Authorship Attribution of E-mail as a Multi-Class Task
Notebook for PAN at CLEF 2011

Kim Luyckx

CLiPS Computational Linguistics Group
University of Antwerp, Belgium
kim.luyckx@ua.ac.be

Abstract In this paper, we describe a multi-class text categorization approach to authorship attribution and test it on sets of e-mail collections. The PAN 2011 competition data consists of e-mails of variable length, written by various candidate authors, with some represented by significantly longer or more e-mails than others. Rather than construct a classifier for each separate author to discriminate it from the others (i.e. binary classification), we adopt a multi-class scheme where all authorship classes are learned simultaneously. We explore the effect of the selection of feature types and of the *C* parameter in the SVM^{multiclass} learning algorithm. Variable-length lexical features showed promising results, nevertheless our authorship attribution approach only scored a mid position amongst the other competitors, for the SMALL as well as the LARGE test sets.

Keywords: authorship attribution, text categorization, SVM multi-class

1 Introduction

Authorship attribution aims at identifying the author of an unseen document given a set of documents of known authorship (i.e. positive and negative instances). The list of candidate authors is typically closed and restricted to the most likely ones (given external circumstances such as time, age, school, forum, *etc.*). The PAN 2011 (5th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse) competition data set – based on the publicly available Enron E-Mail Corpus, a corpus of in-company e-mails – is no exception.

The art of authorship attribution is to find the balance between high-scoring features and discriminative techniques on the one hand and scalability on the other [13]. Applying authorship attribution on a large scale (e.g. in e-mail collections) requires an approach that is robust to large author set sizes, varying data sizes, long and short texts, and a variety of topics and genres [11,14]. The PAN 2011 competition data set is packed with these challenges.

As far as the learning phase is concerned, it is common practice in authorship attribution to combine several binary classifiers – often one-versus-all or one-versus-one learners – to solve a problem that is in fact multi-class. Actual multi-class learning is often avoided, partly as a result of the dominance of (binary) Support Vector Machines (SVMs) in the field. This paper applies SVM^{multiclass} in order to train a single model that distinguishes between all authorship classes simultaneously.

In this paper, the main focus is on authorship attribution in test scenarios with in-training authors only. After the development and evaluation of our authorship attribution system, we submitted test runs for the LARGE and SMALL test scenarios. We will briefly describe our attempts to detect out-of-training authors (for the LARGE+ and SMALL+ scenarios), but preliminary results did not support a test run submission. We will first elaborate on the data set characteristics and preprocessing steps taken and then describe the specifics of our approach. After that, we go into detail on the results obtained during development and on the parameters and performance of the system selected for test run submission.

2 Data set characteristics and preprocessing

The training and development data made available for the PAN 2011 competition are challenging in a number of respects. First of all, working with short texts poses a specific challenge in that it requires reliable and robust representation as well as robust learning with limited data. Some studies have shown promising results with short texts of about 500 characters [15] or 500 words [12], while others suggest 2,500 words as a minimum requirement [4]. The PAN 2011 data set, with an average e-mail length of about sixty words does not come close to those indications. Another aspect is the number of candidate authors – 26 in the SMALL set and 72 in the LARGE set. Author set size has received only limited attention so far, but nevertheless has a significant impact on classification performance as well as on the features in the attribution model [11,14]. A last aspect are skewed class distributions, with some classes being represented by 10,000 words or 200 e-mails and others by only 500 words or 10 e-mails, potentially leading to an advantage in learning for the former classes.

Only limited preprocessing of the data was performed. After tokenization, we removed all information between <*omni*> and <*/omni*> tags because it contains software-specific tags, calendar entries, e-mail addresses, and phone numbers. Although we assume this information to be irrelevant for authorial style, removing the information did cause us to lose training data for two of the authors: x10114697001411515 and 339173 (present in the SMALL and LARGE data sets).

3 Authorship Features and Classification

We adopt a standard text categorization approach [16], previously successful in topic detection, authorship attribution [6,14,17], and gender prediction [10]. We experimented with four types of features during development:

– CHR or character *n*-grams – successions of *n* characters including spaces and punctuation marks – have proven useful for language identification [2], topic detection [3] and style-based text categorization (e.g. authorship attribution) [5,9]. Taking into consideration the limited text length and high number of candidate authors in the competitiondata, character *n*-grams are particularly interesting since they have shown robustness to these effects [13]. One of the downsides of using character *n*-grams is their lack of interpretability. We tested several values for *n*: 2, 3, 4, and 5 and a combination of all (cf. *variable-length n*-grams).

