In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010), pp. 38--42 2010.

Authorship Attribution Using Probabilistic Context-Free Grammars

Sindhu Raghavan Adriana Kovashka Raymond Mooney
Department of Computer Science
The University of Texas at Austin
1 University Station C0500
Austin, TX 78712-0233, USA
{sindhu,adriana,mooney}@cs.utexas.edu

Abstract

In this paper, we present a novel approach for authorship attribution, the task of identifying the author of a document, using probabilistic context-free grammars. Our approach involves building a probabilistic context-free grammar for each author and using this grammar as a language model for classification. We evaluate the performance of our method on a wide range of datasets to demonstrate its efficacy.

1 Introduction

Natural language processing allows us to build language models, and these models can be used to distinguish between languages. In the context of written text, such as newspaper articles or short stories, the author's style could be considered a distinct "language." Authorship attribution, also referred to as authorship identification or prediction, studies strategies for discriminating between the styles of different authors. These strategies have numerous applications, including settling disputes regarding the authorship of old and historically important documents (Mosteller and Wallace, 1984), automatic plagiarism detection, determination of document authenticity in court (Juola and Sofko, 2004), cyber crime investigation (Zheng et al., 2009), and forensics (Luyckx and Daelemans, 2008).

The general approach to authorship attribution is to extract a number of style markers from the text and use these style markers as features to train a classifier (Burrows, 1987; Binongo and Smith, 1999; Diederich et al., 2000; Holmes and Forsyth, 1995; Joachims, 1998; Mosteller and Wallace, 1984). These style markers could include the frequencies of certain characters, function words, phrases or sentences. Peng et al. (2003) build a character-level n-gram model for each author. Stamatatos et al. (1999) and Luyckx and Daelemans (2008) use a combination of word-level statistics and part-of-speech counts or n-grams. Baayen et al. (1996) demonstrate that the use of syntactic features from parse trees can improve the accuracy of authorship attribution. While there have been several approaches proposed for authorship attribution, it is not clear if the performance of one is better than the other. Further, it is difficult to compare the performance of these algorithms because they were primarily evaluated on different datasets. For more information on the current state of the art for authorship attribution, we refer the reader to a detailed survey by Stamatatos (2009).

We further investigate the use of syntactic information by building complete models of each author's syntax to distinguish between authors. Our approach involves building a probabilistic context-free grammar (PCFG) for each author and using this grammar as a language model for classification. Experiments on a variety of corpora including poetry and newspaper articles on a number of topics demonstrate that our PCFG approach performs fairly well, but it only outperforms a bigram language model on a couple of datasets (e.g. poetry). However, combining our approach with other methods results in an ensemble that performs the best on most datasets.

2 Authorship Attribution using PCFG

We now describe our approach to authorship attribution. Given a training set of documents from different authors, we build a PCFG for each author based on the documents they have written. Given a test document, we parse it using each author's grammar and assign it to the author whose PCFG produced the highest likelihood for the document.

In order to build a PCFG, a standard statistical parser takes a corpus of parse trees of sentences as training input. Since we do not have access to authors' documents annotated with parse trees, we use a statistical parser trained on a generic

corpus like the Wall Street Journal (WSJ) or Brown corpus from the Penn Treebank (http://www.cis.upenn.edu/~treebank/) to automatically annotate (i.e. treebank) the training documents for each author. In our experiments, we used the Stanford Parser (Klein and Manning, 2003b; Klein and Manning, 2003a) and the OpenNLP sentence segmenter (http://opennlp.sourceforge.net/). Our approach is summarized below:

Input – A training set of documents labeled with author names and a test set of documents with unknown authors.

1. Train a statistical parser on a generic corpus like the WSJ or Brown corpus.
2. Treebank each training document using the parser trained in Step 1.
3. Train a PCFG G_i for each author A_i using the treebanked documents for that author.
4. For each test document, compute its likelihood for each grammar G_i by multiplying the probability of the top PCFG parse for each sentence.
5. For each test document, find the author A_i whose grammar G_i results in the highest likelihood score.

Output – A label (author name) for each document in the test set.

3 Experimental Comparison

This section describes experiments evaluating our approach on several real-world datasets.

