Paragraph Clustering for Intrinsic Plagiarism Detection using a Stylistic Vector-Space Model with Extrinsic Features
Notebook for PAN at CLEF 2012

Julian Brooke and Graeme Hirst

Department of Computer Science, University of Toronto
jbrooke,gh@cs.toronto.edu

Abstract Our approach to the task of intrinsic plagiarism detection uses a vector-space model which eschews surface features in favor of richer extrinsic features, including those based on latent semantic analysis in a larger external corpus. We posit that the popularity and success of surface *n*-gram features is mostly due to the topic-biased nature of current artificial evaluations, a problem which unfortunately extends to the present PAN evaluation. One interesting of aspect of our approach is our way of dealing with small, imbalanced span sizes; we improved performance considerably in our development evaluation by countering these effect using the expected difference of sums of random variables.

1 Introduction

The task of intrinsic plagiarism detection involves distinguishing portions of a single text which are written by different authors [24]. Key characteristics of the task are the lack of texts written purely by one author or another, as is typically the case in authorship attribution [23], and lack of a database of texts from which the texts are formed, which is the focus of extrinsic plagiarism detection [21]. As such, it has more in common with (to the point of being arguably synonymous with) the task of stylistic inconsistency detection [14,15,18], and our approach in the task is strongly influenced by this work.

Relatively successful approaches to intrinsic plagiarism detection [22,21,17] have often relied exclusively on variation in word or character *n*-gram frequency as the key indicator of stylistic variation, an approach that is clearly effective in the general task of authorship attribution [23]. However, in the context of spans as small as paragraphs, we are somewhat skeptical that these sorts of features capture anything much beyond the topic shifts which are a common artifact of the usually artificially-created test sets. In fact, in the context of another stylistic text classification task, native language identification [19], we found that the effectiveness of character *n*-grams as stylistic features seemed to derive largely from the confounding effects of topic in the corpus [8]; when topic was (partially) controlled for, performance of these features plummeted by over 30%. In the case of real-world plagiarism, we would expect that differences in style, not topic, would be the key indicator of plagiarism, and, although focusing on surface (intrinsic) features may provide superficial improvement in artificial settings, we think it is important to branch out and incorporate stylistic information that reflects underlying dimensions of stylistic variation; a similar approach applied to various text classification tasks has shown promise [2]. In recent work, stylistic segmentation of a real stylistically diverse document (a long poem) [7], we compared the typical surface *n*-gram features to richer extrinsic features, and found that the *n*-gram surface features, though reasonably useful on their own, did not seem to combine well with more-targeted features, and we ultimately discarded them. Therefore, in the present work, which is in most other respects a reasonably straightforward clustering approach based on maximizing vector distance between author spans, we entirely eschew *n*-gram features in favor of the linguistically motivated extrinsic features that we applied to poetry segmentation.[1] In addition, we use a novel approach based on modeling expected random differences to attenuate the effects of variation in span length.

2 Feature Selection and Extraction

The set of features that we explore for this task falls roughly into two categories: surface and extrinsic. The distinction is not entirely clear cut, but we wish to distinguish features that use the basic properties of the words or their PoS, which have traditionally been the focus of automated stylistic analysis, from features which rely heavily on external lexical information such as word sentiment and, in particular, vector-space representations, which are more novel for this task.

2.1 Surface Features

Word length Word length is a common textual statistic in register and readability studies. Readability, in turn, has been used for plagiarism detection [24], and related metrics were consistently among the best performing features for Guthrie [15].

Syllable count Syllable count is reasonably good predictor of the difficulty of a vocabulary, and is used in some readability metrics.

Punctuation frequency The presence or absence of punctuation such as commas, colons, semicolons, and quotes can be very good indicator of style. We also include periods, which offer a measure of sentence length.

Parts of speech Lexical categories can indicate, for instance, the degree of nominalization, which is a key stylistic variable [5]. We collect statistics for the four main lexical categories (noun, verb, adjective, adverb) as well as for prepositions, determiners, and proper nouns.

