Knowledge Discovery

Know-Center @ PAN 2015 Author Identification

Overview

Problem

The PAN 2015 author identification task can be formulated as follows: Given a set of documents by a single known author as well as a document of unknown authorship, determine whether this unknown document was written by that particular author or not.

Approach

Our system for the PAN 2015 authorship verification challenge is based upon a two-step pre-processing pipeline. In the first step we extract different features that observe stylometric properties, grammatical characteristics and pure statistical features. In the second step of our pre-processing we merge all those features into a single meta feature space. We train an SVM classifier on the generated meta features to verify the authorship of an unseen text document. We report the results from the final evaluation as well as on the training datasets.

Implementation

We based our work for the PAN 2015 author identification challenge upon Know-Center submissions of previous years[1]–[3].

We consider authorship verification a supervised classification problem. For each author, we pre-process each document in two steps. In step one we extract different features.

After extracting those features we face a number of different feature spaces with different ranges of values. This introduces a problem for many machine learning algorithms. In step two of our pre-processing pipeline, we tackle this problem by generating meta features from those extracted features. These meta features do all exist in a single meta feature space.

To generate the meta features, we aggregate the extracted feature spaces and compare it to the unseen document using the Kolmogorov-Smirnov test.

Finally, we train an SVM classifier on our meta features.

Technologies

We used our own information extraction pipeline written in Java, the MALLET's [5] implementation of LDA for topic modelling and the open source style and grammar checker LanguageTool (https://languagetool.org/) for grammar and stylistic features. For the evaluation we trained an SVM, using the machine learning framework WEKA [6].

Features

We extract different feature groups in the first step auf our pre-processing pipeline. The feature groups with some of the features belonging to the respective group are:

- **Statistical features**: term frequencies, character n-grams, word n-grams
- **Grammar features**: wrong quotes, unpaired brackets, sentences starting with upper-case letters
- **Stylometric features**: Hapax Legomena, Brunets W, Simpsons D, Sichel S, Honores H [4] and sentence length n-grams
- **Topic features**: topic distribution (topic vectors generated by LDA [7]).

Results

In order to evaluate the performance on the training dataset we used the method *crossValidateModel* provided by the WEKA class *Evaluation* to report the results doing 10-fold cross validation. Evaluations on previously released training datasets scored similar results. Our approach performed best for the English language, which is the only language where all feature-groups are supported by our pre-processing pipeline.

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	AUC
English training set	0.91	0.087	0.912	0.91	0.91	0.882
Spanish training set	0.74	0.287	0.743	0.74	0.741	0.595
Dutch training set	0.72	0.295	0.72	0.72	0.72	0.634
Greek training set	0.74	0.263	0.74	0.74	0.74	0.669

Comparing the results of the PAN 2015 authorship verification evaluations to those on the training datasets it seems our approach is prone to overfit the training dataset.

Dataset	AUC	C1	Final Score	Runtime
English testing set	0.50692	0.506	0.2565	00:07:21
Spanish testing set	0.49	0.49	0.2401	00:04:12
Dutch testing set	0.50815	0.51515	0.26178	00:02:27
Greek testing set	0.48	0.48	0.2304	00:03:57

Publications

O. Pimas, M. Kröll, and R. Kern, "Know-Center at PAN 2015 author identification Notebook for PAN at CLEF 2015," CLEF2015 Work. Notes, 2015.

References

[1] R. Kern, "Grammar Checker Features for Author Identification and Author Profiling," *CLEF 2013 Eval. Labs Work. – Work. Notes Pap.*, 2013.

[2] R. Kern, S. Klampfl, and M. Zechner, "Vote/Veto Classification, Ensemble Clustering and Sequence Classification for Author Identification - Notebook of PAN at CLEF 2012," *Work. Notes Pap. CLEF 2012 Eval. Labs*, pp. 1–15, 2012.

[3] R. Kern, C. Seifert, M. Zechner, and M. Granitzer, "Vote/Veto Meta-Classifier for Authorship Identification - Notebook for PAN at CLEF 2011," *CLEF 2011 Proc. 2011 Conf. Multiling. Multimodal Inf. Access Eval. (Lab Work. Noteb. Pap. Amsterdam, Netherlands*, 2011.

[4] F. J. Tweedie and R. H. Baayen, "How Variable May a Constant be? Measures of Lexical Richness in Perspective," *Comput. Hum.*, vol. 32, no. 5, pp. 323–352, 1998.

[5] A. McCallum, "MALLET: A machine learning for language toolkit." 2002.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software : An Update," *SIGKDD Explor.*, vol. 11, no. 1, pp. 10–18, 2009.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn.*, 2003.

Oliver Pimas
Know-Center GmbH
opimas@know-center.at

Acknowledgements: The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.