Syntactic methods for topic-independent authorship attribution

Johanna Björklund Niklas Zechner

March 5, 2016

Abstract

The efficacy of syntactic features for topic-independent authorship attribution is evaluated, taking a feature set of frequencies of words and punctuation marks as baseline. The features are 'deep' in the sense that they are derived by parsing the subject texts, in contrast to 'shallow' syntactic features for which a part-of-speech analysis is enough. The experiments are conducted on a corpus of novels written around the year 1900 by 20 different authors, and cover two tasks. In the first task, text samples are taken from books by one author, and the goal is to pair samples from the same book. In the second task, text samples are taken from several authors, but only one sample from each book, and the goal is to pair samples from the same author. In the first task, the baseline feature set outperformed the syntax-based feature set, but for the second task, the outcome was the opposite. This suggests that, compared to lexical features such as vocabulary and punctuation, syntactic features are more robust to changes in topic.

1 Introduction

Authorship attribution consists in identifying the author of an anonymously written document, given a set of candidate authors and sample texts for each. There are also variations such as document clustering or author verification that avoids a closed-world assumption with respect to the set of authors, but these will not be considered here. Authorship attribution has a surprisingly number of applications (Stamatos, 2009): For instance, in authorship verification, to establish whether a text was written by a certain author (Koppel & Schler, 2004); in plagiarism detection, to find similarities between texts (zu Eissen, Stein, & Kulig, 2007), (Stein & zu Eissen, 2007); in author profiling, to extract information about the age, education, an gender of the author of a text (Koppel, Argamon, & Shimoni, 2002); and finally, in the detection of stylistic inconsistencies, as may easily happen in collaborative writing (Collins, Kaufer, Vlachos, Butler, & Ishizaki, 2004; Graham, Hirst, & Marthi, 2005).

A more recent use of authorship attribution the detection of so-called "troll armies". These are made up of Internet users that are paid to promote political agendas under the guise of civilians expressing private opinions. As an example, the dutch company "Subvert and Profit" declare themselves to have a payroll of 25,000 users that vote, rate, and post as instructed. Methods for authorship attribution also have applications to other forms of creative output, for example, source code and musical scores (Frantzeskou, Stamatatos, Gritzalis, & Katsikas, 2006).

A central question is the extent to which we can recognize authors as they shift between topics. Previous studies (see Section 1.2) suggest that function words are more effective than part-of-speech (POS) information to separate between authors when the topic varies (Menon & Choi, 2011). This can be seen as a case for lexical features over syntactic features. However, there is more to syntax than POS, so in this article we go one step further and consider what happens when we do a full syntactical analysis and use fragments of the resulting parse trees as our features. As a baseline, we take the most frequently used words and punctuation marks with respect to the training set, which coincides with the standard sets of function words.

To convey the general idea, consider for example the syntactic analysis shown in Figure 1 (the lines are from the poem *Theory* by Parker (1928)). In the parse tree, a pair of repeated patterns have been marked out with bold font. These particular patterns need not be idiosyncratic for the writer, but similar to the case for function words, it is likely that she tends

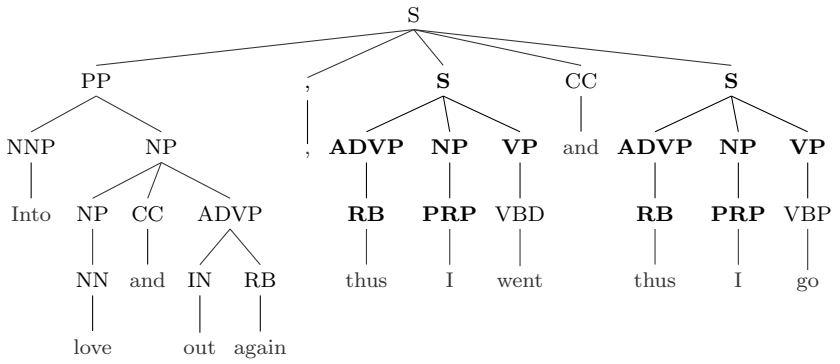

Figure 1: Repeated syntactical fragments in the parse tree of the poem *Theory* by turn-of-the century novelist Dorothy Parker (1928).

towards certain constructions rather than others, and that these show up as a statistical fingerprint that identifies her works.

1.1 Syntactic features

Before we proceed, let us review some previous work that relates to the current effort. We begin with those that concern author attribution in general, and in the next section look at those that address variations in topic. A complete exposition is not possible here, but several surveys are available, for example (Stamatatos, 2009).

One of the first to consider syntactic features in authorship attribution was Baayen, van Halteren, and Tweedie (1996). They were motivated by previous successes using function words, that is, common words like *and*, *or*, *not*, etc., which was argued to tap into the syntactic structure of a sentence. In their study, the classification task consisted in distinguishing between samples from a pair of books written by different authors, by means of principal component analysis with respect to either the 50 most common words or syntactic features. Baayen et al. (1996) argued that the syntactic approach led to a better separation of samples from different authors.

Other early actors were Stamatatos, Kokkinakis, and Fakotakis (2000) who looked at stylistic features in author and topic classification. Their input consisted of online texts, and the features, for example part-of-speech tags, keywords, and overall phrase structure, were derived through automatic analysis without human intervention. Classification was done through multiple regression and discriminant analysis. In the case of topic identification, their system achieved an accuracy of 80 per cent in the case of 10 topics. For authorship attribution, the accuracy was around 70 per cent, again for 10 authors.

