Classifying Author and Topic – a Case Study on Swedish Literature

Niklas Zechner
Språkbanken
Department of Swedish
University of Gothenburg
niklas.zechner@gu.se

Abstract

Using material from the Swedish Literature Bank, we investigate whether a common method of author identification using word frequencies is actually sensitive to differences in topic. The results show that this is the case, thereby casting doubt on much previous work in author identification. This sets the stage for a broader future study, comparing other methods and generalising the results.

1 Introduction

Author identification is a competitive field, with many studies reporting ever increasing accuracies. Often, the accuracy as reported by the experiment is seen as irrefutable proof that the method works. But there may be reason to be sceptical of the optimistic results. Previous work has shown that there are several things to take into account for text classification generally, before methods can be considered reliable and comparable. The size of the texts has a large impact on the accuracy, and naturally the number of candidate classes also matters (Zechner, 2017). Even minor details in how the test data is handled can lead to significant overestimation of the accuracy (Zechner, 2014).

When it comes to author identification specifically, one of the main pitfalls is neglecting to account for differences in topic, style, or genre (Mikros and Argiri, 2007). If we apply a classification method to texts by several different authors, but each author mainly writes on a particular topic, how do we know if the classification method is detecting authors or topics? If the method is sensitive to topic, the accuracy reported in testing may be far higher than what we would get from a real-life application, where the text to be identified is on a different topic. It is a difficult issue, because we can not easily come by topic-controlled data. Many have tried to get around the problem by basing their methods on features of the text which are assumed to be independent of topic, but few have put that assumption to the test.

In a previous study (Björklund and Zechner, 2017), we investigated this problem by examining a set of novels, using each separate novel as an approximation of topic. In this study, we begin to expand on that work and apply a similar approach to a larger corpus, this time in Swedish.

1.1 The problem

In a typical author identification task, we want to find which of a set of candidate texts is written by the same author as a given target text. To test a method on this task, we would need a number of text samples, at least two of which are by the same author. One of the two acts as the target text, and one is mixed in with texts by other authors to form the candidates. We now have a set of candidate texts with one “true” candidate, the one which is actually by the same author as the target text, and some number of “false” candidates, which are by other authors. If the method correctly identifies the true candidate, it is considered successful. By repeating the experiment, we can estimate the accuracy of the method, that is, the probability of successful identification.

Commonly, when we test a method like this, we only have access to an unstructured text or set of texts by each of a number of authors. This could be articles or letters, or internet data such as forum messages or blog texts. This causes a problem when evaluating the test results. If the methods can reliably identify text samples from the same source, is that because they are written by the same author, or is it because they are on a similar topic? There is a risk that the methods look very accurate in a test setting, but are actually much less so when we apply them to a real-life problem.

1.2 The approach

To address this issue, we use text samples from books, under the hypothesis that each book can be seen as a separate topic. This allows us to try three variants of the identification task. In the first case, the true candidate (the one we are hoping to identify) comes from the same book as the target text, and the false candidates come from books by other authors. This corresponds to the commonly seen experiment, where we are effectively identifying a combination of author and topic. In the second case, the true candidate is again from the same book as the target text, but the false candidates are now from other books by the same author. This way, we are identifying only topic, without the influence from author. In the third case, the true candidate is from the same author as the target text, but not from the same book, and the false candidates are again texts from other authors, so that we are identifying author without the influence of topic. By comparing the results, we hope to see if the method is more sensitive to author or topic.

2 Data

We use data from the Swedish Literature Bank (litteraturbanken.se), a collection of old novels, from which we include only the ones that have been manually digitised. We restrict the data to works in Swedish, by a single known author, and leave out works that contain duplicate text, such as multiple editions of the same book. This leaves 481 books by 140 authors.

Each book is cut up into pieces of 40 000 words, leaving out any trailing words. One reason for this is so that the texts are all the same lengths, making the results meaningful and reproducible. Previous work has found that the accuracy of classification varies greatly with the length of texts, so that if we were to include entire books of varying length, the experiments would have little predictive value (Zechner, 2017). Another reason is that we want to compare texts from the same book, so it is necessary to divide at least some of the books into parts. We get 825 pieces in total.

3 Method

We use a feature set consisting of just ten (relative) word frequencies, specifically those words that are the most common in the data generally. “Words” here also include punctuation, and are counted independent of capitalisation. The words in this case are: comma, full stop, “och”, “i”, “att”, “det”, “en”, “som”, “han”, “jag”.

For each text (that is, for each piece of 40 000 words), we create a profile of its frequencies for these ten words. As a distance measure, we calculate the (absolute) difference in each feature value, and sum over all features; in vector terms, this is the Manhattan distance, without any normalisation. Using these profiles, it is easy to compare any pair of text and calculate the distance. That can then be applied to the identification problem as described above, by comparing the target text to each of the candidate texts, and choosing the one with the smallest distance measure.

