Proving ownership: the case of ‘wag in a bag’

Paul Clough

http://ir.shef.ac.uk/cloughie/

Information School
University of Sheffield (UK)
How this talk came about

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
How this talk came about

Inspiration, Innovation, or Infringement: Multidisciplinary Perspectives on Piracy and Copyright - Emmanuel College

An interdisciplinary conference for invited scholars from a variety of disciplines to consider piracy and copyright infringement from the perspectives of their fields. Several lines of inquiry will be pursued: piracy and moral opprobrium; appropriation of "expressions" (as opposed to "ideas"); justifications for appropriation of expression (fair use; fair dealing). The fields represented include literature, music, history of the book, criminology, anthropology, information technology, comparative law, legal history, linguistics, and economics. Each non lawyer presenter will be paired with a legal commentator.

Workshop leaders: Professor John Crystal, Reader in Linguistics, University of Cambridge (Professor of Information Law; Professorial Fellow, Emmanuel College) and Dr Jennifer Doudna, Professor of Molecular and Cell Biology and Director of the Joint Center for Structural Genomics, University of Cambridge (Cambridge) and Professor Jane Tolmie, Director of the Centre for Intellectual Property and Information Law (CIPIL) at the University of Cambridge

For further details please contact:

Dr Aplin, Reader in Intellectual Property Law

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
On the Ownership of Text

YORICK WILKS
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, S1 4DP Sheffield, UK
E-mail: yorick@dcs.shef.ac.uk

Abstract. The paper explores the notions of text ownership and its partial inverse, plagiarism, and asks how close or different they are from a procedural point of view that might seek to establish either of these properties. The emphasis is on procedures rather than on the conventional subject division of authorship studies, plagiarism detection etc. We use, as a particular example, our research on the notion of computational detection of text rewriting, in the benign sense of a standard journalist’s adaptation of the Press Association newsfeed. The conclusion is that, whatever may be the case in copyright law, procedural detection and establishment of the ownership is a complex and vexed matter. Behind the paper is an unspoken appeal to return to an earlier historical phase, one where texts were normally
Outline

• Text reuse and plagiarism detection
• Past experiences with the Measuring Text Reuse (METER) project
• (Helping to) prove ownership
• Summary

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Text reuse and plagiarism
The notion of text reuse

• The activity whereby pre-existing written material is reused or recycled during the creation of a new text
  – Involves the rewriting of one text to create another
• Don’t have to start with editing an existing text; could include sub-conscious reuse
  – The point is that you can trace it back to specific source(s) which is important in the context of proving reuse
• As old as storytelling itself; but technology has caused unease in ownership (Wilks, 2004)
• Examples include summarisation, translation and the ‘classic’ case plagiarism
The notion of **text reuse**

- **From the author’s perspective**
  - “Reuse involves finding the relevant material, modifying it as needed and stitching the pieces together.” (Levy, 1993)

- **From the reader’s perspective can be cast as a text attribution problem**
  - “Given two texts is it possible to determine, within acceptable levels of certainty, whether one text is derived from the other?” (Wilks, 2004)


11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Text reuse involves *rewriting*

- **Basic rewriting operations**
  - Insertion of words
  - Deletion of words
  - Substitution of words
- **These enable changes, such as**
  - Restyling the texts, e.g. technical to non-technical
  - Re-ordering words within a sentence, or sentences within a discourse
  - Changes in tense and voice (e.g. passive to active voice)
  - Making abstract ideas more concrete and vice-versa
  - Merging or splitting sentences

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Original (news agency):
A Chief Constable's daughter who assaulted two officers in her father's force after drinking a litre of strong cider was today sentenced to 150 hours community service.

Rewrite (The Sun - popular press):
A Top Cop's daughter who assaulted two of her Dad's officers after downing a litre of cider was sentenced to 150 hours' community service yesterday.

Rewrite (The Independent - quality press):
The daughter of the Chief Constable of Sussex was sentenced to 150 hours' community service yesterday.
Examples of text reuse

Fragmentary Texts
Quotations and Text Re-uses of Lost Authors and Works

About

Fragmentary Texts is a project directed by Monica Berti and devoted to methodologies and tools for collecting and representing quotations and text re-uses of classical sources.

In the field of textual criticism, "fragments" are the result of a work of extraction and interpretation of information pertaining to lost works that is embedded in surviving texts. These fragments of information derive from a great variety of text re-uses that range from verbatim quotations to vague allusions and translations.

