White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/89031/

Version: Accepted Version

Proceedings Paper:
Clough, P.D., Willett, P. and Lim, J. (2015) Unfair Means: Use Cases beyond Plagiarism. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 6th International Conference of the CLEF Association, CLEF '15, 08-11 Sep 2015, Toulouse, France. Lecture Notes in Computer Science, 9283. Springer International Publishing, pp. 229-234. ISBN: 978-3-319-24026-8.

https://doi.org/10.1007/978-3-319-24027-5_21

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Unfair Means: Use Cases beyond Plagiarism

Paul Clough, Peter Willett and Jessie Lim

Information School, University of Sheffield, UK
{p.d.clough, p.willett}@sheffield.ac.uk

Abstract. The study of plagiarism and its detection is a highly popular field of research that has witnessed increased attention over recent years. In this paper we describe the range of problems that exist within academe in the area of 'unfair means', which encompasses a wider range of issues of attribution, ownership and originality. Unfair means offers a variety of problems that may benefit from the development of computational methods, thereby requiring appropriate evaluation resources. This may provide further areas of focus for large-scale evaluation activities, such as PAN, and researchers in the field more generally.

1. Introduction

Plagiarism[1] and its detection has been a popular area of research for the past 25 years, particularly within academia [1, 2, 4, 12]. Factors, such as the increased availability of freely-accessible digital content, the emergence of online essay banks and writing services, and technological developments are resulting in a rise in plagiarism, particularly in education[2]. However, within the educational context plagiarism is just one example of cheating, which may also include [14]: (i) *collusion*: collaboration among students; (ii) *falsification*: student presenting another work as his own; and (iii) *replication*: student submitting same work again (i.e., self-plagiarism) inter alia. Plagiarism is also not restricted to students, but has also surfaced amongst academics [9]. For example, Citron & Ginsberg [3] analyze text reuse within ArXiv.org and Errami et al. [5] identify duplication in PubMed abstracts. In addition, plagiarism can, and does, occur in documents beyond text[3]. In this paper we discuss the range of problems commonly encountered in Higher Education that extend beyond typical examples of plagiarism and that may provide further case studies for research into computational methods for determining authorship, attribution and originality.

2. Related Work

Two aspects commonly discussed in the literature are *intent* and *extent*. The former captures whether the (re-)use of existing sources was intentional or unintentional

[1] Joy and Luck [8] (p. 129) define plagiarism as *"unacknowledged copying of documents or programs"*

[2] http://www.bbc.co.uk/news/magazine-12613617 (site visited: 25/06/2015)

[3] https://www.plagiarismtoday.com/2015/03/05/plagiarism-is-more-than-just-text/ (site visited: 25/06/2015)

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

(e.g., the result of poor academic writing skills); the latter notion captures the extent or degree to which an original source is edited and modified, ranging from verbatim cut-and-paste to substantial rewriting and obfuscation (e.g., paraphrasing). Two further aspects commonly discussed are 'type' of plagiarism and detection methods.

2.1 Types of plagiarism

Several authors have attempted to categorize plagiarism into different forms or types [2, 7, 10, 12]. For example, Martin [10] lists several distinct forms of plagiarism: word-for-word plagiarism, paraphrasing plagiarism, plagiarism of secondary sources, plagiarism of the form of a source, plagiarism of ideas and plagiarism of authorship. These types get progressively harder to detect, both manually and automatically, as they go on. Harris [7] also categorizes different types of plagiarism, including buying a paper from a commercial paper mill, translating foreign language articles into English or another language, cutting and pasting text from several sources (patchwork plagiarism or *quilting*) and faking citations. Potthast et al. [11] propose a more formalized taxonomy of plagiarism types that also includes approaches to detect them. In this paper we provide a similar categorization scheme for unfair means, although at this stage less detailed and formalized than existing schemes.

2.2 Plagiarism detection

Detecting plagiarism and making decisions about originality is a human process; however, automated tools can assist with this process [4]. Various factors can signal plagiarism, such as inconsistencies in writing style, unexpected use of advanced vocabulary, incorrect references and shared similarities with existing materials. In discussing problems of *text attribution,* Wilks [13] describes four general problems: identifying inconsistencies within a text that suggest it is unlikely written by the claimed author, finding the likely sources of an unoriginal text, identifying collaboratively-written texts (i.e., collusion) and identifying copying between texts (i.e., plagiarism). The notions of *intrinsic* and *extrinsic* have also been used as plagiarism detection tasks at the PAN series of evaluation activities: intrinsic relates to identifying stylistic inconsistencies within a text; extrinsic relates to identifying the possible sources of a plagiarized document [6]. Alzahrani et al. [1] summarize the range of approaches commonly used to detect plagiarism, ranging from simpler lexical methods to more complex authorship- and semantic-based methods. Potthast et al. [11] also describe different types of approaches for producing exact and modified copies.