[1] Nonetheless, we believe that vocabulary-based features do have a role to play in intrinsic plagiarism detection, albeit a restricted one. Differences in (near-)synonym choices may signal plagiarism; the student writes *rain and snow*, but plagiarizes some text that instead says *precipitation*. But incorporating this feature into an intrinsic plagiarism detection system would require it to be appropriately circumscribed in order to prevent confounds of topic. We did not implement this feature here, as it did not seem applicable to the kind of data used in the PAN task.

Pronouns We count the frequency of first-, second-, and third-person pronouns, which can indicate the interactiveness and narrative character of a text [5].

Verb tense Past tense is often preferred in narratives, whereas present tense can give a sense of immediacy.

Type-token ratio A standard measure of lexical diversity.

Lexical density Lexical density is the ratio of the count of tokens of the four substantive parts of speech to the count of all tokens.

Contextuality measure The contextuality measure of Heylighen and Dewaele [16] is based on PoS tags (e.g. nouns decrease contextuality, while verbs increase it), and has been used to distinguish formality in collaboratively built encyclopedias [13].

2.2 Extrinsic features

For those lexicons which include only lemmatized forms, the words are lemmatized before their values are retrieved.

Presence in Dale-Chall Word List A list of 3000 basic words that is used in the Dale-Chall Readability metric [12].

Unigram count in 1T Corpus Another metric of whether a word is commonly used. We use the unigram counts in the 1T 5-gram Corpus [6]. Here and below, if a word is not included, it is given a zero.

Sentiment polarity The positive or negative stance of a span could be viewed as a stylistic variable. We use a hand-built lexicon for the SO-CAL sentiment analysis system, which has shown superior performance in lexicon-based sentiment analysis [25].

Sentiment extremity Some sentiment lexicons provide a measure of the degree to which a word is positive or negative. Instead of summing sentiment scores, we sum their absolute values, to get a measure of how extreme (subjective) the span is. In our previous work[7], we found that an high coverage automatic resource, SentiWordNet (SWN) [3] was preferable to a low-coverage manual resource, and so we use that resource instead of the manual resource used for the polarity feature.

Formality Formality score uses a lexicon of formality that we created in previous work [10]. The values vary between -1 and $+1$, with negative values indicating informal, colloquial words (e.g. *damn*), and positive values indicating formal terms (e.g. *therefore*). The lexicon was built by comparing, using cosine similarity, the latent semantic analysis (LSA) [20] vectors derived from a binary word-document matrix built from a filtered version of ICWSM Spinn3r weblog corpus [11], similar to the sentiment lexicon creation of Turney and Littman [26]. For stylistic concerns, binary matrices appear to be preferable to the *tf-idf* weighted matrices that are typically used for topic variation [10].

LSA vector features In Brooke et al. [9], we posited that, in highly diverse register/genre corpora, the lowest dimensions of word vectors derived using LSA (or other dimensionality reduction techniques) often reflect stylistic concerns; we found that using the first 20 dimensions to build our formality lexicon provided the best results in a near-synonym evaluation. Early work by Biber [5] in the Brown Corpus using a related technique (factor analysis) resulted in the discovery of several identifiable dimensions of register. In the poem segmentation of Brooke et al. [7], we investigated using these LSA-derived vectors directly, with each of the first 20 dimensions corresponding to a separate feature, and found that, taken as a whole, they were superior to any other feature. We use the same word-document matrix, from the ICWSM blog corpus, as in formality creation. One technical note: the length of the LSA vector depends greatly on the frequency of the word, which would mean that common words would essentially drown out rarer words; so before considering each LSA dimension as an individual feature, we first normalize the entire vector to the unit circle.

Normalization For all the features, we also normalize at the level of the word, and then sum across the span; for instance, to calculate a noun frequency metric, we first assign each noun in the text a 1, and all other tokens a 0, normalize this distribution to a mean of 0 and a standard deviation of 1, and then average the normalized values across the span that we're interested in. When applying unsupervised techniques with features of vastly different ranges, normalization is a very important step [15], and calculating everything at the word level gives the system the flexibility to easily consider many different possible spans.

