An Empirical Evaluation of SVM on Meta Features for Authorship Attribution of Online Texts

Hongwei Yao[1], Tieyun Qian[1], Li Chen[2], Manyun Qian[2], and Xueyu Mo[2]

[1] State Key Laboratory of Software Engineering,
Wuhan University, Wuhan, China
[2] Department of Computer Science,
Central China Normal University, Wuhan, China
yaohongwei2012@126.com, qty@whu.edu.cn,
{ccnuchenli,ccnuqianmanyun,guilinmoxueyu}@163.com

Abstract. Authorship attribution (AA) has been studied by many researchers. Recently, with the widespread of online texts, authorship attribution of online texts starts to receive a great deal of attentions. The essence of this problem is to identify a set of features that can capture the writing styles of an author. However, previous studies on feature identification mainly used statistical methods and conducted out experiments on small data sets, i.e., less than 10. This scale is distance from the real application of AA of online texts. In addition, due to the special characteristics of online texts, statistical approaches are rarely used for this problem. As the the performance of authorship identification depends highly on the the combination of the features used and classification methods, the feature sets for traditional authorship attribution needs to be re-examined using machine learning approaches. In this paper, we evaluate the effectiveness of six types of meta features on two public data sets with SVM, a well established machine learning technique. The experimental results show that lexical and syntactic features are the most promising features for AA of online texts. Furthermore, a number of interesting findings regarding the impacts of different types of features on authorship attribution are discovered through our experiments.

Keywords: authorship attribution of online texts, meta features, comparative evaluation.

1 Introduction

Authorship attribution (AA) , also known as authorship classification, has been studied by many researchers [11,27,8,12,6]. It was originally proposed to classify Shakespeare plays, Bronte Sisters' novels, etc. Later on, it was applied to other literary works such as American and English literature and news articles. Recently, AA was applied to various types of online texts such as emails [29,16], blogs [20], forum posts [26] and reviews [17]. We call this theme of studies authorship attribution of online texts. The problem of AA of online texts is useful in many applications, e.g., fake reviewers detection, internet plagiarism and cybercrime investigation.

The number of classes or authors used in the traditional AA is often very small. For example, the corpus used by Escalante et al. [6] comprises documents of 10 authors.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 28–37, 2013.

The number of classes used by Kim et al. [17] is only 8. Recently, Solorio et al. showed that the classification results deteriorated quickly as the number of authors increase from 5 to 100 [26]. However, the number of classes (authors) for AA of online texts can be much larger. For example, in online reviews, a large number of reviewers have written reviews about products and services. The complexity level of AA of online texts is thus becoming extremely high.

The research effects on authorship attribution mainly focus on (1) developing advanced techniques, and (2) extracting effective features. Early studies mainly use statistical univariate methods such as Naïve Bayes (NB) classifier and principle component analysis (PCA). Most modern AA approaches are based on machine learning techniques like decision trees, neural networks, and support vector machine (SVM). Generally, machine learning approaches show better performance than statistical methods. On the other hand, the extraction of writing features has been the essence of AA ever since the earliest work. Almost 1000 features, including length features, richness features, character and lexical features, syntactic features, stylistic features, have been explored in the literature. However, there exist lots of controversies on the effectiveness of different features. The main reason can be due to the various application circumstances and the lack of publicly available data sets. In addition, the performance of authorship analysis depends highly on the the combination of the features used and analytical techniques [32]. It is hard to reach agreements on a best set of features for different approaches. Although Grieve and Halteren conducted quantitative evaluations on a number of features [9,10], these two studies have a number of shortcomings.

- They used the statistical analysis [9] and distance comparison [10] in authorship analysis. Due to the sensitivity to noises, the incapability to deal with the large number of features, and the strong requirements on mathematical assumptions, these techniques are rarely used in modern authorship attribution.
- They carried out experiments on data sets of a relatively small size. The number of authors classified by Grieve [9] and Halteren [10] is 40 and 8, respectively. An analysis on this scale is far from the real application of authorship attribution of online texts.

Given the fact that machine learning techniques achieve high accuracy in AA of online texts [6,26], it is desirable to investigate how the features perform with machine learning approaches for large AA problems. Furthermore, the machine learning technique allows to process a very large number of features. This provides the opportunity to use meta features which enable a closer look at the same type of features on a high level. In this paper, we re-examine the effectiveness of six types of commonly used meta features for AA of online texts. We use two public data sets for online texts. Each of them has a large number of authors, i.e., 62 and 100. The learning technique is support vector machine (SVM). We aim to seek answers to the following questions with empirical evidence.