Now we can run the three tests we want to compare: identifying a book among a set of books by other authors, identifying a book among a set of books by the same author, and identifying an author among others by comparing with a different book by that author. By repeating the process, we can find an estimated accuracy for each case.

But it is possible to go a step further. We can think of each of the possible pairs of texts as being of one of three types: Same book, same author (but different book), and different author. From the 825 chunks analysed, we get in total 537 same-book pairs, 16 356 same-author pairs, and 323 007 different-author pairs. Since the method is simple and fast, we can easily go though all the possible pairs, and find the distribution of distance measures for each type of pair.

Knowing this distribution has great value in a practical application, because it allows us to calculate the probability that a pairing is of a particular type, and thus the probability that two texts are by the same author, or from the same book. But we can also use it to get a better estimate on the accuracy of the identification problem.

Suppose we want to identify the author of a given text out of 100 candidates, using one other text by that same author and 99 texts by other (not necessarily distinct) authors. This will mean one same-author comparison, and 99 different-author comparisons. Using the simplifying assumption that the similarity between a given text and a random text by the same author does not correlate with the average similarity between that given text and a random text by a different author, we do not need to investigate specific text samples one by one. Instead, we can think of it as a simpler statistical problem: For a given same-author pair, how likely is it that it will have a lower distance measure than each of a set of 99 different-author pairs?

To find out, we do not need to choose 99 random different-author pairs. Instead, we keep a sorted list of the different-author pairs. Choosing one same-author pair, we can use a simple binary search to see what fraction f of the different-author pairs have a higher distance measure. Then, the probability of 99 of them having a higher distance measure is just f^{99}; this is the probability of this same-author pair being correctly identified. This is simple enough that we can repeat it for all the same-author pairs, and calculate the average accuracy, without having used any random subset.

If we look closer at this corpus, we find that there is one author who is far more prolific than the others: August Strindberg. Our sample contains no less than 64 of his works, far more than any other author. Since the number of same-author pairs for an author increases approximately as the square of the number of works by that author, that means that he has a very large impact on the results – about three quarters of the same-author pairs are from Strindberg. This might skew the results, so we run the tests twice, with and without Strindberg.

4 Results

The distributions of distance values for the three types of pairs are shown in Figure 1. We can see that the values for same-author pairs are lower than those for different-author pairs, as can be expected, but also that the values for same-book pairs are lower still. This immediately tells us that methods like this one would be strongly topic-dependent.

The same-book and same-author distributions for Strindberg have been separated out. We can see that they have much higher distance measures, meaning that his works would be much more difficult to identify. Evidently, Strindberg has a more diverse writing style than most; further speculation is beyond the scope of this study.

As outlined in the previous section, we can use the distributions to calculate what would be the average accuracy of an identification test. We choose an identification task with 100 candidates, and try the three different cases: Identifying a book among books by other authors (identification based on both author and topic), identifying a book among other books by the same author (only topic), and identifying an author among others while using a different book as reference (only author). The resulting accuracies are 52%, 17%, and 8%. If we leave out Strindberg, we get 67%, 11%, and 20%.

The distributions can also be used to calculate the probability that a pair is of a given type. For example, suppose we know that a text sample is either from book A, book B or book C. The three books are by different authors (neither of whom is Strindberg) and we have another sample of book A, but not of book B or C. We compare the unknown sample and the one from book A, and get a distance measure of 0.04. How likely is it that the unknown sample is from book A? Since there are three candidates, and we have no further information, the a priori probability is 1/3, or in other words, the a priori probability of a different-author pair is twice as high as that of a same-book pair. Looking at Figure 1, we see that at 0.04, the same-book curve is at 9, and the different-author curve at 6. The final probability for a same-book pair (and therefore, the probability that the unknown sample is from book A) is $1 * 9/(1 * 9 + 2 * 6) = 43\%$.

5 Discussion

We can see directly from the distribution curves that this method is not topic-independent. The accuracy calculations verify this, and indicate that the method may be at least as sensitive to topic as it is to author. This means that similar methods may not be reliable for author identification; even if experiments show positive results, the accuracy in a real-world application might be far lower.

We should keep in mind that this is not meant as a tool for topic identification. Whether this is an accurate representation of topic is also irrelevant; we are interested in separating out any traits not related to the author. Furthermore, authors may well write several books on the same topic. But that would only mean that we have underestimated the problem. If we have only partially separated topic from author – as is likely the case – the decrease in accuracy for a real application would be even greater. Future studies may be able to test this using data from more diverse sources.

