A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector.

Caroline Lyon, Ruth Barrett and James Malcolm

{c.m.lyon, r.barrett, j.a.malcolm} @herts.ac.uk

Computer Science Department, University of Hertfordshire

Abstract

Fundamental features of natural language can be exploited to produce an effective system for the automated detection of plagiarism and collusion. Independently written texts can be effectively identified as they have markedly different characteristics to those that include passages that have been fully or partially copied. This paper describes the implementation of the Ferret plagiarism and collusion detector, and its use in the University of Hertfordshire and other institutions. The difference between human and machine analysis is examined, and we conclude that an approach using machine processing is likely to be necessary in many situations.

Introduction

This paper examines the theoretical background to electronic detection of similar passages of text, and shows how machine processing followed by human scrutiny can be a most effective approach in many situations. We examine the underlying concepts, the implementation of an automated plagiarism detector, the difference between machine and human analysis, and see why an approach using machine processing is likely to be successful. The paper describes the use of the Ferret system within the University of Hertfordshire and other institutions, and practical issues that have been addressed.

Our discussion is mainly based on this plagiarism detector. It takes in a set of students' work, submitted electronically, and determines whether any members of this set are suspiciously similar to each other or to articles off the Web. This is a standalone local system designed to run on any lecturer's computer, and it requires no more technical expertise than the Turnitin system, which it complements. In effect, it compares each document with each other, and produces a ranked table of texts with a resemblance measure for each pair. Any pair of texts can be displayed side by side with similar passages highlighted. Passages do not have to match exactly: any similarity is picked up. Finally, human scrutiny is needed to decide whether matching passages indicate plagiarism or collusion, or whether, for instance, a source has been correctly cited and thus no offence has been committed [Lyon *et al*., 2001].

Characteristics of independently written texts compared to plagiarised texts

Any text can be characterised by the set of short word sequences of which they are composed, typically taken as three-word sequences or trigrams, as shown in Figure 1. This might be called a fingerprint, except that the set of trigrams is larger than the original document.

Fig 1 Example of decomposition into trigrams:

String of words:

plagiarism is a common problem in universities

Decomposed into trigrams:

plagiarism is a ***is a common*** ***a common problem***
common problem in ***problem in universities***

The operation of Ferret is based on the empirical fact is that independently written texts have a comparatively low level of matching trigrams: for texts of 1000 – 5000 words the proportion of matching trigrams is not more than 8%. This is the case even when the same person writes on a similar subject on different occasions. Experiments were carried out on the well-known Federalist Papers, an exhaustively analysed set of essays, the foundation of the American constitution. In this corpus the same subjects are addressed repeatedly, and we examined 81 texts. The aim of the experiment was to establish a threshold up to which independently written texts might resemble each other [Lyon *et al*., 2001]. Above this threshold copying or collusion is suspected.

The phenomenon of low levels of matching trigrams is the result of the characteristic zipfian distribution of words in English and other languages. A small number of words are common, but a significant number of words occur infrequently. [Shannon, 1951; Manning and Schutze, 1999]. For instance, in the Brown corpus of 1 million words, 40% of the words occur only once [Kupiec, 1992]. This characteristic is more marked for bigrams and even more pronounced for trigrams. This is illustrated by the statistics (taken from [Lyon *et al.*, 2001]) shown in Table 1, showing the predominance of unique trigrams. Note that even after 38 million words of the Wall Street Journal have been seen, a new article (even in this limited domain of financial journalism) will on average have 77% of its trigrams differing from those already in the corpus [Gibbon, 1997].

However, if there has been plagiarism or collusion a higher proportion of trigrams than expected will match.