1. How do the various meta features perform with the widely used machine learning technique like SVM? Are there any meta features consistently better than other features for the large AA problems?
2. Can the attribution task benefit from the combinations of different types of meta features? If this is the case, what is the most promising combination?

2 Related Works

Authorship attribution has received a great deal of attentions in recent years. A variety of approaches have been developed for this problem. Existing methods can be categorized into two main themes. One focuses on finding appropriate features for quantifying the authors' writing style, and the other focuses on developing efficient and effective techniques to perform the classification task.

There is a body of literature examining the effects of different features. The use of function words could date back about half a century ago [22]. Since then, various features have been proposed for modeling writing styles. Existing studies show that the function words [1,3] and rewrite rules [11] might be promising in AA problems. Other features that have been investigated include length features [7,8], richness features [11,19], punctuation frequencies [8], character n-grams [9,12], word n-grams [4,23], POS n-grams [7,13], *k*-ee subtree [17] and topic models [24].

There are also a number of works that study the attribution methods. Mosteller used the Bayesian statistical analysis on function words and obtained good discrimination results [22]. Additional research in recent years focuses exclusively on classification or categorization methodologies, including discriminant analysis [27], PCA [14], exponentiated gradient algorithm [2], neural networks [8,32], multi-layer perceptrons [8], clustering [23], decision trees [28,31], and SVM [5,7,19,12]. In general, machine learning approaches have better performance than statistical methods [32]. In particular, SVM is regarded as one of the best approaches [21,17].

Overall, current surveys on different feature sets are carried out on a small number of authors. There is very limited work on the evaluation of features for the large scale AA problems of online texts. More importantly, previous studies that conduct comparative evaluation on features for AA problem use the statistical analysis [9] and distance comparison [10]. Thorough evaluations using the well established SVM method for large AA problems are still missed in the literature. Therefore, it becomes necessary to make a comprehensive study on how different features affect large authorship attribution of online texts within a SVM framework. That is why we conduct this study.

3 Meta Features

Let $A = \{a_1, a_2, ..., a_k\}$ be a set of *k* authors and $D = \{D_1, D_2, ..., D_k\}$ be *k* sets of documents with D_i being the document set of author $a_i \in A$. Supervised AA builds a model or classifier from the training data and applies it to the test set to determine the author a of each test document d, where a is from A ($a_i \in A$). Each author is treated as a class, and each document is represented as a vector of features. In this paper, we extracted six types of meta features (see below). We do not use any application-specific features such as structural layouts [29] and signature [26] because these features are domain dependant and thus not universally adopted in most of AA problems.

3.1 Length Meta Features

We compute the average document length in terms of word count in one document, the average sentence length in terms of word count in one sentence, and the average word length in terms of character count in one word, which give us three average length features.

3.2 Character Meta Features

Character n-grams is simple and easily available for any natural language [9]. In this paper, we extract frequencies of n-grams ($n = 1..2$) on the character-level.

3.3 Lexical Meta Features

It is straightforward to view an text article as a bag-of-words, like that has been widely used in topic-based text classification. We represent each article by a vector of word frequencies.

3.4 Syntactic Meta Features

We use four typical content-independent structures including n-grams of POS tags ($n = 1..3$) and rewrite rules [7,13]. The syntactic features are extracted from the parsed syntactic trees. We use the Stanford PCFG parser [18] to generate the grammar structure of sentences in each document.

3.5 Stylistic Meta Features

We derive three stylistic features directly from the raw data. (1) Function words: We use a list of *157* function words in this paper, which is downloaded from $www.flesl.net/Vocabulary/Single-word\_Lists/function\_word\_list.php$. (2) Punctuation frequency [8]: We use *32* common punctuation marks in our experiments. (3) The frequency of each word-length for each article: we get a distribution for k-length word ($1 \leq k \leq 15$).

3.6 Richness Meta Features

Originally, the vocabulary richness functions are used to quantify the diversity of the vocabulary of a text [11,30]. In this paper, we apply the richness metrics to counts of word unigrams, POS tags ($n = 1..3$), and rewrite rules.

4 Experimental Evaluation

All our experiments use the $SVM^{multiclass}$ classifier [15] with default parameter settings. We report classification accuracy as the evaluation metric.

