Overview of the 1st International Competition on Plagiarism Detection

Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso
Bauhaus-Universität Weimar & Universidad Polytécnica de Valencia

Outline

· Introduction
· Plagiarism Corpus
· Detection Performance Measures
· Competition on Plagiarism Detection
Introduction

- Plagiarism is ...
- To define plagiarism, you must first select a definition to plagiarize.
Introduction

- Plagiarism is ...
- To define plagiarism, you must first select a definition to plagiarize.
- “Plagiarism detection” refers to the automatic identification of plagiarism.
- Plagiarism detection divides into two problem classes:
  (a) External plagiarism detection.
  (b) Intrinsic plagiarism detection.
- The distinguishing property is the (un-)availability of a reference collection.

Introduction

Terminology:

- $d_q$ Suspicious document
- $d_x$ Source document
- $s$ Plagiarized section of text in a document
- $r$ Detection of plagiarized text in a document

A plagiarism case refers to $(d_q, d_x, s_q, s_x)$,
where $s_q \in d_q$, $s_x \in d_x$, and $s_q$ is the plagiarized version of $s_x$. 
Plagiarism Corpus

PAN Plagiarism Corpus 2009 (PAN-PC-09)

The PAN-PC-09 is a new large-scale resource for the controlled evaluation of plagiarism detection algorithms. [1]

Corpus overview:

- 41,223 text documents (obtained from 22,874 books from the Project Gutenberg [2])
- 94,202 plagiarism cases
- 70% is dedicated to external plagiarism detection,
  30% is dedicated to intrinsic plagiarism detection
- Types of cases: monolingual with and without obfuscation, and cross-lingual
- Authenticity of cases: real, emulated, and artificial


Plagiarism Corpus

Plagiarism Obfuscation Synthesis

Plagiarists often “modify” the text they plagiarize in order to obfuscate their offense.

- **Obfuscation synthesis task:**
  Given a section of text $s_x$, create a section $s_q$ which has a high content similarity to $s_x$ under some retrieval model but with a different word order or wording than $s_x$.

- **Optimal obfuscation synthesizer:**
  $s_x = “The quick brown fox jumps over the lazy dog.”$
  $s_q^* = “Over the dog which is lazy jumps quickly the fox which is brown.”$
  $s_q^* = “Dogs are lazy which is why brown foxes quickly jump over them.”$
  $s_q^* = “A fast bay-colored vulpine hops over an idle canine.”$

- **Obfuscation Synthesis Strategies:**
  (a) Random text operations
  (b) Semantic word variation
  (c) POS-preserving word shuffling
Plagiarism Corpus

Plagiarism Obfuscation Synthesis

Random text operations:
Given $s_x$, $s_q$ is created by shuffling, removing, inserting, or replacing words or short phrases at random.

Examples:

$s_x = “The quick brown fox jumps over the lazy dog.”$

$s_q = “over The. the quick lazy dog context jumps brown fox”$
$s_q = “over jumps quick brown fox The lazy. the”$
$s_q = “brown jumps the. quick dog The lazy fox over”$
Plagiarism Corpus

Plagiarism Obfuscation Synthesis

Semantic word variation:
Given $s_x$, $s_q$ is created by replacing each word by one of its synonyms, antonyms, hyponyms, or hypernyms, chosen at random.

Examples:

$s_x = \text{“The quick brown fox jumps over the lazy dog.”}$

$s_q = \text{“The quick brown dodger leaps over the lazy canine.”}$

$s_q = \text{“The quick brown canine jumps over the lazy canine.”}$

$s_q = \text{“The quick brown vixen leaps over the lazy puppy.”}$
POS-preserving word shuffling:
Given $s_x$ its sequence of parts of speech (POS) is determined. Then, $s_q$ is created by shuffling words at random while the original POS sequence is maintained.

Examples:

$s_x = \text{“The quick brown fox jumps over the lazy dog.”}$

POS = “DT JJ JJ NN VBZ IN DT JJ NN .”

