Unsupervised Anomaly Detection

David Guthrie, Louise Guthrie, Ben Allison, Yorick Wilks
University of Sheffield
Natural Language Processing Group
Regent Court, 211 Portobello St., Sheffield, England S1 4DP
{dguthrie, louise, ben, yorick}@dcs.shef.ac.uk

Abstract

This paper describes work on the detection of anomalous material in text. We show several variants of an automatic technique for identifying an 'unusual' segment within a document, and consider texts which are unusual because of author, genre [Biber, 1998], topic or emotional tone. We evaluate the technique using many experiments over large document collections, created to contain randomly inserted anomalous segments. In order to successfully identify anomalies in text, we define more than 200 stylistic features to characterize writing, some of which are well-established stylistic determiners, but many of which are novel. Using these features with each of our methods, we examine the effect of segment size on our ability to detect anomaly, allowing segments of size 100 words, 500 words and 1000 words. We show substantial improvements over a baseline in all cases for all methods, and identify the method variant which performs consistently better than others.

1 Introduction

Anomaly detection refers to the task of identifying documents, or segments of text, that are unusual or different from *normal* text. Previous work on anomaly detection (sometimes called novelty detection) has assumed the existence of a collection of data that defines "normal" [e.g. Allison and Guthrie, 2006; Markou and Singh, 2003], which was used to model the normal population, and methods were developed to identify data that differs significantly from this model. As an extension to some of that work, this paper describes a more challenging anomaly detection scenario, where we assume that we have no data with which to characterize "normal" language. The techniques used for this task do not make use of any training data for either the normal or the anomalous populations and so are referred to as unsupervised.

In this unsupervised scenario of anomaly detection, the task is to find which parts of a collection or document are most anomalous with respect to the rest of the collection. For instance, if we had a collection of news stories with one fictional story inserted, we would want to identity this fictional story as anomalous, because its language is anomalous with respect to the rest of the documents in the collection. In this example we have no prior knowledge or training data of what it means to be "normal", nor what it means to be news or fiction. As such, if the collection were switched to be mostly fiction stories and one news story then we would hope to identify the news story as anomalous with respect to the rest of the collection because it is a minority occurrence.

There is a very strong unproven assumption that what is artificially inserted into a document or collection will be the most anomalous thing within that collection. While this might not be true in the general case, every attempt was made to ensure the cohesiveness of the collections before insertion to minimize the chance of finding genuine, unplanned anomalies. In preliminary experiments where a genuine anomaly did exist (for example, a large table or list), it was comforting to note that these sections were identified as anomalous.

The focus of this paper is the identification of segments in a document that are anomalous. Identifying anomalies in segments is difficult because it requires sufficient repetition of phenomena found in small amounts of text. This segment level concentration steered us to make choices and develop techniques that are appropriate for characterizing and comparing smaller segments.

There are several possibilities for the types of anomaly that might occur at the segment level. One simple situation is an off-topic discussion, where an advertisement or spam is inserted into a topic specific bulletin board. Another possibility is that one segment has been written by a different author from the rest of the document, as in the case of plagiarism. Plagiarism is notoriously difficult to detect when the source of the plagiarism can be found (using a search engine like Google or by comparison to the work of other students or writers.) In addition, the plagiarized segments are likely to be on the same topic as the rest of the document, so lexical choice often does not help to differentiate them. It is also possible for a segment to be anomalous because of a change in tone or attitude of the writing. The goal of this work is to develop a technique that will detect an anomalous segment in text without knowing in advance the kind of anomaly that is present.

Unsupervised detection of small anomalous segments can not depend on the strategies for modeling language that are employed when training data is available. With a large amount of training data, we can build up an accurate characterization of the words in a document. These language-modeling techniques make use of the distribution of the vocabulary in a document and, because language use and vocabulary are so diverse, it is necessary to train on a considerable amount of data to see the majority of cases (of any specific phenomenon) that might occur in a new document. If we have a more limited amount of data available, as in the segments of a document, it is necessary to characterize the language using techniques that are less dependent on the actual distribution of words in a document and thus less affected by the sparseness of language. In this paper we make use of techniques that employ some level of abstraction from words and focus on characterizing style, tone, and classes of lexical items.

