Guess again and see if they line up: Surrey's runs at plagiarism detection

Notebook for PAN at CLEF 2013

Lee Gillam

Department of Computing, University of Surrey, UK
l.gillam@surrey.ac.uk

Abstract. This paper briefly describes the approaches taken to the two subtasks of Source Retrieval and Text Alignment, in the Plagiarism Detection track at PAN 13. For the first of these, we reuse our PAN12 approach – which combines frequency and a contrastive corpus measure to select keywords for querying the ChatNoir search system; for the second, we reuse software that had previously featured in PAN11 and PAN12. We comment on how effective both approaches were, and what steps should be taken if the competition remains substantially similar next time.

1 Introduction

The PAN activity first appeared in 2007. The external detection part of the plagiarism detection task in PAN changed markedly in 2012 from prior moderate-sized index comparison to two tasks: i) a retrieval of documents from a search engine as might be useful in match; ii) a matching between given pairs of documents. Offering a search engine for the first of these avoids the need for those who have struggled to construct an efficient index with a few gigabytes of text to struggle further with terabytes. However, common search engines work best for plagiarism detection with long quoted phrases, and the same cannot yet be said of the offered search system.

In PAN2012, the Candidate Document Retrieval subtask was specified as "retrieve a set of candidate source documents that may have served as an original to plagiarize from". In 2013, the title of the subtask changed to Source Retrieval, and a different specification was used: "retrieve all plagiarized sources". While it is clear that the former indicates a desire to deal with *coverage* of a suspect document, the latter seems to require comprehensive recall. The fact that the former was also intended by this latter description only became apparent later through email correspondence with the organizers, after quite some time had been spent looking for high recall – which would include duplicates. This, and an apparently serious instability of the search system during a key 3 day preparation period, in which various client side workarounds were attempted but ultimately not actually needed, subtracted significant effort from exploring the most effective approach. The combination offers one explanation for Surrey's low performance in Source Retrieval this year as efforts available became absorbed in such issues. The combination of these factors will likely discourage early participation next year, and it is vital that the (sub)task and system are both stabilized such that focus is on the (sub)task at hand; also so that it is possible to compare results year-on-year. Since it is also not clear whether the core of the search system is functionally identical to the previous year, participants cannot know whether what worked well should work equally well with the same data in the next year, and in contrast to alignment there is not a ready built means to measure performance with which to assess year-on-year similarity or determine better strategies. That is, unless each participant builds their own approach to doing this. This, however, relies on the existence of an evaluation framework that addresses "duplicates" rather than one which offers idealized results.

In this paper, we briefly outline the approaches taken at the University of Surrey to these two subtasks of Source Retrieval and Text Alignment in the Plagiarism Detection track at PAN 13. First, in section 2, we offer an overview of a project called IPCRESS, sponsored by the UK government-funded Technology Strategy Board and collaborative with a major automotive company, which at the time of writing was just starting up, but within which a private search style of operation is under construction and should be instructive for our future participation in PAN. In section 3, we reprise our use of a combination of frequency and a contrastive corpus measure to select keywords with which to make queries to the PAN search system, described more fully in our PAN 2012 paper (Gillam, Newbold and Cooke 2012). Section 4 gives a brief overview of re-used software from PAN 2011 and PAN 2012. Section 5 concludes with considerations for future work, and recommendations.