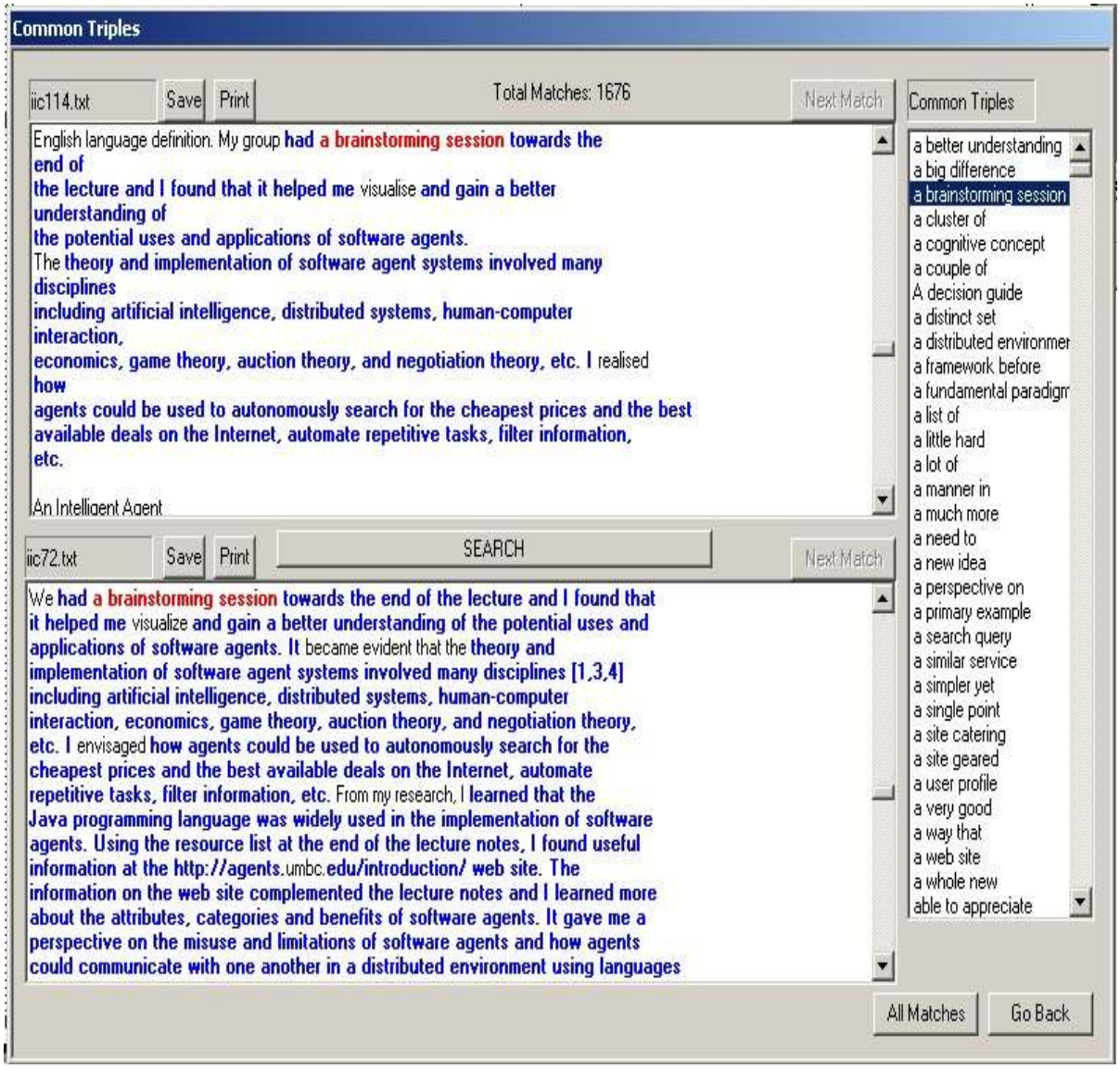
Table 1 Statistics from a TV news corpus, the Federalist papers and the Wall Street Journal corpora:

Source	Number of words in corpus	Distinct trigrams	Unique trigrams (occur only once)	% of trigrams that are unique
TV News corpus	985,316	718,953	614,172	85%
Federalist Papers (part)	183,372	135,830	118,842	87%
Wall Street Journal [Gibbon, 1997, p258]	972,868 4,513,716 38,532,517	648,482 2,420,168 14,096,109	556,185 1,990,507 10,907,373	86% 82% 77%

Ferret is a spin off from research in Automated Speech Recognition, where the frequency of unique trigrams is a fundamental problem – the sparse data issue. But the phenomenon that is the bane of speech recognition systems can be turned on its head and exploited to detect copied text.

This characteristic distribution of trigrams is immediately apparent visually using the Ferret, where matching word sequences are highlighted. When two independently written documents are displayed side by side there will be scattered highlighted matching word sequences. However, if there has been plagiarism or collusion, then there are solid patches of matching text, possibly with insertions or deletions, but with an overall visual impact of blocks of similar text. Figure 2 gives an example, where the two documents are not identical, but most of the text is the same, despite some insertions and deletions (words that are not in bold).

Fig 2 Example of two pieces of work where students had colluded:



Implementing Ferret

The principle underlying the Ferret system is based on matching short strings of words, exploiting the non-linear distribution of words in English and other European languages. Each text is converted into its set of characteristic trigrams, and these are compared for each pair. A resemblance metric, based on set theoretic principles, is used. If the resemblance measure exceeds a certain level, copying is suspected.

For a cohort of student work, each text is compared to each other, and also to a limited number (50) of pages downloaded from the Web. Ferret is capable of handling 300 documents of 10,000 words each on a standard laptop or desktop computer. It can process files in .doc, .rtf, .pdf and .txt format. Files are converted to .txt, and figures are omitted. After the file comparison process is completed, a ranked table is produced, showing each pair of files. Then any pair can be selected and displayed side by side, as in Figure 2. Processing time is measured in seconds rather than minutes.