We approach the unsupervised anomaly detection task slightly differently than we would if we were carrying out unsupervised classification of text [Oakes, 1998; Clough, 2000]. In unsupervised classification (or clustering) the goal is to group similar objects into subsets; but in unsupervised anomaly detection we are interested in determining which segments are most different from the majority of the document. The techniques used here do not assume anomalous segments will be similar to each other: therefore we have not directly used clustering techniques, but rather developed methods that allow many different types of anomalous segments within one document or collection to be detected.

2 Characterizing Segments of Text

The use of statistical methods with simple stylistic measures [Milic, 1967, 1991; Kenny, 1982] has proved effective in many areas of natural language processing such as genre detection [Kessler *et al*., 1997; Argamon *et al*., 1998; Maynard *et al.*, 2001], authorship attribution [McEnery and Oakes, 2000; Clough *et al.*, 2002; Wilks 2004], and detecting stylistic inconsistency [Glover and Hirst 1996; Morton, 1978, McColly, 1978; Smith, 1998]. Determining which stylistic measures are most effective can be difficult, but this paper uses features that have proved successful in the literature, as well as several novel features thought to capture the style of a segment of text.

For the purposes of this work, we represent each segment of text as a vector, where the components of the vector are based on a variety of stylistic features. The goal of the work is to rank each segment based on how much it differs from the rest of the document. Given a document of *n* segments, we rank each of these segments from one to *n* based on how dissimilar they are to all other segments in the document and thus by their degree of anomaly.

Components of our vector representation for a segment consist of simple surface features such as *Average word and average sentence length, the average number of syllables per word*, together with a range of Readability Measures [Stephens, 2000] such as *Flesch-Kincaid Reading Ease, Gunning-Fog Index, and SMOG Index*, some vocabulary richness measures such as: *the percentage of words that occur only once, percentage of words which are among the most frequent words in the Gigaword newswire corpus* (10 years of newswire), as well as the features described below.

All segments are passed through the RASP (Robust and Accurate Statistical Parser) part-of-speech tagger developed at the Universities of Sussex and Cambridge. Words, symbols and punctuation are tagged with one of 155 part-of-speech tags from the CLAWS 2 tagset. We use this mark-up to compute features that capture the distribution of part of speech tags. The representation of a segment includes the

i) Percentages of words that are articles, prepositions, pronouns, conjunction, punctuation, adjectives, and adverbs

ii) The ratio of adjectives to nouns

iii) *Percentage of sentences that begin with a subordinating or coordinating conjunctions* (but, so, then, yet, if, because, unless, or…)

iv) *Diversity of POS tri-grams* - this measures the diversity in the structure of a text (number of unique POS trigrams divided by the total number of POS trigrams)

Texts are also run through the RASP morphological analyzer, which produces words, lemmas and inflectional affixes. These are used to compute the

i) Percentage of passive sentences

ii) Percentage of nominalizations.

2.1 Rank Features

Authors can often be distinguished by their preference for certain prepositions over others or their reliance on specific grammatical constructions. We capture these preferences by keeping a ranked list sorted by frequency of the:

i) Most frequent POS tri-grams list
ii) Most frequent POS bi-gram list
iii) Most frequent POS list
iv) Most frequent Articles list
v) Most frequent Prepositions list
vi) Most frequent Conjunctions list
vii) Most frequent Pronouns list

For each segment, the above lists are created both for the segment and for the complement of that segment. We use 1 – r (where r is the Spearman Rank Correlation coefficient) to indicate the dissimilarity of each segment to its complement.

2.2 Characterizing Tone

The General Inquirer Dictionary (http://www.wjh.harvard.edu/~inquirer/) developed by the social science department at Harvard, contains mappings from words to social science content-analysis categories. These content-analysis categories attempt to capture the tone, attitude, outlook, or perspective of text and can be an important signal of anomaly. The Inquirer dictionary consists of 7,800 words mapped into 182 categories with most words assigned to more than one category. The two largest categories are *positive* and *negative* which account for 1,915 and 2,291 words respectively. Also included are all *Har-*

vard IV-4 and *Lasswell* categories. We make use of these categories by keeping track of the percentage of words in a segment that fall into each category.

