N-gram Distributions in Texts as Proxy for Textual Fingerprints

Carl VOGEL
Centre for Computing and Language Studies, O'Reilly Institute, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2 Ireland (vogel@tcd.ie)

Abstract. Recent experiments using mainly character unigram distributions in authorship attribution tasks are discussed. Results so far indicate efficacy in similarity judgements seemingly good enough for 'balance of probabilities' standards, but not yet for proof 'beyond reasonable doubt'.

Keywords. Authorship attribution, text classification, forensic linguistics

Introduction

This paper describes an area of text classification applied to the problem of very fine-grained categories, authorship attribution, in particular. The connection to verbal communication is fairly obvious in most applications of statistically driven text classification techniques. The possibility of authorship as a category that can be reliably classified on the basis of corpora of texts is where the issue of nonverbal communication inheres. Texts can be reliably classified to the level of granularity required by authorship attribution to the extent that authors encode in their texts, consciously or unconsciously, details that make the tests automatically identifiable as such.

The work here is driven by hypotheses within forensic linguistics that there are valid and reliable methods which can be applied to authorship attribution tasks. Behind these hypotheses is the claim that individuals essentially unconsciously fingerprint themselves in the texts that they produce. Many metrics are proposed in the literature. Research analyzes consciously manipulated author style, and other aspects of text that are extremely difficult for an author to manipulate. Orthography is one such aspect of texts. An author may deliberate a great deal on selection of lexical items from open-class categories, somewhat less reflection tends to be involved in closed class categories, and therefore it is interesting to explore distributions of words in closed class categories used by an author across texts and genre, and in comparison with other authors as a complement to any analysis of lexical richness or quirkiness. However, orthographic analysis crosses both categories,[1] and is basically driven by the fact that while one can choose one's words, one does not generally choose how words are spelled.

Along these lines I and my research group have been experimenting with character n-gram analyses of texts and using similarity of character n-gram distributions to guide

[1] That orthography is open-class requires a moment's reflection since it is usually dealt with in terms of a finite alphabet of anticipated characters. However, new characters (and fonts) are always on the horizon.

the assessment of similarity among texts.[2] The research involves comparative analysis, varying both the value of n and the scale of tokenization—that is, character n-grams, word n-grams, n-grams of part of speech tags—as well as accounting for stop-lists, etc. Some of the experiments have been on closed systems of texts in which authorship is actually known, others involve partly open systems of texts in which the claim of single authorship is disputed, and still more open ended explorations in sentiment analysis (§2).

An assumption is that through the sequences of words that constitute them, and aside from physical evidence like handwriting or signatures, texts communicate information about authorship that is unintended by the author, yet which can potentially be used to identify the author. It must be acknowledged that some authors do directly manipulate orthography for effect—*lipograms* provide a relevant example: [20] is a 50,000 word English novel written without a single instance of the letter E, quite a task given the frequency in English of that letter. The next section describes the method we have been exploring; §2 details some of the results so far; and, §3 outlines ongoing work.

1. The Method

The first step is to prepare the electronic corpus with an initial classification of texts, ideally balanced for size, concatenating or splitting files as appropriate, and balancing numbers of files in each category. Next is choice of a unit to count and its size. The unit measure could be letters, alphanumeric characters including spaces and punctuation, words, or part-of-speech tags assigned to words. For size of n and token choice, the experiments reported here will address mainly letter unigrams, following suggestion from forensic linguistics [2,3,4]. The perl scripts that implement the method are parameterized for both choices, but letter unigrams have proven most reliable in experiments so far.

Obviously, any number of other properties of texts can and have been examined, and word level analyses are quite popular (for example [1,9]). It is also possible to take a number of linguistic 'habits' into account at once: average word length, average sentence length, proportion of open class and closed class categories, and so on [5]. Many of these other possibilities are more clearly stylistic than letter unigrams. However, the interesting thing about letter n-gram distributions inherent in text is that they are so difficult for an author to consciously manipulate as they are only a consequence of the word choices that an author makes, as noted above. This is of separate interest from being objectively individuated in a way that 'average sentence complexity' cannot. Note that even reproducible measures such as Sampson's [18] average depth of words depend on theoretically unresolved issues: node depth depends on the constituent analysis assumed, for example, for coordination. The research reported here follows the suggestion of Chaski [2] that letter n-grams are actually the most reliable thing to count if forensic analysis of texts using corpus linguistic techniques is to satisfy the Daubert test of admissibility of expert testimony in criminal court. Certainly, letter n-gram statistics are reproducible. They also have validity in that they depend on the words selected, and exhibit zipfian distributions inherent in other aspects of language use.

