Predicting the Date of Authorship of Historical Texts

Andrew Tausz

March 11, 2011

Abstract

In this document, we address the following question: Given a piece of text, can we predict its year of authorship? This question is of interest in situations where there is a dispute about the author or origin of a literary work. We note that this investigation is performed using a corpus of approximately 10,000 highly heterogeneous texts that vary significantly in genre and content.

1 Introduction

The task of inferring authorship information is of obvious importance in the humanities. A common example is when someone discovers a piece of historical text. Ideally, we would like to determine who wrote it and when it was written. Additionally, we can ask other questions such as what region of a country it originated from, and what were the demographic characteristics of its author.

In this project, the main consideration is the year of authorship, rather than the identity of the author. The main goal is to determine to what extent language usage can help predict the date of production of a work.

Note that the process of obtaining the year of authorship of a work is inexact (even when the author and title are known), and the relation between the year of production and the language might be weak for the following reasons:

- Historians do not actually know when many books were written. For example although we know the first year of publication of Shakespeare's works, the actual years in which he wrote many of his works are unknown. It is possible that an author waited many years before publishing a work, or maybe it was published posthumously.
- Newer editions of books often add to or change parts of the text.
- An editor of book might provide extensive footnotes or translations into the English used at the time of editing. Thus a work might actually contain text from different eras.
- Translations may distort the relationship between year of authorship and the language use. For example, suppose one considers a modern day translation (2011) of the Old English work Beowulf. Then the translation is neither representative of Old English nor the language that we use today. It is obviously not Old English since it is a translation. However, it is also not really genuine English from 2011 either, since it might be written in a historically sounding style and it will probably contain many words that we almost never use today. One remedy to this is to just omit translations and only consider original works.

2 Previous Work

The existing literature for inferring authorship dates is quite sparse. However, there has been substantial work done in related areas - notably genre and authorship classification, as discussed in [KNS97] and [KKWH10]. By far the largest research endeavor that is related to our current investigation is the Science article [MSA+10] accompanied by the Culturomics project. In it, the authors create an n-gram database from many millions

of books from the Google Books website. In the article [MSA+10], they describe various cultural inferences based on language usage.

From a statistical perspective, in [TE87] Thisted and Efron analyze word usage by Shakespeare to answer the question of whether he was the author of a newly discovered poem. Their approach is based on nonparametric empirical Bayes methods. Along similar lines, in [TR09], the authors consider a work by C.S. Lewis. Lastly, a similar topic (literary period classification of fictional works) was explored in [CCW09]. However, we note that our assumptions about the data are substantially weaker than in [CCW09]. For example we do not pre-select representative authors for different time-periods, and are not restricted to fictional works.

3 Obtaining and Processing the Data

The primary data source was the Project Gutenberg website. This website contains many thousands of electronic books (ebooks) which are in the public domain. The data processing tasks were completed as follows:

1. **Obtain texts:** In accordance to Project Gutenberg's crawling policies, the wget command was used to download english language files with the ".txt" extension, with a break of 2 seconds in between each file. The files were then decompressed and organized into appropriate folders.

2. **Construct index:** Due to the inconsistency of the names of the downloaded files, it was necessary to construct an index which maps the unique etext number to the file location. Duplicate files for the same book (with different encodings) were eliminated.

3. **Extract authorship information:** The text files themselves lacked sufficient structure which would enable one to extract information such as the title, author, publisher, etc. However, Project Gutenberg provides an RDF catalog file which contains all of this information in structured XML form. In this step we retrieve the title, and author from the RDF file and store it in the index.

4. **Infer additional information:** In this step we try to find the most likely year of authorship for a work. Unfortunately the catalog file does not contain the year of publication as a field. However, many of the books contained years in either the title, the author's name, or the translator's name. Additionally, authorship years for a subset of books (e.g. all of the works of Shakespeare) were compiled from external sources. The procedure for extracting the year of authorship is thus as follows:

    - If the title of the work is in the set of already known bibliographical sources, then use the authorship date from that. This dataset contains authorship information for all of the works of Shakespeare and Dickens.
    - If the external index does not contain the work, then look in the title for a date. For example, the number 1922 was extracted from the title "The Best British Short Stories of 1922". However, care must be taken in doing this - an example being the Jules Verne book "Ticket No. 9672". This is remedied by rejecting all dates of authorship that are before 1000 or after 2011.
    - Look at the author string. Many of the author entries (but not all), contained the date of birth and death for the author. For example "Mitford, Mary Russell, 1787-1855". A reasonable value for the year of authorship is the minimum of the author's date of death and their date of birth plus 30. Note that the dates of the author can also be used to correct the date extracted from the title (in the case of books about historical events from the past).
    - Look to see whether the book has a translator. For example, one of the books was the Divine Comedy by Dante. The author string was "Dante Alighieri, 1265-1321" and there was an additional field for the translator which was "Cary, Henry Francis, 1772-1844". Clearly in this situation the dates of the author are irrelevant, and it is only the translation date that matters.