4.1 Experiment Setup

We use two different kinds of online texts. The first one consists of posts from the Chronicle of Higher Education (CHE) [26]. This data set has 100 authors with 16,171 documents. The second one is IMDB data set [25] which contains the IMDB reviews in May 2009. This data set has 62,000 reviews by 62 users (1,000 reviews per user). Both of the data sets are publicly available upon the request to authors. For CHE data, we conduct experiments on its fixed partition of training (80%) and testing (20%) for all collections. For IMDB, we randomly split training and test documents 5 times, 70% of one author's documents are used for training and the rest 30% for testing. The results are averaged over the 5 splits.

We extract and compute the length, character, lexical, stylistic, and richness meta features directly from the raw data, and we use the Stanford PCFG parser [18] to generate the grammar structure of sentences in each document for extracting syntactic features. We do not remove stop words as some of them are actually function words. We normalize each feature's value to [0, 1] interval by dividing by the maximum value of this feature in the training set. Table 1 shows some statistics on these two data sets.

Table 1. Vocabulary size for different features

Meta Features	Features	Vocabulary Size	
		CHE	IMDB
Length	Avg Doc Len	1	1
	Avg Sent Len	1	1
	Avg Word Len	1	1
Character	Char 1-Gram	1476	2094
	Char 2-Gram	6286	13805
Lexical	Bag of words	34840	195274
Syntactic	POS 1-Gram	63	63
	POS 2-Gram	1575	1917
	POS 3-Gram	12967	21950
	Rewrite Rules	7916	19240
Stylistic	Function Words	157	157
	Punctuation	32	32
	Len k Words	15	15
Richness	Word	6	6
	POS 1-Gram	6	6
	POS 2-Gram	6	6
	POS 3-Gram	6	6
	Rewrite Rules	6	6
	Function Words	6	6

From Table 1, we can see that the two data sets have their own characteristics, and the corresponding classification tasks differ significantly in their difficulty. For example, the number of words in IMDB is greatly larger than that in CHE. In addition, there are more syntactic features in IMDB, indicating the authors intend to use more flexible syntactic structures when writing.

4.2 Results with Single Meta Features

Table 2 presents the results in terms of accuracy with single meta features. It is clear that the lexical and syntactic meta features have the best performance. Especially on IMDB data, their improvements over other meta features are very significant. On the other hand, both the length and richness meta features are the least successful of all the features. Contradiction to the past research [9] which reported that character n-gram are some of the most accurate techniques in their test, the character n-grams do not have a big enough impact on authorship attribution of online texts. It only ranks the third and the forth among six types of meta data for CHE and IMDB, respectively. Overall, the results for IMDB are much better than those for CHE, which is naturally because CHE contains 100 authors (classes) while IMDB only has 62.

Table 2. Results on single meta features

Meta Features	Acc. (CHE)	Acc. (IMDB)
Length	3.29	1.97
Character	7.32	17.92
Lexical	**17.57**	**47.37**
Syntactic	**12.88**	**50.80**
Stylistic	11.78	12.68
Richness	3.69	2.58

The Performance Ranking of Single Meta Feature

The performance ranking of single meta feature for CHE and IMDB in descending order is {Lexical, Syntactic, Stylistic, Character, Richness, Length}, and {Syntactic, Lexical, Character, Stylistic, Richness, Length}, respectively.

4.3 Results with the Combination of Two Types of Meta Features

Results for Combo_length

It is clear from Table 3 that the combination of length meta feature and any other features performs better than the single length feature. This is intuitive because the length meta feature consists only three dimensions, which are a bit too less for SVM classifier. If we take a closer look, we can also find that the two performance rankings are totally consistent with those for single meta feature. This strongly indicates that the length meta feature has a very slight impact on this task.

Results for Combo_richness

In Table 4, one can see that most of the combinations of richness and other meta features are more successful than richness itself. Almost all of them have a positive change. The only exception is the combination of richness and length on IMDB. It has a negative change. However, the same combination on CHE data set does achieve a significant improvement over its corresponding single richness feature. We also observe that rich_syntac is very close to rich_lex for IMDB but it is higher than rich_lex for CHE. This shows that richness is not a stable combination factor. Its performance varies with the data set and the counterpart meta feature.

