Text Reuse with ACL: (Upward) Trends

Parth Gupta and Paolo Rosso

Technical University of Valencia (UPV), Spain

July 10, 2012

Outline

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 straight forward 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.

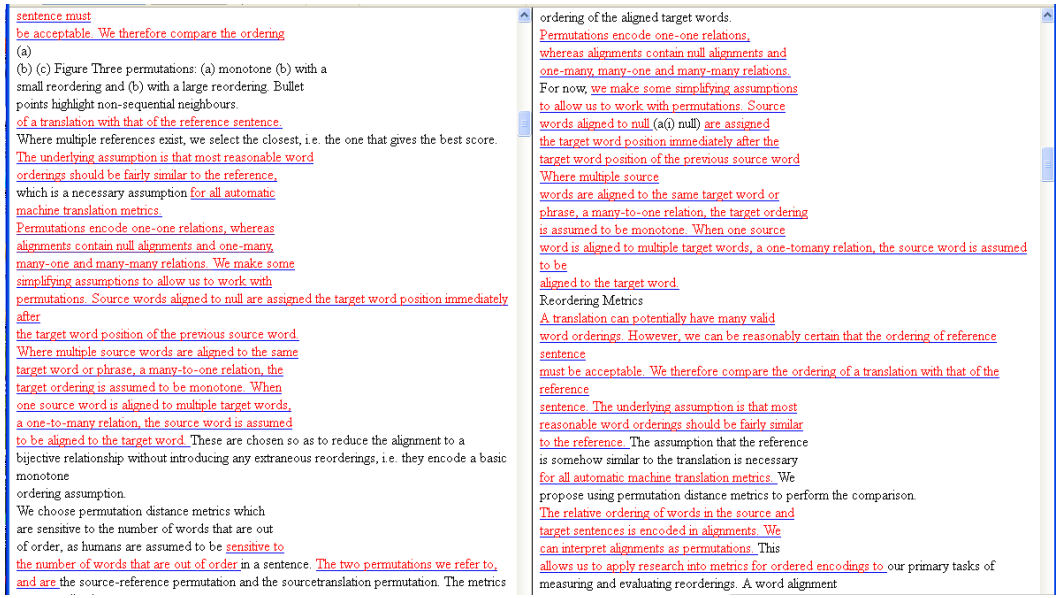
Detection Method

- 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

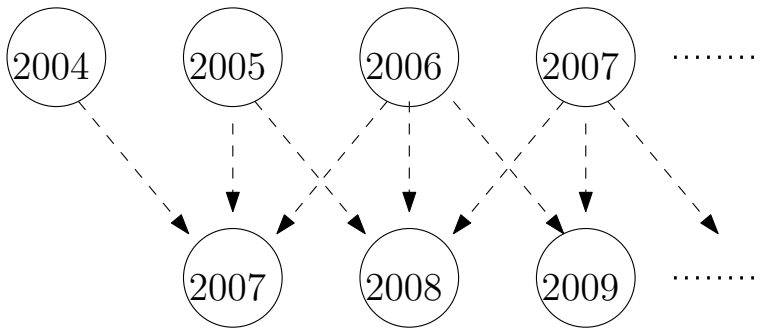

sentence must
be acceptable. We therefore compare the ordering
(a)
(b) (c) Figure Three permutations: (a) monotone (b) with a
small reordering and (b) with a large reordering. Bullet
points highlight non-sequential neighbours.
of a translation with that of the reference sentence.
Where multiple references exist, we select the closest, i.e. the one that gives the best score.
The underlying assumption is that most reasonable word
orderings should be fairly similar to the reference,
which is a necessary assumption for all automatic
machine translation metrics.
Permutations encode one-one relations, whereas
alignments contain null alignments and one-many,
many-one and many-many relations. We make some
simplifying assumptions to allow us to work with
permutations. Source words aligned to null are assigned the target word position immediately
after
the target word position of the previous source word.
Where multiple source words are aligned to the same
target word or phrase, a many-to-one relation, the
target ordering is assumed to be monotone. When
one source word is aligned to multiple target words,
a one-to-many relation, the source word is assumed
to be aligned to the target word. These are chosen so as to reduce the alignment to a
bijective relationship without introducing any extraneous reorderings, i.e. they encode a basic
monotone
ordering assumption.
We choose permutation distance metrics which
are sensitive to the number of words that are out
of order, as humans are assumed to be sensitive to
the number of words that are out of order in a sentence. The two permutations we refer to,
and are the source-reference permutation and the sourcetranslation permutation. The metrics
ordering of the aligned target words.
Permutations encode one-one relations,
whereas alignments contain null alignments and
one-many, many-one and many-many relations.
For now, we make some simplifying assumptions
to allow us to work with permutations. Source
words aligned to null (a(i) null) are assigned
the target word position immediately after the
target word position of the previous source word
Where multiple source
words are aligned to the same target word or
phrase, a many-to-one relation, the target ordering
is assumed to be monotone. When one source
word is aligned to multiple target words, a one-tomany relation, the source word is assumed
to be
aligned to the target word.
Reordering Metrics
A translation can potentially have many valid
word orderings. However, we can be reasonably certain that the ordering of reference
sentence
must be acceptable. We therefore compare the ordering of a translation with that of the
reference
sentence. The underlying assumption is that most
reasonable word orderings should be fairly similar
to the reference. The assumption that the reference
is somehow similar to the translation is necessary
for all automatic machine translation metrics. We
propose using permutation distance metrics to perform the comparison.
The relative ordering of words in the source and
target sentences is encoded in alignments. We
can interpret alignments as permutations. This
allows us to apply research into metrics for ordered encodings to our primary tasks of
measuring and evaluating reorderings. A word alignment