At the time (Baayen et al., 1996) and (Stamatatos et al., 2000) were written, a complete syntactic analysis was not considered practical. For this reason, Luyckx and Daelemans (2005) looked at syntactic features that could be obtained through shallow parsing, for example, part-of-speech tags and verb forms, and used these to separate between two authors. The test samples were chosen to be on the same topics for both authors and in the same genre. Luyckx and Daelemans (2005) found that lexical and syntactic features lead to comparable performances, and that the combination was better than either. Another example of this approach is the work of Argamon and Shimoni (2003) who study combinations of simple lexical and syntactic features to infer the gender of the author of a literary text.

Improvement in natural-language parsing technology led to a renewed interest in deeper syntactic features. Such an application of syntactic features is exemplified in the case of authorship attribution by Gamon (2004), who improve an existing algorithm based on shallow features by information about rewrite rules and semantic relations. In (Lučić & Blake, 2015), the authors look closer at how the local syntactic dependencies that an author uses when referring to a named entity can be used for recognition.

Ayala, Pinto, Gómez-Adorno, León, and Castillo (2013) compare lexical and shallow syn-

tactic features, with a slightly deeper, graph-based, form of analysis in combination with the data-mining tool Subdue (Olmos, Gonzalez, & Osorio, 2005), and evaluate these on different test sets covering between 3 and 14 authors. On the 8 datasets considered, the former method performed better in 3 cases, and the latter in 5 cases, but the authors acknowledge that the results are inconclusive and suggest further study.

In (Tschuggnall & Specht, 2014), the authors mine so-called *pq*-grams (Augsten, Böhlen, & Gamper, 2005) from parse trees and use these as features. Here, p and q are a pair of natural numbers, and a *pq*-gram is a subtree whose shape is parametrized by p and q. The notion is often presented as a generalization of n-grams to trees. Tschuggnall and Specht (2014) evaluate their system on various data sets, for example, for a data set written by 4 authors they achieve an accuracy rate of 72 per cent. Hollingsworth (2012) evaluate various forms of lexical and syntactic features on a corpus of novels written by 3 authors and get similar results across the board, where the main difference appears to be the number of features, rather than the kind.

Raghavan, Kovashka, and Mooney (2010) train probabilistic context-free grammars (PCFG) to recognize the works of target authors. This method achieves, for example., a classification accuracy of 78 per cent for a data set with 6 authors. Feng, Banerjee, and Choi (2012) acknowledge the novelty of their work, but question how much deep syntactic features really contribute, compared to lexical productions that contain all lexical information. For this reason, they explore the usefulness of different deep syntactic features, in combination with an SVM classifier, for authorship attribution. Evaluations are made on scientific writing and classic literature. The first data set spans 10 authors, each represented by at least 8 scientific articles. The second data set spans 5 authors, this time represented by 60 documents each, created by selecting the 3,000 first words from one of the author's novels and dividing them into blocks of 50 sentences each. Feng et al. (2012) obtain a per-author accuracy exceeding 90 per cent even when disregarding lexical productions (and with them all words).

Fuller, Maguire, and Moser (2014) combine the approaches of Raghavan et al. (2010) and Feng et al. (2012), in that they work with PCFGs stripped from lexical information, so as to emulate traditional stylistic analysis. Fuller et al. argue that syntactic features are particularly interesting in the study of literature. Their method performs significantly below an SVM classifier for some of the smaller data sets tested, but as the sample size grows, the relative difference becomes smaller. In the case of a set of 10 contemporary authors of suspense novels, the accuracy is 83.2 per cent for the SVM and 79.6 per cent for the PCFGs. In the case of a set of 10 authors of classical novels, the difference disappears altogether: both classifiers gives an average per-author accuracy of 84.4 per cent.

Most studies in statistical or machine learning based authorship attribution focus on two or a few authors. This leads to an overestimation of the importance of the features extracted from the training data and found to be discriminating for these small sets of authors. Most studies also use sizes of training data that are unrealistic for situations in which stylometry is applied, and thereby overestimate the accuracy of their approach in these situations. A more realistic interpretation of the task is as an authorship verification problem, and this can be approximated by pooling data from many different authors as negative examples.

A measure of syntactic difference is developed by Wiersma, Nerbonne, and Lauttamus (2011), which makes it possible to identify syntactic differences between documents in a collection. With this method, they identify under- and overuse of specific constructs, which is shown useful for distinguishing between the English spoken by learners as compared to natives.

1.2 Topic-independent classification

Many publications on authorship attribution recognize the impact of topic and register on the results. The topic is the information that the text is meant to convey, and the register the variety of language used to express it, such as a formal or informal manner of writing. A common way to avoid the difficulties caused by changes in topic and register is to restrict experiments to a relatively homogeneous corpus. We also do that to some extent, as we only compare results within a particular register, but allow the topics to vary.

A promising branch of topic-independent authorship attribution originates from an article by Blei, Ng, and Jordan (2003) on latent Dirichlet allocation (LDA), a generative model suited for text corpora. LDA are based on hierarchical Bayesian networks and include information about low-level features such as stylistic markers, topics, and documents. The LDA was later extended by Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004) to include authorship informa-

tion, and applied by Seroussi, Bohnert, and Zukerman (2012) to authorship attribution. In doing so, they showed that by including information in the classification process about the *a priori* likelihood that an author will write on a particular topic, and when he or she does so, how this influences the style markers, the classification accuracy goes up. The difference between this approach and ours, is that we do not need a closed world assumption when it comes to topics.

Mikros and Argiri (2007) investigate how robust different types of features are to changes in topic. For this purpose, they assembled a corpus of Modern Greek newswire articles written by two authors. Using a two-way ANOVA test, they found that features such as lexical richness, variations in sentence and word lengths, character frequencies and function words correlated considerably with the topic, and must therefore be used with caution. The study does not cover syntactic features, neither shallow nor deep.

