Text Reuse with ACL: (Upward) Trends

Parth Gupta and Paolo Rosso

Technical University of Valencia (UPV), Spain

July 10, 2012
Outline

Introduction
Experimental Setup
Analysis
Concluding Remarks
Acknowledgements

Parth Gupta and Paolo Rosso
Technical University of Valencia (UPV), Spain

Text Reuse with ACL: (Upward) Trends
Introduction

- Text reuse refers to using original text again in different work.
- Text reuse in its most general form has two types: *verbatim* and *modified*.
- There is a fuzzy line between text reuse and plagiarism and often this line is legislative.
- Usually the cases of valid or unvalid text reuse are better judged by human judges and detection systems assist the judge by providing potential cases.
Introduction contd..

- There are not straightforward measures to declare a case of text reuse as plagiarism and therefore, publishing houses deploy their own rules and definitions.

- For instance, IEEE and ACM both consider a case of reuse as plagiarism in case of:
  1. unaccredited reuse of text;
  2. accredited large portion of text without proper delineation or quotes to the complete reused portion.

- IEEE\(^1\) does not allow reusing large portion of own previous work, generally referred as self reuse or self plagiarism, without delineation, while ACM\(^2\) allows it provided the original source being *explicitly* cited.

\(^{1}\)http://www.ieee.org/publications_standards/publications/rights/ID_Plagiarism.html

\(^{2}\)http://www.acm.org/publications/policies/plagiarism_policy
Introduction contd..

- We do not do any kind of citation analysis.
- We try to capture the verbatim reuse of text that too in large amount.
Experimental Setup

▶ Catering the need of high reproducibility, we used a publicly available system rather than using any state-of-the-art system tested in PAN$^3$ (Uncovering Plagiarism, Authorship and Social Software Misuse) at CLEF.

WCopyFind$^4$

▶ Preprocess the text using user defined variables
  ▶ e.g. Ignore punctuation, letter case, numbers, etc
▶ Prepare the 32 bit hash codes of text units (n-grams)
▶ Highlight the matching chunks in the corresponding documents

$^3$http://pan.webis.de
$^4$WCopyFind is freely available under GNU public license at http://plagiarism.bloomfieldmedia.com/z-wordpress/software/wcopyfind/.
Version 4.1.1 is used.
Detection Method

- The WCopyFind displays the results based on the hashcode match between the target and source documents.
- Also generates a report file with matching number of words.

```
110  110  110  x.txt  y1.txt
111  111  111  x.txt  y2.txt
...
169  169  169  x.txt  y4.txt
1067 1067 1067  x.txt  y6.txt
199  199  199  x.txt  y7.txt
201  201  201  x.txt  y9.txt
```
Detection Method contd..

- High overlap of text between papers in reference section
- To avoid this overlap, a threshold of 500 words was chosen

- We manually analysed a small set of cases in order to check the reliability of the threshold

Parth Gupta and Paolo Rosso
Technical University of Valencia (UPV), Spain
Text Reuse with ACL: (Upward) Trends
Experimental Setup contd..

Dataset

- Type: Long, Short and Workshop papers
## Dataset contd..

<table>
<thead>
<tr>
<th>Year</th>
<th>Long</th>
<th>Short</th>
<th>Workshop</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>47</td>
<td>0</td>
<td>68</td>
<td>115</td>
</tr>
<tr>
<td>1994</td>
<td>52</td>
<td>0</td>
<td>56</td>
<td>108</td>
</tr>
<tr>
<td>1995</td>
<td>56</td>
<td>0</td>
<td>15</td>
<td>71</td>
</tr>
<tr>
<td>1996</td>
<td>58</td>
<td>0</td>
<td>73</td>
<td>131</td>
</tr>
<tr>
<td>1997</td>
<td>73</td>
<td>0</td>
<td>232</td>
<td>305</td>
</tr>
<tr>
<td>2007</td>
<td>131</td>
<td>57</td>
<td>340</td>
<td>528</td>
</tr>
<tr>
<td>2008</td>
<td>119</td>
<td>68</td>
<td>363</td>
<td>550</td>
</tr>
<tr>
<td>2009</td>
<td>121</td>
<td>93</td>
<td>740</td>
<td>954</td>
</tr>
<tr>
<td>2010</td>
<td>160</td>
<td>70</td>
<td>772</td>
<td>1002</td>
</tr>
<tr>
<td>2011</td>
<td>164</td>
<td>128</td>
<td>783</td>
<td>1075</td>
</tr>
</tbody>
</table>
Trend Analysis

We analyse three types of trends

1. Source of Text Reuse
   1.1 Text reuse in the papers from the past years papers as source
   1.2 Text reuse among the papers accepted in the same year
   1.3 The type of the papers involved in the reuse

2. The comparison with the past

3. Author analysis
At Present

1. Text reuse in the papers from the previous year submissions

Past 3 years as source

Immediate past year as source
At Present contd..

