Overview of the 6th International Competition on Plagiarism Detection

Martin Potthast, Matthias Hagen, Anna Beyer,
Matthias Busse, Martin Tippmann, and Benno Stein

Bauhaus-Universität Weimar
www.webis.de

Paolo Rosso
Universitat Politècnica de València

Outline

- Introduction
- Source Retrieval
- Text Alignment
- Summary

Plagiarism Detection

Source Retrieval

Given

- suspicious document
- web search engine

Task

- retrieve plagiarized sources
- minimize retrieval costs

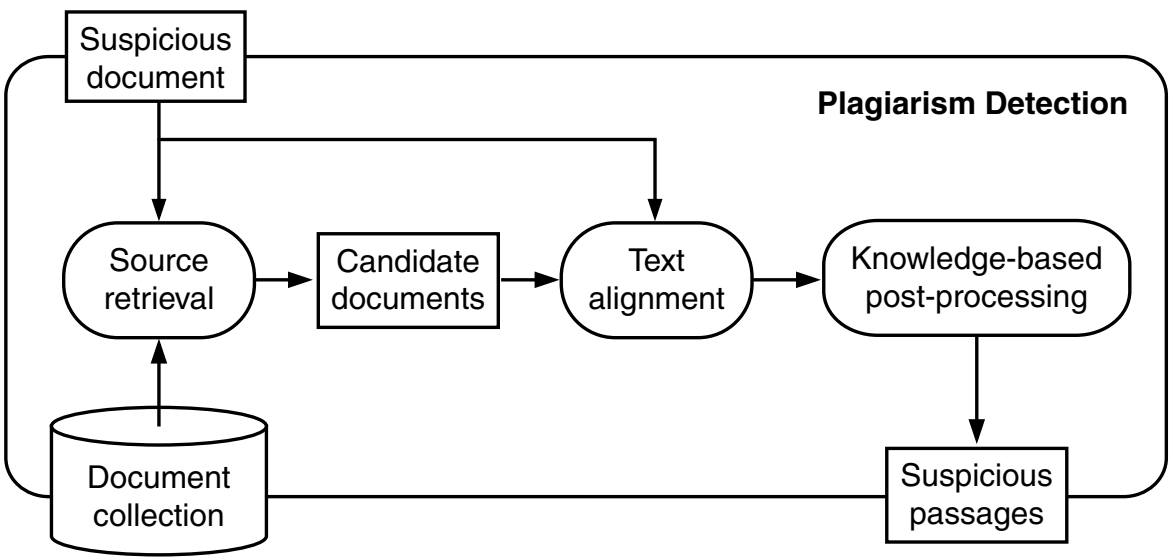
Text Alignment

Given

- pair of documents

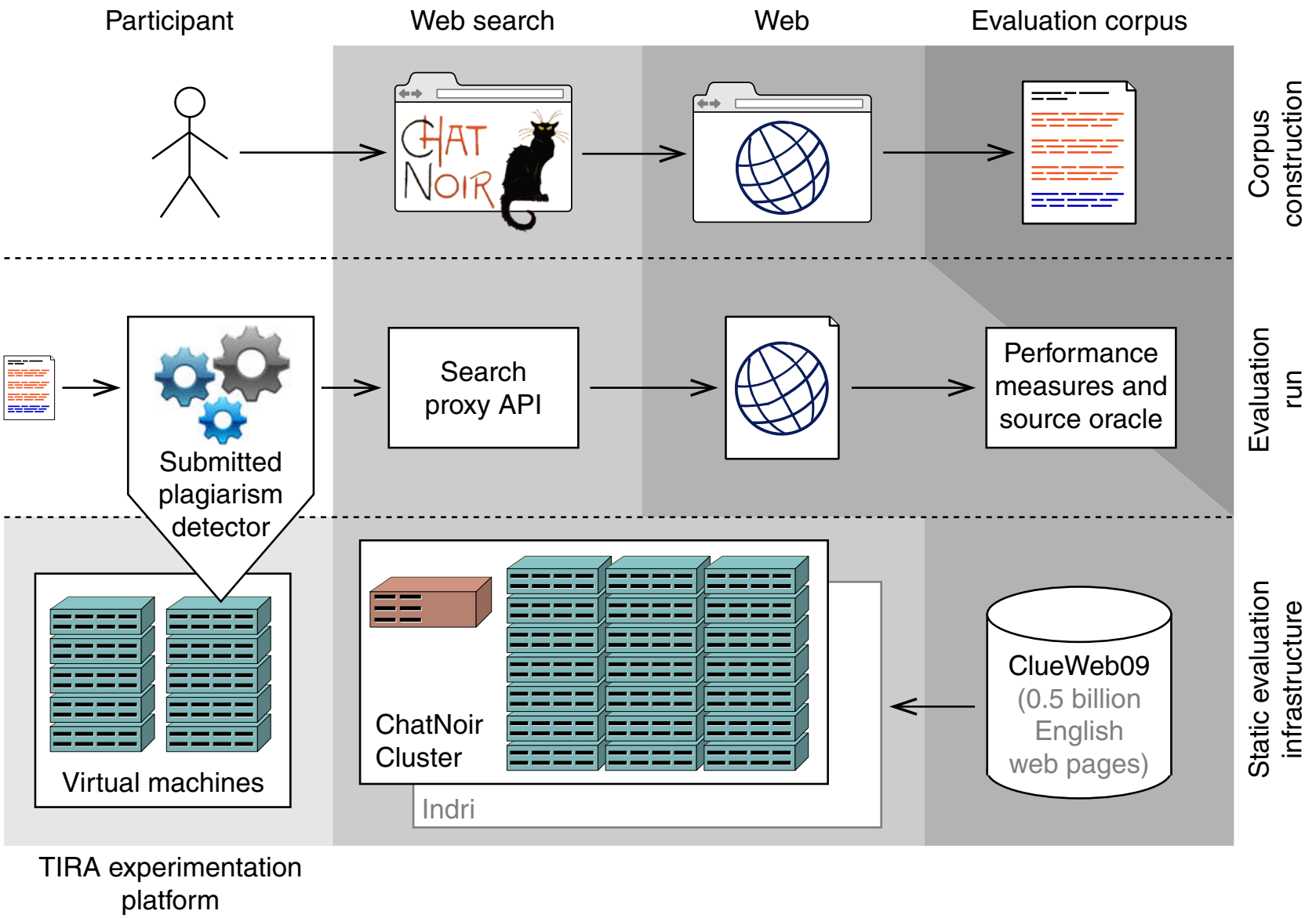
Task

- extract passages of reused text



Source Retrieval

Source Retrieval



Source Retrieval

Corpus & Performance Measures

Corpus

- Webis Text Reuse Corpus 2012
- 297 manually written essay-length documents
- Each consists of up to 70 passages of resued text from the ClueWeb
- Training: 98 documents
- Test: 99 documents

Retrieval performance measures

- Precision, recall, and F_α

Cost-effectiveness measures

- Workload as counts of queries and downloads
- Workload until 1st detection
- Runtime

Source Retrieval

Survey of Approaches

An analysis of the participants' notebooks reveals a source retrieval process:

1. Chunking
   Given a suspicious document, it is divided into (possibly overlapping) passages of text. Each chunk of text is then processed individually.
2. Keyphrase Extraction
   Given a chunk (or the entire suspicious document), keyphrases are extracted from it in order to formulate queries with them.