Another evaluation of different features across topics was made by (Menon & Choi, 2011). In their experiments, they use an SVM classifier and work with a corpus of classical novels written by 14 authors. The features considered include n-grams of words and parts-of-speech, mood words, and function words. The experiments are varied by the disjointness of topics and the likelihood of a certain topic. In all experiments, function words achieve the best results, though they sometimes fall short of the combination of all considered features.

2 Method

As mentioned in the introduction, we are interested in the robustness of deep syntactic features to changes in topic. The aim of our experiments is not novelty, but a better understanding of what we already have, and therefore our choices of data sets and classifiers tend towards the conservative.

2.1 Data sets

We work with data from three different corpora, which we refer to as Boards, Blogs and Novels; see Table 1 for an overview, and Figure 2 for samples. The Boards corpus[1] has been collected from an Irish web forum, and covers 10 years of discussions between 130,000 users. In our experiments, we used the data from 2007 and 2008, and removed users with fewer than 60 posts, leaving 5,450 authors. The Blogs corpus[2] contains the collected productions of 19,320 bloggers, all written in 2004. For lack of better information, we take each blogger to be a unique author. The Boards and Blogs corpora are similar in that (i) they are taken from contemporary web-based sources, (ii) they contain a substantial number of authors, and (iii) the amount of data for each individual author varies greatly. The usefulness of Boards and Blogs in this study is to benchmark the performance of word- and syntax-based classification in the absence of topic information.

The third corpus is different. It consists of 290 English novels, written between 1880 and 1920, by 25 different authors. To investigate how the accuracies of the different methods depend on the amount of training data, we want to be able to run our tests on samples of varying sizes. Since the size of the greatest test set is limited by the size of the smallest data set for a single author, we omit the five least productive authors. The amount of training data thus used in the larger experiments is probably unrealistic for applications such as forensics, but lets us understand how classification performance relates to data size. A practically motivated study on authorship attribution with limited data found in (Luyckx & Daelemans, 2008).

In order to simulate the effect of varying topics, we treat each novel as a separate topic. This is not ideal – some novels may of course be on the same topic, and they are in the same medium and style – but large corpora marked for topic as well as author are difficult to come by, particularly since topic is much more vague than author, so we consider this a reasonable place to start. To our advantage, separate novels are at least much more well-defined than topics.

[1] Collected from `boards.ie`, available at `http://data.sioc-project.org/`

[2] Collected from `blogger.com`, available at `http://u.cs.biu.ac.il/~koppel/`

Doesn't that club have a reputation for being an absolute shambles of a setup?

Name and shame imo. Oh, and - so is your face. So there.

You are allowed to lie about your hand. But you cannot declare your hand while there is still action to take place. Simple really.

i didnt say it was a rule i thought was good.

huhwah? I think it's a silly rule. Read my last post again!

Hold on everyone before you get your knickers in a twist. Irrespective of the rights and wrongs of this ruling (and I agree the rule is ridiculous) the fact remains that it is in the rulebook for GJP events (and in other clubs and events around the country).

(a) Boards

Every day should be a half day. Took the afternoon off to hit the dentist, and while I was out I managed to get my oil changed, too. Remember that business with my car dealership this winter? Well, consider this the epilogue. The friendly fellas at the Valvoline Instant Oil Change on Snelling were nice enough to notice that my dipstick was broken, and the metal piece was too far down in its little dipstick tube to pull out. Looks like I'm going to need a magnet. Damn you, Kline Nissan, daaaaaaammmnnn yooouuuu....

Today I let my boss know that I've submitted my Corps application. The news has been greeted by everyone in the company with a level of enthusiasm that really floors me.

(b) Blogs

When Mary Lennox was sent to Misselthwaite Manor to live with her uncle everybody said she was the most disagreeable-looking child ever seen. It was true, too. She had a little thin face and a little thin body, thin light hair and a sour expression. Her hair was yellow, and her face was yellow because she had been born in India and had always been ill in one way or another. Her father had held a position under the English Government and had always been busy and ill himself, and her mother had been a great beauty who cared only to go to parties and amuse herself with gay people.

(b) Novels

Figure 2: Sample extracts from each of the corpora Boards, Blogs, and Novels

Table 1: Data characteristics for the Boards, Blogs, and Novels corpora.

	Boards	Blogs	Novels
Authors	5,450	19,320	20
Mean words per author	22,372	8,719	106,117
Written	1998–2008	2004	1880–1920

Table 2: The ten most common words and punctuation marks for each corpus.

	Boards	Blogs	Novels
1	.	.	,
2	the	,	the
3	,	i	.
4	to	the	”
5	a	to	and
6	i	and	of
7	and	’	to
8	of	a	-
9	in	of	a
10	it	it	i

2.2 Features

With the web-based corpora in particular, it is possible to consider various paratextual data, such as posting time and markup, but we exclude those things and extract only the text itself from each corpus. We also ignore the information that could be extracted from looking at the posts separately. It is likely that average length of posts would have been a useful feature, and one might also use algorithms that count distributions of feature values over posts rather than looking at a single feature value for the entire text, but we will use neither in this study.

In order to simulate unknown identities, we divide each candidate's texts chronologically in two parts, *a* and *b*, with equal amounts of data. There are studies in which the texts are divided by taking alternate sentences or posts, but in our experience this leads to inflated and hence unrealistic results (Zechner, 2015). The algorithm then compares each *a* part to each *b* part, calculating a similarity rating for each pair. We count for each *a* part which *b* part has the highest similarity. If it is the one which actually is by the same candidate, that is considered a successful match. The percentage of successful matches is considered the accuracy. This way, we have a method which can answer questions like *which of these authors wrote this text*, but since we also have similarity measures as an intermediate step, we can also potentially ask *how likely are these two texts to be written by the same author*.