– LEX or *n*-grams of words are tested in our experiments without limitation. In cross-topic authorship attribution, we would normally avoid topic-specific words as they affect performance when transferred to other topics. However, the Enron E-mail Corpus is a very homogeneous data set topic-wise, so we retained the full list.
– DISC represents a set of 124 preselected discourse features, such as *while, whereas, however, nevertheless* and *on the contrary*. Argamon [1] used a set of functional lexical features to represent the semantic function of each clause in a sentence and text (e.g. conjunction, elaboration, extension). The MOD feature type represents a set of 33 preselected modal verbs, such as *can, could, must, might, should, may, shall, would* and their negated counterparts.

The relative frequency of each feature (normalized for text length) in every e-mail is calculated and represented numerically. We restricted the number of features in CHR and LEX to a thousand by applying chi-square as a feature selection metric and select the top-*n*. This metric has been used in several studies in text categorization in general [19], and in authorship attribution specifically [5,13].

We experimented with SVM^{multiclass}, a multi-class SVM algorithm developed by Joachims [7,18] for learning and classification. SVMs are the method of choice in many studies in text categorization, and in authorship attribution in particular [1,10]. Examining the various SVM parameters was not the scope of our work, but we did explore the effect of the soft margin parameter *C*, a parameter that needs to be re-established for every data set. The other parameters are kept at their default value. According to Joachims [8], the *C* parameter "is a parameter that allows one to trade off training error vs. model complexity. A small value for C will increase the number of training errors, while a large C will lead to a behavior similar to that of a hard-margin SVM." [p. 40].

4 Experimental results and evaluation

During the development phase, we trained an authorship attribution system on all available training material and tested it on the validation set. Table 1 presents the results of a series of experiments, exploring the effect of the type of feature on the overall performance of the system on the training sets of the competition. We used the evaluation metrics as used for test run evaluation: micro- and macro-averaged precision, recall, and F_1 The result tables below are restricted to F_1 scores.

Results on the SMALL set (cf. Table 1) show that the overall performance of character *n*-grams is higher than that of lexical features. Only when a combination of lexical features of various lengths is used (in LEX), lexical features outperform character *n*-grams. Modality and discourse markers fail to score well, and combining character with lexical features does not increase performance either. Highest performance is obtained by character trigrams, a feature type we will use for the test run. In the LARGE set, character *n*-grams are slightly outperformed by word unigrams, and even more so by a combination of variable-length lexical features (in LEX). For both SMALL and LARGE, we use the top and second-best scoring feature type for the competition test run.

In these results, the *C* parameter was set relatively high, at 5,000. We explored the effect of using lower *C* values, but in most cases, the difference with the original results was not significant, so we decided to stick to *C*=5,000 for the test run.

Table 1. Experimental results on training corpus for PAN-AA-2011 with *C*=5,000. Exploration of the effect of feature type selection.

SMALL			LARGE		
Feature type	Macro F_1	Micro F_1	Feature type	Macro F_1	Micro F_1
CHR2	28.54	50.12	CHR2	22.70	35.38
CHR3	**37.10**	**59.43**	CHR3	27.34	40.59
CHR4	34.07	57.20	CHR4	23.70	36.94
LEX1	33.13	54.88	LEX1	**28.81**	**42.24**
LEX2	28.95	50.06	LEX2	23.45	37.38
LEX3	22.37	40.34	LEX3	14.28	26.40
LEX4	16.16	28.70	LEX4	10.07	21.98
LEX5	16.07	31.32	LEX5	10.54	21.73
DISC	4.51	8.58	DISC	1.66	3.42
MOD	2.04	6.54	MOD	1.74	4.44
CHR	26.93	49.74	CHR	22.04	35.56
LEX	**34.00**	**57.25**	LEX	**31.17**	**46.14**
CHR+LEX	31.38	54.12	CHR+LEX	24.45	38.21
MOD+DISC	7.20	13.71	MOD+DISC	2.20	4.03

For the SMALL+ and LARGE+ cases in the competition data, out-of-training authors were included in the test set. Detecting out-of-training authors either requires negative instances labelled 'NoneOfTheAbove' in training or a learning algorithm that is able to make a 'NoneOfTheAbove' decision. Since we apply SVM^{multiclass} for hard classification, we tested two naive strategies to create artificial negative instances on the basis of the positive instances we had already created for the SMALL and LARGE cases. A first strategy ('class average') was to add for each in-training class an instance representing the average values for all *positive* instances of that class (and for each feature). A second strategy ('negative class average') was to add for each in-training class an instance representing the average values for all *negative* instances of that class. Results shown in Table 2 indicate that the negative class average strategy does indeed influence some decisions, but we decided against submitting a test run for the cases with out-of-training authors.