3 Clustering

Our general approach to both of the paragraph-clustering subtasks of the intrinsic plagiarism detection task is to assign paragraphs into author groups that maximize the (average) distance between authors. Following Guthrie's work in stylistic outlier detection [15] and our own previous conclusions [7], we use L_1 or city block distance as the distance metric. Another important insight of Guthrie that is it is desirable to use spans as large as possible, i.e. we consider the distance between the spans suspected to be written by a single author, rather than the distances between individual paragraphs (e.g. a graph-based approach). In particular, for the single-intruder task, we considered all possible start/end pairs for second author intrusion, and calculated the difference between the main and intruder spans, choosing the pair that produced a maximal distance. For the multi-author task, we began by assigning all paragraphs to a single author and none to the three other authors. We then iteratively moved spans from one author group to another, each step being the one that provided the maximum increase in average distance, until no further improvement was possible.

However, there is a serious flaw in this kind of approach: all other things being equal, shorter spans have more random variation and thus are, on average, more distant (sometimes *much* more distant) from any given span than a longer, more homogeneous span. Fortunately, this effect can be modeled. We did this by calculating the expected distance of sums of random variables. Supposing a span of some basic length (we used

50 tokens) to have a random component of normal distribution (with a mean and standard deviation of 1), we can estimate the expected influence of this randomness on the distance measure between any pair of spans — for instance, spans of length 400 and 100 — by looking at the sum of random variables corresponding to the *n* basic spans that make it up — in this example, the expected difference between the sum of 8 random variables and the sum of 2 random variables with the same distribution. We ran 100 trials using a random number generator and computed a table of such expected differences, and then divided our calculated distance by the corresponding number in the table to get a new distance that takes expected difference into account.[2]

4 Evaluation

Even before we reached the final version described above, our approach had perfect performance on the two example texts provided by the PAN organizers, so we created some additional corpora for testing, collecting a few different types of texts (early modern novels, translated Russian novels, and political treatises) from Project Gutenberg, and automatically creating mixed texts of various difficulty. Here, we present results using two relatively easy corpora which consist of texts of paragraphs randomly pulled from novels by Fyodor Dostoyevsky (in English translation) and Thomas Hardy, and political texts by Thomas Paine and Jean-Jacques Rousseau. For each text in the corpus for the mixed-author task, we first choose the number of authors (between 2 and 4), then randomly selected the authors, the number of paragraphs (between 10 and 30) and then the paragraphs themselves from random locations in the text. For the two-author insertion task, we randomly choose two authors, a total number of paragraphs, and two non-equal indices within that range; for each author, a random starting location was randomly selected and consecutive paragraphs from the first author were randomly selected for paragraphs before the first index, and then after the second, and consecutive paragraphs from the second author were inserted between the two indices. Both corpora have 30 texts created in this fashion.

There are various metrics for extrinsic cluster evaluation; Amigó et al. [1] review various options and select the BCubed precision and recall metrics [4] as having all of a set of key desirable properties. BCubed precision is a calculation of the fraction of item pairs in the same cluster which are also in the same category, whereas BCubed recall is the fraction of item pairs in the same category which are also in the same cluster. The harmonic mean of these two metrics is BCubed F-score, which served as our metric for development.

We compare our algorithms with task-specific random baselines (with 50 trials) and two related alternatives: one which excludes our expected difference corrector and another that is based purely on the maximizing distances between individual paragraphs in the spans, rather than treating each cluster as a whole. The results for each of the two tasks are in Table 1.

There is little doubt that our expected difference adjustment has an overall positive effect, and, in the multi-author task, this provides the best result by a reasonably large

[2] There may be a closed-form solution to this problem, but in our case it was easier to derive it empirically.

Table 1. Clustering results with BCubed metrics on our test data.

