Fourth International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse

Benno Stein
Martin Potthast
Paolo Rosso
Alberto Barrón-Cedeño
Efstathios Stamatatos
Moshe Koppel

Email: pan@webis.de

Abstract

The Fourth International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 10) was held in conjunction with the 2010 Conference on Multilingual and Multimodal Information Access Evaluation (CLEF-10) in Padua, Italy. The workshop was organized as a competition covering two tasks: plagiarism detection and Wikipedia vandalism detection. This report gives a short overview of the plagiarism detection task. Detailed analyses of both tasks have been published as CLEF Notebook Papers [3, 6], which can be downloaded at `www.webis.de/publications`.

1 Introduction

Altogether 18 groups from all over the world developed plagiarism detectors for PAN 10, which is 5 more than in the PAN 09 competition [5]; 5 groups attended for the second time. The effectiveness of the detectors was analyzed with the recently published evaluation framework, which consists of the PAN plagiarism corpus PAN-PC-10 and three performance measures [4]. During the construction of PAN-PC-10 a number of different parameters have been varied in order to create a high diversity of plagiarism cases. Table 1 gives an overview: the corpus is divided into documents suspicious of plagiarism and potential source documents. Note that only a subset of the suspicious documents actually contains plagiarism cases, and that for some cases the sources are unavailable. Also, the fraction of plagiarism per document and the document lengths have been varied. An important property of the plagiarism cases is their degree of *obfuscation*, which can be understood as a kind of paraphrasing to disguise the plagiarism attempt: plagiarists often rewrite their source passages in order to render the detection more difficult. Different paraphrasing strategies have been employed in the corpus including paraphrasing by machine translation from German and Spanish to English, automatic paraphrasing, and manual paraphrasing. Besides the obfuscation type also the length of the plagiarism cases was varied, as well as the fact whether or not the topic of a plagiarized document matches that of the source document.

In plagiarism detection one distinguishes between external and intrinsic detection situations: within the external situation the source document for a plagiarized document can

Table 1: Corpus statistics for 27 073 documents and 68 558 plagiarism cases in PAN-PC-10.

Document Statistics				
Document Purpose		*Plagiarism per Document*		
source documents	50%	hardly	(5%-20%)	45%
suspicious documents		medium	(20%-50%)	15%
– with plagiarism	25%	much	(50%-80%)	25%
– w/o plagiarism	25%	entirely	(>80%)	15%
Detection Task		*Document Length*		
external detection	70%	short	(1-10 pp.)	50%
intrinsic detection	30%	medium	(10-100 pp.)	35%
		long	(100-1000 pp.)	15%

Plagiarism Case Statistics	
Obfuscation	
none	40%
automatic	
– low obfuscation	20%
– high obfuscation	20%
manual	6%
translated ({de,es} to en)	14%
Case Length	
short (50-150 words)	34%
medium (300-500 words)	33%
long (3000-5000 words)	33%
Topic Match	
intra-topic cases	50%
inter-topic cases	50%

be found in a document collection D that is at the detector's disposal; within the intrinsic situation only the plagiarized document itself is given, and the detector looks for conspicuous writing style changes. The performance of a plagiarism detector is quantified by the well-known measures precision and recall, supplemented by a third measure called *granularity*, which accounts for the fact that detectors sometimes report overlapping or multiple detections for a single plagiarism case.

2 Selected Results

Figure 1 shows the detectors' precision and recall, depending on the paraphrasing strategy. In each plot the detectors are ordered with respect to their final rank, which has been computed over all tasks (see [3] for details). With respect to precision the detectors roughly divide into two groups: these with a high precision (> 0.7) and those without. The recall correlates with the ranking, whereas the top 3 detectors are set apart from the rest. Some detectors achieve a higher recall than their ranking suggests, for instance the detector of Muhr et al. [2], which outperforms even the winning detector.

Compared to PAN 09 the detectors have matured and specialized to the problem domain. However, most of the PAN 10 submissions focused on plagiarism detection in local document collections and could not be applied easily for plagiarism detection against the web. A novelty at PAN 10 was that many participants addressed cross-language plagiarism cases by automatically translating all non-English documents to English. Most of the retrieval models for candidate retrieval employed "brute force" fingerprinting instead of selecting few n-grams from a document—as it is done in near-duplicate detection algorithms like shingling and winnowing [1, 7]. With about 4.2 words the average n at PAN 10 compared to that of PAN 09; the winning approach used 5 words. Some participants put more effort into text preprocessing, e.g., by performing synonym normalization. Such and similar heuristics can be considered as "counter-obfuscation" heuristics. Fingerprinting cannot be applied easily when retrieving source candidates from the web, so some participants employed standard keyword