One of the main challenges when looking for traces of lost works is the reconstruction of the complex relationship between the fragment and its source of transmission. Pursuing this goal means dealing with three main tasks: 1) weighing the level of interference played by the author who has reused and transformed the original context of the fragment; 2) measuring the distance between the source text and the derived text; 3) trying to perceive the degree of text re-use and its effects on the final text.

Projects

demo.fragmentarytexts
Digital Athenaeus
Digital Humanities Leipzig
I Frammenti degli Storici Greci (DSHG)
Perspecta – Fragmentary Text Deriva
Perspecta – Fragmentary authors
The Digital Fragments Historionum. Stoczekan (DH12) Project
The Leipzig Open Fragmentary Texts Series (LOFTS)
Working with Text in a Digital Age

http://www.fragmentarytexts.org

Legislative origins of Obamacare*
Policy Ideas from previous bills, by party and chamber

*Patient Protection and Affordable Care Act (PPACA)

Economist.com/graphicdetail

http://www.economist.com/blogs/graphicdetail/2013/10/daily-chart-1

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
The notion of plagiarism

• Plagiarism is unacknowledged or unethical text reuse
  – Text reuse becoming easier (‘CTRL+C’ ‘CTRL+V’)
• Plagiarism is a “recent” term (1800’s)
  – We wouldn’t call Shakespeare a plagiarist even though he reused plots from Ovid

“Generally, borrowing is a tradition in literature and other art forms and more than a tradition: creativity feeds on what has gone before, new work is formed out of old.” During the renaissance and romantic eras of literary writing, even the “great” authors would reuse the ideas, storylines and plots of others in their own literary creations. It was not considered immoral or unethical; rather it was seen as a stimulus for creativity. Text reuse was (and is) the epitome of literary recognition.”
Angélil-Carter (2000:23)
The notion of plagiarism

• Plagiarism is unacknowledged or unethical text reuse
  – Text reuse becoming easier (‘CTRL+C’ ‘CTRL+V’)

• Plagiarism is a “recent” term (1800’s)
  – We wouldn’t call Shakespeare a plagiarist even though he reused plots from Ovid

• Probably most publicised form is student plagiarism
  – Includes plagiarism of software code as well as text

• Plagiarism detection has received considerable attention over the last 25 years (software code and natural language)

• In industry plagiarism is known as copyright infringement
  – “If plagiarism is the bane of the academic world, copyright infringement is the scourge of the legal one.” Osen (1997)

Forms of plagiarism

• Plagiarism can take several distinct forms (Martin, 1994)
  – **Word-for-word plagiarism**: direct copying of phrases or passages from a published text without quotation or acknowledgement.
  – **Paraphrasing plagiarism**: when words or syntax are changed (rewritten), but the source text can still be recognised.
  – **Plagiarism of secondary sources**: when original sources are referenced or quoted, but obtained from a secondary source text without looking up the original.
  – **Plagiarism of the form of a source**: the structure of an argument in a source is copied (verbatim or rewritten).
  – **Plagiarism of ideas**: the reuse of an original thought from a source text without dependence on the words or form of the source.
  – **Plagiarism of authorship**: the direct case of putting your own name to someone else’s work

Detecting plagiarism

• Multiple forms of plagiarism detection exist
• For a single text
  – Identify inconsistencies that indicate a text is unlikely written by the claimed author (*intrinsic*)
  – Find likely sources of plagiarised text (*extrinsic*)
• For multiple texts
  – Identify collaboratively-written texts (*collusion*)
  – Identify copying between texts (*detailed analysis*)


11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Manual plagiarism detection


11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Signals of plagiarism

- Common indicators of plagiarism in text include
  - Use of advanced or technical vocabulary beyond that expected of the writer
  - A large improvement in writing style compared to previous submitted work
  - Inconsistencies within the written text itself, e.g. changes in vocabulary, style or quality
  - Incoherent text where the flow is not consistent or smooth, which may signal that a passage has been cut-and-pasted from an existing electronic source
  - A large degree of similarity between the content of two or more submitted texts. This may include similarity of style as well as content
  - Shared spelling mistakes or errors between texts
  - Dangling references, e.g. a reference appears in the text, but not in the bibliography
  - Use of inconsistent referencing in the bibliography suggesting cut-and-paste
The goal of an automatic plagiarism detection system is to assist manual detection by

- Reducing amount of time spent comparing texts (makes comparison between large numbers of multiple texts feasible)
- Finding possible source texts from resources available to the system