[2] When I use "we" in this paper, it refers to myself and past and present members of my research group involved in this research who are named in the acknowledgements section, and to a some extent folks who will join this activity within the group in the near future.

For each file, a list of n-grams and their absolute and relative frequencies is extracted. Ultimately the relative-frequency distributions of particular n-grams in texts will be compared among files to identify the files with most similar distributions. The idea is simple and can be tested using any number of statistical tests in order to associate confidence intervals with judgements of similarity. In work designed to locate idiosyncrasies in non-native English produced by Finnish emigrants to Australia, [15] use a vector of relative frequencies of POS trigrams and calculation of cosine to gauge similarity, with permutation tests to assess statistical significance (this amounts to assessing the differences between texts using some measure δ and then estimating the proportion of times random samples from the compared texts behave the same with respect to δ). We combine χ^2 testing and the Mann-Whitney rank ordering test for confidence intervals.

Using statistics for hypothesis testing in this context requires comment. Kilgarriff, among others, has pointed out the risks of using hypothesis testing in corpus linguistics [7,8]. The null hypothesis, that the tested samples are randomly drawn from the same population, is only loosely applicable to corpus linguistics since the underlying phenomena are not at all random. Texts are structured, and random sampling from them will reveal distributions of letters, words, parts of speech, that are far from random, and the structure will be revealed increasingly with the total number of n-grams sampled and the value of n. With large enough sample sizes, the χ^2 critical value will be exceeded. However, the test is still interesting for the relative magnitude of the χ^2 value. If one wanted to spot locations of specific differences between two texts one would examine large values of the χ^2, inspecting the exact n-grams for which large values obtain. We calculate the χ^2 value for each n-gram presented by two compared files on the basis of the different observed and expected frequencies in each. Highly influenced by [7], the cumulative sum of the χ^2 value for each of the n-gram tests is then divided by the number of the n-grams tested, yielding a symmetric measure of similarity between the two files. Normally each cell in the contingency table for an individual test must contain at least five observed instances; thus, in allowing a zero (an n-gram not even shared by the files), the analysis we have been exploring amplifies the differences between files that don't have tokens in common. It seems best to make it as hard as possible for the tests to conclude that the files are similar enough to have been written by the same author.

Any authorship attribution task involves in principle comparing more than two texts. The task typically compares a text of known provenance and a questioned text. The metric described above provides a measure of similarity between them. However, there is typically more than one candidate author, and comparisons must be made with known texts of the contending authors. All of the pairwise comparisons of files are made to obtain the cumulative χ^2 value as a similarity ranking index in which smaller values indicate greater similarity, as described above. The resulting rank list is valuable for direct inspection of similarity of categories of files. It makes sense to proceed in two ways simultaneously. The first way is to consider the texts as each constituting a unit category,[3] identifying the pairs of files that are close to each other according to the similarity score. This effects a sense of clustering since it is done without reference to the actual categories the individual files might be members of. All files about one theme might well cluster together more than all files by one author. Thus, where clustering is by author, it is quite interesting. The second method is to gather the texts into categories according to their

[3] A unit category is a category with one instance.

provenance and compare the questioned texts to the categories corresponding to the texts of each of the authors. This brings more external knowledge to bear on the question than considering the entire corpus of texts as each constituting a category.[4] In either method one has rankings of a file according to its similarity to some file(s) and a set of other files. One can then use the Mann-Whitney test to assess whether there is statistical significance associated with the consequent categorization. Where one has external knowledge about the categories the files belong to naturally (e.g. "all files written by author X" or "not spam") it is best to use it, thus preferring the assessments provided by the second method, but making note of the ranked similarity of the first method as well.