3. Query Formulation
   Given sets of keywords extracted from chunks, queries are formulated which are tailored to the API of the search engine used.
4. Search Control
   Given a set of queries, the search controller schedules their submission to the search engine and directs the download of search results.
5. Download Filtering
   Given a set of downloaded documents, all documents are removed that are not worthwhile for detailed comparison to the suspicious document.

Source Retrieval

Evaluation Results

Software Submission		Downloaded Sources			Total Workload		Workload to 1st Detection		No Detect.	Runtime
Team (alphabetical order)	Year	F_1	Prec.	Rec.	Queries	Dwlds	Queries	Dwlds		
Elizalde	2013	0.16	0.12	0.37	41.6	83.9	18.0	18.2	4	11:18:50
Elizalde	2014	0.34	0.40	0.39	54.5	33.2	16.4	3.9	7	04:02:00
Foltynek	2013	0.11	0.08	0.26	166.8	72.7	180.4	4.3	32	152:26:23
Gillam	2013	0.06	0.04	0.15	**15.7**	86.8	16.1	28.6	34	**02:24:59**
Haggag	2013	0.38	**0.67**	0.31	41.7	**5.2**	13.9	**1.4**	12	46:09:21
Kong	2013	0.01	0.01	**0.59**	47.9	5185.3	**2.5**	210.2	**0**	106:13:46
Kong	2014	0.12	0.08	0.48	83.5	207.1	85.7	24.9	6	24:03:31
Lee	2013	0.40	0.58	0.37	48.4	10.9	6.5	2.0	9	09:17:10
Prakash	2014	0.39	0.38	0.51	60.0	38.8	8.1	3.8	7	19:47:45
Suchomel	2013	0.05	0.04	0.23	17.8	283.0	3.4	64.9	18	75:12:56
Suchomel	2014	0.11	0.08	0.40	19.5	237.3	3.1	38.6	2	45:42:06
Williams	2013	**0.47**	0.60	0.47	117.1	12.4	23.3	2.2	7	76:58:22
Williams	2014	**0.47**	0.57	0.48	117.1	14.4	18.8	2.3	4	39:44:11
Zubarev	2014	0.45	0.54	0.45	37.0	18.6	5.4	2.3	3	40:42:18