3.1 Data

We collected a variety of documents with known authors including news articles on a wide range of topics and literary works like poetry. We downloaded all texts from the Internet and manually removed extraneous information as well as titles, author names, and chapter headings. We collected several news articles from the New York Times online journal (http://global.nytimes.com/) on topics related to business, travel, and football. We also collected news articles on cricket from the ESPN cricinfo website (http://www.cricinfo.com). In addition, we collected poems from the Project Gutenberg website (http://www.gutenberg.org/wiki/Main_Page). We attempted to collect sets of documents on a shared topic written by multiple authors. This was done to ensure that the datasets truly tested authorship attribution as opposed to topic identification. However, since it is very difficult to find authors that write literary works on the same topic, the Poetry dataset exhibits higher topic variability than our news datasets. We had 5 different datasets in total – Football, Business, Travel, Cricket, and Poetry. The number of authors in our datasets ranged from 3 to 6.

For each dataset, we split the documents into training and test sets. Previous studies (Stamatatos et al., 1999) have observed that having unequal number of words per author in the training set leads to poor performance for the authors with fewer words. Therefore, we ensured that, in the training set, the total number of words per author was roughly the same. We would like to note that we could have also selected the training set such that the total number of sentences per author was roughly the same. However, since we would like to compare the performance of the PCFG-based approach with a bag-of-words baseline, we decided to normalize the training set based on the number of words, rather than sentences. For testing, we used 15 documents per author for datasets with news articles and 5 or 10 documents per author for the Poetry dataset. More details about the datasets can be found in Table 1.

Dataset	# authors	# words/auth	# docs/auth	# sent/auth
Football	3	14374.67	17.3	786.3
Business	6	11215.5	14.16	543.6
Travel	4	23765.75	28	1086
Cricket	4	23357.25	24.5	1189.5
Poetry	6	7261.83	24.16	329

Table 1: Statistics for the training datasets used in our experiments. The numbers in columns 3, 4 and 5 are averages.

3.2 Methodology

We evaluated our approach to authorship prediction on the five datasets described above. For news articles, we used the first 10 sections of the WSJ corpus, which consists of annotated news articles on finance, to build the initial statistical parser in

Step 1. For Poetry, we used 7 sections of the Brown corpus which consists of annotated documents from different areas of literature.

In the basic approach, we trained a PCFG model for each author based solely on the documents written by that author. However, since the number of documents per author is relatively low, this leads to very sparse training data. Therefore, we also augmented the training data by adding one, two or three sections of the WSJ or Brown corpus to each training set, and up-sampling (replicating) the data from the original author. We refer to this model as "PCFG-I", where I stands for *interpolation* since this effectively exploits linear interpolation with the base corpus to smooth parameters. Based on our preliminary experiments, we replicated the original data three or four times.

We compared the performance of our approach to bag-of-words classification and n-gram language models. When using bag-of-words, one generally removes commonly occurring "stop words." However, for the task of authorship prediction, we hypothesized that the frequency of specific stop words could provide useful information about the author's writing style. Preliminary experiments verified that eliminating stop words degraded performance; therefore, we did not remove them. We used the Maximum Entropy (MaxEnt) and Naive Bayes classifiers in the MALLET software package (McCallum, 2002) as initial baselines. We surmised that a discriminative classifier like MaxEnt might perform better than a generative classifier like Naive Bayes. However, when sufficient training data is not available, generative models are known to perform better than discriminative models (Ng and Jordan, 2001). Hence, we chose to compare our method to both Naive Bayes and MaxEnt.

We also compared the performance of the PCFG approach against n-gram language models. Specifically, we tried unigram, bigram and trigram models. We used the same background corpus mixing method used for the PCFG-I model to effectively smooth the n-gram models. Since a generative model like Naive Bayes that uses n-gram frequencies is equivalent to an n-gram language model, we also used the Naive Bayes classifier in MALLET to implement the n-gram models. Note that a Naive-Bayes bag-of-words model is equivalent to a unigram language model.