Distance calculation	Multi-author			Insertion		
	Prec.	**Rec.**	**F-score**	**Prec.**	**Rec.**	**F-score**
Random baseline	0.411	0.378	0.386	0.754	0.694	0.704
Individual paragraph	**0.589**	0.620	0.582	0.818	**0.969**	**0.879**
Combined span	0.426	0.794	0.543	0.749	0.923	0.818
Combined span with expected adjustment	0.533	**0.871**	**0.645**	**0.905**	0.866	**0.879**

margin. For the insertion task, which has a much higher random baseline, the individual span distance comparison was found to be roughly equivalent to our combined span approach with the adjustment.

5 Discussion

Our linguistically motivated, vector-space clustering approach shows promise, particularly with our expected difference adjustment. There is, however, obviously more work to do in this regard; for instance, using this adjustment our multi-author method never, in practice, predicts more than two authors, probably because the differences between short spans are now being underestimated rather than overestimated, meaning that two relatively short author spans (e.g. 3rd and 4th authors) are now highly dispreferred under our distance-maximizing algorithm. This may partially explain our relatively poor performance on the multi-author intrinsic plagiarism task, but in fact there is a more obvious reason. For instance, here are two paragraphs from different authors in the multi-author task evaluation data (text1):

> John did not dream about the deli. He had nightmares of Douglas falling onto swords of knights on horseback, and woke several times throughout the night sweating and breathing heavily.
>
> But in an empty house, surrounded by evidence of Caroline's long absence, Hillie's words plagued him, and he was forced to accept that his mind might be capable of the cruelest of tricks. He felt desperately, hopelessly alone.

Stylistically, we find the two authors nearly indistinguishable. There are small differences (the second author prefers longer sentences and hyperbole), but the easiest way, for either human or computer, to identify the two is by the names of the characters. All but two of the paragraphs contains a proper name that appears in several other paragraphs (one author talks about Geoff, Hillie, and Caroline, the other about Douglas, John, and Mrs. Cumberland). Beyond proper names, there are also recurring topics: mail in one story, a job at a deli in another.[3] Any model that uses word or character

[3] There are even two sets of repeated paragraphs in this particular text: paragraphs 14 and 21 are the same, as are paragraphs 24 and 30!

n-grams should be able to take easy advantage of these regularities. Our model, conversely, was specifically designed not to do so; rather, it was developed to detect significant stylistic differences. In fact, in the task evaluation data there are quite clearly more obvious stylistic differences between some excerpts from the same novel than between some excerpts from different novels:

> On his departure, Hillie had pressed her business card into his hand. “My number’s on there,” she told him. “Call me, all right? I want you to promise.” “I’m sorry,” Geoff said. “I don’t see the point.” “You’ve suffered a shock. You can’t be expected to cope at home on your own.” Geoff had simply smiled at her. “I won’t be on my own,” he said. “I keep telling you. Caroline will look after me.”

Because of the presence of dialogue, this passage is radically different, stylistically, from the other passage from the same novel that was shown above. However, it is clear which of the two novels it comes from, since there are several proper-noun indicators. Given the range of subgenres within the novel genre, i.e. narration, description, and dialogue, this genre is a particularly bad choice for the purposes of simulating intrinsic plagiarism, since those stylistic features which exist and might be useful for distinguishing authors will be ultimately be drowned out by this confounding variation. Instead, topic-related features, which would be highly unreliable in the real world (for reasons that are obvious; what is the purpose of a student plagiarizing something that is topically distinct from the matrix text in which it is embedded?), are strongly preferred.

Thus, we argue that for the PAN evaluation to be a useful reflection of the real-world task of intrinsic plagiarism detection, it requires data that are well-controlled for topic. This might include changing proper nouns (and other highly topical elements) in the data so that they are matched across authors. Insofar as clustering by authorship (as opposed to merely detecting intrusions by one or more authors into a matrix text) is taken as an interesting research problem that mirrors aspects of intrinsic plagiarism detection, the evaluation needs to be constructed so that the focal task is not confounded by orthogonal issues such as subgenre detection.