Could a different set of features do better? Word frequencies are one of the more common types of features used in author identification. One common approach to avoiding the influence of topic is to use function words, that is, words whose meaning is mainly grammatical rather than semantic, based on the assumption that such words are not topic-dependent (Mosteller and Wallace, 1964). There

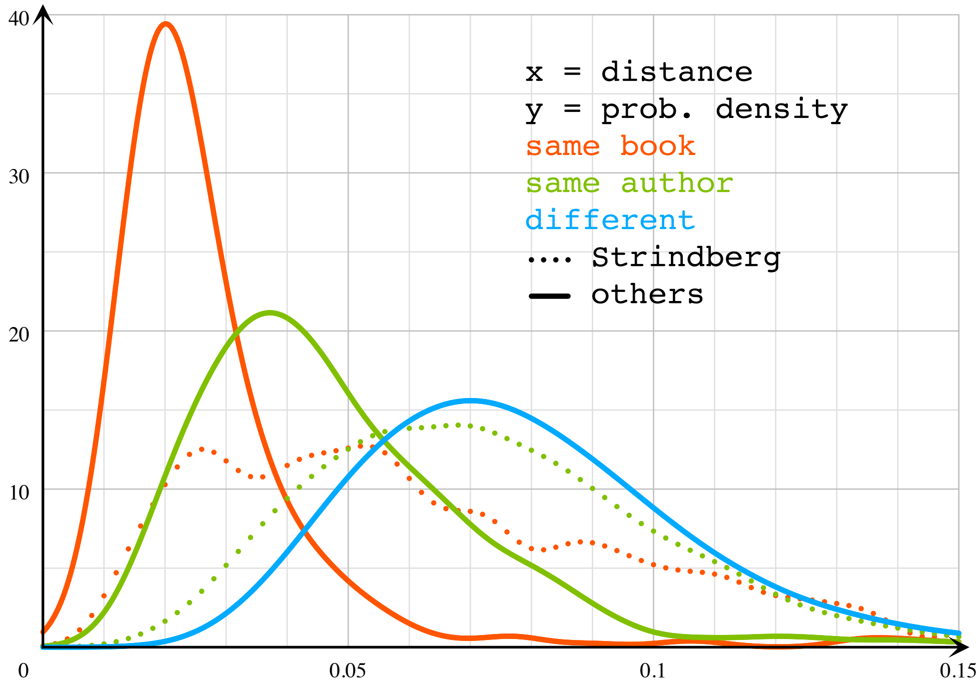

Figure 1: Distributions of distance measures for types of pairs. Distributions sum to one, and have been smoothed with a Gaussian blur, sd = 0.005. The different-author curve also includes Strindberg.

is no clear rule for what should be considered a function word, but generally, the most common words tend to fall in that category, and that is also the case here. Clearly, using function words was not enough to ensure topic independence. Word or character n-grams are likely to suffer from the same issue. As for whether syntactic features avoid topic-dependence better than function words, different studies have found them to be worse (Menon and Choi, 2011), equally good (Luyckx and Daelemans, 2005), or better (Björklund and Zechner, 2017). Testing that for larger datasets, and trying other options, is left for future work.

It should also be noted that the methods used here are not intended to be as accurate as possible. We could very likely improve the accuracy by using a larger set of features, or by using some form of normalisation on the feature values, or by using a more advanced classifier. It is also clear from tests not shown here that the accuracy depends heavily on the size of the samples; samples significantly smaller than these would drastically lower the accuracies, and larger samples would improve them.

We hope to build on this small experiment towards a larger study of classification on this type of corpus. The large amount of data and clear metadata may be useful for other types of classification, including gender and year of writing. A more comprehensive study of different feature sets might also reveal which types of features are best for identifying authors, which are better for topic, and which are better for identifying something else entirely.

References

Johanna Björklund and Niklas Zechner. 2017. Syntactic methods for topic-independent authorship attribution. *Natural Language Engineering*, 23(5):789–806.

Kim Luyckx and Walter Daelemans. 2005. Shallow text analysis and machine learning for authorship attribution. *LOT Occasional Series*, 4:149–160.

Rohith Menon and Yejin Choi. 2011. Domain independent authorship attribution without domain adaptation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 309–315.

George K Mikros and Eleni K Argiri. 2007. Investigating topic influence in authorship attribution. In *PAN*.

Frederick Mosteller and David L Wallace. 1964. *Inference and disputed authorship: The Federalist*. Addison-Wesley.

Niklas Zechner. 2014. Effects of division of data in author identification. In *Proceedings of the fifth Swedish language technology conference*.

Niklas Zechner. 2017. *A novel approach to text classification*. Ph.D. thesis, Umeå universitet.