While the PCFG model captures the author's writing style at the syntactic level, it may not accurately capture lexical information. Since both syntactic and lexical information is presumably useful in capturing the author's overall writing style, we also developed an ensemble using a PCFG model, the bag-of-words MaxEnt classifier, and an n-gram language model. We linearly combined the confidence scores assigned by each model to each author, and used the combined score for the final classification. We refer to this model as "PCFG-E", where E stands for *ensemble*. We also developed another ensemble based on MaxEnt and n-gram language models to demonstrate the contribution of the PCFG model to the overall performance of PCFG-E. For each dataset, we report *accuracy*, the fraction of the test documents whose authors were correctly identified.

3.3 Results and Discussion

Table 2 shows the accuracy of authorship prediction on different datasets. For the n-gram models, we only report the results for the bigram model with smoothing (Bigram-I) as it was the best performing model for most datasets (except for Cricket and Poetry). For the Cricket dataset, the trigram-I model was the best performing n-gram model with an accuracy of 98.34%. Generally, a higher order n-gram model (n = 3 or higher) performs poorly as it requires a fair amount of smoothing due to the exponential increase in all possible n-gram combinations. Hence, the superior performance of the trigram-I model on the Cricket dataset was a surprising result. For the Poetry dataset, the unigram-I model performed best among the smoothed n-gram models at 81.8% accuracy. This is unsurprising because as mentioned above, topic information is strongest in the Poetry dataset, and it is captured well in the unigram model. For bag-of-words methods, we find that the generatively trained Naive Bayes model (unigram language model) performs better than or equal to the discriminatively trained MaxEnt model on most datasets (except for Business). This result is not suprising since our datasets are limited in size, and generative models tend to perform better than discriminative methods when there is very little training data available. Amongst the different baseline models (MaxEnt, Naive Bayes, Bigram-I), we find Bigram-I to be the best performing model (except for Cricket and Poetry). For both Cricket and Poetry, Naive Bayes

Dataset	MaxEnt	Naive Bayes	Bigram-I	PCFG	PCFG-I	PCFG-E	MaxEnt+Bigram-I
Football	84.45	86.67	86.67	**93.34**	80	91.11	86.67
Business	83.34	77.78	90.00	77.78	85.56	91.11	**92.22**
Travel	83.34	83.34	**91.67**	81.67	86.67	**91.67**	90.00
Cricket	91.67	**95.00**	91.67	86.67	91.67	**95.00**	93.34
Poetry	56.36	78.18	70.90	78.18	83.63	**87.27**	76.36

Table 2: Accuracy in % for authorship prediction on different datasets. Bigram-I refers to the bigram language model with smoothing. PCFG-E refers to the ensemble based on MaxEnt, Bigram-I, and PCFG-I. MaxEnt+Bigram-I refers to the ensemble based on MaxEnt and Bigram-I.

is the best performing baseline model. While discussing the performance of the PCFG model and its variants, we consider the best performing baseline model.

We observe that the basic PCFG model and the PCFG-I model do not usually outperform the best baseline method (except for Football and Poetry, as discussed below). For Football, the basic PCFG model outperforms the best baseline, while for Poetry, the PCFG-I model outperforms the best baseline. Further, the performance of the basic PCFG model is inferior to that of PCFG-I for most datasets, likely due to the insufficient training data used in the basic model. Ideally one would use more training documents, but in many domains it is impossible to obtain a large corpus of documents written by a single author. For example, as Luyckx and Daelemans (2008) argue, in forensics one would like to identify the authorship of documents based on a limited number of documents written by the author. Hence, we investigated smoothing techniques to improve the performance of the basic PCFG model. We found that the interpolation approach resulted in a substantial improvement in the performance of the PCFG model for all but the Football dataset (discussed below). However, for some datasets, even this improvement was not sufficient to outperform the best baseline.

The results for PCFG and PCFG-I demonstrate that syntactic information alone is generally a bit less accurate than using n-grams. In order to utilize *both* syntactic and lexical information, we developed PCFG-E as described above. We combined the best n-gram model (Bigram-I) and PCFG model (PCFG-I) with MaxEnt to build PCFG-E. For the Travel dataset, we find that the performance of the PCFG-E model is equal to that of the best constituent model (Bigram-I). For the remaining datasets, the performance of PCFG-E is better than the best constituent model. Furthermore, for the Football, Cricket and Poetry datasets this improvement is quite substantial. We now find that the performance of some variant of PCFG is always better than or equal to that of the best baseline. While the basic PCFG model outperforms the baseline for the Football dataset, PCFG-E outperforms the best baseline for the Poetry and Business datasets. For the Cricket and Travel datasets, the performance of the PCFG-E model equals that of the best baseline. In order to assess the statistical significance of any performance difference between the best PCFG model and the best baseline, we performed the McNemar's test, a non-parametric test for binomial variables (Rosner, 2005). We found that the difference in the performance of the two methods was not statistically significant at .05 significance level for any of the datasets, probably due to the small number of test samples.

