Automatic Categorization of Author Gender via N-Gram Analysis

Jonathan Doyle Vlado Kešelj

Faculty of Computer Science, Dalhousie University,
6050 University Avenue, Halifax, Nova Scotia, Canada
e-mail : {doyle,vlado}@cs.dal.ca

Abstract

We present a method for automatic categorization of author gender via n-gram analysis. Using a corpus of British student essays, experiments using character-level, word-level, and part-of-speech n-grams are performed. The peak accuracy for all methods is roughly equal, reaching a maximum of 81%. These results are on par with other, established techniques, while retaining the simplicity and ease-of-generalization inherent in n-gram techniques.

1 Introduction

There are different subtasks of text classification and they can be divided into topic-based and non-topic-based classification. The traditional text classification is topic-based and a typical example is news classification. Recently, there has been an increasing activity in the area of non-topic classification as well, e.g., in sub-tasks such as

1. genre classification (Finn and Kushmerick, 2003), (E. Stamatatos and Kokkinakis, 2000),
2. sentiment classification,
3. spam identification,
4. language and encoding identification, and
5. authorship attribution and plagiarism detection (Khmelev and Teahan, 2003).

Many algorithms have been invented for assessing the authorship of a given text. These algorithms rely on the fact that authors use linguistic devices at every level—semantic, syntactic, lexicographic, orthographic and morphological (Ephratt, 1997)—to produce their text. Typically, such devices are applied unconsciously by the author, and thus provide a useful basis for unambiguously determining authorship. The most common approach to determining authorship is to use stylistic analysis that proceeds in two steps: first, specific *style markers* are extracted, and second, a *classification procedure* is applied to the resulting description. These methods are usually based on calculating lexical measures that represent the richness of the author's vocabulary and the frequency of common word use (Stamatatos et al., 2001). Style marker extraction is usually accomplished by some form of non-trivial NLP analysis, such as tagging, parsing and morphological analysis. A classifier is then constructed, usually by first performing a non-trivial feature selection step that employs mutual information or Chi-square testing to determine relevant features.

However, there are several disadvantages of this standard approach. First, techniques used for style marker extraction are almost always language dependent, and in fact differ dramatically from language to language. For example, an English parser usually cannot be applied to German or Chinese. Second, feature selection is not a trivial process, and usually involves setting thresholds to eliminate uninformative features (Scott and Matwin, 1999). These decisions can be extremely subtle, because although rare features contribute less signal than common features, they can still have an important cumulative effect (Aizawa, 2001). Third, current authorship attribution systems invariably perform their analysis at the *word level*. However, although word level analysis seems to be intuitive, it ignores the fact that morphological features can also play an important role, and moreover that many Asian languages such as Chinese and Japanese do not have word boundaries explicitly identified in text. In fact,

word segmentation itself is a difficult problem in Asian languages, which creates an extra level of difficulty in coping with the errors this process introduces. Additionally, the number of authors is small in all reported experiments, so the size of author-specific information is not an issue. If the number of authors, or classes in general, is large, we have to set a limit on the author-specific information, i.e., on the author profile.

In this paper, we propose a simple method that avoids each of these problems.

Two important operations are:

1. choosing the optimal set of n-grams to be included in the profile, and

2. calculating the similarity between two profiles.

The approach does not depend on a specific language, and it does not require segmentation for languages such as Chinese or Thai. There is no any text preprocessing or higher level processing required for character or word n-grams, while the most complicated NLP tool used being a part-of-speech tagger used in two of the experiments.

The small profile size is not important only for efficiency reasons, but it is also a natural mechanism for over-fitting control.

2 N-Gram Analysis

The term 'N-gram' refers to a series of sequential tokens in a document. The series can be of length 1 ('unigrams'), length 2 ('bigrams'), *etc*, towards the generalized term "N-gram". The tokens used can be words, letters, or any other unit of information present throughout the document. This versatility allows N-gram analysis techniques to be applied to other media: both images (Rickman and Rosin, 1996) and music (Doraisamy and Ruger, 2003) have been the focus of N-gram research.

N-grams have been used in a wide variety of situations, including optical character recognition (Harding et al., 1997) and author attribution (Keselj et al., 2003). The technique involves the construction of a 'profile' — essentially a listing of the relative proportions of each potential N-gram. When an item is to be classified, its profile is compared with known ones to determine the best match. The basic method of comparison is an N-dimensional distance measurement.

The use of n-gram probability distribution and n-gram models in NLP is a relatively simple idea, but it has been found to be effective in many applications. For example, character level n-gram language models can be easily applied to any language, and even non-language sequences such as DNA and music. Character level n-gram models are widely used in text compression—e.g., the PPM model (T. Bell and Witten, 1990)—and have recently been found to be effective in text mining problems as well (I. Witten and Teahan, 1999). Text categorization with n-gram models has also been attempted by (Cavnar and Trenkle, 1994).

