Overview of the 5th International Competition on Plagiarism Detection

Martin Potthast, Matthias Hagen, Tim Gollub,
Martin Tippmann, Johannes Kiesel, and Benno Stein

Bauhaus-Universität Weimar
www.webis.de

Paolo Rosso
Universitat Politècnica de València

Efstathios Stamatatos
University of the Aegean

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

©www.webis.de 2013

Plagiarism Detection

Source Retrieval

Given

- suspicious document
- web search engine

Task

- retrieve plagiarized sources
- minimize retrieval costs

Overview

- Plagiarism corpus:
  Webis Text Reuse Corpus 2012
- Web corpus: ClueWeb 2009
- Web search: Indri and ChatNoir
- New text alignment oracle
- Software submissions

Text Alignment

Given

- pair of documents

Task

- extract passages of reused text

Overview

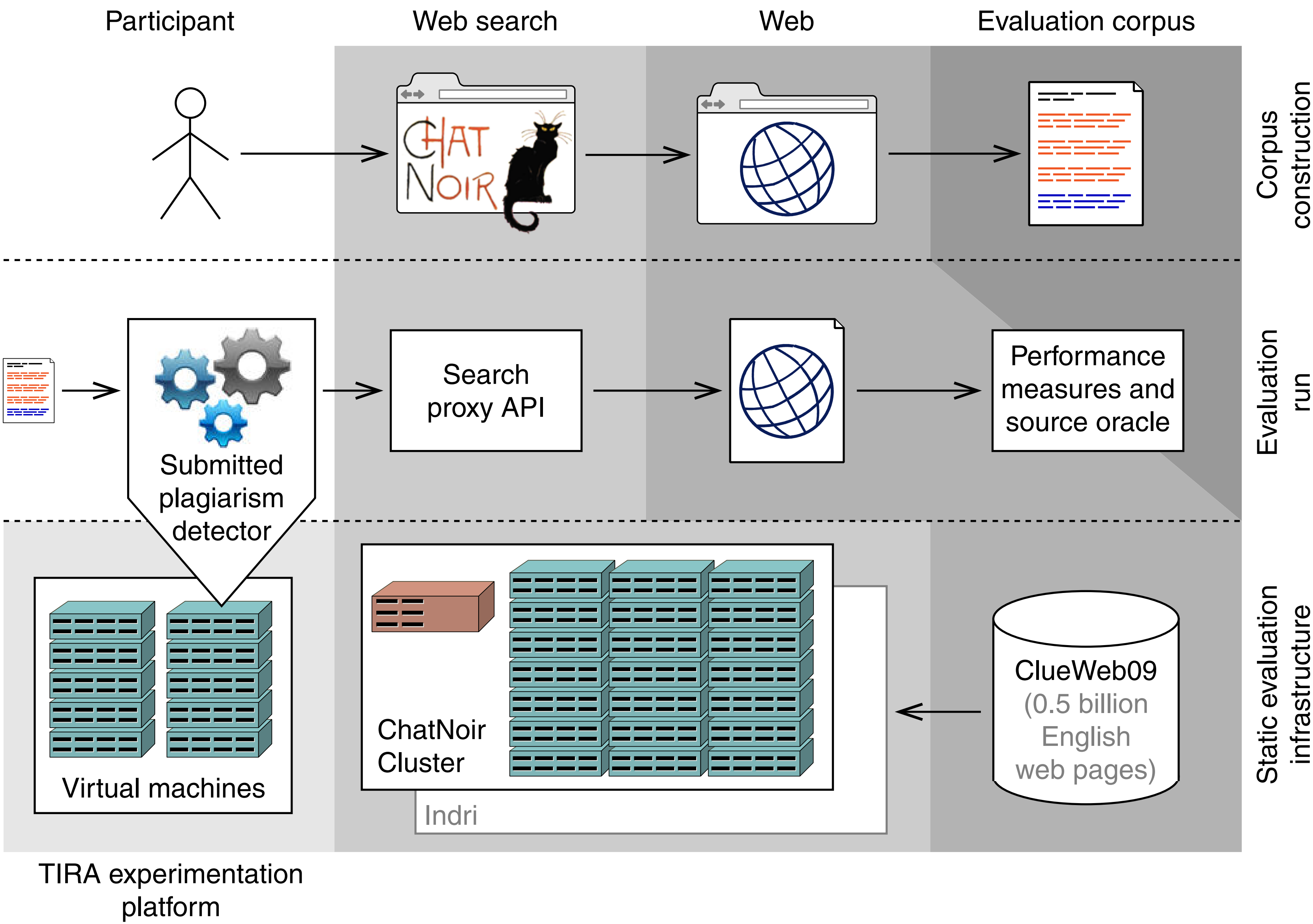
- Plagiarism corpus:
  PAN Plagiarism Corpus 2013
- New obfuscation:
  Cyclic translation and summaries
- Software submissions
- Cross-year evaluation