The system must

- Minimise the number of false positives and false negatives
- Maximize the number of true positives and true negatives

“The task [of automatic plagiarism detection] may be simplified by finding a distinctive characteristic such as a misspelled identifier or paraphrased comment, though such a capability is hard to build into any automated plagiarism detection system” (Whale, 1990)
Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods

Salha M. Alzahrani, Naomie Salim, and Ajith Abraham, *Senior Member, IEEE*

<table>
<thead>
<tr>
<th>Technique</th>
<th>Tasks</th>
<th>IR</th>
<th>Language(s)</th>
<th>Plagiarism Type(s)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>extrinsic</td>
<td>intrinsic</td>
<td>mono-lingual</td>
<td>cross-lingual</td>
<td></td>
</tr>
<tr>
<td>Char-Based (CNG)</td>
<td>✓</td>
<td>✓</td>
<td>any</td>
<td>copy</td>
<td>[1-6]</td>
</tr>
<tr>
<td>Vector-Based (VEC)</td>
<td>✓</td>
<td>✓</td>
<td>any</td>
<td>copy, near copy</td>
<td>[7-11]</td>
</tr>
<tr>
<td>Syntax-Based (SYN)</td>
<td>✓</td>
<td>✓</td>
<td>specific</td>
<td>copy, near copy</td>
<td>[6, 12, 13]</td>
</tr>
<tr>
<td>Semantic-Based (SEM)</td>
<td>✓</td>
<td>✓</td>
<td>specific</td>
<td>copy, near copy</td>
<td>[14, 15]</td>
</tr>
<tr>
<td>Fuzzy-Based (FUZZY)</td>
<td>✓</td>
<td>✓</td>
<td>specific</td>
<td>copy, near copy</td>
<td>[16-19]</td>
</tr>
<tr>
<td>Structural-Based (STRUC)</td>
<td>✓</td>
<td>✓</td>
<td>specific</td>
<td>copy, near copy</td>
<td>[21, 29]</td>
</tr>
<tr>
<td>Stylometric-Based (STYLE)</td>
<td>✓</td>
<td>✓</td>
<td>specific</td>
<td>copy, near copy</td>
<td>[22, 23, 32-35]</td>
</tr>
<tr>
<td>Cross-Lingual (CROSS)</td>
<td>✓</td>
<td>✓</td>
<td>cross</td>
<td>copy</td>
<td>[31, 36-38]</td>
</tr>
</tbody>
</table>

The notions in the table indicate the following: ✓ means include/support by evidence from research stated in the references column, ☐ means possibility to include/support but need further research for proof.

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Recommended reading

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
The METER project

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Text reuse - where it all started for me...... the METER project

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Text reuse and churnalism

Checking for churn in:


'Silver spoons are ever more firmly clamped in the mouths of those who were born with them,' said Frances O'Grady, general secretary of the TUC, during her key note speech at the annual congress in Liverpool. Photograph: Lynne Cameron/PA The leader of the trade union movement has warned that Britain risks creating a "Downton Abbey-style" society in which social mobility has gone into reverse. Frances O'Grady told the TUC's annual congress that under the coalition blame for the country's ills had been heaped on the vulnerable while

Results:

2 News articles matching:

1. TUC congress: We are heading for a Downton Abbey-style society
   The Guardian, 6th September 2014
   92% cut, 73% pasted, 3923 characters overlap

2. TUC leader claims Conservative party is sliding Britain back to 'Downton Abbey...
   The Independent, 6th September 2014
   22% cut, 37% pasted, 922 characters overlap

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Text reuse in the news

• Example of daily reuse is the newswire-newspaper scenario
  – Newswires provide pre-fabricated source (called copy) for journalists
  – Newswires provide a critical role in news reporting
• In most cases text reuse completely legitimate and the norm
• Theoretical and practical interest
  – In how best to conceptualise the problem
  – In how best to detect and measure text reuse
Typical news production cycle


11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Editing text

- Common editing operations include
  - Insertion of new material
  - Deletion of unwanted text
  - Lexical substitution
  - Changes of syntax

- These enable changes, such as
  - Re-styling texts, e.g. from PA-speak to tabloidese
  - Re-ordering text, e.g. changing order of events
  - Changes in tense and voice (e.g. from active to passive)
  - Making abstract ideas more concrete (and vice-versa)

- Newspapers adopt house style (guidelines for writing)
Original (PA)  A drink-driver who ran into the Queen Mother’s official Daimler was fined 700 and banned from driving for two years.