3 Method

All experiments presented here are performed by characterizing each segment as well as characterizing the complement of that segment (the union of the remaining segments). This involves constructing a vector of features for each segment and a vector of features for each segment's complement as well as a vector of lists (see Rank Features section) for each segment and its complement. So, for every segment in a document we have a total of 4 vectors:

V1 - feature vector characterizing the segment

V2 - feature vector characterizing the complement of the segment

V3 - vector of lists for all rank features for the segment

V4 - vector of lists for all rank features for the complement of the segment

We next create a vector of list scores called *V5* by computing the Spearman rank correlation coefficient for each pair of lists in vectors *V3 and V4*. (All numbers in *V5* are actually 1-Spearman rank coefficient so that higher numbers mean more different). We next sum all values in V5 to produce a **Rank Feature Difference Score**.

Finally, we compute the difference between two segments by taking the average difference in their feature vectors plus the Rank Feature Difference Score.

3.1 Standardizing Variables

While most of the features in the feature vector are percentages (% of adjectives, % of negative words, % of words that occur frequently in the Gigaword, etc.) some are on a different scale, such as the readability formulae. To test the impact of different scales on the performance of the methods we also perform all tests after standardizing the variables. We do this by scaling all variables to values between zero and one.

4 Experiments and Results

In each of the experiments below all test documents contain exactly one anomalous segment and exactly 50 "normal" segments. Whilst in reality it may be true that multiple segments are anomalous within a document, there is nothing implicit in the method which looks for a single anomalous piece of text; for the sake of simplicity of evaluation, we insert only one anomalous segment per document.

The method returns a list of all segments ranked by how anomalous they are with respect to the whole document. If the program has performed well, then the truly anomalous segment should be at the top of the list (or very close to the top). Our assumption is that a human wishing to detect anomaly would be pleased if they could find the truly anomalous segment in the top 5 or 10 segments marked most likely to be anomalous, rather than having to scan the whole document or collection.

The work presented here looks only at fixed-length segments with pre-determined boundaries, while a real application of such a technique might be required to function with vast differences between the sizes of segments. Once again, there is nothing implicit in the method assuming fixed-length sizes, and the choice to fix certain parameters of the experiments is to better illustrate the effect of segment length on the performance of the method. One could then use either paragraph breaks as natural segment boundaries, or employ more sophisticated segmentation techniques.

We introduce a baseline for the following experiments that is the probably of selecting the truly anomalous segment by chance. For instance, the probability of choosing the single anomalous segment in a document that is 51 segments long by chance when picking 3 segments is 1/51 + 1/50 + 1/49 or 6%.

4.1 Authorship Tests

For these sets of experiments we examine whether it is possible to distinguish anomaly of authorship at the segment level. We test this by taking a document written by one author and inserting in it a segment written by a different author. We then see if this segment can be detected using our unsupervised anomaly techniques.

We create our experimental data from a collection consisting of 50 thousand words of text written by each of 8 Victorian authors: Charlotte Bronte, Lewis Carroll, Arthur Conan Doyle, George Eliot, Henry James, Rudyard Kipling, Alfred Lord Tennyson, H.G. Wells.

Test sets are created for each pair of authors by inserting a segment from one author into a document written by the other author. This creates 56 sets of experiments (one for each author inserted into every author.) For example we insert a segment, one at a time from Bronte into Carroll and anomaly detection is performed. Likewise we insert segments one at a time from Carroll into Bronte and perform anomaly detection. Our experiment is always to test if we can detect this inserted paragraph.

For each of the 56 combinations of authors we insert 30 random segments from one into the other, one at a time. We performed 56 pairs * 30 insertions each = 1,680 sets of insertion experiments. For each of these 1,680 insertion experiments we also varied the segment size to test its effect on anomaly detection. We then count what percentage of the time this paragraph falls within the top 3, top 5, top 10, and top 20 segments labeled by the program as anomalous. The results shown here report the average accuracy for each segment size (over all authors and insertions).