For each half, we extract two different sets of features: words and syntax. From these we will be able to choose subsets to compare. The first feature set consists of word frequencies, starting with the most common words. As words, we include punctuation marks, but disregard capitalization.

Table 2 shows the ten most common words for each corpus. As we can see, there are many similarities, but also a few differences: In Blogs, we have a high number of *I*, as expected, since blogs often contain texts about the author, much like a diary. It is also worth noting that these blogs are from a time when the concept of blogs was quite new; perhaps today they have branched out to other subjects with fewer *I*. In Novels, we see more commas than full stops. It is expected that these more formal texts have longer sentences, and the writing style of the time period may also contribute to the greater number of commas. The lack of apostrophes can be similarly explained, and more quotation marks are also expected in a novel. In Boards, we might note that *a* is more prominent than in the others; perhaps in a setting where many people write together, more new entities are introduced, which could increase the frequency of the indefinite forms.

As a second feature set, we extract syntactic patterns. We use the well-known Stanford parser (Klein & Manning, 2003) and look at parts-of-speech for parent-child pairs in the syntax tree. E.g., in the sentence “White mice eat green apples”, we get:

1x plural noun as subject of verb; (*mice*, *eat*)

1x plural noun as object of verb; (*apples*, *eat*)

2x adjective describing plural noun; (*white*, *mice*) and (*green*, *apples*)

1x sentence root; (*eat*)

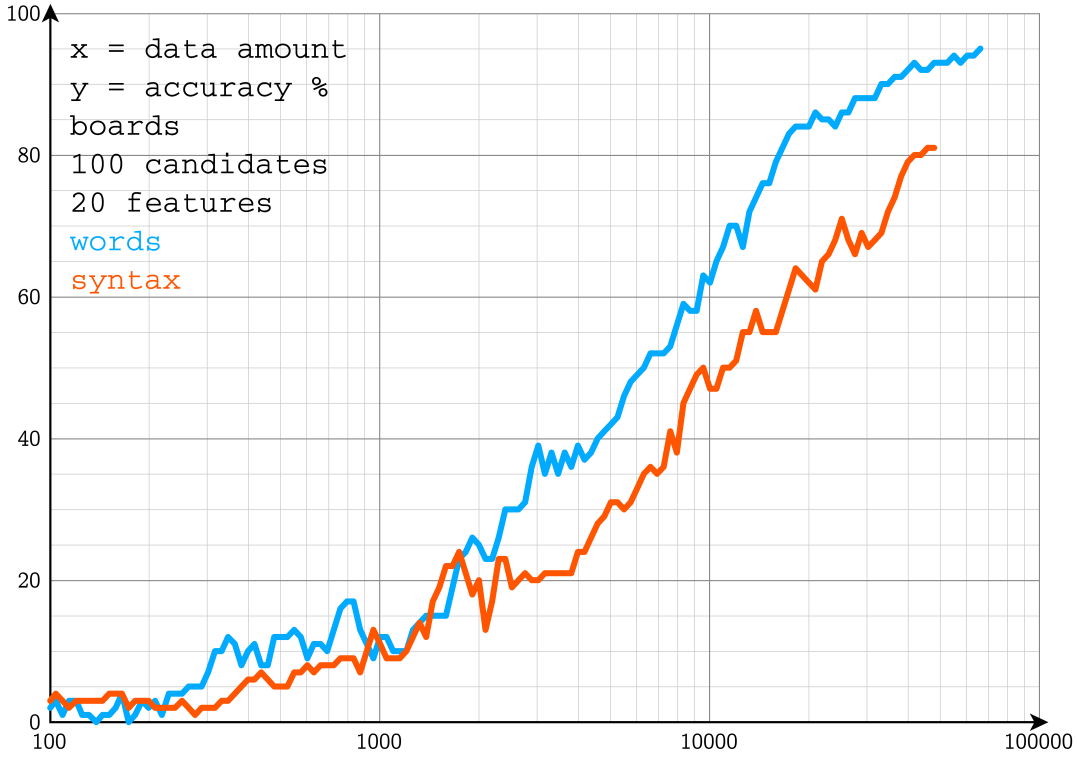

Figure 3: The accuracy of word-based and syntax-based classifiers, applied to the Boards corpus, as a function of amount of data.

2.3 Classifiers

For the actual classification, we use cosine similarity on normalized frequency distributions. For example, suppose that we use two word counts as our features, *the* and *to*. We extract the relative frequencies of each feature for each candidate from the corpus; that is, how many percent of a candidate's words are *the*, and how many percent are *to*. Since we would like all the feature values to be treated similarly, regardless of how common the word is, we adjust the distributions of each feature so that the average is 0 and the variance is 1.

3 Results

In this section, we compare the classification accuracy obtained by working with words and syntactical features, respectively. Figures 3 and 4 show the accuracy for both methods, applied to the Boards and Blogs corpora and given as a function of the number of words in the sample. We see that the word-based classifier performs better than the syntax-based classifier on both corpora. This is in line with the previous body of work, see for example (Fuller et al., 2014), in which purely syntactic approaches perform worse than more traditional lexical approaches.

One explanation for why authorship attribution appears to be an easier problem for Boards than for Blogs is that the texts are less formal and more personal, whereas the authors have made greater efforts in Blogs to write in a literary style. According to Stamatatos (Stamatatos, 2009), this should be less of an advantage for the syntactical classifier, which has difficulties with slang and loose grammar. However, it is not clear from our results whether the difference between the two methods increase when we go from Blogs to Boards.

In lieu of a corpus annotated with topics, we use the previously mentioned approach, taking each book of the Novels corpus to be a separate topic. On the one hand, we can attempt to identify the author of some text from one novel, by comparing it only with text from another novel. That way, we have hopefully isolated author difference, not topic differences. On the other hand, we can pick several books by the same author, and attempt to identify whether two texts are from the same book, thus isolating topic differences instead.