The performance of PCFG and PCFG-I is particularly impressive on the Football and Poetry datasets. For the Football dataset, the basic PCFG model is the best performing PCFG model and it performs much better than other methods. It is surprising that smoothing using PCFG-I actually results in a drop in performance on this dataset. We hypothesize that the authors in the Football dataset may have very different syntactic writing styles that are effectively captured by the basic PCFG model. Smoothing the data apparently weakens this signal, hence causing a drop in performance. For Poetry, PCFG-I achieves much higher accuracy than the baselines. This is impressive given the much looser syntactic structure of poetry compared to news articles, and it indicates the value of syntactic information for distinguishing between literary authors.

Finally, we consider the specific contribution of the PCFG-I model towards the performance of

the PCFG-E ensemble. Based on comparing the results for PCFG-E and MaxEnt+Bigram-I, we find that there is a drop in performance for most datasets when removing PCFG-I from the ensemble. This drop is quite substantial for the Football and Poetry datasets. This indicates that PCFG-I is contributing substantially to the performance of PCFG-E. Thus, it further illustrates the importance of broader syntactic information for the task of authorship attribution.

4 Future Work and Conclusions

In this paper, we have presented our ongoing work on authorship attribution, describing a novel approach that uses probabilistic context-free grammars. We have demonstrated that both syntactic and lexical information are useful in effectively capturing authors' overall writing style. To this end, we have developed an ensemble approach that performs better than the baseline models on several datasets. An interesting extension of our current approach is to consider discriminative training of PCFGs for each author. Finally, we would like to compare the performance of our method to other state-of-the-art approaches to authorship prediction.

Acknowledgments

Experiments were run on the Mastodon Cluster, provided by NSF Grant EIA-0303609.

References

H. Baayen, H. van Halteren, and F. Tweedie. 1996. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132, September.

Binongo and Smith. 1999. A Study of Oscar Wilde's Writings. *Journal of Applied Statistics*, 26:781.

J Burrows. 1987. Word-patterns and Story-shapes: The Statistical Analysis of Narrative Style.

Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2000. Authorship Attribution with Support Vector Machines. *Applied Intelligence*, 19:2003.

D. I. Holmes and R. S. Forsyth. 1995. The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, 10:111–127.

Thorsten Joachims. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML)*, pages 137–142, Berlin, Heidelberg. Springer-Verlag.

Patrick Juola and John Sofko. 2004. Proving and Improving Authorship Attribution Technologies. In *Proceedings of Canadian Symposium for Text Analysis (CaSTA)*.

Dan Klein and Christopher D. Manning. 2003a. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*, pages 423–430, Morristown, NJ, USA. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003b. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15 (NIPS)*, pages 3–10. MIT Press.

Kim Luyckx and Walter Daelemans. 2008. Authorship Attribution and Verification with Many Authors and Limited Data. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 513–520, August.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Frederick Mosteller and David L. Wallace. 1984. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer-Verlag.

Andrew Y. Ng and Michael I. Jordan. 2001. On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems 14 (NIPS)*, pages 841–848.

Fuchun Peng, Dale Schuurmans, Viado Keselj, and Shaojun Wang. 2003. Language Independent Authorship Attribution using Character Level Language Models. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Bernard Rosner. 2005. *Fundamentals of Biostatistics*. Duxbury Press.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 1999. Automatic Authorship Attribution. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 158–164, Morristown, NJ, USA. Association for Computational Linguistics.

E. Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Rong Zheng, Yi Qin, Zan Huang, and Hsinchun Chen. 2009. Authorship Analysis in Cybercrime Investigation. *Lecture Notes in Computer Science*, 2665/2009:959.