At the end of the preprocessing steps one obtains an index containing the unique etext number, the filename for the actual text, as well as the title, author, date of birth, date of death, and year of authorship. In total, approximately 10,000 books were obtained, totalling 3.7 gigabytes. Note that there were some texts for which it was not possible to obtain the year of authorship or publication. These were removed from consideration, leaving a corpus of 8022 books. By looking through this index, we note the following observations:

- There was significant variety in the types, genres and topics of the books collected. Although many of the works were fictional, there were a significant number of non-fiction works.
- The data is quite "messy". For example, there is no standard format for the header and footer. There is often editorial notes sprinkled in throughout many books. There is also no standard format for the title, author, table of contents, editor, etc.
- Some books are actually in multiple languages - an interesting example being gospel hymns written in both English and Ojibway.
- Some of the books that are labelled as being in "English" have somewhat surprising content. For example, the book zeta310.txt was claimed to have been written in English. However, upon closer inspection it actually contains the first 1 million digits of the Riemann zeta function evaluated at $s = 3$. Similarly, other non-book texts include the human genome and other mathematical constants.
- The distribution of the authorship years are far from uniform. The most popular 20-year period was 1900-1920 which contained 1918 books, wheras the period 1960-1980 only contained 34 books, for example. Additionally, the years were sparse since many years did not contain any books.
- Many of the books contained various editorial comments, and notes that are not part of the actual content of the book. These annotations are likely to have not been written in the same year as the text itself.
- Due to the heterogeneity, variety, messiness and impurity (with comments and annotations) of the corpus, we do not expect extremely high classification performance. It is unlikely that we will be able to obtain fine grained estimates of the authorship year.

Since all of the books contained both headers and footers containing legal and usage notes, various heuristic techniques were used to ignore these portions. The first 400 lines of each file were ignored. Reading was stopped after any mention of "Project Gutenberg", which most likely signals the beginning of the header.

4 Classification Results and Error Analysis

Due to the sparsity of the data (even with close to 10,000 books), we do not have coverage of all years in the range of the data. Thus a year-by-year answer is somewhat unrealistic if we take the approach of discrete classification. Furthermore, formulating it as a classification problem (as opposed to a regression problem) ignores the fact that years are linearly ordered - suppose that we predict a book was written in 1700, then we should also give a high probability to the event that it was written in 1701. Nevertheless, a straightforward way of proceeding is to divide up the years into bins - possibly by century, decade, or other groups.

4.1 The Simplest Possible Attempt

In order to start off, we illustrate an example of basic naive Bayes algorithm with only unigram counts as features. Since, this is a discrete classifier, and there is considerable sparsity among the class labels (the years of authorship) as discussed before, it does not make sense to use the years of authorship directly. Instead, to start off we divided the years into 3 bins: {Before 1700, 1700-1900, After 1900}. The classification performance is shown below. Note that the performance measures seem to be quite good since we are dealing only with 3 bins.

Label	Precision	Recall	F1
After 1900	78.0	63.4	69.9
Before 1700	85.2	87.2	86.2
1700 - 1900	63.2	74.0	68.2

As a sanity check, we can make sure that the classifier correctly ranks words that should have an obvious classification. For example, for the word "internet", we get the following numerator log probabilities: {Before 1700: -15.5, 1700 - 1900: -15.5, After 1900: -11.0} indicating an overwhelming preference for 1900+. Similarly, for the word "thou", we get the values {Before 1700: -7.5, 1700 - 1900: -8.8, After 1900: -10.4}.

4.2 Feature Types

The first approach was to use word-level n-grams as the features in the classification algorithms. However, this is definitely not the only possibility. The feature sets that were investigated include:

- Word level n-grams. We looked at unigrams, bigrams and trigrams. Note that punctuation was completely ignored for these.
- Character level n-grams. Here, the text was divided up into character sequences of length n, with the space being considered as a regular character. For example, the phrase "the dog" would give the following trigrams: {the, he_, e_d, _do, dog}. Again punctuation was ignored.
- Punctuation use. Only punctuation marks were used as features.
- Number use. Only numbers were added as features (just like word unigrams), and all alphabetic words were ignored.