The average percent of time we can detect anomalous segments in the top *n* segments varies according to the segment size, and as expected, the average accuracy increases as the segment size increases. For 1000 word segments, anomalous segment was found in the top 20 ranked segments about 95% of the time (81% in the top ten, 74% of the time in the top 5 and 66% of the time in the top three segments). For 500 word segments, average accuracy ranged from 76%

down to 47% and for 100 word segments it ranged from 65% down to 27%.

Top n segments	Percentage of the time found	Percentage of the time found (standardized features)	chance
Segment size: 100 words			
3	26.22	27.03	6.00
5	34.59	32.71	10.21
10	50	44.41	21.59
20	64.73	62.8	49.16
Segment size: 500 words			
3	47.49	43.94	6.00
5	51.9	51.88	10.21
10	59.71	64.1	21.59
20	71.62	76.38	49.16
Segment size: 1,000 words			
3	58.92	66.04	6.00
5	69.63	74.22	10.21
10	79.94	81.47	21.59
20	94.83	92.77	49.16

Table 1: Average Results for Authorship Tests

4.2 Fact versus Opinion

In another set of experiments we tested whether opinion can be detected in a factual story. Our opinion text is made up of editorials from 4 newspapers making up a total of 28,200 words.

Our factual text is newswire randomly chosen from the English Gigaword corpus [Graff, 2003] and consists of 4 different 50,000 word segments one each from one of the 4 news wire services.

Each opinion text segment is inserted into each newswire service one at a time for at least 25 insertions on each newswire. Tests are performed like the authorship tests using 3 different segment sizes.

Results in this set of experiments were generally better and the average accuracy for 1000 word segments, top 20 ranking was 91% (70% in the top 3 ranked segments.) Small segment sizes of 100 words also yielded good results and found the anomaly in the top 20, 92% of the time (although only 40% of the time in the top 3.

Top n segments	Percentage of the time found	Percentage of the time found (standardized features)	chance
Segment size: 100 words			
3	43.14	40.2	6.00
5	49.02	66.18	10.21
10	63.73	77.45	21.59
20	81.37	92.16	49.16
Segment size: 500 words			
3	40.2	38.24	6.00
5	66.18	51.47	10.21
10	77.45	68.14	21.59
20	92.16	88.73	49.16
Segment size: 1,000 words			
3	70	68	6.00
5	81	74	10.21
10	88	81	21.59
20	91	87	49.16

Table 2: Average results for fact versus opinion

4.3 Newswire versus Chinese Translations

In this set of experiments we test whether English translations of Chinese newspaper segments can be detected in a collection of English newswire. We used 35 thousand words of Chinese newspaper segments that have been translated into English using Google's Chinese to English translation engine. English newswire text is the same randomly chosen 50,000 word segments from the Gigaword corpus. Again, tests are run using the 3 different segment sizes.

For 1000 word segments, the average accuracy for detecting the anomalous document among the top 3 ranked segments is 93% and in the top 10 ranked segments 100% of the time.

Top n segments	Percentage of the time found	Percentage of the time found (standardized features)	chance
Segment size: 100 words			
3	43.14	31.37	6.00
5	52.94	33.33	10.21
10	68.63	56.86	21.59
20	74.51	68.63	49.16
Segment size: 500 words			
3	84.31	76.47	6.00
5	88.24	80.39	10.21
10	90.2	86.27	21.59
20	94.12	94.12	49.16
Segment size: 1,000 words			
3	92.86	89.29	6.00
5	96.43	92.86	10.21
10	100	92.86	21.59
20	100	96.43	49.16

Table 3: Average Results for Newswire versus Chinese Translations

5 Conclusions

There experimental results are very positive and show that, with a large segment size, we can detect the anomalous segment within the top 5 segments very accurately. In the author experiments where we inserted a segment from one author into 50 segments by another, we can detect the anomalous author's paragraph in the top 5 anomalous segments returned 74% of the time (for a segment size of 1000 words), which is considerably better than chance. If you randomly choose 5 segments out of 51 then you only have a 10.2% chance of correctly guessing the segment. Other experiments yielded similarly encouraging figures. The task with the best overall results was detecting when a Chinese translation was inserted into a newswire document, followed surprisingly by the task of detecting opinion articles amongst facts.