Table 3. Results on combo-length meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Len_Char	6.20	+49.94%	17.55	+88.77%
Len_Lex	15.94	+79.36%	40.08	+95.08%
Len_Rich	4.98	+33.54%	2.30	+14.35%
Len_Style	6.45	+48.99%	12.62	+84.39%
Len_Syntac	13.26	+75.19%	50.65	+96.11%

Table 4. Results on combo-richness meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Rich_Char	6.47	+42.97%	15.57	+12.99%
Rich_Len	4.98	+25.90%	2.30	-12.17%
Rich_Lex	12.85	+71.28%	48.37	+94.67%
Rich_Style	7.13	+48.25%	13.42	+80.77%
Rich_Syntac	16.24	+77.28%	48.98	+94.73%

Results for Combo_stylistic

In Table 5, we can see that that the combination of stylistic and other meta features sometimes deteriorates the performance. Another interesting finding is the Style_Lex combo performs significantly better than the Style_Syntac combo on IMDB. However, CHE does not show the same pattern. Indeed, the performance of Style_Lex combo is much worse than lexical meta feature on CHE. Similarly, the Style_Rich combo feature shows the opposite performance change than its single richness feature on CHE and IMDB. These finding infer that the improvement of combination of stylistic with lexical and richness feature is dependent on the characteristic of data sets.

Table 5. Results on combo-stylistic meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Style_Char	6.85	-71.97%	21.89	+42.07%
Style_Len	6.45	-82.64%	12.62	-0.48%
Style_Lex	16.46	+28.43%	54.17	+76.59%
Style_Rich	7.13	-65.22%	13.42	+5.51%
Style_Syntac	21.01	+43.93%	43.98	+71.17%

Results for Combo_character

Table 6 shows the results on combo-character meta features. While other combinations are useful to some extent, the performances of Char_Len and Char_Rich decrease on both the CHE and IMDB data sets. Hence it is not good to combine the character meta feature with either length or richness meta feature.

Table 6. Results on combo-character meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Char_Len	6.20	-18.06%	17.55	-2.10%
Char_Lex	24.45	+70.06%	52.31	+65.74%
Char_Rich	6.47	-13.14%	15.57	-15.09%
Char_Style	6.85	-7.30%	21.89	+18.14%
Char_Syntac	16.49	+55.61%	51.61	+65.28%

Results for Combo_lexical

Table 7 show the results for combo_lexical. The Lex_Syntac combo achieves the most significant improvement on IMDB dataset. This is intuitive since the single lexical and syntactic meta features rank the first and the second by themselves. However, it is a bit surprising that it is Lex_Char rather than Lex_Syntac to be the best for CHE. This hints that some weak meta features may have a very positive impact on AA if it is combined with a strong one.

Table 7. Results on combo-lexical meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Lex_Char	24.45	+28.14%	52.31	+9.44%
Lex_Len	15.94	-10.23%	40.08	-18.19%
Lex_Rich	12.85	-36.73%	48.37	+2.07%
Lex_Style	16.46	-6.74%	54.17	+12.55%
Lex_Syntac	21.80	+19.40%	68.03	+30.37%

Results for Combo_syntactic

Table 8 shows the results for combo_syntactic. All the combo syntactic meta features reach an improvement over the single syntactic meta feature on CHE. However, on IMDB, we note there are some decreases. The reason can be due to that syntactic is already the best single meta feature for IMDB. The combination with a weak meta feature may be harmful to the performance.

The Performance Ranking of Combo Meta Feature

The performance ranking of combo meta features for CHE and IMDB in descending order is {Lex_Char, Lex_Syntac, Syntac_Char, Style_Lex, Syntac_Rich}, and {Lex_Syntac, Style_Lex, Lex_Char, Syntac_Char}, respectively. Note that we only list the combos which perform better than both of the corresponding single meta features.

Table 8. Results on combo-syntactic meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Syntac_Char	16.49	+21.89%	51.61	+1.57%
Syntac_Len	13.26	+2.87%	50.65	-0.30%
Syntac_Lex	21.80	+40.92%	68.03	+25.33%
Syntac_Rich	16.24	+20.69%	48.98	-3.72%
Syntac_Style	21.01	+38.70%	43.98	-15.51%

5 Conclusion

In this paper, we adopt a machine learning algorithm, i.e., SVM, to examine the impacts of different meta features on authorship attribution of online texts. We conduct extensive comparative studies in authorship recognition using single and combo meta features on two real world data sets with a large number of classes. We have the following interesting findings. Firstly, the lexical and syntactic meta features are the most promising for AA of online texts, and the the effects of length and richness are trivial. Secondly, the performance of the combination of two types of meta features is dependant on the data and the single meta feature. As some of the combinations deteriorate the performance, one should carefully examine the characteristics of data and the performance of single meta feature before the combination is conducted. Thirdly, our results show that the combination of two strong meta features outperform any of their corresponding individual features, and thus this kind of combination is generally more applicable than that of two weak ones.

Acknowledgements. This research was supported in part by the NSFC project (61272275, 61202036, 61272110, U1135005), and the 111 project(B07037).