3. Types of Unfair Means Problems

As previously stated plagiarism, particularly in Higher Education, is one of many problems encountered within a wider area known as *unfair means* or *unfair practice*. This refers to a student attempting to gain advantage over another student in assessment, or assisting someone else to gain an unfair advantage or qualification. This

paper seeks to provide an initial review of the area, with the immediate aim of identifying the range of types of miscreant behavior that can occur. To this end, we have examined publicly available guidelines on academic misconduct provided by ten universities in the UK (Section 3.1), and conducted interviews with faculty involved in handling unfair means in departments in the University of Sheffield (Section 3.2).

3.1 Review of university guidelines

Careful reading of the academic misconduct guidelines for the universities of Bangor, Cambridge, Lincoln, Manchester, Northumbria, Sheffield, Swansea, Brunel, Hull and York has identified the types of behavior summarized in Table 1. It will be seen that they have been divided into two main types, depending upon whether they are practiced by students in a formal unseen examination context, or in, e.g., coursework assignments that are carried out in the student's own time. The list contains the most frequently mentioned behaviors, but is certainly not fully inclusive; for example, the Bangor guidelines consider the presentation of false evidence of extenuating circumstances to an examination board, or failing to obtain informed consent from participants in research projects as examples of unfair practice. Further examples of academic misconduct in one of the ten include the selling, distributing, website posting, or publishing information provided by instructors (e.g., lecture notes), or using them for any commercial purpose without permission of the instructor.

3.2 Interviews with staff in the University of Sheffield

Face-to-face interviews were conducted with the unfair means officers (hereafter UMOs) in ten departments that encompassed all five of the faculties (Arts & Humanities, Engineering, Medicine, Pure Science, and Social Science) in the University of Sheffield. The questions covered the responsibilities of their role, the procedures they followed when unfair means was suspected, the types of unfair means and of material with which they had to deal, the tools available to assist them, and the scale of the problem in their department. The interviews typically lasted about 30 minutes and were recorded for subsequent analysis. In a short paper such as this it is not possible to go into any great depth, and some of the responses are of only local interest but it is possible to draw some more general conclusions as to the type of problem that might benefit from the development of new types of computational support tool.

Straight-forward plagiarism was by far the most common type of occurrence and mentioned by all of the respondents, with collusion being the next-most common occurrence. All of the other types of behavior listed in Table 1 were mentioned at least once, with the sole exception of impersonation during an examination. Perhaps surprisingly, since it is arguably the most serious type of academic offence, the submission of bought or commissioned work was mentioned four times; that said, one respondent noted that this was very difficult to detect, with the implication that it might have happened more frequently than it had been identified. Also mentioned was what might be described as *translation plagiarism*: loading a plagiarized piece of coursework into an online translation program that converts the input English to an-

other language, and then back-translating the resulting text to yield a submission typical of that which might be expected from a student whose first language was not English. Several comments suggested that the use of unfair means in general was more common with such students.

Table 1 A summary of types of unfair means behaviour

Non-examination conditions	Plagiarism (either intentional or unintentional)	Copying text or images without acknowledging the source, passing off someone else's work or ideas as the author's own
	Double submission or self-plagiarism	Work may have been previously submitted for a different assessment
	Collusion	Submitting assessed work meant to be your own on which others have collaborated
	Fabrication of data /results	Making up research data, presentation or inclusion in a piece of work of figures or data which have been made up or altered and which have no basis in verifiable sources
	Falsification of data /results	Falsifying signatures of others, e.g. on consent forms or transcripts, misrepresentation of results
	Submitting bought or commissioned work	Submitting work which has been produced by someone else, e.g. another student, an essay bank or a ghost writing service
Examination conditions	Impersonation	Impersonation of a student during an examination or allowing oneself to be impersonated
	Cheating	Cheating in an examination by possessing or using materials prohibited in the examination room, copying from others (or communicating) during an exam, being in possession of notes or text books during exam, unfair use of electronic devices, presentation of an examination script as one's own work when the script includes material produced by unauthorised means including collusion

All of the UMOs dealt with cases of textual unfair means; examples of other types of material included software code, database or website designs, and architectural drawings (although the person mentioning this noted that it could only be detected manually by the person marking the student submissions). The Turnitin system developed by iParadigms is used in all departments as the principal tool to support UMOs in their work. This can only handle textual material; therefore, the MOSS (Measure Of Software Similarity) tool is used when software plagiarism has been suspected. As one would expect, frequent mention was made of the fact that tools such as these should only be used as a precursor to a detailed inspection by the UMO. There were several comments on the time-consuming nature of these inspections, with the implication that substantial benefits in terms of time and effort could be achieved if more effective tools could be developed; that said, much of the time requirement relates to the administrative procedures necessary to ensure that students are treated fairly when

the use of unfair means is suspected, especially in the case of more serious offences such as essay purchase or plagiarism in multiple assignments.