References

1. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retreval 12, 461–486 (August 2009), http://dl.acm.org/citation.cfm?id=1555682.1555686
2. Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., Levitan, S.: Stylistic text classification using functional lexical features. Journal of the American Society for Information Science and Technology 58, 802–822 (2007)
3. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th conference on Language Resources and Evaluation (LREC’10). Valletta, Malta (2010)
4. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING ’98). pp. 79–85 (1998), http://dx.doi.org/10.3115/980845.980859
5. Biber, D.: Variation Across Speech and Writing. Cambridge University Press (1988)

6. Brants, T., Franz, A.: Web 1T 5-gram Corpus Version 1.1. Google Inc. (2006)
7. Brooke, J., Hammond, A., Hirst, G.: Unsupervised stylistic segementation of poetry with change curves and extrinsic features. In: Workshop on Computational Linguistics for Literature at the 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '12) (2012)
8. Brooke, J., Hirst, G.: Native language detection with 'cheap' learner corpora. In: Proceedings of the 2011 Conference on Learner Corpus Research (LCR2011) (2011)
9. Brooke, J., Wang, T., Hirst, G.: Automatic acquisition of lexical formality. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10): Posters. pp. 90–98 (2010)
10. Brooke, J., Wang, T., Hirst, G.: Inducing lexicons of formality from corpora. In: Proceedings of the Language Resources and Evaluation Conference (LREC '10), Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods. pp. 17–22 (2010)
11. Burton, K., Java, A., Soboroff, I.: The ICWSM 2009 Spinn3r Dataset. In: Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009). San Jose, CA (2009)
12. Dale, E., Chall, J.: Readability Revisited: The New Dale-Chall Readability Formula. Brookline Books, Cambridge, MA (1995)
13. Emigh, W., Herring, S.C.: Collaborative authoring on the web: A genre analysis of online encyclopedias. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS '05). Washington, DC (2005)
14. Graham, N., Hirst, G., Marthi, B.: Segmenting documents by stylistic character. Natural Language Engineering 11(4), 397–415 (2005)
15. Guthrie, D.: Unsupervised Detection of Anomalous Text. Ph.D. thesis, University of Sheffield (2008)
16. Heylighen, F., Dewaele, J.M.: Variation in the contextuality of language: An empirical measure. Foundations of Science 7(3), 293–340 (2002)
17. Kestemont, M., Luyckx, K., Daelemans, W.: Intrinsic plagiarism detection using character trigram distance scores. In: Proceedings of the PAN 2011 Lab: Uncovering Plagiarism, Authorship, and Social Software Misuse (2011), http://www.webis.de/research/events/pan-11
18. Koppel, M., Akiva, N., Dershowitz, I., Dershowitz, N.: Unsupervised decomposition of a document into authorial components. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11) (2011)
19. Koppel, M., Schler, J., Zigdon, K.: Determining an author's native language by mining a text for errors. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05). pp. 624–628 (2005)
20. Landauer, T.K., Dumais, S.: A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240 (1997)
21. Oberreuter, G., L'Huillier, G., Ríos, S.A., Velásquez, J.D.: Approaches for intrinsic and external plagiarism detection. In: Proceedings of the PAN 2011 Lab: Uncovering Plagiarism, Authorship, and Social Software Misuse (2011), http://www.webis.de/research/events/pan-11
22. Stamatatos, E.: Intrinsic plagiarism detection using character *n*-gram profiles. In: Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and, Social Software Misuse (PAN-09). pp. 38–46. CEUR Workshop Proceedings, volume 502 (2009), http://ceur-ws.org/Vol-502/

23. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009), http://dx.doi.org/10.1002/asi.21001
24. Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic plagiarism analysis. Language Resources and Evaluation 45(1), 63–82 (2011)
25. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics 37(2), 267–307 (2011)
26. Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21, 315–346 (2003)