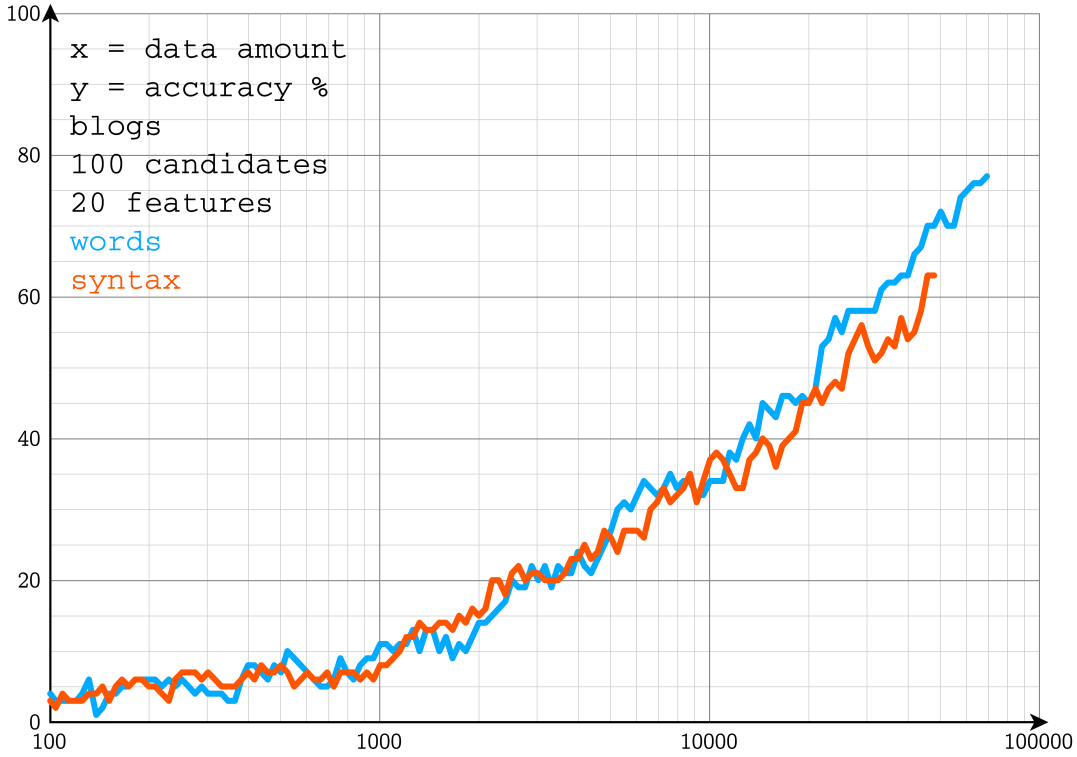

Figure 4: The accuracy of word-based and syntax-based classifiers, applied to the Blogs corpus, as a function of amount of data.

If we can see that some methods work better in the first case, it suggests that that method is more author-dependent than topic dependent.

We first look at the case where we pick one book per author, extract two text samples from each of those books, and then try to match up samples from the same book. We thus classify with respect to author and topic at the same time, which should be easy compared to the tasks that we will consider later on. The outcome of this experiment is shown in Figure 5 and suggests that the word-based classifier outperforms the syntax-based classifier, albeit by a small margin.

We continue by looking at the case where we have a single author, and try to identify individual books. This means that we try to identify topics without help from author idiosyncrasies. The results are shown in Figure 6. The accuracy is now lower, which is to be expected since the data is more homogeneous, and again words seem to be more effective than syntax.

Finally, we look at the case where we try to identify the author of text from one book by comparing with texts from other books. In other words, we try to identify the author without help from the topic. The results are given in Figure 7. Compared to the previous graphs, syntax rises considerably faster, whereas words fall behind. This is very interesting news. Up until now, syntax has seemed like a poor option, being computational demanding to extract and having worse performance than a simple word count. These results now show that there are situations where syntactic features give a better result. More precisely, they suggest that syntax is less topic-dependent than word counts, and therefore more reliable for identifying authors. Since topics are so vaguely defined, controlling for topic is virtually impossible, so this effect has bearing on any kind of author identification.

It should be noted here that since the sample only includes 20 candidates, the expected accuracy of a random guess is 5 per cent, so we can also expect random fluctuations of at least that much. Looking at the first part of the graph, it would appear that the curves are both equal, but in fact all we can see is that they are both too low to get meaningful data. In the second half of the graph, the syntax curve is markedly higher.

To statistically verify the hypothesis that deep syntactical features outperform lexical features, we use the non-parametric Wilcoxon signed-rank test (Wilcoxon, 1945). The required