2 The IPCRESS project

In collaboration with Jaguar Land Rover and GeoLang Ltd and funding from the UK government-backed Technology Strategy Board for 18 months, the University of Surrey have formulated the **I**ntellectual Property **P**rotecting **C**loud Se**r**vic**e**s in **S**upply Chain**s** (IPCRESS) project to addresses industry barriers to Cloud adoption related to data security and resilience. The focus for IPCRESS is on the difficulty of entrusting valuable Intellectual Property (IP) to third parties, through the Cloud, as is necessary to allow for the construction of components in the supply chain – such information needs to be readily readable and usable by suppliers, and so encryption-based approaches become, at best, inconveniences. IPCRESS is developing the capability for tracking IP through supply chains, built around Surrey's private search approach to plagiarism detection which is suited to tracking IP without revealing IP (US patent filed November 2011; PCT filed November 2012). Such tracking is suited to the tasks of (i) preventing IP leakage; (ii) detecting IP leakage or theft; and (iii) identifying retention beyond allowed review periods Although at the time of writing the project is just weeks into starting, the inherent deperimeterisation of both Cloud and supply chains – which makes an 'insider' hard to define although loss of £9.2bn of IP in the UK is reportedly greatly assisted by such a person each year - means that the project already has significant user interest. Discussions around such a system have been presented previously (Cooke and Gillam 2011).

Timing of project activities should be interesting with respect to subsequent PAN iterations, provided that task stability is assured, and the alignment subtask has helped in identifying priorities for match – although the purpose of including random obfuscation remains unclear and is unlikely to be useful in our context.

3 Source Retrieval, the subtask formerly known as Candidate Document Retrieval

Candidate Document Retrieval at PAN 2012 involved creating a set of queries for a text that might retrieve texts from a search engine that match that text. The extent of match requires subsequent analysis post-retrieval. In PAN 2012, we offered enhanced weirdness (*ew*, eqn.1), obtained by squaring the relative frequency in our scaled weirdness equation (e.g. Gillam, Tariq and Ahmad, 2005):

$$ew = \frac{N_{GL} f_{SL}{}^{2}}{(1 + f_{GL}) N_{SL}{}^{2}} \qquad (1)$$

where f_{SL} is the frequency of a word in the (split) text, f_{GL} is its frequency in the 100m tokens of the British National Corpus (BNC), and N_{SL} and N_{GL} are the token counts of the (split) text and the BNC respectively. This was used in the approach briefly outlined below:

For each suspicious text, ***T***:

1. Split to sub-texts ***S*** by number of lines *l*.
2. For each sub-text in ***S***, generate queries ***Q*** by:
   a. Rank by ***ew***.
   b. Select the top 10 terms, and re-rank by frequency
   c. top frequency-ranked word paired with the next ***m*** words
3. Retrieve texts for each query in ***Q***.

As mentioned previously, in 2013 the title of the subtask changed to Source Retrieval and a different specification was used: "retrieve all plagiarized sources". The implication was that comprehensive recall was desirable –although the organizers had actually intended the task to be the same as in 2012, and so the change of title and specification seem unusual.

When the search system eventually became stable, for a subset of the training documents it was possible to find some 187 sources (l=10, m = 5) requiring some 5533 downloads. However, sufficient subsequent information regarding reporting of duplicates was unhelpful in determining strategy suitability. Having also discovered the different intent of the organizers, and now constrained by time, a strategy was selected which reduced the number of downloads substantially, but would likely sacrifice recall. Final results obtained were consistent with that. The approach is rapid in runtime, but ends up with recall figures massively different to those in 2012 (assuming these figures are even comparable, 0.1 vs 0.5567).

3 Detailed Comparison

In Cooke et al (2011) we described various aspects of our system as used for the external plagiarism detection task, which we stated could process the entire PAN11 collection within relatively short timescales, and which was still able to produce a reasonable degree of matching performance (4th place, with PlagDet=0.2467329, Recall=0.1500480, Precision=0.7106536, Granularity=1.0058894). In 2012, we showed how good granularity was achieved, with high recall and precision for non-obfuscated text, the approach achieves high precision, but lacks recall in the face of obfuscation.

Test	**Plagdet Score**	**Recall**	**Precision**	**Granularity**
02_no_obfuscation	**0.92530**	**0.90449**	**0.94709**	**1.0**
03_artificial_low	0.09837	0.05374	**0.93852**	**1.04688**
04_artificial_high	0.01508	0.00867	**0.96822**	**1.20313**
06_simulated_paraphrase	0.11229	0.05956	**0.97960**	**1.0**

In 2013, apart from for non-obfuscated data, descriptions of the nature of data used seem also to have shifted from the previous year. Our precision and granularity figures remain high, but it is difficult to conclude anything with regard to performance comparison for the other tasks – and prior examples of random obfuscation suggest that this is unlikely to be a realistic problem worth focusing on.