3 Corpus

We used a collection of student essays from the British Academic Written English (BAWE) corpus (Nesi et al., 2004). Only the pilot data for this corpus was available; it nominally consisted of 500 essays, though not all of these were suitable for inclusion. The metadata included for each essay consisted of information such as author gender, first language, the grade received *etc.*

Two essays were simply not present; others did not have metadata present indicating author gender. After these unacceptable essays were excluded, 495 were left in the set. Within these, the average document length was 2,812 words or 17,994 characters, with 1,391,710 words and 8,907,064 characters total.

4 Methodology

4.1 Profile Generation

For each experiment, an individual profile was created for each document in the test set using the Perl module `Text::NGrams`. The cutoff point for each individual profile was 100,000 N-grams; as no document had this number of unique N-grams, this implies that the profile for each document was complete. Profiles were created using character, word, and part-of-speech tags as the tokens to be profiled. In the latter case, an additional experiment was performed after replacing non-function tags with an asterisk. Profiles were generated for N-grams of size 1 through 5 inclusive, with that

size being the limit of computational feasibility. No pre-processing of the data was performed; the documents were left as found in the corpus.

4.2 Part-of-Speech Tagging

Additional copies of the text were generated with words replaced by their part-of-speech tag. The tagging was performed automatically using the Perl module `Lingua::EN::Tagger`, a second-order Hidden Markov Model-based tagger. A further copy of the text was made with all non-function words removed, under the assumption that treating content-bearing words would not little meaning with respect to style. They were arbitrarily replaced by an asterisk. The speech categories considers as function words were: prepositions, pronouns, conjunctions, question adverbs (*e.g.* 'when'), interjections, and determiners.

4.3 Training and Testing Sets

20% each of the male and female lists, rounded up, were randomly selected; these documents would constitute the test set. There were 42 male and 58 female-authored texts in this set, for a total of 100 essays. The remaining essays were taken as the training set.

The 'male' and 'female' essays within this set were listed, and for each list, the profiles of that list's members were combined. The combined profiles were then normalized so as to have a sum N-gram count equal to 1. See Table 1 for a sample of the data produced. This step was performed for all N-gram sizes for which profiles had been generated.

Table 1: Top five character bigrams from the female training set, showing both normalized and unnormalized values. Data has been truncated for presentation.

	Normalized	Unnormalized
E_	0.03274	121221
_T	0.02743	101542
S_	0.02480	91827
TH	0.02277	84306
_A	0.01945	72001

4.4 Determining Closest Profile

For each of the 100 documents in the test set, a 'distance' measurement was calculated to the trained 'male' and 'female' profiles. The distance between two profiles was calculated as in (Keselj et al., 2003); the exact formula is given in equation 1.

$$\sum_{n \,\epsilon\, \text{profiles}} \left(\frac{2(p_1(n) - p_2(n))}{p_1(n) + p_2(n)} \right)^2 \qquad (1)$$

Lower distances indicate a closer match; for each essay, the lower distance was printed as the system's guess. The output was recorded and tested for accuracy, the results of which can be found in the next section.

The experiment was repeated for various profile cutoff lengths; in each case, the merged test profile was simply truncated after a given number of entries and the distance measurements re-run. Note that this will have no effect once the cutoff length exceeds the maximum profile length, as there will be no items to be truncated at that point.

5 Results

5.1 Character N-Grams

Both male and female authors had spaces as their most frequently-used character, followed by *e*,*t*,*i*, and *a*. Only minor differences followed thereafter — the profiles were very similar. This is to be expected, as are the poor results for unigrams in this category.

An increase in the n size provided a noticeable improvement, reaching a peak accuracy of 76% is reached for an N of 4 and an L of 20,000.

Table 2: Results using character-based extraction

Profile Length	N-Gram Size				
	1	2	3	4	5
100	51%	67%	58%	59%	58%
1000	51%	69%	64%	63%	68%
5000	51%	69%	74%	73%	70%
10000	51%	69%	42%	74%	71%
20000	51%	69%	42%	76%	72%

5.2 Word N-Grams

The female authors appeared to have a slightly higher vocabulary than the male authors, with unique word counts of 31734 and 30186 respectively. The different rises for word pairs, with 277769 unique word pairs within the female training set, compared to 237417 in the male. This effect may be partially explained by the larger number of female-authored documents.

In general, the word-based categorization was highly successful, achieving a peak accuracy of 81% is reached for an N of 4 and an L of 10,000–20,000.