Rewrite (The Times)  Eamon Reidy, 32, a drink-driver who rammed into Queen Elizabeth the Queen Mother’s Daimler was fined 700 and banned from driving for two years. *(Quality Press)*

Rewrite (The Sun)  A DRUNK driver who ploughed into the Queen Mother’s limo was fined 700 and banned for two years yesterday. *(Popular Press)*

Rewrite (The Mirror)  A BOOZY driver who smashed into the Queen Mums’s chauffeur-driven Daimler minutes after she had been dropped off was banned for two years and fined 700 yesterday. *(Popular Press)*

Rewrite (Daily Star)  A DRUNK driver who crashed into the back of the Queen Mum’s limo was banned for two years yesterday. *(Popular Press)*
Text reuse in the British Press

• The Press Association (PA): national newswire for UK and Ireland
  – Provides regional and national news to customers in Britain and abroad
  – Daily PA outputs 1,500 news, sport and feature stories
  – A pre-fabricated documentary source for journalists

• The PA forms a critical function for the British Press
  – Widely regarded as a credible, authoritative and trustworthy source

• PA text is widely reused
  – Directly: cut-and-paste or paraphrased
  – Indirectly: fact-checking, background and ‘copy tasting’

“News agencies provide most of the copy on any newspaper. Most agency news stories will run almost verbatim” (Bell, 1991)
Why measure text reuse?

• Like most newswires, the PA does not monitor the uptake or dissemination of their copy because
  – Lack of tools and technologies
  – Lack of conceptual framework

• Potential applications of reliably measuring reuse include
  – Monitoring uptake to identify unused or little-used stories
  – Identifying the most reused stories within the British media
  – Identifying dependent customers
  – Devising new methods of charging based on pay-per-use

• Large volume of text generated daily makes manual analysis infeasible
Identifying derived texts

WORLD OF WRITTEN-TEXTS

Press Conference

A

PA version

Times version

B

C

Daily Mail

Derivation relations:

(A,C) = derived
(B,C) = assumed to be independent
(A,B) = independent

Question:

Is C derived from A? (YES)
Is B derived from A? (NO)

Possible derivation relation
Non-derived thematic similarity relation

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Conceptualising the problem

• PA wanted to identify (likely) reuse of their copy
  – Distinguish cases of derived vs. non-derived
  – For derived cases attempt to identify cases where PA is used as the only source vs. when used amongst many

• Resulted in a simple ternary classification scheme driven by pragmatic concerns
  – Practical requirements of the PA
  – Ability of human annotators carrying out the task

• Newspapers classified at
  – Document level: coarse-grained indication of text reuse
  – Lexical-level: fine-grained indication of text reuse

• Ternary document-level categorisation
  – Wholly, partially or non-derived
Classification at the document level

• Document-level scheme consists of three relations between newswire-newspaper text pair:
  – **Wholly-derived (WD):** it is likely that the newswire text has been used to create the derived text and is the only source
  – **Partially-derived (PD):** it is likely that the newswire text has been used to create the derived text, but is one of many sources
  – **Non-derived (ND):** it is unlikely that the newswire text has not been used in the production of the derived text

• Judgments based on experience of trained journalists
  – Typically decision is first derived vs. not derived (*derivation*)
  – If derived then assessment of *degree* of text reuse (all or part)
Manually identifying reuse

• Key discriminators between derived and non-derived texts included
  – Differences between key facts (e.g. dates, names)
  – Order in which the story unfolds
  – Degree of lexical similarity and length of matching sequences
  – Existence of key facts in newspaper and not in PA

• Certain *differences expected* between derived texts
  – Those resulting from linguistic variations (e.g. register, tense)
  – Substitution of semantically-equivalent words/phrases
  – Re-ordering of news events
  – Also the application of the house style guide

• Certain *similarities expected* between non-derived texts:
  – Matches due to domain (e.g. commonly occurring phrases)
  – Direct and indirect quotes

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
The METER corpus

- Followed standard guidelines for creating representative corpus during construction (e.g. Atkins et al., 1992)
- Collection of 1,716 texts from PA and 9 British national dailies
  - Tabloids, middle-road tabloids and broadsheets
- Scope of corpus constrained to 2 domains
  - Law and court reporting (769 stories)
  - Showbiz (175 stories)
- Temporal extent of corpus constrained to 1 year
- Newspaper texts annotated with conceptual scheme
- Used to analyse text reuse and evaluate proposed algorithms

