Authorship Attribution in Law Enforcement Scenarios[1]

Moshe KOPPEL[2], Jonathan SCHLER and Eran MESSERI
Department of Computer Science, Bar-Ilan University, Israel

Abstract. Typical authorship attribution methods are based on the assumption that we have a small closed set of candidate authors. In law enforcement scenarios, this assumption is often violated. There might be no closed set of suspects at all or there might be a closed set containing thousands of suspects. We show how even under such circumstances, we can make useful claims about authorship.

Keywords. Authorship attribution, text categorization, machine learning, law enforcement.

Introduction

We are going to discuss authorship attribution in law enforcement scenarios. The problem is quite straightforward. You get an anonymous text, presumably from some assailant, if it's a law enforcement situation, and you want to know as much as you possibly can about the writer of the text. Ideally, what you'd like to know is whether a particular suspect wrote the text. That is sometimes hard to do. If we can't do that, we'd like to at least be able to say something about the guy. Or it might not be a guy. So we'd like to know: is it a man or a woman? How old is the person? What is his or her native language? Can we say anything about their personality? The question is how much can we know, just given a little piece of text (no handwriting, it's electronic text). The answer is, we can know a lot more than you might think.

Let's first consider the vanilla authorship attribution problem, the kind of problem definition that you give if you are a researcher in Computer Science, who doesn't care much about the real world but wants to have a well-defined problem that submits well to the kind of tools that you've got. In the vanilla problem you've got a small closed set of candidate authors for each of whom you've got lots of texts. And you want to be able to take some new anonymous text and say which one of your handful of authors – in the ideal situation, two authors – it is. Was *Edward, the Third* written by Shakespeare or Marlowe? That's the perfect authorship attribution problem for a researcher.

But the vanilla problem is not especially hard nor does it often occur in real life. In law enforcement you will typically have no suspects. You just have a text and you have

[1] This paper is a lightly edited transcript of a talk given by Moshe Koppel at the June 2007 NATO meeting in Beer Sheva on security informatics and terrorism. Thanks to Bracha Shapira and Paul Kantor for organizing the meeting, to our collaborator, Shlomo Argamon, and to Cecilia Gal for preparing the transcript of the talk.

[2] Corresponding Author: Moshe Koppel, Dept. of Computer Science, Bar-Ilan University, Ramat Gan, Israel; E-mail: moishk@gmail.com.

no clue who wrote it. In that case the question is, can we profile the author? What can we say about the person who wrote this text? Alternatively, you might have thousands of suspects, and then the question is, can we find the needle in the haystack? Ideally, you want to say, of the tens of thousands of people who might have written this, exactly which one it is. Incredibly enough you can do this a lot of the time, as we'll see.

And the key thing is that the texts might be very short. Unless the assailant is the Unabomber, he doesn't send a 50,000 word tract for analysis. He's more likely to send some short little note that says, “give me the money or we'll shoot you”. So, the question is, given a short note, how much can we say. (Well, that example was a little bit *too* short, but we will see that even a couple of hundred words is quite useful.)

1. Solving the Vanilla Attribution Problem

Let's first discuss how we solve the vanilla problem with a small number of authors. The general picture is shown in Figure 1. Suppose you've got texts by A and texts by B. First, you clean them up, removing whatever junk is totally inappropriate. Then you translate them into numerical vectors that capture measures (say, frequency) of features that you think might be relevant to this problem. Now that you have two sets of vectors, some of type A and some of type B, you use your favorite learning algorithm to build a classifier that distinguishes A vectors from B vectors. Once you've got your classifier, you put new texts into it for attribution. That's pretty much how the game works.