$s_q = \text{“The brown lazy fox jumps over the quick dog.”}$

$s_q = \text{“The lazy quick dog jumps over the brown fox.”}$

$s_q = \text{“The brown lazy dog jumps over the quick fox.”}$
Plagiarism Corpus

Critical Remarks

- Accidental similarities between suspicious and source documents.
- Anomalies in the plagiarized text produced by the obfuscation synthesizers.
- Inaccurate simulation of Web retrieval.
Intrinsic / external ratio

30% for intrinsic plagiarism detection

70% for external plagiarism detection
Suspicious / source ratio

- 50% source documents
- 25% without plagiarism
- 25% with plagiarism
- 50% suspicious documents
Document length

50% short (1-10 pages)

35% medium (10-100 pages)

15% large (100-1000 pages)
Case length

- 250-750 chars (50-150 words)
- 1500-5000 chars (300-1000 words)
- 15000-25000 chars (3000-5000 words)
Fraction of plagiarism per document

θ: 5 75 50 25 100%

- External corpus
- Intrinsic corpus
Linguality

90% monolingual cases (English)

10% cross-lingual cases (German and Spanish)
Detection Performance Measures

Terminology

- $S_i \in S$: Plagiarized section from the set of all plagiarized sections.
- $r_i \in R$: Detected section from the set of all detected sections.
Detection Performance Measures

Micro-averaged Recall and Precision

- Micro-averaged recall and precision compute straightforward:
  \[
  \text{rec}_{PDA} = \frac{8}{13}, \quad \text{prec}_{PDA} = \frac{8}{16}
  \]

+ Simple to understand and simple to compute by counting char overlaps.
  - Rewards the detection of long sections which are typically easier to detect.
Macro-averaged Recall computes straightforward:

\[
rec_{PDA}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|s \cap \bigcup_{r \in R} r|}{|s|},
\]

where \(\cap\) computes the positionally overlapping characters.

But macro-averaged precision is undefined!
Detection Performance Measures

Macro-averaged Recall and Precision

Problem: Given \( s_i \), which \( r_i \in R \) are attempts to detect \( s_i \)?

Each \( s_i \) defines a query \( q_i \) for which one gets results from \( R \).

However, the mapping of detections to sections is ambiguous.
Therefore we define precision in a new way:

\[
\text{prec}_PDA(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|r \cap \bigcup_{s \in S} s|}{|r|},
\]

where \( \cap \) computes the positionally overlapping characters.

The reference basis is switched, and the detections \( R \) become the targets.

Precision computes as if \( R \) were plagiarized sections and \( S \) were detections, i.e., as recall of \( R \) under \( S \).
PDAs often report the same $s_i$ with multiple detections.

We therefore define the granularity of a PDA as follows:

$$\text{gran}_{\text{PDA}}(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |C_s|,$$

where

- $S_R = \{ s \mid s \in S \land \exists r \in R : s \cap r \neq \emptyset \}$ denotes the detected subset of $S$, and
- $C_s = \{ r \mid r \in R \land s \cap r \neq \emptyset \}$ denotes the subset of $R$ that detect a given $s$. 

Detection Performance Measures

Overall Score

- Recall, precision and granularity do not allow for a total order of PDAs.
- Hence, they are combined to an overall score:

\[
overall_{PDA}(S, R) = \frac{F}{\log_2(1 + gran_{PDA})},
\]

where \( F \) denotes the harmonic mean of recall and precision.
- The granularity is logarithmized to smooth its impact on the overall score.
Competition on Plagiarism Detection

1st International Competition on Plagiarism Detection 2009

1st Actually, plenty of firsts!
International 13 working groups from 14 countries participated.
Competition [on] First large-scale comparison of detection algorithms.
Plagiarism First large-scale corpus of artificial plagiarism.
Detection New plagiarism detection performance measures.
2009 13 weeks from March till June.
Competition on Plagiarism Detection

1st International Competition on Plagiarism Detection 2009

1st Actually, plenty of firsts!
International 13 working groups from 14 countries participated.
Competition [on] First large-scale comparison of detection algorithms.
Plagiarism First large-scale corpus of artificial plagiarism.
Detection New plagiarism detection performance measures.
2009 13 weeks from March till June.

Competition tasks and phases:

- **External Plagiarism Detection Task.** Given suspicious and source documents the task is to identify the plagiarism cases between them.
- **Intrinsic Plagiarism Detection Task.** Given only suspicious documents the task is to identify the plagiarized sections.
- **Training phase.** 10 weeks of development based on a training corpus.
- **Competition phase.** 3 weeks competition based on a test corpus.
## Competition on Plagiarism Detection