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Features identified as suitable discriminators of text reuse
- The degree of lexical similarity
- The length and distribution of matching verbatim sequences
- The existence of new information in the newspaper version
- The ordering of content between texts

Concentrated on three “simple” lexical approaches
- N-gram matching (plagiarism detection)
- Sequence comparison (sequence comparison)
- Sentence alignment (translation)

All approaches make use of minimal NLP
Provides initial baseline for further algorithm development

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Modelling text reuse

• All approaches capture similarities/differences
  – Used to automatically classify texts as WD, PD and ND
  – Assumption: longer verbatim matches and higher similarity indicate derivation
• N-gram matching used in plagiarism detection
  – Find matches between texts of length $N$ and measure similarity
• Sequence comparison (Greedy String Tiling)
  – Automatically find longest matching substrings between texts
  – e.g. used in biological sequence comparison and UNIX diff
• Sentence alignment used in translation (TESAS)
  – Treat newspaper as “translation” of newswire text
  – Automatically align sentences between texts


11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Classification task

• Similarities/differences used to discriminate text reuse
• Problem cast into a supervised learning problem
  – Where concept to be learned is one of derivation
  – Similarity/difference measures are attributes
  – Similarity/difference scores are attribute values
  – Each newswire-newspaper pair is an instance
  – Concept to learn for text pair is WD, PD or ND
• Automatic classification at document level allows
  – Finding extracted features which are good discriminators of reuse
  – Classify new cases and therefore the PA to practically quantify text reuse

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Some results

• Different methods give similar classification performance
• A combination of scores from each approach works best
  – Average of 70-75% accuracy for 3-way classification
  – Average of 80-90% accuracy for derived (WD+PD) vs. non-derived
• General observations
  – On average WD texts are easiest to classify
  – Most WD texts follow similar ordering to newswire
  – Most WD and ND instances misclassified as PD
  – Showbiz and tabloid texts exhibit more lexical variation
  – Limitations with using only “simple” methods and measures
  – Lexical overlap, match length and ordering all useful features
Visualising text reuse

This part will show self-reuse (of A)

PA duplicated text

This part will show any reuse between A and B

Quite obvious diagonals - implies verbatim reuse

• UNRELATED TEXTS
• NON-DERIVED TEXTS
• DERIVED TEXTS
Modelling rewriting

• Goal: can we model the edits between two texts?
• Simple approach based on combining Greedy String Tiling and dynamic programming (implementation of Unix Diff)
• Consider how PA text can be transformed into newspaper text using four simple edit operations (edit costs=1)
  – The *insertion* of tokens into the newspaper text
  – The *deletion* of tokens from the PA text
  – The *swap* of adjacent tiles
  – The *move* of tiles that are non-adjacent
• Result is edit script and quantities for edit operations
Modelling rewriting

Apply GST

[PA source]: Today the BBC sacked Noel Edmonds.
[Newspaper]: Noel Edmonds was sacked today by the BBC.

Apply Unix

INSERT (move): _noel_edmonds
INSERT: was-false
INSERT (move): _sacked
---match---[today]
INSERT: by-false
---match---[the BBC]
DELETE: _sacked
DELETE: _noel_edmonds

Diff

1. Noel Edmonds [was] Today the BBC sacked Noel Edmonds (insert “was”)
2. Noel Edmonds was [sacked] Today the BBC sacked Noel Edmonds (insert “sacked” - a move)
3. Noel Edmonds was sacked Today [by] the BBC sacked Noel Edmonds (insert “by”)
4. Noel Edmonds was sacked Today [by] the BBC Noel Edmonds (delete “sacked” – a move)
5. Noel Edmonds was sacked Today [by] the BBC (delete “Noel Edmonds” – a move)

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
(Helping to) prove ownership
Inspiration, Innovation, or Infringement: Multidisciplinary Perspectives on Piracy and Copyright - Emmanuel College

An interdisciplinary conference for invited scholars from a variety of disciplines to consider piracy and copyright infringement from the perspectives of their fields. Several lines of inquiry will be pursued: piracy and moral opprobrium; appropriation of "expressions" (as opposed to "ideas"); justifications for appropriation of expression (fair use; fair dealing). The fields represented include literature, music, history of the book, criminology, anthropology, information technology, comparative law, legal history, linguistics, and economics. Each non lawyer presenter will be paired with a legal commentator.