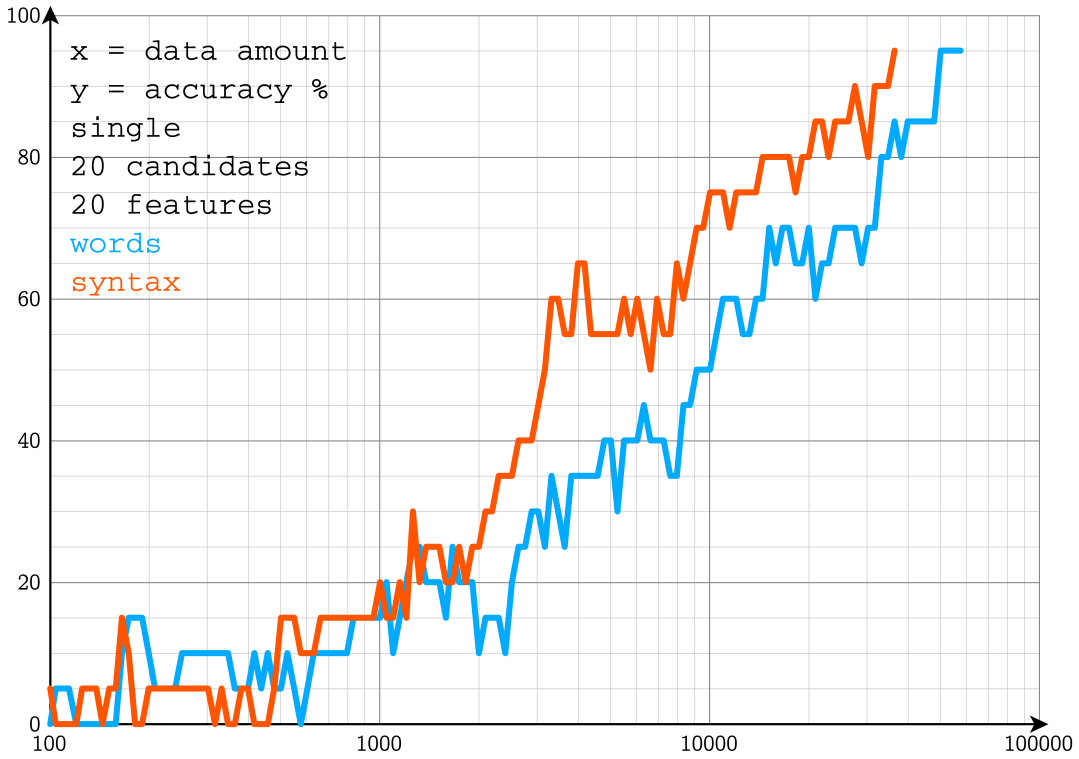

Figure 5: The accuracy of word-based and syntax-based classifiers as a function of amount of data, when the corpus contains exactly one book per author.

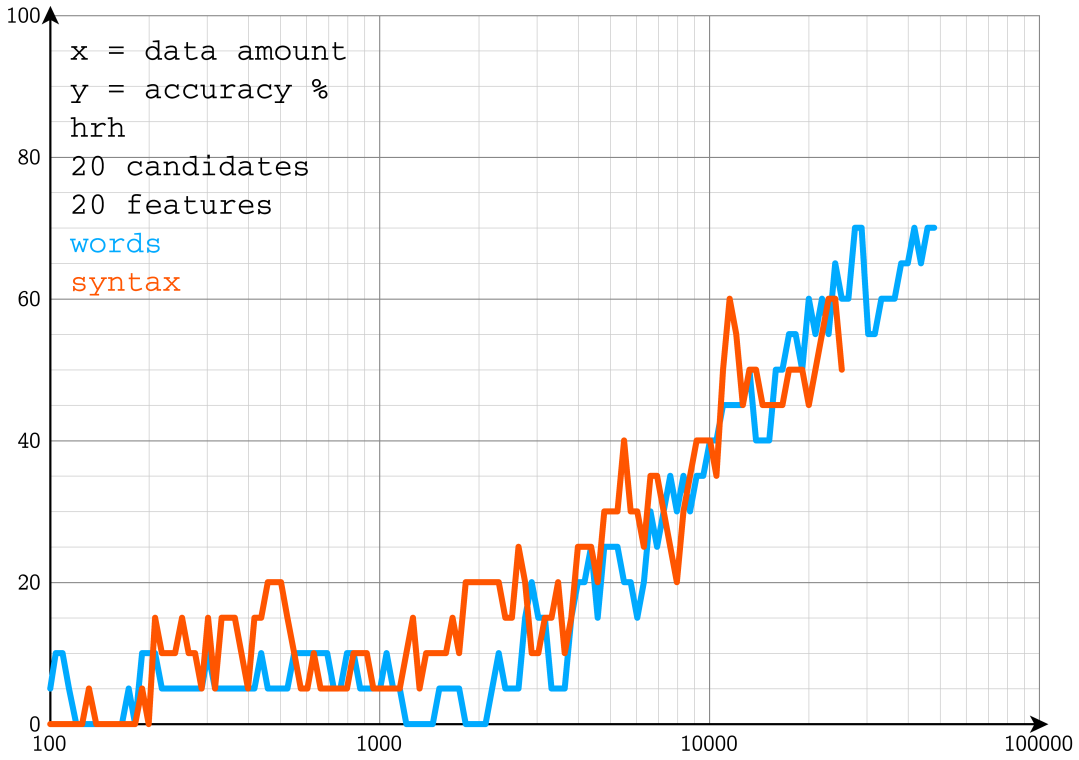

Figure 6: The accuracy of word-based and syntax-based classifiers as a function of amount of data, when the corpus only contains books by one author.