II. Text reuse in the papers among the same year submissions

![Graph showing the trend of text reuse from 2007 to 2012]

Parth Gupta and Paolo Rosso
Technical University of Valencia (UPV), Spain

Text Reuse with ACL: (Upward) Trends
At Present contd..

### III. Type of the papers involved in text reuse

**Past year papers as source**

**Same year papers as source**

![Graph for past year papers as source](image1)

![Graph for same year papers as source](image2)
In Retrospect

1. Text reuse in the papers from the previous year submissions

Past 3 years as source

Immediate past year as source
In Retrospect contd..

II. Text reuse in the papers among the same year submissions
In Retrospect contd..

III. Type of the papers involved in text reuse

Parth Gupta and Paolo Rosso
Technical University of Valencia (UPV), Spain
Text Reuse with ACL: (Upward) Trends
Relative Comparison

<table>
<thead>
<tr>
<th>Year</th>
<th>Tot. Cases</th>
<th>Tot. Accepted</th>
<th>% Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>1</td>
<td>115</td>
<td>0.87</td>
</tr>
<tr>
<td>1994</td>
<td>2</td>
<td>108</td>
<td>1.85</td>
</tr>
<tr>
<td>1995</td>
<td>0</td>
<td>71</td>
<td>0</td>
</tr>
<tr>
<td>1996</td>
<td>0</td>
<td>131</td>
<td>0</td>
</tr>
<tr>
<td>1997</td>
<td>15</td>
<td>305</td>
<td>4.92</td>
</tr>
<tr>
<td>2007</td>
<td>27</td>
<td>528</td>
<td>5.11</td>
</tr>
<tr>
<td>2008</td>
<td>22</td>
<td>550</td>
<td>4.00</td>
</tr>
<tr>
<td>2009</td>
<td>49</td>
<td>954</td>
<td>5.14</td>
</tr>
<tr>
<td>2010</td>
<td>70</td>
<td>1002</td>
<td>6.99</td>
</tr>
<tr>
<td>2011</td>
<td>104</td>
<td>1075</td>
<td>9.67</td>
</tr>
</tbody>
</table>

eventhough the difference between the number of papers accepted in the last three years is not so big, the same of the text reuse cases is.
Author Analysis

- Self reuse - at least one author is common in the papers involved in text reuse.
- Cross reuse - otherwise

<table>
<thead>
<tr>
<th>Reuse Type</th>
<th>No. of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self</td>
<td>232</td>
</tr>
<tr>
<td>Cross</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>249</td>
</tr>
</tbody>
</table>
Author Analysis contd..

- We analysed the frequency of a particular author being involved in text reuse - surprisingly follows Zipf’s law

![Graph showing frequency distribution of authors involved in text reuse]

- Statistics for span 2007-11
  - Total No. of authors = 8855
  - (Text reuse) At least once = 635 (<10%)
  - More than 5 cases = 80 (~1%)
Remarks

▶ These cases are reported based on the verbatim copy of the text in the ACL proceedings only. We did not aim to detect any text reuse that is paraphrased, which in reality can not be neglected.

▶ Including the other major conferences and journals of the field, the number of reported cases may increase

▶ Manual analysis revealed, in some cases, the related work section is completely copied from another paper

▶ In many cases, two papers shared a large portion of the text and differ mostly in the experiments and results
Remarks contd..

- Self reuse is more prominent in the ACL papers compared to the cross reuse even though the latter does not amount to zero.
- The ethicality and the acceptability of the self text reuse is arguable.
- Note that the aim of this paper is not to judge the acceptability of the text reuse cases but to advocate the need of such systems to help in the review process.
- Text reuse in the same year submissions turned out to be an eye opener because in such cases the text is novel but is used to publish in multiple formats and can stay unnoticed from the reviewers.
Remarks contd..

- In order to uphold the quality and the novelty of the work accepted in ACL, it is essential to implement a clear policy for text reuse and the technology to handle such reuse cases
- We hope this work will help the ACL research community to consider handling the text reuse for the upcoming editions
Thank You! 😊

Acknowledgements

- We thank Rafael Banchs for his valuable suggestions and discussions