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

Experimental Setup contd..

Dataset

- Papers from ACL Anthology for years: 1990-97 and 2004-2011.
- Analysis Years: 1993-97 and 2007-11 - span of five years for the Past and Present.
- Type: Long, Short and Workshop papers



Dataset contd..

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

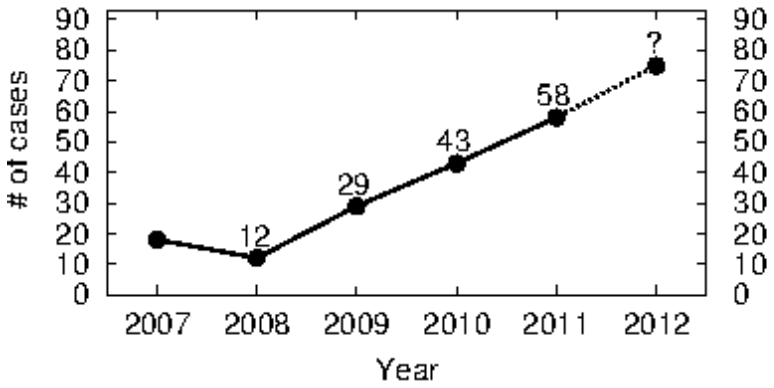
Trend Analysis

We analyse three types of trends

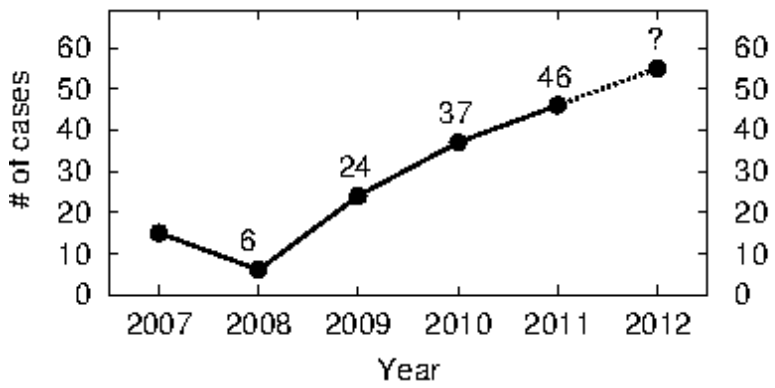
1. Source of Text Reuse
   - 1.1 Text reuse in the papers from the past years papers as source
   - 1.2 Text reuse among the papers accepted in the same year
   - 1.3 The type of the papers involved in the reuse
2. The comparison with the past
3. Author analysis