On the whole, our experiments show that standardizing the scores on a scale from 0 to 1 does indeed produce better results for some types of anomaly detection but not for all the tasks we performed. We believe that in all cases where it performed worse than the standard raw scores, were cases where the genre distinction was very great. We performed an additional genre test not reported in this paper where Anarchist's Cookbook segments were inserted into newswire. Many of the readability formulas, for instance, distinguish these genre differences quite well but their effects on anomaly detection are greatly reduced when we standardize these scores.

References

[Allison and Guthrie, 2006] Ben Allison and Louise Guthrie. *Detecting Anomalous Documents*, Forthcoming.

[Argamon *et al.*, 1998] Shlomo Argamon, Moshe Koppel, Galit Avneri. Routing documents according to style. In *Proceedings of the First International Workshop on Innovative Internet Information Systems* (IIIS-98), Pisa, Italy, 1898.

[Biber, 1988] Douglas Biber. *Variation across speech and writing*. Cambridge University Press, Cambridge, 1998.

[Clough, 2000] Paul Clough. *Analyzing style – readability*. University of Sheffield, http://ir.shef.ac.uk/cloughie/papers.html, 2000.

[Clough *et al.*, 2002] Paul Clough, Rob Gaizauskas, Scott Piao, Yorick Wilks. Measuring Text Reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, 2002

[Glover and Hirst 1996] Angela Glover and Graeme Hirst. Detecting stylistic inconsistencies in collaborative writing. In *Sharples, Mike and van der Geest, Thea (eds.), The new writing environment: Writers at work in a world of technology,* Springer-Verlag, London, 1996.

[Graff, 2003] David Graff. English Gigaword. Linguistic Data Consortium, catalog number LDC2003T05, 2003

[Kenny, 1982] Anthony Kenny. *The computation of style: An introduction to statistics for students of literature and humanities*, Pergamon Press, Oxford, 1982.

[Kessler *et al*., 1997] Brett Kessler, Geoffrey Nunberg, Hinrich Schütze. Automatic Detection of Text Genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 32-38, 1997.

[Maynard *et al.*, 2001] Diana Maynard, Valentin Tablan, Cristian Ursu, Hamish Cunningham, Yorick Wilks. Named Entity Recognition from Diverse Text Types. *Recent Advances in Natural Language Processing Conference*, Tzigov Chark, Bulgaria, 2001.

[Markou and Singh, 2003] Markos Markou and Sameer Sing. Novelty Detection: a review- parts 1 and 2. *Signal Processing,* 83(12):2481-2521, 2003.

[McColly, 1987] William B. McColly. Style and structure in the Middle English poem Cleanness. *Computers and the Humanities*, 21:169-176.

[McEnery and Oakes, 2000] Tony McEnery and Michael Oakes. Authorship Identification and Computational Stylometry. In *Robert Dale, Hermann Moisl and Harlod Somers (eds.) Hanbook of Natural Language Processing*, 545-562, Marcel Dekker, New York, 2000.

[Milic, 1967] Louis Tonko Milic. A quantitative approach to the style of Johnathan Swift. *Studies in English literature*, 23:317, The Hague: Mouton, 1967.

[Milic, 1991] Louis Tonko Milic. Progress in stylistics: Theory, statistics, computers. *Computers and the Humanities*, 25:393-400, 1991.

[Morton, 1978] Andrew Queen Morton. Literary detection: *How to prove authorship and fraud in literature and documents*. Bowker Publishing Company, Bath, England, 1978.

[Oakes, 1998] Mickael Oakes. *Statistics for Corpus Linguistics*, Edinburgh Textbooks in Empirical Linguistics, Edinburgh, 1998.

[Stephens, 2000] Cheryl Stephens. All about Readability. *Plain Language Network*, http://www.plainlanguagenetwork.org/stephens/readability.html, 2006.

[Smith, 1998] MWA Smith. The Authorship of Acts I and II of Pericles: A new approach using first words of speeches. *Computers and the Humanities*, 22:23-41, 1998.

[Wilks, 2004] Yorick Wilks. On the Ownership of Text. *Computers and the Humanities*. 38(2):115-127, 2004.