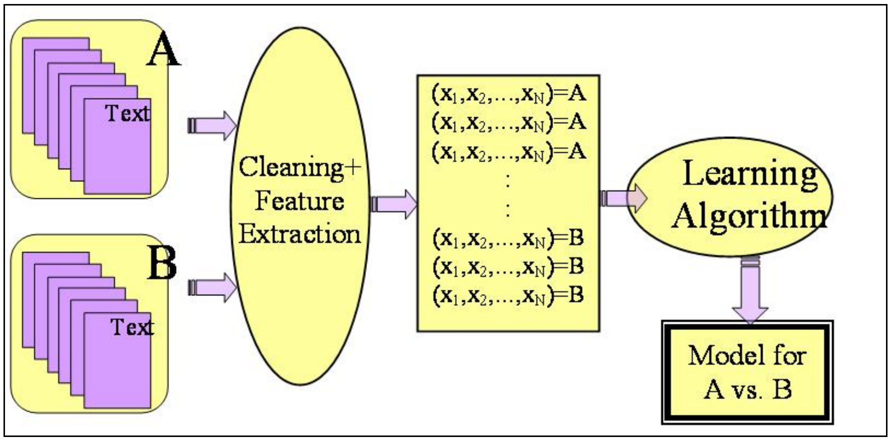

Figure 1. The categorization process

The most important question is what kind of features we should use. One of the dirty secrets of machine learning is that, although researchers generally spend more time talking about better learning algorithms than about what features to use, in real life, it doesn't really matter all that much which learning algorithm you use. The good ones all give pretty much the same results. What really matters is what feature sets you use; if you do some good feature engineering, you can often improve the results quite a bit.

What kind of features do you want for authorship attribution? Ideally, you want the kind of features that stay constant within a given author but vary between authors. And what kind of features might those be? Well, they are not topic words, because a given author is going to use different topic words depending on whether he's writing about a given topic or not. So, it's got to be stylistic information. The main stylistic features are the following:

- *Function words*: The ancestor of all stylistic features for authorship analysis are the function words, those little words like "and" and "it" and "of" and "if" and "the" that don't really tie in very strongly to content. Those were used by Mosteller and Wallace back in 1964 in their work on the Federalist Papers [1], which is the seminal work in this area.
- *Syntax*: Different authors use syntax differently, so we might try to pick up on syntactic habits of individual authors. Parsing is slow and unreliable. A better approach is to consider the frequency with which authors use particular sequences of parts-of-speech (POS); that information indirectly gives you hints to syntax.
- *Systemic Functional Linguistics (SFL) trees*: A more general approach that gives us elements of both of the above feature types is to use systemic functional linguistics features. Essentially, those are glorified parts-of-speech but, as can be seen in Figure 2, instead of very general parts of speech, like conjunctions, we'll talk about specific kinds of conjunctions, and then get even more fine-grained than that, until finally at the bottom of this tree we have actual function words. This subsumes both function words, which are down at the leaves, and parts-of-speech which are up at the roots.

Conjunctions	
ConjExtension	and, or, but, yet, however,…
ConjElaboration	for_example, indeed,…
ConjEnhancement	
ConjSpatiotemporal	then, beforehand, afterwards, while, during…
ConjCausalConditional	if, because, since, hence, thus, therefore,…

Figure 2. SFL tree sample

- *Morphology*: The frequency of use of various grammatical suffixes and prefixes can sometimes be useful clues for authorship. In English, these tend not to be very useful because there are just not that many of them, but in languages like Hebrew or Arabic, morphology is crucial. In such languages, many of the function words that we use in English don't exist as separate words and only show up in the morphology.

- *Complexity measures*: Historically, the first features explored as possible markers of authorial style were complexity measures such as average word length, average sentence length, all kinds of entropy measures, type/token ratio, hapax legomena and so on [2].
- *Idiosyncrasies*: Researchers like Donald Foster, who identified the anonymous author of the roman-a-clef *Primary Colors* as Joe Klein, rely mostly on authorial idiosyncrasies, including neologisms, exotic syntax and word construction and so forth.

Once we have built vectors based on the frequency with which features of these sorts are used, we use some learning algorithm to distinguish between authors. We have run many vanilla authorship experiments using a variety of learning algorithms including linear SVM [3], which is almost a *de facto* standard, Bayesian Regression (using software kindly made available by the Rutgers group [4]) and real-valued balanced Winnow [5,6], a kind of exponential gradient algorithm. And they all work. If you need to decide between two candidate authors and you've got a reasonable amount of known text for each author, I can pretty much guarantee you can attribute a not-too-short anonymous text with accuracy well above 90%. The amount of text I'm discussing is not especially large: maybe a few tens of documents of 500 words and up.

So, the vanilla attribution problem is definitely solvable and, in fact, function words and single parts of speech are generally enough. Systemic functional linguistics trees by themselves are enough, because they subsume the previous two. Idiosyncrasies are the best thing to use if you've got unedited text (like email); obviously, if you're dealing with edited documents, this is useless. And morphology is useful for particular languages, such as Hebrew, which is rich in morphology.