For each file, the method examines it with respect to each of the possible categories in turn. This means examining the file's similarity with respect to each file of some category as opposed to each of the files in the complement of that category. The Mann-Whitney test is used independently for each pair of category checks. The null hypothesis is that the similarity with the category considered and its complement are indistinguishable in similarity rank. So, the null hypothesis is rejected when the file is more like one category than it is like the complement, subject to the confidence intervals supplied by the Mann-Whitney. Thus, when assigning a unique category to each file as in the clustering method one is essentially asking what other file is the file most similar to (which one already knows from the χ^2 divided by the number of degrees of freedom) and whether, taking ties into account, that is statistically significant. The method using unit categories is defined in a way that makes it arithmetically difficult to obtain actual statistical significance (by construction, a file is compared to a category with one file in it in relation to its similarity with the complement category, which is constituted by the remainder of the files). For the first method, the actual similarity score is the most interesting statistic. The second method allows one to examine each file within a category to consider its proximity to that category in relation to each other possible category. This means that one can test the homogeneity of a category.

Because the Mann-Whitney test is conducted independently for matching a file between all of the files in one category and the complement of that category, one is not guaranteed a unique match even at a strict significance threshold. A particular file might be like the undisputed works of Shakespeare (as opposed to all other files) with confidence $p \leq .02$ and also like the works of Bacon with the same confidence. However, if, for example, a letter unigram analysis and an analysis based on part of speech tags agree, then one might have increased confidence in the homogeneity of the questioned file with respect to its attributed category, but false positives are not eliminated. Nonetheless, it is clear that given the sorts of questions that one approaches texts with in authorship attribution problems (i.e. many texts and many candidate authors, where each author constitutes a category), a test which admits several categories at once rather than independent tests of a category vs. complement would also be worth adapting and exploring.[5]

2. Representative Experiments

The method has been applied in a number of experiments in search of conditions that invalidate the approach. Using binary categories, spam vs. nonspam, the letter unigram

[4] Balancing the file sizes and amalgamating and splitting the files has analogous impact.

[5] The Mann-Whitney test is a special case of the Kruskal-Wallis test where there are just two categories.

method scored perfectly in an evaluation as a spam filter, against a Bayesian method based on words [16]. Admittedly, this is not a hard test given a fixed training set, since the problem with spam email is that the training set requires constant updating. The methods have also been explored in the context of aligning political parties on a left-right spectrum using only political manifesto texts [19] (this is not an invented problem but one that is topical within political science research [11,10]).

It has also been tested on speeches by prominent politicians of this century and the last [6]. Using the actual authorship identification criterion, results that were relievingly noisy emerged. Texts of Eamon de Valera, Franklin Roosevelt, Gerry Adams, George W. Bush, Huey P. Long and Margaret Thatcher were all correctly assigned using letter unigrams, but only Dick Cheney and de Velara had that status using word unigrams. Bertie Aherne's texts were assigned entirely incorrectly to John Hume, and the texts of George H.W. Bush were assigned to Dick Cheney, Bill Clinton and G.W. Bush. Clearly the tests are noisy on political speeches. This is good because politicians rarely author their own speeches and many have multiple speech writers.

A study intended to determine conditions of applicability of the method are reported by [13]. For example, in work on the Federalist papers, blind classification using the 'genre' category (e.g. letter vs. essay, etc.) were more successful than actual authorship attribution. In another experiment, midi recordings of Lennon and McCartney music were textualized into ABC encoding.[6] There were 141 resulting Beatles files, and a control of 72 files with ABC encodings of Mozart. Again, mixed results obtained: each of the Beatles was correctly discriminated from Mozart, but not from each other.

Another set of experiments used literature archived by Project Gutenberg [14]. Authors sampled were Yeats, Wilde, Shaw, Edgeworth, Gregory. Letter unigrams and bigrams proved best for clustering the texts correctly by their authorship. In many cases where correct categorization was made, the statistical significance reported by the Mann-Whitney test was insufficient to credit the method with providing a correct assessment. In a related test the textual contributions of each character from four plays extracted into files, and clustered characters by author. Results here were quite noisy, except that Yeats' characters highly significantly ($p < .001$) classified as Yeats' using letter unigrams and word unigrams. Some authors evidently have deeper fingerprints than others.