Specific types of example where new or improved tools could assist UMOs include translation plagiarism, the copying of images, cases of collusion where exactly the same material is presented but in different wordings, materials purchased from on-demand essay-writing services, and the citation of sources that on close inspection appear to have little or nothing to do with the content of the assignment. Although some of the activities at PAN deal with these issues (e.g., the plagiarism detection task at PAN@CLEF addresses intrinsic and extrinsic plagiarism detection and translation plagiarism; the author identification task at PAN@CLEF focuses on author identification and verification that is common in ghostwriting; and PAN@FIRE deals with source code plagiarism), there are still areas that could be explored in the area of unfair means as well as developing tools for UMOs. Understanding the domain and identifying areas for deploying new technologies are vital in developing realistic use cases to frame the development and evaluation of new tools.

4. Discussion

Further to our discussion of the findings in Section 3 further areas within Higher Education that may benefit from the use of computational methods include the following:

- **Identification of fabricated or falsified data/results**: for example identifying statistical anomalies within quantitative data, or identifying falsified documents, transcripts or language certificates.
- **Supporting 'proving' plagiarism**[4]: educators must be able to prove that unfair means has occurred. This could include, for example, developing techniques to compute deviations from 'normal' language distributions.
- **Citation and referencing analysis**: helping to identify fake (i.e., non-existent) citations, referencing inconsistencies or the use of incorrect references (i.e., references that do not match the context of the citation text).
- **Analysis of authorship style of contract services**: identifying whether coursework was likely produced by third-party services would be highly useful. This could include profiling the authorship style of commonly-used essay banks and online translation systems.
- **Plagiarism detection beyond text**: although much focus has been English text, there are many other forms of resource that are dealt with by UMOs. For example, non-English texts, program code, HTML and web pages, designs (e.g., database designs), images, drawings, presentations, and music.
- **Online learning environments**: increasingly institutions are offering distance learning courses and using online learning environments. This presents challenges around establishing the identity and authorship of students.
- **Discipline-specific plagiarism detection**: although there are elements of unfair means that are common across disciplines, there are clearly unique as-

[4] https://www.plagiarismtoday.com/2015/04/29/the-challenge-of-proving-plagiarism/

pects too that may require the use of bespoke tools. For example, plagiarism detection of laboratory notebooks within biomedical sciences or equations within mathematics.

5. Summary

This paper discusses the notion of unfair means in Higher Education, a wider issue than plagiarism that deals with various types of academic misconduct, including falsification and fabrication. Through a preliminary review of university guidelines and interviews with staff responsible for handling academic misconduct at the University of Sheffield, we highlight the range of problems encountered in Higher Education today. In the longer term, we hope that the study will encourage researchers to develop new computational tools that can assist in the detection not just of plagiarism, but also of the other types of unfair means. Future work will include developing a more detailed and formalized framework or taxonomy for categorizing unfair means.

References

1. Alzahrani, S. et al. 2012. Using structural information and citation evidence to detect significant plagiarism cases in scientific publications. *Journal of the Association for Information Science and Technology,* 63(2), 286-312.
2. Bretag, T., Mahmud, S. 2009. A Model for Determining student plagiarism: electronic detection and academic judgement. *Journal of University Teaching & Learning Practice*, 6(1). Available at: http://ro.uow.edu.au/jutlp/vol6/iss1/6
3. Citron D.T. and Ginsparg P. 2015. Patterns of text reuse in a scientific corpus. *PNAS*, 112(1), 25–30.
4. Culwin, F., Lancaster, T. 2001. Plagiarism issues for higher education. *Vine*, 31(2), 36-41.
5. Errami, M. et al. 2008. Déjà vu: A study of duplicate citations in Medline, *Bioinformatics*, 24(2), 243–249.
6. Gollub, T. et al. 2013. Recent Trends in Digital Text Forensics and Its Evaluation - Plagiarism Detection, Author Identification, and Author Profiling. In Proceedings of 4th International Conference of the CLEF Initiative, LNCS 8138, 282-302.
7. Harris, R.A. 2001. *The plagiarism handbook: Strategies for preventing, detecting, and dealing with plagiarism*. California: Eyrczak Publishing.
8. Joy, M., Luck, M. 1999. Plagiarism in programming assignments. *IEEE Transactions on Education*, 42(1), 129-133.
9. Lesk, M. (2015) How many scientific papers are not original? *PNAS*, 112(1), 6-7.
10. Martin, B. 1994. Plagiarism: a misplaced emphasis. *Journal of Information Ethics*, 3(2), 36-47.
11. Potthast, M. et al. 2011. Cross-language plagiarism detection. *Language Resources and Evaluation,* 45(1), 45-62.
12. Pupovac, V., Fanelli, D. 2014. Scientists Admitting to Plagiarism: A Meta-analysis of Surveys. Journal of Science and Engineering Ethics, 1-22.
13. Wilks, Y. 2004. On the Ownership of Text. *Computers and the Humanities*, 38(2), 115-127.
14. Wood, G., Warnken, P. 2004. Academic original sin: plagiarism, the Internet and librarians. *Journal of Academic Librarianship*, 30(3), 237–242.