- Ranked by recall, the 2014 approaches outperform all except two from 2013
- Ensemble recall: 0.85; only 14 topic with ensemble recall less than 0.6
- Some returning participants improve (Elizalde, Suchomel) [TIRA]

Text Alignment

Text Alignment

Corpus & Performance Measures

Corpus

- The evaluation corpus has been reused from last year
- A supplemental corpus serves as baseline
- Problems / Criticism of “corpus reuse”
  - Gives rise to overfitting
  - Some participants found out about it
  - ➜ In the future, we’ll be more open about this

Performance measures

- Plagdet, precision, recall, granularity, and runtime as usual
- New measures that capture more abstract aspects of detection performance

Text Alignment

Survey of Approaches

An analysis of the participants' notebooks reveals a detailed comparison process:

1. Seeding
   Given a suspicious document and a source document, matches (also called "seeds") between the two documents are identified using some seed heuristic. Seed heuristics either identify exact matches or *create* matches by changing the underlying texts in a domain-specific or linguistically motivated way.
2. Extension
   Given seed matches identified between a suspicious document and a source document, they are merged into aligned text passages of maximal length between the two documents which are then reported as plagiarism detections.
3. Filtering
   Given a set of aligned passages, a passage filter removes all aligned passages that do not meet certain criteria.

Text Alignment

Survey of Approaches (continued)

New trend: obfuscation prediction

> Given a pair of documents, predict the most likely type of obfuscation of reused passages between them.

Approaches

- Decide a priori, before aligning the documents, which alignment strategy/parameters to apply
- Decide a posteriori, after aligning the documents using multiple alignment strategies/parameters, which result is best

- A priori decisions are machine-learning based
- A posteriori decisions are rule-based

Classification schemes

- no obfuscation vs. rest
- summaries vs. rest
- no obfuscation, random, summaries, rest

Text Alignment

Evaluation Results

Team	PlagDet	Recall	Precision	Granularity	Runtime
Sanchez-Perez	0.88	0.88	0.88	1.00	00:25:35
Oberreuter	0.87	0.86	0.89	1.00	00:05:31
Palkovskii	0.87	0.83	0.92	1.01	01:10:04
Glinos	0.86	0.79	0.96	1.02	00:23:13
Shrestha	0.84	0.84	0.86	1.01	69:51:15
R. Torrejón	0.83	0.77	0.90	1.00	00:00:42
Gross	0.83	0.77	0.93	1.03	00:03:00
Kong	0.82	0.81	0.84	1.00	00:05:26
Abnar	0.67	0.61	0.77	1.02	01:27:00
Alvi	0.66	0.55	0.93	1.07	00:04:57
Baseline	0.42	0.34	0.93	1.28	00:30:30
Gillam	0.28	0.17	0.87	1.00	00:00:55

- The top performers are Sanchez-Perez, Oberreuter, and Palkovskii
- Performances are very close together; further improvements may be difficult
- Summary obfuscation still most difficult; Glinos outperforms Sanchez-Perez

- PlagDet combines recall, precision, and granularity
- Granularity measures the number of times a plagiarism case is detected

Text Alignment

Performance Measures Revisited

Text Alignment

Performance Measures Revisited

- Currently, detection performance is measured at character level
- We introduce two more abstract levels: case level, and document level
- The new measures build upon the character level ones

Text Alignment

Performance Measures Revisited

- Currently, detection performance is measured at character level
- We introduce two more abstract levels: case level, and document level
- The new measures build upon the character level ones

Terminology (simplified)

- s denotes a plagiarism case; S a set of plagiarism cases
- r denotes a plagiarism detection; R a set of plagiarism detections
- They refer to two text passages (suspicious and source) in two documents
- We say "r detects s" iff source passage and suspicious passage overlap
- $|s|$ and $|r|$ denote the sum character lengths of the passages of s and r
- $|r \cap s|$ denotes the length of detection if r detects s in characters