At Present

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



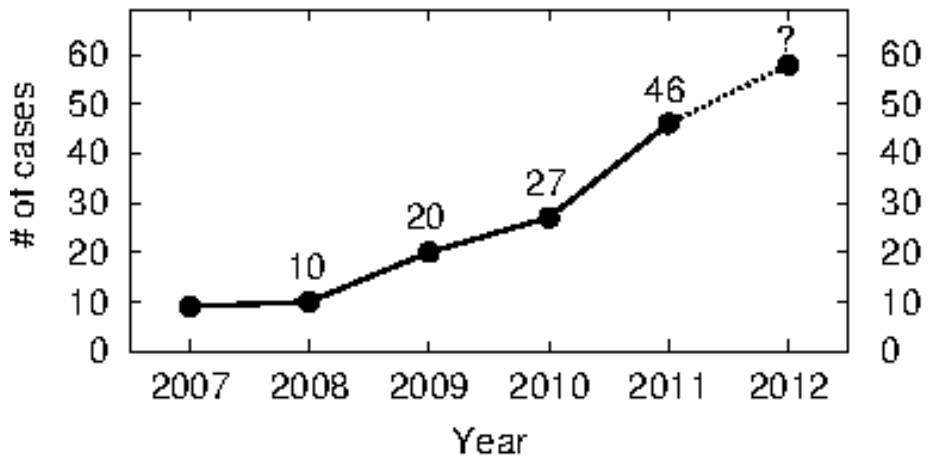
Past 3 years as source



Immediate past year as source

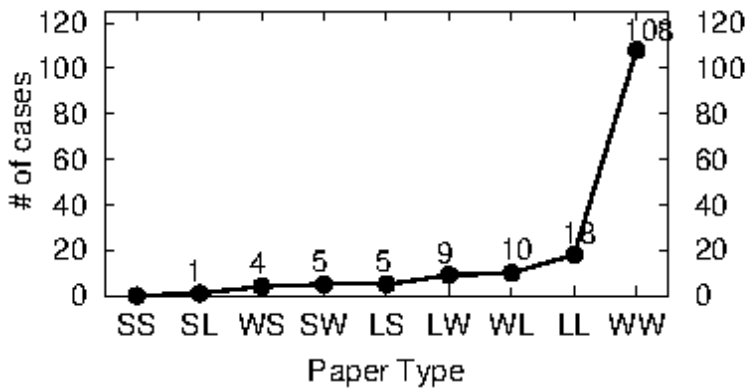
At Present contd..

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

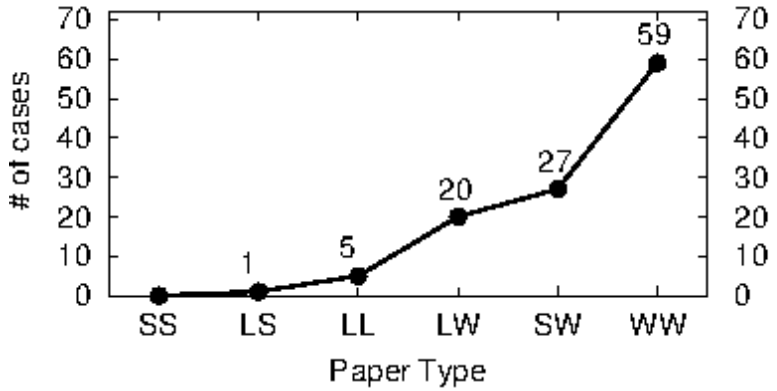


At Present contd..