2. Profiling

All the above was just to provide background. What about real life? In real life, you might have a text written by some anonymous assailant, but without any specific suspects at all. In such cases you'd be satisfied to extract some general information about the gender, age, native language and personality of the author.

So let's consider the problems of gender, which is binary (this point is apparently subject to debate but not for now) and age, which we divide into three categories, teenagers, people in their twenties, and people in their thirties or above. As luck would have it, for the purpose of running systematic experiments, we needed people to write electronic texts about anything they wanted and to also tell us their gender and their age, and we got about 100 million volunteers. They call themselves bloggers. We took tens of thousands of blogs labeled for gender and age and randomly threw out enough of them so that we had the same number of male and female writers in each age group. (You may be interested to know that, as of several years ago, a large majority of bloggers below the age of 18 were females, while a large majority of those above 22 were males. The numbers may have changed since then.).

We also ran experiments on native speakers of five different languages (Russian, Czech, Bulgarian, Spanish and French) writing in English [10]; our objective was to determine an author's native language. For this we used the International Corpus of

Learner English [7] and selected 258 essays from speakers of each of the five languages.

For each of the experiments, we used two feature sets, SFL trees for capturing style and frequent non-function words for content [11,12]. We used Bayesian regression as our learning algorithm and ran ten-fold cross-validation experiments to estimate accuracy of attribution on out-of-sample documents. The results for each experiment are shown in Figure 3.

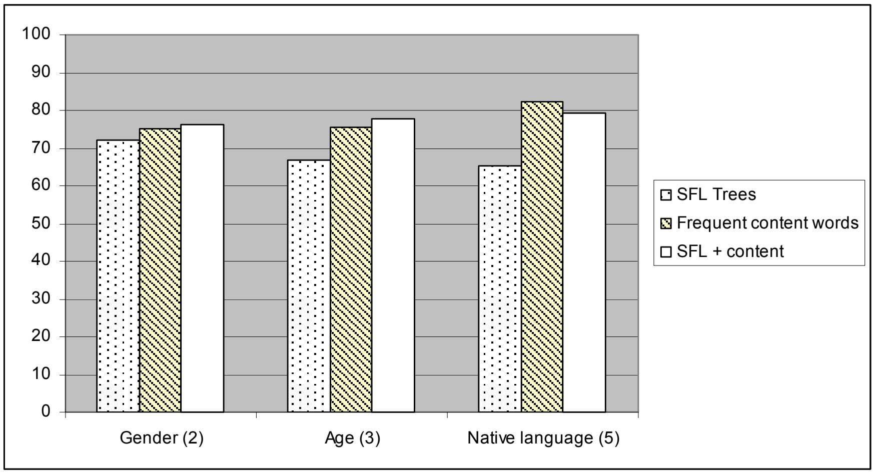

Figure 3. Experimental results

As can be seen, for the gender problem, which has a natural baseline of 50%, we obtain accuracy of 76% using both feature sets. In fact, when we also use blog jargon as features – "LOL" and "ROTFL" and all that – we get accuracy over 80%. The features that are used most differently by males and females (as measured using information gain [8]) are shown in Table 1. Note that among style features, personal pronouns are used more by females, particularly the words "I, me, him" and "my", while males make more frequent use of the determiners "a" and "the" and certain kinds of prepositions. In fact, there are hundreds of features that are used very differently by males and females. This is true in a variety of genres: blogs, fiction, and non-fiction, including academic articles in professional journals. The numbers vary but the differences between males and females are consistent in all of them. In all those genres, females use more pronouns and males use more determiners. (As for content features that are used most differently by male and female bloggers, there must be a message in there somewhere but it's not likely to help us in typical law enforcement scenarios.)

Table 1. Most distinguishing features (information-gain) for gender

Class	SFL Features	Content Features
Female	**personal pronoun**, I, me, him, my	cute, love, boyfriend, mom, feel
Male	**determiner**, the, of, **preposition-matter**, as	system, software, game, based, site

As can be seen, for the age problem, which has a natural baseline of 42.7% (the size of the largest of the three classes), we obtain accuracy of 77.7% using both feature sets. Unsurprisingly, when we also use blog jargon as features, we get accuracy over 80%. The content features that are used most differently by bloggers of different ages are shown in Table 2. It is amusing to imagine this as representing the changing interests through the lifespan of the bloggers in our sample. As teens, they are concerned with matters that are either "boring" or "awesome", in their 20's they are mostly concerned with "college", "bar", "apartment", and "dating", and eventually are preoccupied with running the whole world, which is apparently neither boring nor awesome.