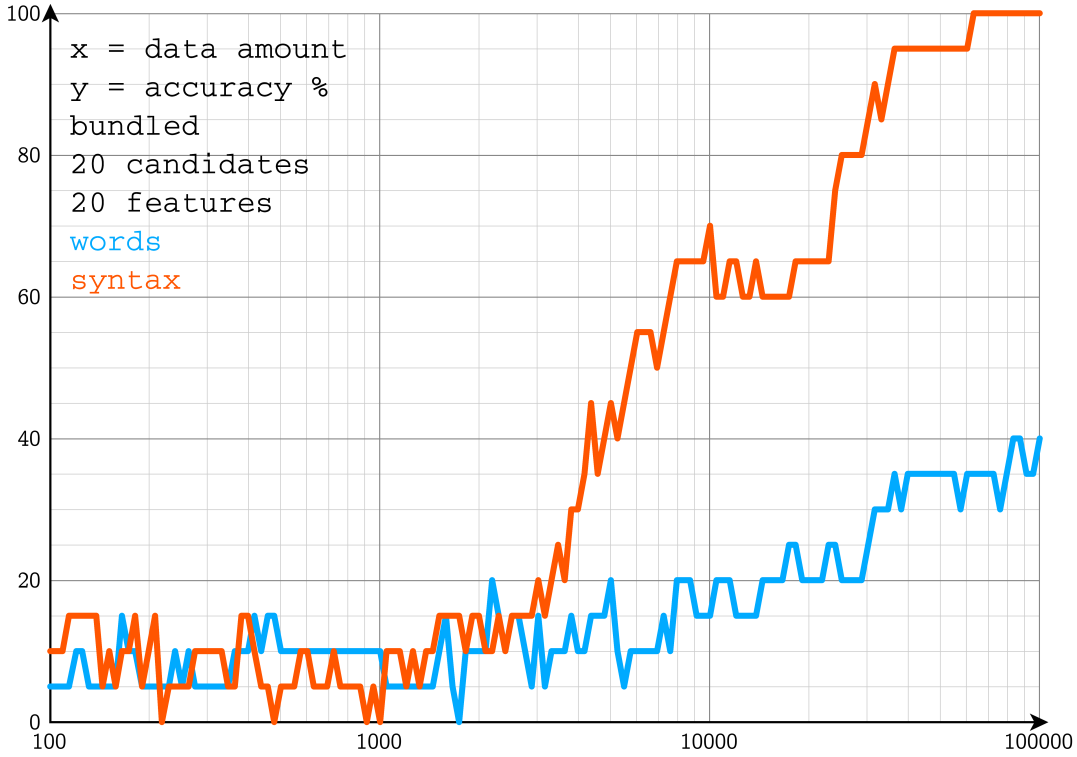

Figure 7: The accuracy of word-based and syntax-based classifiers, when evaluation is made on books not represented in the training set.

assumptions are that data is paired and from the same population, the pairs are chosen randomly and independently, and that data is measured on an ordinal scale. It does not require any assumptions about the data being normally distributed. The test is applied to the interval $10,000 \leq x < 100,000$, giving us 50 data points per curve. The outcome is that the difference is statistically significant for $p < 0.0001$.

We can see this effect from a different perspective, by going through the books in the Novels corpus, and measuring similarities between the first and second halves of the books in terms of words and syntax. For each first half, we compare with each second half - the one from the same book, the ones from other books by the same author, and the ones from books by other authors. The corresponding plot is shown in Figure 8. The highest curves are from pairs consisting of the two halves of the same book. The curves in the middle are for pairs from different books by the same author, and the bottom curves are for pairs of texts by different authors. This gives us a nice overview of what the similarities might look like; we see for example that if the similarity score is negative, we can be fairly certain the texts are not from the same book. What is interesting here is the small by clearly visible difference between words ans syntax: For texts from the same book, the words methods show a higher similarity, but for texts from different books by the same author, syntax shows a higher similarity. This strengthens the conjecture that a syntax-based approach is less topic-dependent.

4 Conclusion

Our experiments suggest that deep syntactic features are more robust than features based on words and punctuation against changes in topics, and are therefor lead to more accurate predictions. This is in line with the intuition that the choice of topic has greater impact on the actual words than how they are combined into sentences.

The corpus used in our experiments consisted of turn-of-the-century English novels, and it remains to verify how well the results hold for other registers. Another open question is the relative performance of syntactic and lexical features as the number of authors grows. In absence of manually annotated data, we took each novel to denote a particular topic. The soundness of this approach is open for discussion, and it would be interesting to repeat the

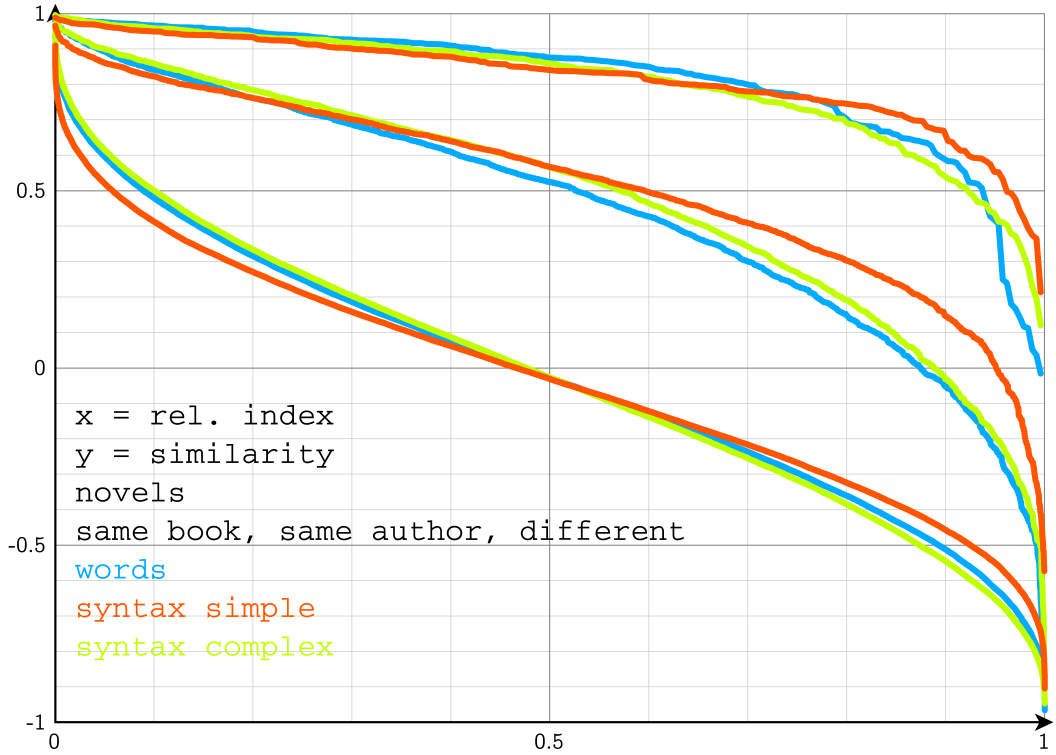

Figure 8: The distributions of similarities for the different methods and pair types.

experiments on corpora with gold-standard topic information.