III. Type of the papers involved in text reuse



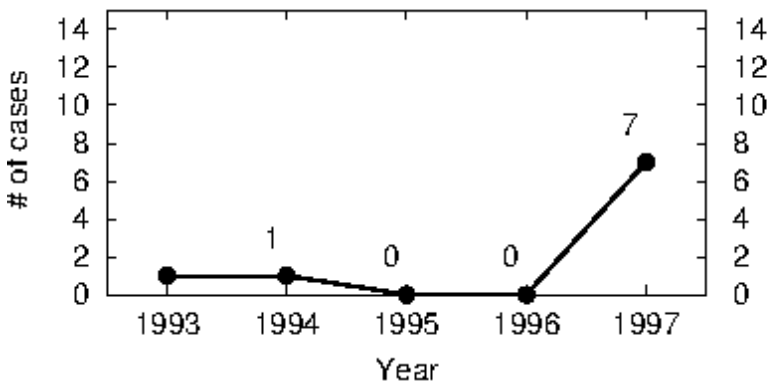
Past year papers as source



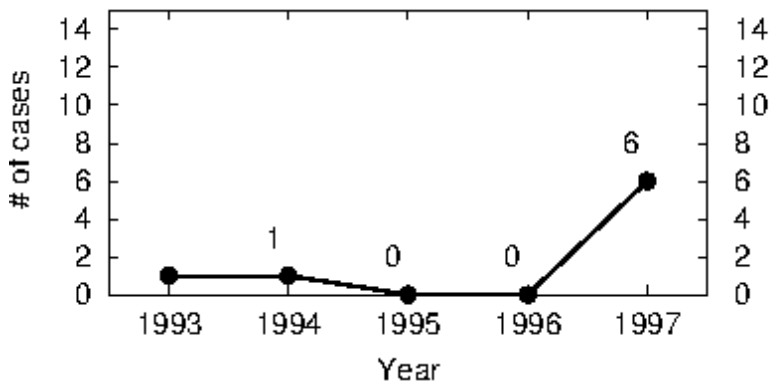
Same year papers as source

In Retrospect

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



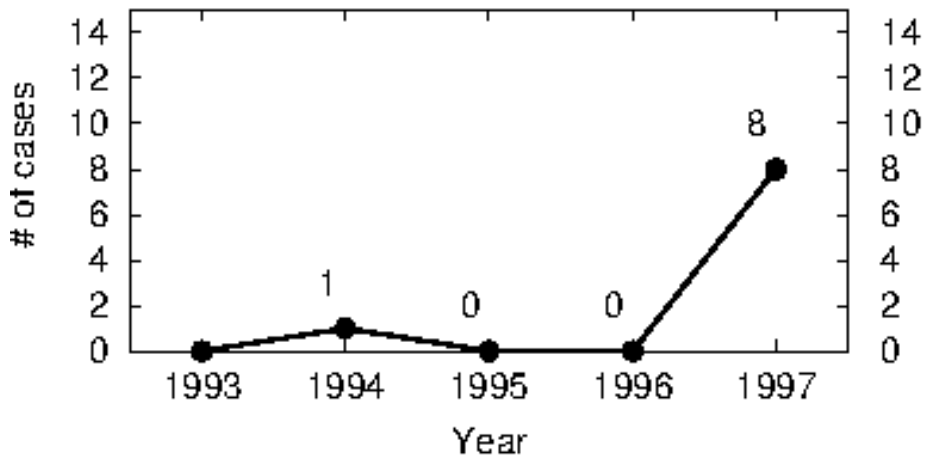
Past 3 years as source



Immediate past year as source

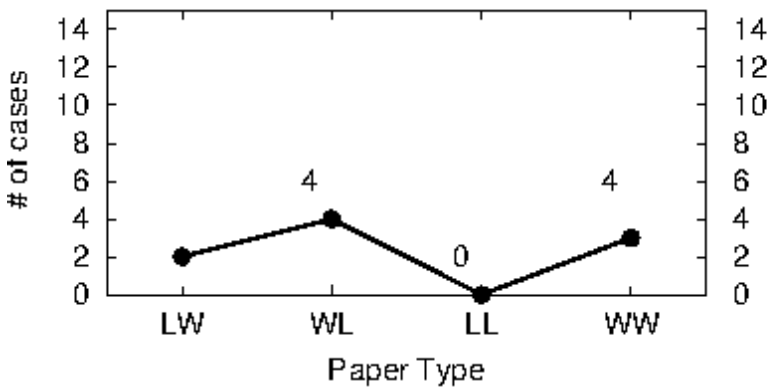
In Retrospect contd..

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

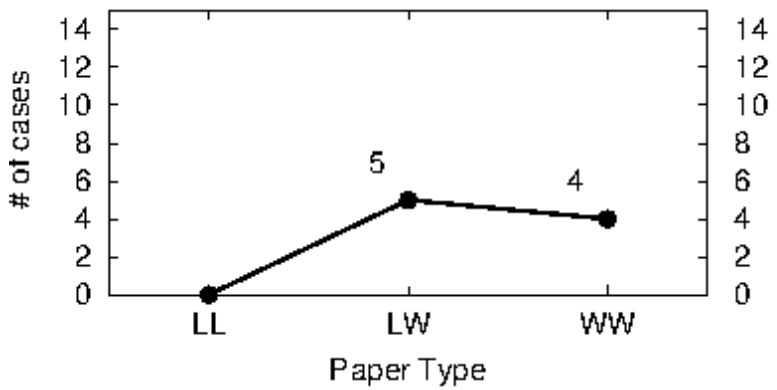


In Retrospect contd..

III. Type of the papers involved in text reuse



Past year papers as source



Same year papers as source

Relative Comparison

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

- 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.

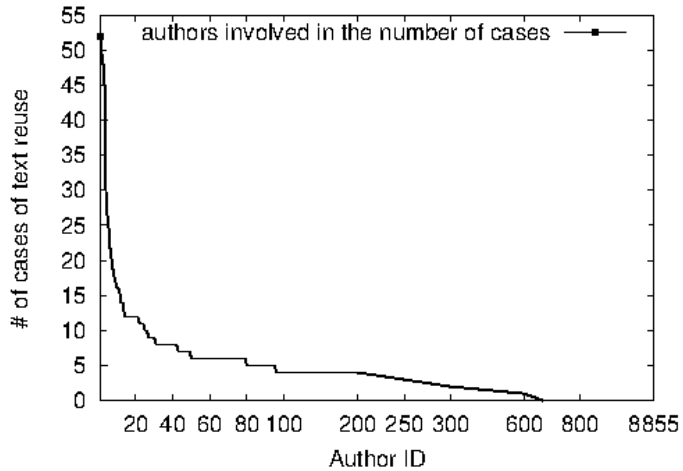
Author Analysis

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

Reuse Type	No. of Cases
Self	232
Cross	17
Total	249

Author Analysis contd..

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



- 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 eld, 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 eventhough 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

Tag cloud of Lab's publications