Text Alignment

Performance Measures Revisited

- Currently, detection performance is measured at character level
- We introduce two more abstract levels: case level, and document level
- The new measures build upon the character level ones

Terminology (simplified)

- s denotes a plagiarism case; S a set of plagiarism cases
- r denotes a plagiarism detection; R a set of plagiarism detections
- They refer to two text passages (suspicious and source) in two documents
- We say "r detects s" iff source passage and suspicious passage overlap
- $|s|$ and $|r|$ denote the sum character lengths of the passages of s and r
- $|r \cap s|$ denotes the length of detection if r detects s in characters

We measure character level precision and recall as follows (simplified):

$$\mathit{prec}_{\text{char}}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{\sum_{s \in S} |s \cap r|}{|r|}, \qquad \mathit{rec}_{\text{char}}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{\sum_{r \in R} |s \cap r|}{|s|},$$

Text Alignment

Performance Measures Revisited: Case Level

We measure character level precision and recall as follows (simplified):

$$\mathit{prec}_{\text{char}}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{\sum_{s \in S} |s \cap r|}{|r|}, \qquad \mathit{rec}_{\text{char}}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{\sum_{r \in R} |s \cap r|}{|s|},$$

Text Alignment

Performance Measures Revisited: Case Level

We measure character level precision and recall as follows (simplified):

$$\mathit{prec}_{\text{char}}(S, r) = \frac{\sum_{s \in S} |s \cap r|}{|r|}, \qquad \mathit{rec}_{\text{char}}(s, R) = \frac{\sum_{r \in R} |s \cap r|}{|s|},$$

Text Alignment

Performance Measures Revisited: Case Level

We measure character level precision and recall as follows (simplified):

$$\textit{prec}_{\text{char}}(S, r) = \frac{\sum_{s \in S} |s \cap r|}{|r|}, \qquad \textit{rec}_{\text{char}}(s, R) = \frac{\sum_{r \in R} |s \cap r|}{|s|},$$

Based on these formulas, we define subsets of S and R:

$S' = \{s \mid s \in S \text{ and } \textit{rec}_{\text{char}}(s, R) > \tau_1 \text{ and } \exists r \in R\text{: } r \text{ detects } s \text{ and } \textit{prec}_{\text{char}}(S, r) > \tau_2\}$,
$R' = \{r \mid r \in R \text{ and } \textit{prec}_{\text{char}}(S, r) > \tau_2 \text{ and } \exists s \in S\text{: } r \text{ detects } s \text{ and } \textit{rec}_{\text{char}}(s, R) > \tau_1\}$,

where τ_1 and τ_2 determine the least desired character level detection quality:

- $\tau_1, \tau_2 \to 1$ require perfect detection quality
- $\tau_1, \tau_2 \to 0$ allow for poor detection quality
- τ_1 and τ_2 should be set to the top perceptible detection quality

Text Alignment

Performance Measures Revisited: Case Level

We measure character level precision and recall as follows (simplified):

$$\textit{prec}_{\text{char}}(S, r) = \frac{\sum_{s \in S} |s \cap r|}{|r|}, \qquad \textit{rec}_{\text{char}}(s, R) = \frac{\sum_{r \in R} |s \cap r|}{|s|},$$

Based on these formulas, we define subsets of S and R:

$S' = \{s \mid s \in S$ and $\textit{rec}_{\text{char}}(s, R) > \tau_1$ and $\exists r \in R$: r detects s and $\textit{prec}_{\text{char}}(S, r) > \tau_2\}$,
$R' = \{r \mid r \in R$ and $\textit{prec}_{\text{char}}(S, r) > \tau_2$ and $\exists s \in S$: r detects s and $\textit{rec}_{\text{char}}(s, R) > \tau_1\}$,

where τ_1 and τ_2 determine the least desired character level detection quality:

- $\tau_1, \tau_2 \to 1$ require perfect detection quality
- $\tau_1, \tau_2 \to 0$ allow for poor detection quality
- τ_1 and τ_2 should be set to the top perceptible detection quality

We measure case level precision and recall as follows:

$$\textit{prec}_{\text{case}}(S, R) = \frac{|R'|}{|R|}, \qquad \textit{rec}_{\text{case}}(S, R) = \frac{|S'|}{|S|}$$

Text Alignment

Performance Measures Revisited: Document Level

- D_{plg} denotes the set of suspicious documents
- D_{src} denotes the set of source documents
- $D_{\text{pairs}} = D_{\text{plg}} \times D_{\text{src}}$