Workshop leaders: Professor Lionel Dently (Herschel Smith Professor of Intellectual Property Law; Director of the Centre for Intellectual Property and Information Law; Professorial Fellow, Emmanuel College, Cambridge), Dr Jennifer Davis (Centre for Intellectual Property and Information Law, University of Cambridge) and Professor Jane Ginsburg (Morton L. Janklow Professor of Literary and Artistic Property Law, Columbia Law School).

For further details please contact Geenor Moore
Copyright lawyer commented on technologies used in the METER project from a legal perspective
- Clear similarities/differences between text reuse and UK copyright law

Two areas of similarity with copyright law
- Notion of derivation
- Copying of substantial part (and ideas)

Notion of derivation
- Independently creating same or similar works not infringement
- Necessary to show that alleged infringement is copy or derived
- Debates on whether similarity substantial and beyond coincidence
- Similar to METER with respect to showing probable reuse

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Copyright law protects *expression* rather than ideas

Literal copying (similar to verbatim or ‘cut and paste’)
- Also recognises probable variants (e.g. likely rewrites)

Threshold (derived or not) occurs whether a *substantial part* of the original work has been copied
- Qualitative not quantitative decision (i.e. human judgment)
- Not just about amount of material ‘copied’ but also based also on the skill involved in creating original work

Role of text reuse technologies
- *Assisting* with proving copyright infringement
- Gathering use (or reuse) of copyrighted materials
In the case of METER

• Core parts of project were gaining an understanding of the domain and working with trained journalists to identify discriminating features between derived and non-derived

• Considering similarities
  – Similarities that indicate beyond coincidence relationship
  – Similarities expected even between independently written texts

• Reconciling the differences
  – What differences can be expected in the case of derived texts?

• But difficulties in capturing features identified manually

“The task [of automatic plagiarism detection] may be simplified by finding a distinctive characteristic such as a misspelled identifier or paraphrased comment, though such a capability is hard to build into any automated plagiarism detection system” (Whale, 1990)
Possible discriminators

Can use data such as Google n-grams and to model language and establish ‘norms’

http://googleresearch.blogspot.co.uk/2006/08/all-our-n-gram-are-belong-to-you.html

“Unless it is a very formulaic sentence (such as those appearing as part of a legal disclaimer at the beginning of a book), it is deeply unlikely that you will find it repeated in its exact form in any book, in any library, anywhere” (McEnery and Wilson, 1996:7).

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Possible discriminators

- Assumption: highly unlikely one will find matches above a certain threshold in common between texts unless derived
  - Do derived texts share more longer n-grams than non-derived?

<table>
<thead>
<tr>
<th>N-gram length (words)</th>
<th>Wholly-derived</th>
<th>Partially-derived</th>
<th>Non-derived</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>99.6%</td>
<td>99.7%</td>
<td>100%</td>
</tr>
<tr>
<td>3</td>
<td>99.2%</td>
<td>98.5%</td>
<td>97.6%</td>
</tr>
<tr>
<td>4</td>
<td>96.9%</td>
<td>96.2%</td>
<td>89.0%</td>
</tr>
<tr>
<td>5</td>
<td>90.1%</td>
<td>86.8%</td>
<td>72.7%</td>
</tr>
<tr>
<td>6</td>
<td>80.2%</td>
<td>75.1%</td>
<td>55.8%</td>
</tr>
<tr>
<td>7</td>
<td>74.8%</td>
<td>60.5%</td>
<td>36.4%</td>
</tr>
<tr>
<td>8</td>
<td>70.2%</td>
<td>50.9%</td>
<td>26.0%</td>
</tr>
<tr>
<td>9</td>
<td>65.2%</td>
<td>40.6%</td>
<td>20.6%</td>
</tr>
<tr>
<td>10</td>
<td>58.0%</td>
<td>36.3%</td>
<td>13.3%</td>
</tr>
</tbody>
</table>

Statistically Improbable Phrases

Statistically Improbable Phrases, Statimprophrases or SIPS constitute a system developed by Amazon.com to compare all of the books they index in the Search Inside! program and find phrases in each that are the most unlikely to be found in any other book indexed. The system is used to find the most nearly unique portions of books for use as a summary or keyword.
Possible discriminators