References

Argamon, S., & Shimoni, A. R. (2003). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*, 401–412.

Augsten, N., Böhlen, M. H., & Gamper, J. (2005). Approximate matching of hierarchical data using *pq*-grams. In *Proceedings of the 31st international conference on very large data bases, Trondheim, Norway, 2005* (pp. 301–312).

Ayala, D. V., Pinto, D., Gómez-Adorno, H., León, S., & Castillo, E. (2013). Lexical-syntactic and graph-based features for authorship verification notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, D. Tufis, & N. Ferro (Eds.), *Working notes for CLEF 2013 conference, Valencia, Spain, 2013* (Vol. 1179). CEUR-WS.org.

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Collins, J., Kaufer, D., Vlachos, P., Butler, B., & Ishizaki, S. (2004). Detecting collaborations in text: Comparing the authors' rhetorical language choices in the federalist papers. *Computers and the Humanities*, *38*, 15–36.

Feng, S., Banerjee, R., & Choi, Y. (2012). Characterizing stylistic elements in syntactic structure. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, 2012, Jeju island, Korea* (pp. 1522–1533).

Frantzeskou, G., Stamatatos, E., Gritzalis, S., & Katsikas, S. (2006). Effective identification of source code authors using byte-level information. In *Proceedings of*

the 28th international conference on software engineering (pp. 893–896). New York, NY, USA: Association of Computing Machinery.

Fuller, S., Maguire, P., & Moser, P. (2014). A deep context grammatical model for authorship attribution. In N. Calzolari et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14).* Reykjavik, Iceland: European Language Resources Association.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Computational linguistics.* Association for Computational Linguistics.

Graham, N., Hirst, G., & Marthi, B. (2005). Segmenting documents by stylistic character. *Natural language engineering*, *3*(11), 397–415.

Hollingsworth, C. (2012). Using dependency-based annotations for authorship identification. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), (Vol. 7499, pp. 314–319). Springer Berlin Heidelberg.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics* (pp. 423–430).

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *4*(17), 401–412.

Koppel, M., & Schler, J. (2004). Authorship verification as a one-class classification problem. In *Proceedings of the 21st international conference on machine learning* (p. 62). New York, NY, USA: ACM Press.

Lučić, A., & Blake, C. L. (2015). A syntactic characterization of authorship style surrounding proper names. *Digital Scholarship in the Humanities*, *30*(1), 53–70.

Luyckx, K., & Daelemans, W. (2005). Shallow text analysis and machine learning for authorship attribution. In *Computational linguistics in the Netherlands 2004: Selected papers from the fifteenth CLIN meeting* (pp. 149–160).

Luyckx, K., & Daelemans, W. (2008). Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd international conference on computational linguistics* (pp. 513–520). Stroudsburg, PA, USA: Association for Computational Linguistics.

Menon, R., & Choi, Y. (2011). Domain independent authorship attribution without domain adaptation. In G. Angelova, K. Bontcheva, R. Mitkov, & N. Nicolov (Eds.), *Recent advances in NLP* (pp. 309–315). Sofia, Bulgaria: Bulgarian Academy of Sciences.

Mikros, G. K., & Argiri, E. K. (2007). Investigating topic influence in authorship attribution. In B. Stein, M. Koppel, & E. Stamatatos (Eds.), *Proceedings of the international workshop on plagiarism analysis, authorship identification, and near-duplicate detection* (Vol. 276). Aachen, Germany: CEUR-WS.org.

Olmos, I., Gonzalez, J. A., & Osorio, M. (2005). Subgraph isomorphism detection using a code based representation. In *FLAIRS conference* (pp. 474–479).

Parker, D. (1928). *Sunset gun.*

Raghavan, S., Kovashka, A., & Mooney, R. (2010, July). Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 conference short papers* (pp. 38–42). Uppsala, Sweden: Association for Computational Linguistics.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on uncertainty in artificial intelligence* (pp. 487–494). Arlington, Virginia, United States: AUAI Press.

Seroussi, Y., Bohnert, F., & Zukerman, I. (2012). Authorship attribution with author-aware topic models. In *Proceedings of the 50th annual meeting of the association for computational linguistics* (pp. 264–269). Stroudsburg, PA, USA: Association for Computational Linguistics.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556.

Stamatatos, E., Kokkinakis, G., & Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. *Computional Linguistics*, *26*(4), 471–495.

Stamatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556.

Stein, B., & zu Eissen, S. M. (2007). Intrinsic plagiarism analysis with meta learning. In *Proceedings of the SIGIR workshop on plagiarism analysis, authorship attribution, and near-duplicate detection* (pp. 45–50).

Tschuggnall, M., & Specht, G. (2014). Enhancing authorship attribution by utilizing syntax tree profiles. In *Proceedings of the 14th conference of the European chapter of the association for computational linguistics* (pp. 195–199). Association for Computational Linguistics.

Wiersma, W., Nerbonne, J., & Lauttamus, T. (2011). Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing*, *26*(1), 107-124.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Zechner, N. (2015). *Formal foundations of authorship attribution.* Umeå University, Umeå, Sweden. (Licentiate Thesis)

zu Eissen, S. M., Stein, B., & Kulig, M. (2007). Plagiarism detection without reference collections. *Advances in Data Analysis*, 359–366.