It is interesting to observe the classification performance with various baseline feature sets. In the table below, we show the classification performance using character n-grams. Note that when $n = 1$, we are essentially just doing character counts. Due to the crudeness of this, it is not expexted to yield impressive results. However, as we increase n, the classification performance becomes somewhat more reasonable. In the tables below, we show classification performance using 70% of the corpus as training data, and the remaining as test data. For each file, we take the first 10,000 words in the file. The results are computed using naive Bayes classification - similar results were obtained with the maximum entropy model, but we do not include them here to prevent the proliferation of tables.

n	Label	Precision	Recall	$F1$
1	20-th Century	37.14	38.24	37.68
1	19-th Century	37.84	63.64	47.46
1	18-th Century	43.75	35.90	39.44
1	17-th Century	44.12	46.88	45.45
1	16-th Century	90.00	25.00	39.13
2	20-th Century	60.00	16.67	26.09
2	19-th Century	32.50	72.22	44.83
2	18-th Century	54.05	47.62	50.63
2	17-th Century	40.00	58.82	47.62
2	16-th Century	62.50	13.51	22.22
3	20-th Century	68.75	59.46	63.77
3	19-th Century	41.67	42.86	42.25
3	18-th Century	48.98	64.86	55.81
3	17-th Century	56.60	73.17	63.83
3	16-th Century	100.00	42.86	60.00
4	20-th Century	62.50	64.10	63.29
4	19-th Century	48.89	56.41	52.38
4	18-th Century	63.27	77.50	69.66
4	17-th Century	63.89	71.88	67.65
4	16-th Century	86.67	37.14	52.00

Similarly, we observe the performance with word n-grams.

n	Label	Precision	Recall	$F1$
1	20-th Century	51.11	82.14	63.01
1	19-th Century	69.23	39.13	50.00
1	18-th Century	63.89	65.71	64.79
1	17-th Century	42.59	69.70	52.87
1	16-th Century	87.50	48.84	62.69
2	20-th Century	58.33	60.00	59.15
2	19-th Century	66.67	37.84	48.28
2	18-th Century	73.91	38.64	50.75
2	17-th Century	36.47	88.57	51.67
2	16-th Century	90.00	52.94	66.67
3	20-th Century	65.83	77.63	71.24
3	19-th Century	70.67	43.64	53.96
3	18-th Century	72.31	40.87	52.22
3	17-th Century	55.85	86.11	67.75
3	16-th Century	82.49	56.40	66.99

Note that for n greater than 3, the performance degrades. It is interesting to consider whether non-alphabetic characters have any predictive power. To test this, a feature set was constructed as follows. Each text was divided into words, and all alphabetic characters were removed from the words. This leaves only numbers and punctuation marks. Classification results are shown in the table below. It is somewhat surprising that this works at all, however, we must consider that this reflects punctuation use - for example the frequency of commas.

Label	Precision	Recall	$F1$
20-th Century	58.62	45.95	51.52
19-th Century	40.00	43.75	41.79
18-th Century	42.31	57.89	48.89
17-th Century	60.61	57.14	58.82
16-th Century	68.75	56.41	61.97

If we preserve only numbers and ignore other non-alphabetic characters, the classification performance decreases to the point where it is no longer acceptable:

Label	Precision	Recall	$F1$
20-th Century	21.62	25.00	23.19
19-th Century	30.23	39.39	34.21
18-th Century	26.92	35.00	30.43
17-th Century	62.50	23.81	34.48
16-th Century	36.36	35.29	35.82

4.3 Maximum Entropy Methods

Similarly, it is possible to use a maximum entropy classifier for prediction. In practise, it was observed that maximum entropy classification did not provide a significant improvement in performance on the task at hand. Additionally, the training process was significantly longer than for naive Bayes. For example, for the identical classification task as described in the previous section using word unigrams, the following measures were obtained:

Label	Precision	Recall	$F1$
20-th Century	68.97	52.63	59.70
19-th Century	47.50	50.00	48.72
18-th Century	52.50	60.00	56.00
17-th Century	56.60	75.00	64.52
16-th Century	89.47	56.67	69.39

It is evident that the performance is similar to that of the naive Bayes model - the only slight difference is that the $F1$ score is lower for the 18-th century and higher for the 17-th century.

4.4 Continuous Linear Model

A slightly different approach to authorship date classification is to treat the year as a continuous variable, and then to fit a linear model. Essentially, this makes the assumption that if y is the year of production and x_i are feature counts, then

$$y = x_1\beta_1 + \ldots + x_N\beta_N + \epsilon$$

for Gaussian noise, ϵ. The parameters β_i are estimated using least squares. Although this seems to make sense, in practice it didn't perform as well as expected. For example, using word unigrams as features, and by training on a dataset of 6000 books and testing on 1000, the mean absolute deviation between the predicted year and the actual year was 112.4. This suggests that a linear model does not appropriately capture the relationship between proportional word unigram counts and the year of authorship. However, this does not say anything about whether there is a non-linear relationship, or whether other feature sets have stronger linear predictive power.