As can be seen, for native language, which has a natural baseline of 20%, we get 65% accuracy using SFL features alone. (The results using content are uninteresting for this experiment since differences between the groups are almost certainly artifacts of the experimental setup.) Interestingly, using only SFL features and a variety of idiosyncrasy-based features, we get accuracy above 80%. The best features for distinguishing native speakers of each language are shown in Figure 4.

- Russian –*over*, *the* (infrequent), number_reladverb
- French – *indeed, Mr* (no period), misused *o (*e.g. *outhor)*
- Spanish – *c-q* confusion (e.g., *cuality*), *m-n* confusion (e.g., *confortable*), undoubled consonant (e.g., *comit*)
- Bulgarian – *most*_ADVERB, *cannot* (uncontracted)
- Czech – doubled consonant (e.g. *remmit*)

Figure 4. Best features for distinguishing native speakers

Table 2. Most distinguishing features (information-gain) for different ages. Numbers are uses per 1000 words.

feature	10s	20s	30s
bored	**3.84**	1.11	0.47
boring	**3.69**	1.02	0.63
awesome	**2.92**	1.28	0.57
mad	**2.16**	0.8	0.53
homework	**1.37**	0.18	0.15
mum	**1.25**	0.41	0.23
maths	**1.05**	0.03	0.02
dumb	**0.89**	0.45	0.22
sis	**0.74**	0.26	0.10
crappy	**0.46**	0.28	0.11
college	1.51	**1.92**	1.31
bar	0.45	**1.53**	1.11
apartment	0.18	**1.23**	0.55
beer	0.32	**1.15**	0.70
student	0.65	**0.98**	0.61
drunk	0.77	**0.88**	0.41
album	0.64	**0.84**	0.56
dating	0.31	**0.52**	0.37
semester	0.22	**0.44**	0.18
someday	0.35	**0.40**	0.28
son	0.51	0.92	**2.37**
local	0.38	1.18	**1.85**
marriage	0.27	0.83	**1.41**
development	0.16	0.50	**0.82**
tax	0.14	0.38	**0.72**
campaign	0.14	0.38	**0.70**
provide	0.15	0.54	**0.69**
democratic	0.13	0.29	**0.59**
systems	0.12	0.36	**0.55**
workers	0.10	0.35	**0.46**

3. Finding a Needle in a Haystack

Suppose we've got 10,000 authors and someone gives us one text and we've got to say who wrote it. We began with 10,000 blogs and removed from each one the last post or, if it was too short, enough posts to add up to 500 words [9]. Let's call the removed texts "snippets". Now, given a random snippet, we need to decide which of these 10,000 bloggers wrote it. (We don't have any hints in terms of formatting; we only have the actual text, without even distinguishing quoted text from integral text.)

We begin with a naive information retrieval approach. Let's just assign the snippet to whichever blog looks most similar, using standard information retrieval measures of similarity: cosine of *tf.idf* representations. This method does not work very well. Using three different versions of the *tf.idf* representation (style features only, content features with raw term frequency, content features with binary term appearance), we find that the best of them only assigns the snippet to the actual author 36% of the time.

But, here is what we can do. In a typical law enforcement scenario, we can decide we are not allowed to be wrong, but we are allowed to say, "I don't know". We can say that a given snippet just doesn't have enough information in it for us to say anything. But if we say "we know", we had better get the right answer, generally speaking. So we use a meta-learning technique: we consider how strong the similarity is to the top ranked author and how far back the second ranked author is, using each of our similarity measures. Without belaboring the details of the learning techniques and how they are applied, it's enough to say that if there is one stick-out author who is much likelier than the others to be the actual author, we gamble on that author. Otherwise, we just throw up our hands.

Clearly, the more risk-averse we are, the lower the recall we achieve but the higher the precision. The full recall-precision curve is shown in Figure 5 (upper curve). Note that, for example, we can achieve recall of 40% with precision of 84%. But if we can settle for recall of 30%, we can get precision of 90%. To make this more concrete, if we have 10,000 candidate authors and 10,000 snippets to attribute, and we venture a guess for 4,762 of these snippets, we'll be right for 4,000 of them.