Field trials

As well as at our University the Ferret has been tried at the Joint Services Command Staff College, and at the University of Maastricht. It was demonstrated recently at the Natural Computing Applications Forum (January 2004). The Ferret was not included in the JISC trials [Bull 2001] as at that time it was still under development.

The three faculties involved in the University of Hertfordshire trials were Computer Science (94 and 106 papers), Business (485 papers) and Law (10, 11 and 21 papers). Collusion or plagiarism was found in 14 out of 106 texts in one of the Computer Science experiments, done on past work. In Business and Law the students were informed that the work would be submitted to a plagiarism detector and this is likely to have deterred students from submitting plagiarized work.

At the Joint Services Command Staff College 300 essays of 10,000 words each were analysed. No plagiarism or collusion was found. However, in this experiment it was found that some further files in the Microsoft .obd file format could not at present be processed.

The Ferret has been developed using the English language, but the developers were interested in whether the positive findings with the program could be replicated in another language. The Maastricht University contacted the developers to ask if they could try out the Ferret on condition that they shared their experience. It has worked equally well in Dutch.

At the Maastricht University three faculties were involved in the Ferret trials: Law, Health Sciences, and Psychology. In the first year students have to write a paper of five to six pages on a certain topic, the topic depending on the subject area. The number of papers submitted was: Law, 256, Health Sciences, 275, and Psychology, 31. The program ran smoothly in all three runs. Ferret produces a ranking of pairs of papers, which lists the number of matching trigrams and the resemblance measure. The lecturer can then choose a pair of files and display them next to each other.

The software can only guide the teacher to potential plagiarism; then academic judgement must be applied. After the teachers studied the paired papers, they decided that four pairs resembled each other to such a degree that plagiarism was assumed. After consulting the paper writers, the lecturers decided that in three cases plagiarism was found. Six students from the Faculty of Law were penalized according to the University rules and regulations. The lecturers involved were impressed by the user-friendliness and the speed of the output, and also produced some recommendations for improving the interface. These have been implemented in a newer version of the program.

Further details of experiments are described in [Lyon, Barrett and Malcolm, 2003], in which Ferret and Turnitin are compared. The algorithm underlying the Turnitin system is not known.

The Ferret system complements Turnitin. Ferret checks for collusion between students in a cohort, and a limited number of Web pages. It operates on the lecturer's desk, and returns results almost immediately. The lecturer has then to make a subjective decision on plagiarism or collusion as he or she compares similar passages side by side. Turnitin in contrast has a very large database from the Web and other sources against which students' work can be matched. It has a good record of finding plagiarism effectively. However, it is not a local system, and Turnitin currently does not compare each file with each other in the cohort. Also, with Turnitin there are Data Protection issues that are absent with the Ferret.

Human and machine capabilities

Recent technical advances have coincided with great increases in numbers of students, so that classes of 200-300 are common, and work has to be marked by more than one person. Only through automated systems can work be checked for collusion.

Automated methods of plagiarism detection become more necessary as the pervasive influence of the Web has its effect on the garnering of material for reports. Some simple plagiarism detection can be undertaken by manually searching the Web, but Turnitin or the Ferret Web search facility can have the returned pages compared automatically with the students' work.

However, it is not only because of increasing numbers of students and access to the Web that electronic detection is good practice. Even when the quantity of work is small, machines can often detect copying more effectively than humans. The characteristic of copied text is that the number of matching word sequences is higher than expected. However, humans typically do not remember precise word sequences so much as semantic content. The lexical similarities that a machine can pick up may not strike a human, even if the number of pieces of work being compared is limited, as experimental work has shown [Wanner, 1974; Russell and Norvig, 2003]. For example, which of the following two phrases started this paper:

The theoretical background to plagiarism detection by electronic means is investigated in …..

This paper examines the theoretical background to the electronic detection of similar passages ……

Most readers will not recall. As humans, we remember the semantic content, rather than the exact sequence of words, on which plagiarism detection is based.

Technological advances have enabled many speech and language processing achievements, and among these are plagiarism detectors that would not have been possible a decade back. Processing unrestricted natural language on a personal computer has only become possible in recent years as computing power has expanded. The impact of new technology on the development of the Ferret plagiarism detector is evident both in its genesis as a spin off from work in automated speech recognition, and in its implementation. Field trials indicate that, to detect plagiarism or collusion in work that is submitted electronically, machine processing followed by human scrutiny is often the most effective approach.

References

Kupiec, J. (1992) 'Robust part-of-speech tagging using a Hidden Markov Model', *Computer Speech and Language*, 6, pp. 225-242.

Manning , C. & Schutze, H. (1999) *Foundations of Statistical Natural Language Processing*. MIT Press.

Shannon, C. (1993) 'Prediction and Entropy of printed English', in Shannon, C.E. *Collected Papers*, Sloane and Wyner (eds). IEEE Press.