3. Reflections

The experiments demonstrate mixed results. However, when the classifications come up correct, one cannot help but feel there is something real. It is unclear as yet how to merge tests using different units. A sort of intersection may be inappropriate. Also temporal dimension has yet to be thoroughly explored. All of the tests we have conducted effectively assume that the texts were authored at the same time. We have done no classifications that study diachronic effects. Certainly, this adds noise to any synchronic analysis: the texts of early Wittgenstein may be very different from later Wittgenstein not only in their philosophical content. Thus, assessing language change over time within individuals is urgent to explore further. It is important also because if a path of normal change can be established, it may prove useful in research on aging and early detection of neuro-degenerative disorders which impinge on language production [17].

[6] This was done using freely available software by James Allwright of University of Westminster.

Acknowledgements

Support of the NATO Advanced Study Institute and Science Foundation Ireland funding to RFP 05/RF/CMS002. This research has benefitted from collaboration with Sofie Van Gijsel, Lucy Hogan, Cormac O'Brien, Niamh McCombe, Julia Medori, Myriam Mencke and Mary Ronner. The initial perl scripts were implemented by McCombe [12] and subsequently developed further by both me and Medori [13].

References

[1] John Burrows. 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* **17** (2002), 267–287.

[2] Carole Chaski. Who wrote it? Steps toward a science of authorship identification. *National Institute of Justice Journal* **233** (1997), 15–22.

[3] Carole Chaski. Linguistic authentication and reliability. In *Proceedings of National Conference on Science and the Law* (1999), 97–148.

[4] Carole Chaski. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics* **8** (2001), 1–65.

[5] Jill M. Farringdon. *Analysing for Authorship*. With contributions by Morton, A.Q., M.G. Farringdon and M.D. Baker. Cardiff: University of Wales Press, 1996.

[6] Lucy Hogan. A corpus linguistic analysis of American, British and Irish political speeches. Master's thesis, Centre for Language and Communication Studies, Trinity College, University of Dublin, 2005.

[7] Adam Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics* **6** (2001), 97–133.

[8] Adam Kilgarriff. Language is never, ever, ever random. *Corpus Linguistics and Linguistic Theory* **1-2** (2005), 263–275.

[9] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* **17** (2002), 401–412.

[10] Michael Laver, editor. *Estimating the Policy Position of Political Actors*. Routledge, 2001.

[11] Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review* **97** (2003), 311-331.

[12] Niambh McCombe. Methods of author identification. B.A. (Mod) CSLL Final Year Project, TCD, 2002.

[13] Julia Medori. Experiments testing the reliability and validity of author identification methods. Master's thesis, Computational Linguistics Group, Trinity College Dublin, 2005.

[14] Myriam Mencke. Experiments to validate scientifically reliable author identification techniques. MPhil in Linguistics, Centre for Language and Communication Studies, Trinity College, University of Dublin. 2004.

[15] John Nerbonne and Wybo Wiersma. A measure of aggregate syntactic distance. In John Nerbonne and Erhard Hinrichs, editors, *Linguistic Distances Workshop*. COLING/ACL, 2006.

[16] Cormac O'Brien and Carl Vogel. Spam filters: Bayes vs. chi-squared; letters vs. words. In Markus Alesky, et al. (eds), *Proceedings of the International Symposium on Information and Communication Technologies*, (2003), 298–303.

[17] Mary Ronner and Carl Vogel. Iris Murdoch: Writing over a lifetime—a study in form. Presented at the 2006 Meeting of the Iris Murdoch Society, Kingston, UK. September 15-16, 2006.

[18] Geoffrey Sampson. *Empirical Linguistics*. New York: Continuum, 2001.

[19] Sofie Van Gijsel and Carl Vogel. Inducing a cline from corpora of political manifestos. In Markus Aleksy, et al. (eds), *Proceedings of the International Symposium on Information and Communication Technologies*, (2003), 304–310.

[20] Ernest Vincent Wright. *Gadsby*. Wetzel Publishing Co., 1939.