Text Alignment

Performance Measures Revisited: Document Level

- D_{plg} denotes the set of suspicious documents
- D_{src} denotes the set of source documents
- $D_{\text{pairs}} = D_{\text{plg}} \times D_{\text{src}}$

We define subsets of D_{pairs} based on S and R:

$D_{\text{pairs}|S} = \{(d_{\text{plg}}, d_{\text{src}}) \mid (d_{\text{plg}}, d_{\text{src}}) \in D_{\text{pairs}} \text{ and } \exists s \in S : d_{\text{plg}} \in s \text{ and } d_{\text{src}} \in s\}$
$D_{\text{pairs}|R} = \{(d_{\text{plg}}, d_{\text{src}}) \mid (d_{\text{plg}}, d_{\text{src}}) \in D_{\text{pairs}} \text{ and } \exists r \in R : d_{\text{plg}} \in r \text{ and } d_{\text{src}} \in r\}$

Likewise, $D_{\text{pairs}|R'}$ is based on R' instead of R.

Text Alignment

Performance Measures Revisited: Document Level

- D_{plg} denotes the set of suspicious documents
- D_{src} denotes the set of source documents
- $D_{\text{pairs}} = D_{\text{plg}} \times D_{\text{src}}$

We define subsets of D_{pairs} based on S and R:

$$D_{\text{pairs}|S} = \{(d_{\text{plg}}, d_{\text{src}}) \mid (d_{\text{plg}}, d_{\text{src}}) \in D_{\text{pairs}} \text{ and } \exists s \in S : d_{\text{plg}} \in s \text{ and } d_{\text{src}} \in s\}$$
$$D_{\text{pairs}|R} = \{(d_{\text{plg}}, d_{\text{src}}) \mid (d_{\text{plg}}, d_{\text{src}}) \in D_{\text{pairs}} \text{ and } \exists r \in R : d_{\text{plg}} \in r \text{ and } d_{\text{src}} \in r\}$$

Likewise, $D_{\text{pairs}|R'}$ is based on R' instead of R.

We measure document level precision and recall as follows:

$$\textit{prec}_{\text{doc}}(S, R) = \frac{|D_{\text{pairs}|S} \cap D_{\text{pairs}|R'}|}{|D_{\text{pairs}|R}|}, \qquad \textit{rec}_{\text{doc}}(S, R) = \frac{|D_{\text{pairs}|S} \cap D_{\text{pairs}|R'}|}{|D_{\text{pairs}|S}|}.$$

Text Alignment

Performance Measures Revisited: Evaluation Results

Software Submission		Level of Abstraction		
Team	Year	Char *plagdet*	Case F_1	Document F_1
Sanchez-Perez	2014	0.88	0.90	0.91
Oberreuter	2014	0.87	0.87	0.89
Palkovskii	2014	0.87	0.87	0.87
Glinos	2014	0.86	0.87	0.91
Kong	2012	0.84	0.85	0.87
Shrestha	2014	0.84	0.88	0.89
Gross	2014	0.83	0.88	0.89
Oberreuter	2012	0.83	0.80	0.81
R. Torrejón	2014	0.83	0.83	0.86
R. Torrejón	2013	0.83	0.83	0.85
Kong	2013	0.82	0.85	0.87
Kong	2014	0.82	0.85	0.87
Palkovskii	2012	0.79	0.80	0.81

- $\tau_1 = 0.5$, so that 50% of a plagiarism case s must be detected
- $\tau_2 = 0.5$, so that 50% of a plagiarism detection r must be a true detection
- Some differences in ranking can be observed
- However, it is still unclear which settings for τ_1 and τ_2 are to be preferred

Summary

PAN 2014

- Source retrieval approaches outperform those of last year
- Twice as much test data as last year
- Too much focus on saving downloads than on saving queries

- Obfuscation prediction to diversify alignment approaches
- Less rule-based extension approaches
- New performance measures

PAN 2015 and beyond

- New text alignment corpora in progres
- New version of ChatNoir based on Elastic Search
- More TIRA support

Summary

PAN 2014

- Source retrieval approaches outperform those of last year
- Twice as much test data as last year
- Too much focus on saving downloads than on saving queries

- Obfuscation prediction to diversify alignment approaches
- Less rule-based extension approaches
- New performance measures

PAN 2015 and beyond

- New text alignment corpora in progres
- New version of ChatNoir based on Elastic Search
- More TIRA support

Thank you for your attention, and your contributions to PAN!