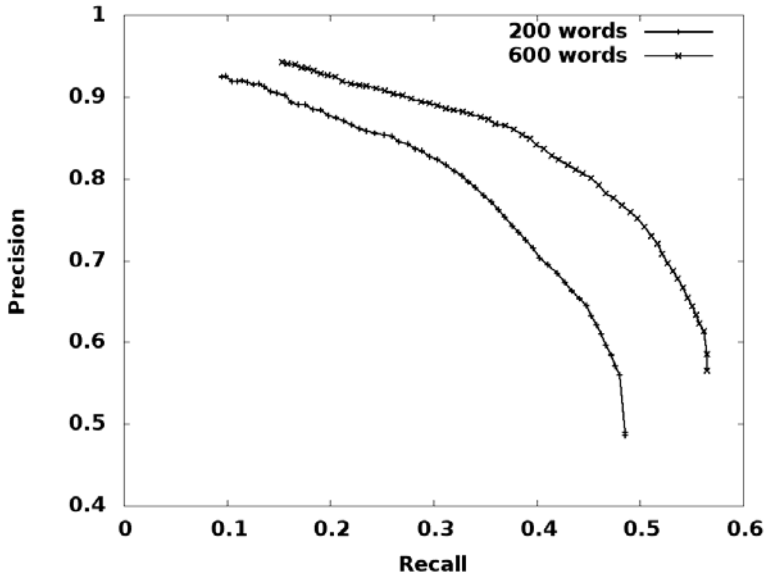

Figure 5. Recall/precision curves in attributing a snippet to one of 10,000 authors. Curves are for snippets of length 600 (upper) and 200 (lower).

Now you might wonder how much text we really need. For these experiments, we used snippets of length between 500 and 600 words. We ran the identical experiment but with the snippets limited to exactly 200 words. As can be seen in Figure 5 (lower curve), at a recall level of 30%, we achieve precision of 82%. So, the results do degrade for very short texts, but they are still quite useful even at very realistic document lengths.

To conclude then, we can use these techniques in order to profile an anonymous author. We can tell you with some reasonable degree of accuracy the author's age, gender, and native language. (We didn't discuss personality, but in fact we can tell if a writer is neurotic or not with the same accuracy as the degree of psychologist agreement on neurosis.) And even with 10,000 candidates, in a fair number of cases, we can confidently and correctly identify the author.

References

[1] F. Mosteller, D. L. Wallace. Inference and Disputed Authorship: The Federalist. Reading, Mass. Addison Wesley. 1964.

[2] G. U. Yule, On Sentence Length as a Statistical Characteristic of Style in Prose with Application to Two Cases of Disputed Authorship, *Biometrika,* **30**, (1938) 363-390.

[3] T. Joachims, Text categorization with support vector machines: learning with many relevant features. *In Proc. 10th European Conference on Machine Learning ECML-98*, (1998) 137-142.

[4] A. Genkin, D. D. Lewis and D. Madigan. Large-scale Bayesian logistic regression for text categorization. Technometrics (to appear). 2006.

[5] I. Dagan, Y. Karov, D. Roth, Mistake-driven learning in text categorization, *In Proc. EMNLP-97: 2nd Conf. on Empirical Methods in Natural Language Processing*, (1997) 55-63.

[6] N. Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, *Machine Learning,* **2**, 4, (1987) 285-318.

[7] S. Granger, E. Dagneaux, F. Meunier, The International Corpus of Learner English. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain, 2002.

[8] T. Mitchell,. Machine Learning. McGraw-Hill , New York, 1999.

[9] M. Koppel, J. Schler, S. Argamon and E. Messeri. Authorship Attribution with Thousands of Candidate Authors, in *Proc. Of 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval*, August 2006.

[10] M. Koppel, J. Schler and K. Zigdon, Determining an Author's Native Language by Mining a Text for Errors, in *Proceedings of KDD 2005*, August 2005, Chicago, IL, USA

[11] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Mining the Blogosphere: Age, gender and the varieties of self–expression , in First Monday, vol 12(9), September 2007.

[12] M. Koppel, J. Schler, S. Argamon and J. Pennebaker. Profiling the Author of an Anonymous Text , to appear in *Communications of the ACM (CACM)*.