Table 2. Experimental results of exploratory strategies to create training instances for out-of-training authors, on the SMALL set with CHR3 and *C*=5,000

Strategy	Macro F_1	Micro F_1
None	20.03	45.11
Class average	20.01	45.07
Negative class average	20.43	45.56

For the test run, we first merged the original training data with the validation set released for development into a larger set of e-mails to be used for training, thus significantly increasing the training set size. Table 3 shows results of both test runs for SMALL

and LARGE. In both cases, performance on the competition test data was very much in line with results on the validation set, which is a good indication of the classifier's robustness and reliability. However, while we expected character trigrams to score best for the SMALL set, they were outperformed by variable-length lexical features. These last also perform best in the LARGE set. We ranked 6th and 9th (out of 17 competitors) for SMALL and 7th and 9th (out of 18) for LARGE. The winning submission for SMALL, by Kourtis *et al.*, scored 47.5% Macro F_1 and 71.7% Micro F_1. The winning submission for LARGE, by Tanguy *et al.* scored 52.0% Macro F_1 and 65.8% Micro F_1.

Looking at these variable-length lexical features, we see – apart from dates and locations – expressions of politeness (*thanks, regards, you soon*), e-mail specifics (*attached is*), pronouns, argumentation elements (*for he*), company names (*Reliant, Dominion, Enpower*), and domain-specific words (*pipeline*). Although they are better interpretable than character *n*-grams, the usefulness of these lexical features is not intuitively clear.

Table 3. Test run evaluation with *C*=5,000. The asterisk indicates the submission that was submitted last and therefore counts for ranking.

Test set	Feature type	Macro			Micro			Position
		Precision	Recall	F_1	Precision	Recall	F_1	
SMALL	LEX	43.5	37.8	37.1	64.2	64.2	64.2	6/17
	CHR3	44.4	35.6	34.3	62.0	62.0	62.0	9/17*
LARGE	LEX	39.1	34.4	34.2	52.2	52.2	52.2	7/18*
	LEX1	34.8	34.5	34.0	50.0	50.0	50.0	9/18

5 Conclusions

In this paper, we described our approach to the authorship attribution task as designed for the PAN 2011 competition. The data set was particularly challenging as it consists of short e-mails written by 26 (for the SMALL set) and 72 authors (for the LARGE set). We took a commonly used text categorization approach and experimented with various types of features. Rather than redefine a multi-class task to several binary tasks, as is often done in the field, we applied a multi-class SVM to ensure all authorship classes are learned simultaneously.

During the development phase, we explored the effect of the feature types and the *C* parameter in SVM^{multiclass}. For the test run, we selected lexical and character *n*-gram features. We also experimented with the data sets where out-of-training classes needed to be identified as such, but decided against submission of a test run for those cases. The actual test run showed that lexical features scored as expected, but this did not lead to a very high ranking.

Acknowledgements

The research presented in this paper is funded through the IWT project AMiCA: Automatic Monitoring for Cyberspace Applications.

References

1. Argamon, S., Whitelaw, C., Chase, P., Dawhle, S., Hota, S., Garg, N., Levitan, S.: Stylistic text classification using functional lexical features. Journal of the American Society of Information Science and Technology 58(6), 802–822 (2007)
2. Cavnar, W., Trenkle, J.: N-gram-based text categorization. In: Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval. pp. 161–175. Las Negas, NV (1994)
3. Clement, R., Sharp, D.: Ngram and Bayesian classification of documents for topic and authorship. Literary and Linguistic Computing 18(4), 423–447 (2003)
4. Eder, M.: Does size matter? Authorship attribution, small samples, big problem. In: Pierrazo, E.e.a. (ed.) Proceedings of Digital Humanities 2010. pp. 132–135. Centre for Computing in the Humanities, King's College London, London, UK (2010)
5. Grieve, J.: Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing 22(3), 251–270 (2007)
6. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: Proceedings of Artificial Intelligence: Methodology, Systems, and Applications (AIMSA). pp. 77–86. Heidelberg: Springer Verlag, Varna, Bulgaria (2006)
7. Joachims, T.: Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods – Support Vector Learning. MIT Press (1999)
8. Joachims, T.: Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms. Kluwer/Springer (2002)
9. Keselj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics. pp. 255–264. Halifax, Canada (2003)
10. Koppel, M., Argamon, S., Shimoni, A.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2003)
11. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis 45(1), 83–94 (2011)
12. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research 8, 1261–1276 (2007)
13. Luyckx, K.: Scalability issues in authorship attribution. Brussels, Belgium: University Press Antwerp (2010)
14. Luyckx, K., Daelemans, W.: The effect of author set size and data size in authorship attribution. Literary and Linguistic Computing 26(1), 35–55 (2011)
15. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 482–491. Syndney, Australia (2006)
16. Sebastiani, F.: Machine learning in automated text categorization. Association for Computing Machinery Computing Surveys 34(1), 1–47 (2002)
17. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic text categorization in terms of genre and author. Computational Linguistics 26(4), 461–485 (2000)
18. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research 6, 1453–1484 (2005)
19. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: Fisher, D. (ed.) Proceedings of the Fourteenth International Conference on Machine Learning (ICML). pp. 412–420. San Francisco, CA, USA: Morgan Kaufmann, Nashville, Tennessee, USA (1997)