<table>
<thead>
<tr>
<th>GST feature</th>
<th>ND</th>
<th>PD</th>
<th>WD</th>
</tr>
</thead>
<tbody>
<tr>
<td>longest tile (words)</td>
<td>11.7 (11.00)</td>
<td>21.80 (17.51)</td>
<td>25.60 (27.65)</td>
</tr>
<tr>
<td>normalised longest tile</td>
<td>0.05 (0.04)</td>
<td>0.07 (0.05)</td>
<td>0.17 (0.11)</td>
</tr>
<tr>
<td>mean tile length (words)</td>
<td>1.58 (0.40)</td>
<td>2.01 (0.93)</td>
<td>3.56 (2.97)</td>
</tr>
<tr>
<td>containment</td>
<td>0.56 (0.14)</td>
<td>0.72 (0.14)</td>
<td>0.87 (0.10)</td>
</tr>
<tr>
<td>mean newspaper gap length</td>
<td>0.38 (0.20)</td>
<td>0.31 (0.17)</td>
<td>0.37 (0.18)</td>
</tr>
<tr>
<td>mean PA gap length</td>
<td>2.56 (1.07)</td>
<td>1.88 (0.83)</td>
<td>15.60 (0.38)</td>
</tr>
<tr>
<td>normalised newspaper longest gap</td>
<td>12.36 (14.90)</td>
<td>15.60 (15.90)</td>
<td>6.60 (29.65)</td>
</tr>
<tr>
<td>normalised PA longest gap</td>
<td>10.10 (6.70)</td>
<td>6.60 (6.00)</td>
<td>9.28 (1.70)</td>
</tr>
</tbody>
</table>

“Even between independently written texts one can expect to find up to 50% overlap” (Finlay, 1999)
## Reconciling differences

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution of synonyms</td>
<td>scalding water (boiling water)</td>
</tr>
<tr>
<td></td>
<td>passed classified information (transferred top-secret data)</td>
</tr>
<tr>
<td></td>
<td>admitted (pleaded guilty to)</td>
</tr>
<tr>
<td></td>
<td>charged with (accused of)</td>
</tr>
<tr>
<td>Use of abbreviations</td>
<td>British Broadcasting Corporation (BBC): <em>compression</em></td>
</tr>
<tr>
<td></td>
<td>DJ (Disc Jockey): <em>expansion</em></td>
</tr>
<tr>
<td></td>
<td>USA (America)</td>
</tr>
<tr>
<td>Temporal changes (time/date)</td>
<td>today (yesterday)</td>
</tr>
<tr>
<td></td>
<td>in 1998 (last year - article written in 1999)</td>
</tr>
<tr>
<td></td>
<td>earlier this year (in March)</td>
</tr>
<tr>
<td>Assumed/implied knowledge</td>
<td>Sheffield Wednesday Football Club (The Owls)</td>
</tr>
<tr>
<td></td>
<td>had drunk X pints of cider (was drunk)</td>
</tr>
<tr>
<td></td>
<td>the 13 – 19 year old (the teenager)</td>
</tr>
<tr>
<td></td>
<td>the Sheffield-born (the Yorkshireman, the Northerner)</td>
</tr>
<tr>
<td>Use of exaggeration</td>
<td>attacked (butchered, slaughtered)</td>
</tr>
<tr>
<td></td>
<td>the defendant said (the defendant insisted)</td>
</tr>
<tr>
<td></td>
<td>executed his lover’s husband (blasted his lover’s husband to death)</td>
</tr>
<tr>
<td></td>
<td>was cleared (was sensationaly cleared)</td>
</tr>
<tr>
<td>Use of pronouns for proper nouns</td>
<td>Johnnie Walker <em>(he)</em></td>
</tr>
<tr>
<td></td>
<td>Leanne and Damien <em>(they, the pair, the couple)</em></td>
</tr>
<tr>
<td></td>
<td>Smith, Jones and Brown <em>(the trio, the gang, the defendants)</em></td>
</tr>
<tr>
<td>Change of tense</td>
<td>he <em>is going</em> to court today <em>(he went</em> to court yesterday)</td>
</tr>
<tr>
<td></td>
<td>Jones <em>will pass</em> sentence today <em>(Jones passed</em> sentence at court yesterday)</td>
</tr>
<tr>
<td></td>
<td>Edmonds said he <em>will quit</em> the BBC <em>(Edmonds quitted</em> the BBC yesterday)</td>
</tr>
<tr>
<td>Nominalisation (verb/noun)</td>
<td>his defection <em>(defected)</em></td>
</tr>
<tr>
<td></td>
<td>X was charged <em>(dealing in cocaine)</em> <em>(a cocaine dealing charge</em> was given to X)</td>
</tr>
<tr>
<td>Change in word order</td>
<td>a policeman’s son <em>(son of a policeman)</em></td>
</tr>
<tr>
<td></td>
<td>charged with dealing in cocaine <em>(a cocaine dealing charge)</em></td>
</tr>
<tr>
<td></td>
<td>a routine vehicle check <em>(S4 - PA)</em> uncovered a cattle lorry that was</td>
</tr>
<tr>
<td></td>
<td><em>(S1 - PA)</em></td>
</tr>
</tbody>
</table>

Clough, P. Measuring text reuse (2003), PhD thesis, University of Sheffield, pp. 87-91
A drink-driver who ran into the Queen Mother’s official Daimler was fined 700 and banned from driving for two years.

