	0.52632	0.85965	0
dutch	reviews	0.85953	0.88	0.84	0
dutch	reviews	0.64029	0.51	0.86	0.0020102
dutch	reviews	0.75232	0.71	0.8	0
greek	articles	0.7483	0.71429	0.78571	0.010809
greek	articles	0.6378	0.50952	0.85238	0.003261
greek	articles	0.85619	0.88571	0.82857	0.0024691
greek	reviews	0.80194	0.75714	0.85238	0.0051692
greek	reviews	0.88102	0.92857	0.8381	0.011905
greek	reviews	0.66584	0.57143	0.79762	0.018752

Table 2. Submission results.

References

1. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf. Retr. 12(4), 461–486 (Aug 2009)
2. He, R.C., Rasheed, K.: Using machine learning techniques for stylometry. In: Proceedings of the International Conference on Artificial Intelligence, IC-AI '04, Volume 2 & Proceedings of the International Conference on Machine Learning; Models, Technologies & Applications, MLMTA '04. pp. 897–903. Las Vegas, Nevada, USA (2004)
3. Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1(3), 233–334 (Dec 2006)
4. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
5. Mendenhall, W., Beaver, R., Beaver, B.: Introduction to Probability and Statistics. Cengage Learning (2012)
6. Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Clustering by Authorship Within and Across Documents. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016)
7. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. pp. 63–70. EMNLP '00, Hong Kong (2000)