### Survey of External Plagiarism Detection Algorithms

<table>
<thead>
<tr>
<th>Heuristic Retrieval</th>
<th>Detailed Analysis</th>
<th>Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Retrieval Model</strong></td>
<td><em>Exact Matches of</em> $d_q$ <em>and</em> $d_x \in D_x$</td>
<td>Grozea et al.</td>
</tr>
<tr>
<td>Character-16-gram VSM</td>
<td>Character-16-grams</td>
<td></td>
</tr>
<tr>
<td>(frequency weights, cosine similarity)</td>
<td><strong>Match Merging Heuristic to get</strong> $(s_q, s_x)$</td>
<td></td>
</tr>
<tr>
<td><strong>Comparison of $D_q$ and $D$</strong></td>
<td>Computation of the distances of adjacent matches.</td>
<td></td>
</tr>
<tr>
<td>Exhaustive</td>
<td>Joining of the matches based on a Monte Carlo optimization.</td>
<td></td>
</tr>
<tr>
<td><strong>Candidates $D_x \subset D$ for a $d_q$</strong></td>
<td>Refinement of the obtained section pairs, e.g., by discarding too small sections.</td>
<td></td>
</tr>
<tr>
<td>The 51 documents most similar to $d_q$.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Heuristic Retrieval</th>
<th>Detailed Analysis</th>
<th>Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Retrieval Model</strong></td>
<td><em>Exact Matches of</em> $d_q$ <em>and</em> $d_x \in D_x$</td>
<td>Kasprzak et al.</td>
</tr>
<tr>
<td>Word-5-gram VSM</td>
<td>Word-5-grams</td>
<td></td>
</tr>
<tr>
<td>(boolean weights, Jaccard similarity)</td>
<td><strong>Match Merging Heuristic to get</strong> $(s_q, s_x)$</td>
<td></td>
</tr>
<tr>
<td><strong>Comparison of $D_q$ and $D$</strong></td>
<td>Extraction of the pairs of sections $(s_q, s_x)$ of maximal size which share at least 20 matches, including the first and the last $n$-gram of $s_q$ and $s_x$, and for which 2 adjacent matches are at most 49 not-matching $n$-grams apart.</td>
<td></td>
</tr>
<tr>
<td>Exhaustive</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Candidates $D_x \subset D$ for a $d_q$</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Documents which share at least 20 $n$-grams with $d_q$.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Competition on Plagiarism Detection

Detection Performance in the External Plagiarism Detection Task

<table>
<thead>
<tr>
<th>Rank</th>
<th>Overall</th>
<th>F</th>
<th>Precision</th>
<th>Recall</th>
<th>Granularity</th>
<th>Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6957</td>
<td>0.6976</td>
<td>0.7418</td>
<td>0.6585</td>
<td>1.0038</td>
<td>Grozea et al.</td>
</tr>
<tr>
<td>2</td>
<td>0.6093</td>
<td>0.6192</td>
<td>0.5573</td>
<td>0.6967</td>
<td>1.0228</td>
<td>Kasprzak et al.</td>
</tr>
<tr>
<td>3</td>
<td>0.6041</td>
<td>0.6491</td>
<td>0.6727</td>
<td>0.6272</td>
<td>1.1060</td>
<td>Basile et al.</td>
</tr>
<tr>
<td>4</td>
<td>0.3045</td>
<td>0.5286</td>
<td>0.6689</td>
<td>0.4370</td>
<td>2.3317</td>
<td>Palkovskii et al.</td>
</tr>
<tr>
<td>5</td>
<td>0.1885</td>
<td>0.4603</td>
<td>0.6051</td>
<td>0.3714</td>
<td>4.4354</td>
<td>Muhr et al.</td>
</tr>
<tr>
<td>6</td>
<td>0.1422</td>
<td>0.6190</td>
<td>0.7473</td>
<td>0.5284</td>
<td>19.4327</td>
<td>Scherbinin et al.</td>
</tr>
<tr>
<td>7</td>
<td>0.0649</td>
<td>0.1736</td>
<td>0.6552</td>
<td>0.1001</td>
<td>5.3966</td>
<td>Pereira et al.</td>
</tr>
<tr>
<td>8</td>
<td>0.0264</td>
<td>0.0265</td>
<td>0.0136</td>
<td>0.4586</td>
<td>1.0068</td>
<td>Vallés Balaguer</td>
</tr>
<tr>
<td>9</td>
<td>0.0187</td>
<td>0.0553</td>
<td>0.0290</td>
<td>0.6048</td>
<td>6.7780</td>
<td>Malcolm et al.</td>
</tr>
<tr>
<td>10</td>
<td>0.0117</td>
<td>0.0226</td>
<td>0.3684</td>
<td>0.0116</td>
<td>2.8256</td>
<td>Allen</td>
</tr>
</tbody>
</table>
Competition on Plagiarism Detection