Eamon Reidy, 32, a drink-driver who rammed into Queen Elizabeth the Queen Mother’s Daimler was fined 700 and banned from driving for two years. (Quality Press)

A DRUNK driver who ploughed into the Queen Mother’s limo was fined 700 and banned for two years yesterday. (Popular Press)

A BOOZY driver who smashed into the Queen Mums’s chauffer-driven Daimler minutes after she had been dropped off was banned for two years and fined 700 yesterday. (Popular Press)

A DRUNK driver who crashed into the back of the Queen Mum’s limo was banned for two years yesterday. (Popular Press)
Summary

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
The case of ‘Wag in a Bag’

- Limited amount of text for authorship analysis therefore hard to create profiles
- Most incriminating evidence was matching sequences of text not likely to have been authored independently
  - Likelihood of co-occurring text segments (i.e. n-grams) appearing in unrelated texts
  - Ordering of the sentences between texts
  - Timestamps of web pages indicating which one created first
  - Similarity of images in web pages

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
The case of ‘Wag in a Bag’

“Get the look of all your fave Celebs with this gorgeous new hair piece” – only occurs in 2 texts when searched online

**Result:** infringing website edited their text (but did not admit plagiarism!)

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Summary

• Text reuse is common activity and detection is an interesting research area

• Considered text reuse in the news domain and how derived texts can be manually discriminated
  – Improbable similarity
  – Probable differences

• Highlighted some example algorithms from different domains

• Simple techniques work well but lots of room for improvement

• Limitation: our understanding shaped by the quality and reliability of human judgments (largely intuition)
Future work

• User studies to better understand processes
  – Human rewriting process (e.g. paraphrasing)
  – Human judgment process (e.g. plagiarism detection)
• Incorporating semantics into the matching process (e.g. paraphrase detection and textual entailment)
• Developing techniques to assist manual detection and with proving text reuse (e.g. visualisations, language models)
• Initiate further collaborations between relevant groups, e.g. linguists, lawyers and computer scientists
• Thanks for PAN for providing evaluation resources and stimulating research activities

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Thank you

• Yorick Wilks
• Robert Gaizauskas
• Jonathan Foster
• John Arundel
• Scott Piao
• Ted Dunning

• Michael Oakes
• Patrick Juola
• Tanya Aplin

11th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2014)
Thank you

Questions?

p.d.clough@sheffield.ac.uk
Modelling text reuse

Textual string:

of all the gin joints in all the towns in all the world she walks into mine

TRIGRAMS:

of all the
the gin
joints in
in all the
gin joints in
joints in all
in all the
towns in
in all the
towns in all
towns in all
in all the
towns in all
in all the
world she
world she walks
she walks into
walks into mine
Modelling text reuse

Original (PA)  A drink-driver who ran into the Queen Mother's official Daimler was fined 700 and banned from driving for two years.

Rewrite (The Sun)  A DRUNK driver who ploughed into the Queen Mother's limo was fined 700 and banned for two years yesterday.

Source Text (ST)  Derived Text (DT)
SS1          DS1
SS2          DS2
SS3          DS3
SS4          DS4
SS5          DS5
SS6
SS7

3:1 mapping
1:1 mapping
1:n mapping

Derived Sentences (DS)

Source Sentences (SS)
Conclusions

• Evaluating search is very important both in academic and commercial contexts
• Evaluation often performed using test collections which provides valuable insights into IR algorithms
  – But need to validate the findings based on test collections with users and in realistic settings
  – System evaluation is part of wider evaluation activities
• ImageCLEF focused on system-oriented evaluation and inherits limitations
  – But created variety of realistic tasks and studied user interaction
• Future work considering evaluating wider IR applications (search is one component) and varying search strategies (e.g. browsing) using controlled lab-based experiments
http://users.dsic.upv.es/~lbarron/plagiarism.html