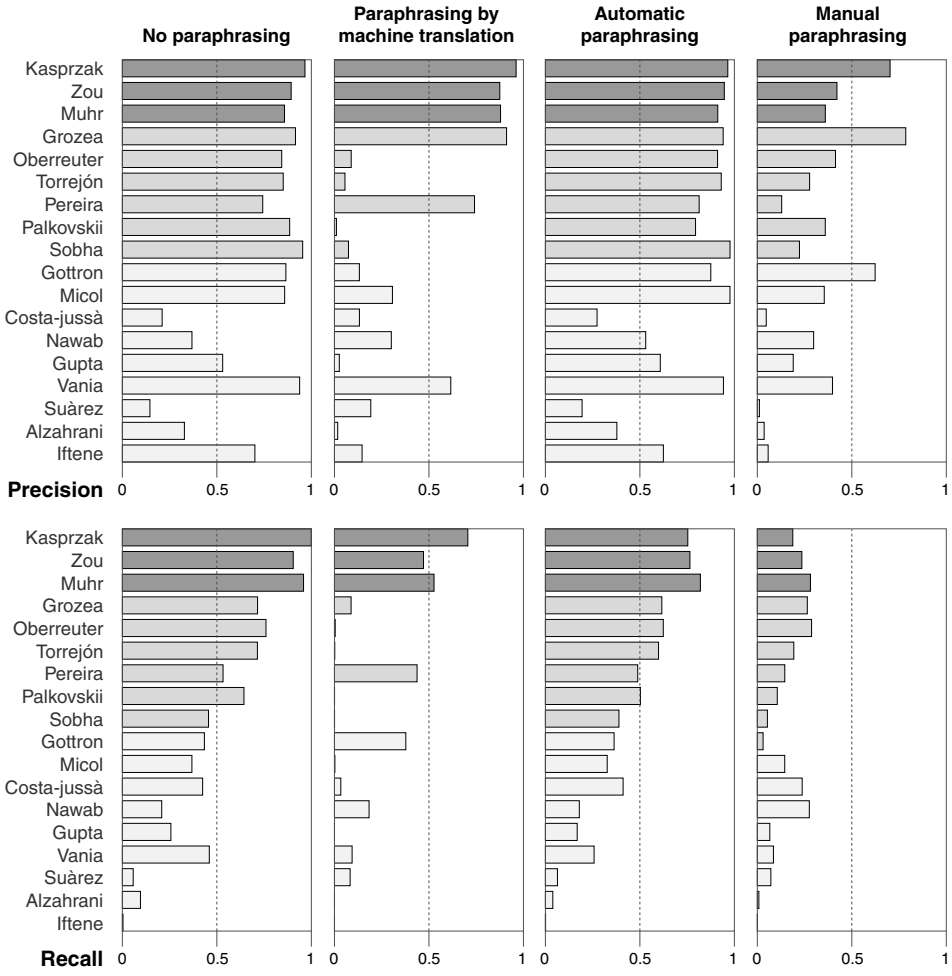

Figure 1: Plagiarism detection performance on the corpus PAN-PC-10.

retrieval technologies such as Lucene and Terrier. Among the submissions it was common practice to chunk the source documents and index the chunks rather than the documents to retrieve plagiarized document portions directly. Anyway, be it fingerprinting or keyword retrieval, the use of inverted indexing to speed up candidate retrieval was predominant at PAN 10; only few participants resorted to a simple comparison of all pairs of suspicious documents and source documents. With regard to the detailed analysis all participants applied sequence alignment heuristics; few noticed the connections to bioinformatics and image processing. In order to minimize the granularity some participants simply discarded overlapping detections with ambiguous sources.

3 PAN 11

PAN 11, again organized in conjunction with the CLEF conference, will cover—aside from plagiarism detection and Wikipedia vandalism detection—an author identification task. Throughout history and especially today, many texts are written anonymously or under pseudonyms, so that readers may not be certain of a text's alleged author. Within author

identification, one of the main challenges is to automatically attribute a text to one of a set of known candidate authors. For the purpose of the evaluation, we have developed a new authorship evaluation corpus. Further information as well as pointers to the previous PAN editions can be found under `pan.webis.de`.

Acknowledgements

Our special thanks go to the participants of the competition for their devoted work. We also thank Yahoo! Research for their sponsorship. This work is partially funded by CONACYT-Mexico and the MICINN project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

References

[1] Andrei Z. Broder. Identifying and Filtering Near-Duplicate Documents. In *COM'00: Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, pages 1–10, London, UK, 2000. Springer-Verlag.

[2] Markus Muhr, Roman Kern, Mario Zechner, and Michael Granitzer. External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System: Lab Report for PAN at CLEF 2010. In Martin Braschler and Donna Harman, editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*, September 2010.

[3] Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. Overview of the 2nd International Competition on Plagiarism Detection. In Martin Braschler and Donna Harman, editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*, September 2010.

[4] Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. An Evaluation Framework for Plagiarism Detection. In Chu-Ren Huang and Dan Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 997–1005, Beijing, China, August 2010. Association for Computational Linguistics.

[5] Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. Overview of the 1st International Competition on Plagiarism Detection. In Benno Stein, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors, *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pages 1–9. CEUR-WS.org, September 2009.

[6] Martin Potthast, Benno Stein, and Teresa Holfeld. Overview of the 1st International Competition on Wikipedia Vandalism Detection. In Martin Braschler and Donna Harman, editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*, September 2010.

[7] Saul Schleimer, Daniel S. Wilkerson, and Alex Aiken. Winnowing: local algorithms for document fingerprinting. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 76–85, New York, NY, USA, 2003. ACM Press.