Detection Performance in the Intrinsic Plagiarism Detection Task

<table>
<thead>
<tr>
<th>Rank</th>
<th>Overall</th>
<th>F</th>
<th>Precision</th>
<th>Recall</th>
<th>Granularity</th>
<th>Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2462</td>
<td>0.3086</td>
<td>0.2321</td>
<td>0.4607</td>
<td>1.3839</td>
<td>Stamatatos</td>
</tr>
<tr>
<td>2</td>
<td>0.1955</td>
<td>0.1956</td>
<td>0.1091</td>
<td>0.9437</td>
<td>1.0007</td>
<td>Hagbi et al.*</td>
</tr>
<tr>
<td>3</td>
<td>0.1766</td>
<td>0.2286</td>
<td>0.1968</td>
<td>0.2724</td>
<td>1.4524</td>
<td>Muhr et al.</td>
</tr>
<tr>
<td>4</td>
<td>0.1219</td>
<td>0.1750</td>
<td>0.1036</td>
<td>0.5630</td>
<td>1.7049</td>
<td>Seaward et al.</td>
</tr>
</tbody>
</table>

* Hagbi and Koppel’s submission is almost the baseline for this task, since they reported practically everything once as plagiarized.
## Competition on Plagiarism Detection

### Detection Performance Overall Tasks

<table>
<thead>
<tr>
<th>Rank</th>
<th>Overall</th>
<th>F</th>
<th>Precision</th>
<th>Recall</th>
<th>Granularity</th>
<th>Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4871</td>
<td>0.4884</td>
<td>0.5193</td>
<td>0.4610</td>
<td>1.0038</td>
<td>Grozea et al.</td>
</tr>
<tr>
<td>2</td>
<td>0.4265</td>
<td>0.4335</td>
<td>0.3901</td>
<td>0.4877</td>
<td>1.0228</td>
<td>Kasprzak et al.</td>
</tr>
<tr>
<td>3</td>
<td>0.4229</td>
<td>0.4544</td>
<td>0.4709</td>
<td>0.4390</td>
<td>1.1060</td>
<td>Basile et al.</td>
</tr>
<tr>
<td>4</td>
<td>0.2131</td>
<td>0.3700</td>
<td>0.4682</td>
<td>0.3059</td>
<td>2.3317</td>
<td>Palkovskii et al.</td>
</tr>
<tr>
<td>5</td>
<td>0.1833</td>
<td>0.4001</td>
<td>0.4826</td>
<td>0.3417</td>
<td>3.5405</td>
<td>Muhr et al.</td>
</tr>
<tr>
<td>6</td>
<td>0.0996</td>
<td>0.4333</td>
<td>0.5231</td>
<td>0.3699</td>
<td>19.4327</td>
<td>Scherbinin et al.</td>
</tr>
<tr>
<td>7</td>
<td>0.0739</td>
<td>0.0926</td>
<td>0.0696</td>
<td>0.1382</td>
<td>1.3839</td>
<td>Stamatatos</td>
</tr>
<tr>
<td>8</td>
<td>0.0586</td>
<td>0.0587</td>
<td>0.0327</td>
<td>0.2831</td>
<td>1.0007</td>
<td>Hagbi et al.</td>
</tr>
<tr>
<td>9</td>
<td>0.0454</td>
<td>0.1216</td>
<td>0.4586</td>
<td>0.0701</td>
<td>5.3966</td>
<td>Pereira et al.</td>
</tr>
<tr>
<td>10</td>
<td>0.0366</td>
<td>0.0525</td>
<td>0.0311</td>
<td>0.1689</td>
<td>1.7049</td>
<td>Seaward et al.</td>
</tr>
<tr>
<td>11</td>
<td>0.0184</td>
<td>0.0185</td>
<td>0.0095</td>
<td>0.3210</td>
<td>1.0068</td>
<td>Vallés Balaguer</td>
</tr>
<tr>
<td>12</td>
<td>0.0131</td>
<td>0.0387</td>
<td>0.0203</td>
<td>0.4234</td>
<td>6.7780</td>
<td>Malcolm et al.</td>
</tr>
<tr>
<td>13</td>
<td>0.0081</td>
<td>0.0157</td>
<td>0.2579</td>
<td>0.0081</td>
<td>2.8256</td>
<td>Allen</td>
</tr>
</tbody>
</table>