References

1. Argamon, S., Levitan, S.: Measuring the usefulness of function words for authorship attribution. In: Literary and Linguistic Computing, pp. 1–3 (2004)
2. Argamon, S., Šarić, M., Stein, S.S.: Style mining of electronic messages for multiple authorship discrimination: First results. In: Proc. of the 9th SIGKDD, pp. 475–480 (2003)
3. Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., Levitan, S.: Stylistic text classification using functional lexical features: Research articles. JASIST 58, 802–822 (2007)
4. Burrows, J.F.: Not unles you ask nicely: The interpretative nexus between analysis and information. Literary and Linguistic Computing 7, 91–109 (1992)
5. Diederich, J., Kindermann, J., Leopold, E., Paass, G., Informationstechnik, G.F., Augustin, D.S.: Authorship attribution with support vector machines. Applied Intelligence 19, 109–123 (2000)
6. Escalante, H.J., Solorio, T., Montes-y Góməz, M.: Local histograms of character n-grams for authorship attribution. In: Proc. of the 49th ACL, pp. 288–298 (2011)
7. Gamon, M.: Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: Proc. of the 20th COLING (2004)

8. Graham, N., Hirst, G., Marthi, B.: Segmenting documents by stylistic character. Natural Language Engineering 11, 397–415 (2005)
9. Grieve, J.: Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing 22, 251–270 (2007)
10. van Halteren, H.: Author verification by linguistic profiling: An exploration of the parameter space. ACM Transactions on Speech and Language Processing 4, 1–17 (2007)
11. van Halteren, H., Tweedie, F., Baayen, H.: Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing 11, 121–132 (1996)
12. Hedegaard, S., Simonsen, J.G.: Lost in translation: authorship attribution using frame semantics. In: Proc. of the 49th ACL, pp. 65–70 (2011)
13. Hirst, G., Feiguina, O.: Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing 22, 405–417 (2007)
14. Hoover, D.L.: Statistical stylistics and authorship attribution: An empirical investigation. Literary and Linguistic Computing 16, 421–424 (2001)
15. Joachims, T.: Making large-scale support vector machine learning practical. In: Advances in Kernel Methods, pp. 169–184. MIT Press (1999)
16. Kern, R., Seifert, C., Zechner, M., Granitzer, M.: Vote/veto meta-classifier for authorship identification - notebook for pan at clef 2011 (2011)
17. Kim, S., Kim, H., Weninger, T., Han, J., Kim, H.D.: Authorship classification: a discriminative syntactic tree mining approach. In: Proc. of the 34th SIGIR, pp. 455–464 (2011)
18. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proc. of the 41st ACL, pp. 423–430 (2003)
19. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Proc. of the 21st ICML (2004)
20. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Lang. Resources & Evaluation 45, 83–94 (2011)
21. Li, J., Zheng, R., Chen, H.: From fingerprint to writeprint. Communications of the ACM 49, 76–82 (2006)
22. Mosteller, F.W.: Inference and disputed authorship: The Federalist. Addison-Wesley (1964)
23. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, markov chains and author unmasking: an investigation. In: Proc. of EMNLP, pp. 482–491 (2006)
24. Seroussi, Y., Bohnert, F., Zukerman, I.: Authorship attribution with author-aware topic models. In: Proc. of ACL, pp. 264–269 (2012)
25. Seroussi, Y., Zukerman, I., Bohnert, F.: Collaborative inference of sentiments from texts. In: Proc. of the 18th UMAP, pp. 195–206 (2010)
26. Solorio, T., Pillay, S., Raghavan, S., y Gomez, M.M.: Modality specific meta features for authorship attribution in web forum posts. In: Proc. of the 5th IJCNLP, pp. 156–164 (2011)
27. Stamatatos, E., Kokkinakis, G., Fakotakis, N.: Automatic text categorization in terms of genre and author. Comput. Linguist. 26, 471–495 (2000)
28. Uzuner, Ö., Katz, B.: A comparative study of language models for book and author recognition. In: Proc. of the 2nd IJCNLP, pp. 969–980 (2005)
29. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining email content for author identification forensics. Sigmod Record 30, 55–64 (2001)
30. Yule, G.U.: The statistical study of literary vocabulary. Cambridge University Press (1944)
31. Zhao, Y., Zobel, J.: Effective and scalable authorship attribution using function words. In: Proceeding of Information Retrival Technology, pp. 174–189 (2005)
32. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. JASIST 57, 378–393 (2006)