Test	**Plagdet Score**	**Recall**	**Precision**	**Granularity**
02_no_obfuscation	0.85884	0.83788	0.88088	**1.0**
03_random_obfuscation	0.04191	0.02142	**0.95968**	**1.0**
04_translation_obfuscation	0.01224	0.00616	**0.97273**	**1.0**
05_summary_obfuscation	0.00218	0.00109	**0.99591**	**1.0**

We also stated that we were unable to disclose too many details about the approach due to a patent application that was in progress. That patent was filed in the US (US13/307,428, 30th November 2011), and at the deadline a PCT filing was made (PCT/GB2012/000883, 30th November 2012). The combination of these two and the commercial opportunity under the IPCRESS project continues to preclude once more.

4 Conclusions

The PAN 2013 plagiarism detection task follows on in certain ways from the task last year. New difficulties emerged with use of the search system (a new API) that took time away from focusing on how to produce a better approach, and a new title and description proved ambiguous when the intention was that the task was identical to the previous year. Further, the evaluation framework of the alignment subtask and of

PAN 2011 has no equivalent for source retrieval. The combination of these and various other decisions apparently made on the fly during task runs offer distractions from the core of the task, and not being certain whether the system on which one is relying is functionally identical to last year does not inspire confidence in retaining or adapting around previously successful strategies. Stability is key for comparability - in being able to determine whether systems are improving year-on-year, or whether performance is about as good as it can get and there is little value in the safe implementation of well-worn approaches. Continuous modification, whilst perhaps technically interesting for those doing it, is not necessarily beneficial for research progress in the area, and it is hoped that the next iteration will involve just such a stability. A framework for evaluating results of source retrieval would also be inherently helpful, as would publically available descriptions of what constitutes a "duplicate" for the so-called oracle, since this seems to refer to duplication of a text in general rather than of specific content. We are keen to participate in strategy development, but keen also that this now becomes the focus.

As expected, our pairwise matching still suffers under obfuscation. With the advent of IPCRESS, this is an issue that we will be addressing during the next year. We are hopeful that recall will be improved by the approaches we are presently considering in our preferred direction of travel towards full-document (private) search – i.e. still without having to reveal the actual content of the documents being matched or using largely uniquely mappable surrogates (e.g. via MD5 hash).

Acknowledgements

The author gratefully recognizes prior contributions of Neil Newbold, Neil Cooke, Peter Wrobel and Henry Cooke to the formulation of the codebase used for this task and by Cooke and Wrobel to the patents generated from these efforts. This work has been supported in part by the EPSRC and JISC (EP/I034408/1) the UK's Technology Strategy Board (TSB, 169201), and also the efforts of the PAN13 organizers in system provision and managing the submissions.

References

Cooke, N. and Gillam, L., 2011, Clowns, Crowds and Clouds: A Cross-Enterprise Approach to Detecting Information Leakage without Leaking Information. In Mahmood, Z. and Hill, R. (eds.) Cloud Computing for Enterprise Architectures. Springer.

Cooke, N., Gillam, L., Wrobel, P., Cooke, H,, Al-Obaidli, F., 2011, A high performance plagiarism detection system. Proc. 3rd PAN workshop.

Gillam, L., Newbold, N. and Cooke, N., 2012, Educated guesses and equality judgements: using search engines and pairwise match for external plagiarism detection. Proc. 4th PAN workshop

Gillam, L., Tariq, M. and Ahmad, K., 2005, Terminology and the Construction of Ontology. Terminology 11(1), pp55-81. John Benjamins Publishing Company. ISSN 0929-9971; E-ISSN 1569-9994