Table 3: Results using word-based extraction

Profile Length	N-Gram Size				
	1	2	3	4	5
100	64%	62%	73%	71%	65%
1000	70%	76%	72%	77%	74%
2000	75%	75%	73%	77%	74%
5000	74%	71%	74%	73%	74%
10000	73%	71%	75%	81%	74%
15000	73%	70%	73%	81%	77%

5.3 Part-of-Speech N-Grams

It has been suggested (Argamon et al., 2003) that part-of-speech N-grams can 'efficiently encode syntactical information', and that this may be of use in style classification. This is not unreasonable; the same source provides evidence for gender-based trends in part-of-speech tags. Specifically, the results for Table 5.3 shows the results for these. A peak accuracy of 76% is reached for an N of 5 and an L of 5,000. This is roughly comparable to the other results in this study.

Table 4: Results using part-of-speech extraction

Profile Length	N-Gram Size				
	1	2	3	4	5
100	42%	64%	61%	52%	66%
500	42%	63%	68%	68%	64%
1000	42%	62%	64%	66%	65%
2000	42%	58%	69%	68%	70%
5000	42%	58%	66%	71%	76%
10000	42%	58%	67%	72%	74%

5.4 Function Word N-Grams

It has been also previously suggested that function words may be a strong determiner of author characteristics (Zhao and Zobel, 2005). To test this, the experiment was run again on profiles for which non-function words had been replaced by an asterisk. The results of our test may be seen in Table 5.4.

As with the full part-of-speech profiles, a peak accuracy of 76% is reached. This time, it is for an N of 4 and an L of 1,000. While the peak is the same, overall accuracy is lower than in Table 5.3.

Table 5: Results using part-of-speech extraction, with non-function words replaced by an asterisk

Profile Length	N-Gram Size				
	1	2	3	4	5
100	42%	60%	58%	62%	63%
500	42%	58%	72%	67%	61%
1000	42%	58%	67%	76%	59%
2000	42%	58%	64%	73%	59%
5000	42%	58%	42%	72%	70%
10000	42%	58%	42%	71%	72%

6 Conclusion

We have presented a method for automatic identification of author gender based on n-gram profiles. The method is successful on this corpus; it would be desirable to try it on others to determine the versatility. Because no information specific to this experiment has been used, it is likely that the techniques would be equally-applicable to other data sets. Further, the technique is not language-specific, suggesting is is probably applicable across languages.

The use of part-of-speech tags had no *substantial* effect on the results, showing only a slight decrease. It is possible that with a more accurate tagger better results would be found.

Although simple, in this case N-gram analysis performs on par with other techniques, achieving a peak accuracy of 81%. For comparative purposes, (Koppel et al., 2002) claim an accurate of 'approximately 80%'.

Acknowledgments

We would like to thank the maintainers of the BAWE corpus for providing access to the pilot

data used in this article. We would also like to acknowledge the contribution of three anonymous reviewers, whose feedback has been helpful.

This research is supported by the Natural Sciences and Engineering Research Council of Canada.

References

A. Aizawa. 2001. Linguistic techniques to improve the performance of automatic text categorization. In *Proceedings 6th NLP Pac. Rim Symp. NLPRS-01.*

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346.

W. Cavnar and J. Trenkle. 1994. N-gram-based text categorization. In *Proceedings SDAIR-94.*

Shyamala Doraisamy and Stefan Ruger. 2003. Robust polyphonic music retrieval with n-grams. *Journal of Intelligent Information Systems*, 21(1):53–70, July.

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

M. Ephratt. 1997. Authorship attribution - the case of lexical innovations. In *Proc. ACH-ALLC-97.*

Aidan Finn and Nicholas Kushmerick. 2003. Learning to classify documents according to genre. In *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis.*

Stephen M. Harding, W. Bruce Croft, and C. Weir. 1997. Probabilistic retrieval of ocr degraded text using n-grams. *Probabilistic Retrieval of OCR Degraded Text Using N-Grams*, 1324:345–359.

M. Mahoui I. Witten, Z. Bray and W. Teahan. 1999. Text mining: A new frontier for lossless compression. In *Proceedings of the IEEE Data Compression Conference (DCC).*

Vlado Keselj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. *Proceedings of the Conference Pacific Association for Computational Linguistics PACLING'03*, August.

D. Khmelev and W. Teahan. 2003. A repetition based measure for verification of text collections and for text categorization. In *SIGIR'2003*, Toronto, Canada.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Hilary Nesi, Gerard Sharpling, and Lisa Ganobcsik-Williams. 2004. Student papers across the curriculum: Designing and developing a corpus of british student writing. *Computers and Composition*, 21(4):439–450.

R. Rickman and P. Rosin. 1996. Content-based image retrieval using colour n-grams. *IEEE Colloquium on Intelligent Image Databases*, pages 15/1–15/6.

S. Scott and S. Matwin. 1999. Feature engineering for text classification. In *Proceedings ICML-99.*

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35:193–214.

J. Cleary T. Bell and I. Witten. 1990. *Text Compression.* Prentice Hall.

Ying Zhao and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. *The 2nd Asia Information Retrieval Symposium.*