Interestingly, with character level bigrams, the mean absolute deviation on the exact same dataset is 82.7, and the mean absolute deviation using character level trigrams is 59.3. Note that as mentioned before, due to the heterogeneity of the corpus, we do not expect extremely precise estimates. Examples of predictions using character trigrams are shown in the following table.

Title	Author	Actual Year	Prediction
The Bent Twig	Fisher, Dorothy Canfield	1914	1899.2
The Pilot	Cooper, James Fenimore	1824	1829.9
The Merry Wives of Windsor	Shakespeare, William	1599	1614.2
What eight million women want	Dorr, Rheta Childe	1901	1909.2
Scientific American Supplem...	Various	1887	1886.6
Othello	Shakespeare, William	1599	1651.7
The Confessions of Harry Lo...	Lever, Charles James	1841	1809.5

From the table we can see that the estimates are quite reasonable. Interestingly, the plays of Shakespeare were consistently estimated to be written several years after they were written. It is not known whether this implies that Shakespeare used relatively advanced vocabulary for his time, or whether this is just due to the characteristics of the corpus. The maximum prediction error using character level trigrams was around 203 years, on the book "Cowleys Essays" by Abraham Cowley. The prediction was 1856, whereas the actual authorship year was stated as 1653.

5 Concluding Remarks and Future Work

In our investigation, we approached the classification problem using three different algorithms: naive Bayes, maximum entropy, and least squares. The least squares estimates gives reasonable continuous estimates of the year of publication, whereas the other two classifiers were used to obtain classifications into bins. Several different feature sets were considered including word and character level n-grams, as well as punctuation use and number use.

Although satisfactory performance was demonstrated in various situations, there are numerous ways in which this work can be extended. Due to the time and computing resource constraints, the scope of the current investigation was somewhat limited. Thus an obvious avenue for further exploration is to increase the amount of data used from approximately 10,000 books to millions. This would have the advantage of producing less sparse data. For example, there were several years in our range of consideration (say 1500-2000) which did not correspond to any authored works. Ideally this would not be the case, and we would have thorough coverage of all of the years in a given range. Additionally, it would be perhaps a good idea to obtain texts from not just one source (Project Gutenberg), but others as well. Possibilities include historical newspaper archives, parliametary transcripts, and perhaps other open source book archives.

The second interesting direction for future investigation is in using more linguistically informed features and models. In our work, the classification was based on relatively "shallow" statistical features - we operated at the bottom row of the Vauquois triangle. It would be interesting to see if features derived from syntactic parsing provide additional classification power. This would that the syntactic structure of the language usage changes over time, not only the word usage.

Lastly, a related investigation (although somewhat different in scope) would be the following. If one constructs a corpus of many newspaper articles, there would be some obvious trends in the word usage of specific terms. For example, there would be spikes for terms relating to terrorism after 2001, and for terms relating to tsunamis after 2005. Additionally, other non-breaking news topics go in and out of fashion. Due to the homogeneity of the corpus (if we are only considering newspaper articles), one would probably be able to obtain a much finer grained estimate of the date of publication of an article. This is in contrast to the much coarser approach taken in this project - since we consider a much wider timespan (going back to the middle ages), and a broad source of significantly diverse texts.

6 Acknowledgements

The author would like to note that the maximum entropy classification code and the l-BFGS optimization code were based on the cs224n codebase.

References

[CCW09] Caitlin Colgrove, Sheldon Chang, and Phumchanit Watanaprakornkul, *Literary period classification (cs224n project)*.

[KKWH10] Sangkyum Kim, Hyungsul Kim, Tim Weninger, and Jiawei Han, *Authorship classification: a syntactic tree mining approach*, Proceedings of the ACM SIGKDD Workshop on Useful Patterns (New York, NY, USA), UP '10, ACM, 2010, pp. 65–73.

[KNS97] Brett Kessler, Geoffrey Numberg, and Hinrich Schütze, *Automatic detection of text genre*, Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (Stroudsburg, PA, USA), EACL '97, Association for Computational Linguistics, 1997, pp. 32–38.

[MSA+10] Jean-Baptiste Michel, Yuan K. Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez L. Aiden, *Quantitative analysis of culture using millions of digitized books*, Science **331** (2010), no. 6014, 176–182.

[TE87] Ronald Thisted and Bradley Efron, *Did shakespeare write a newly-discovered poem?*, Biometrika **74** (1987), no. 3, 445–455.

[TR09] Jeffrey Thompson and John Rasp, *Did c. s. lewis write the dark tower?*, Austrian Journal of Statistics **38** (2009), no. 2, 71–82.