©www.webis.de 2013

Source Retrieval

©www.webis.de 2013

Source Retrieval



©www.webis.de 2013

Source Retrieval

Performance Measures

Retrieval performance

- Precision, recall, and F_α

Cost-effectiveness

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

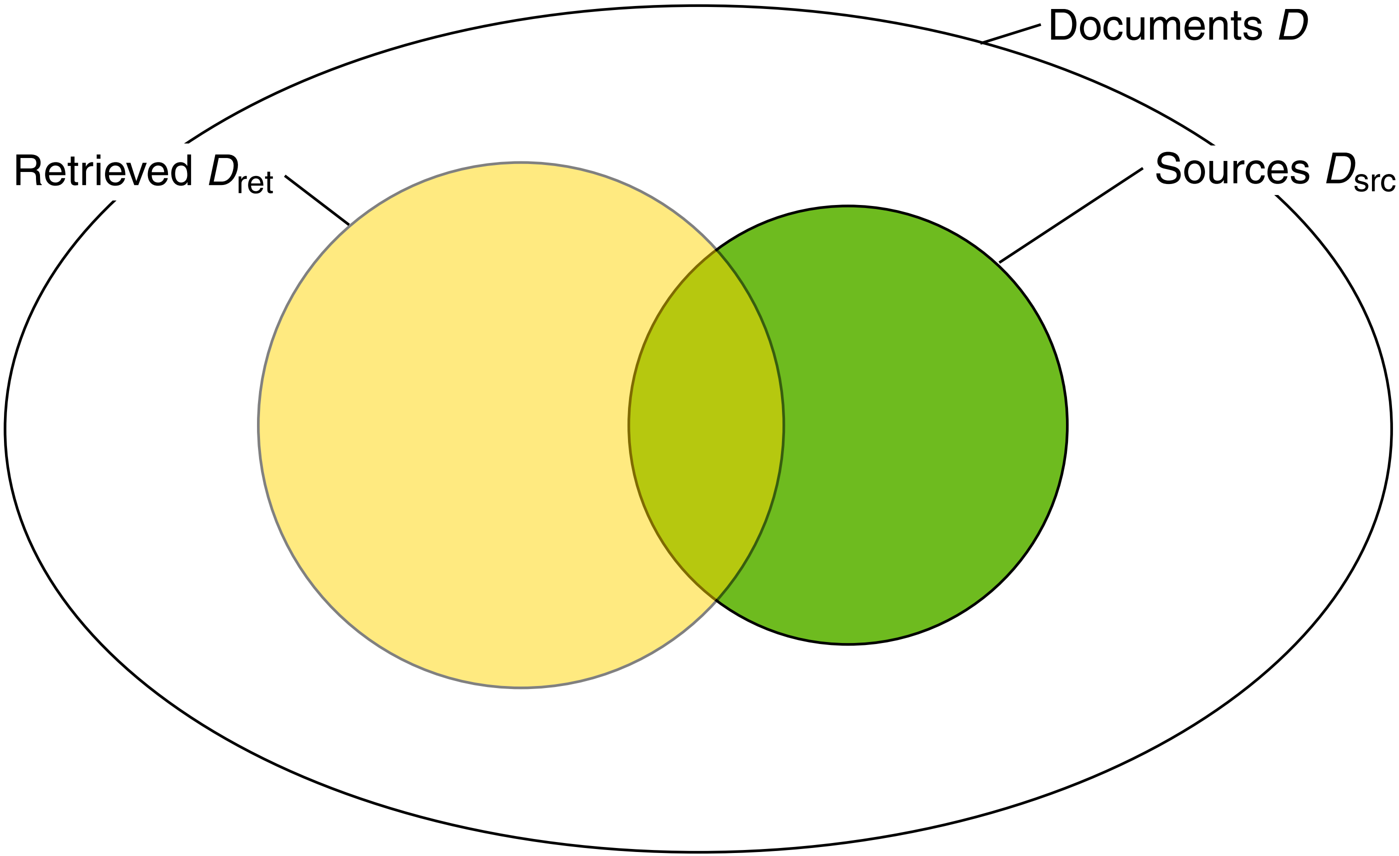
Considerations

- Source retrieval is a recall-intensive task
- Diversity of retrieved documents is important
- Retrieval costs should be minimized
- Weight of each measure still unknown
- No ranking formula as of yet

©www.webis.de 2013

Source Retrieval

Retrieval Performance



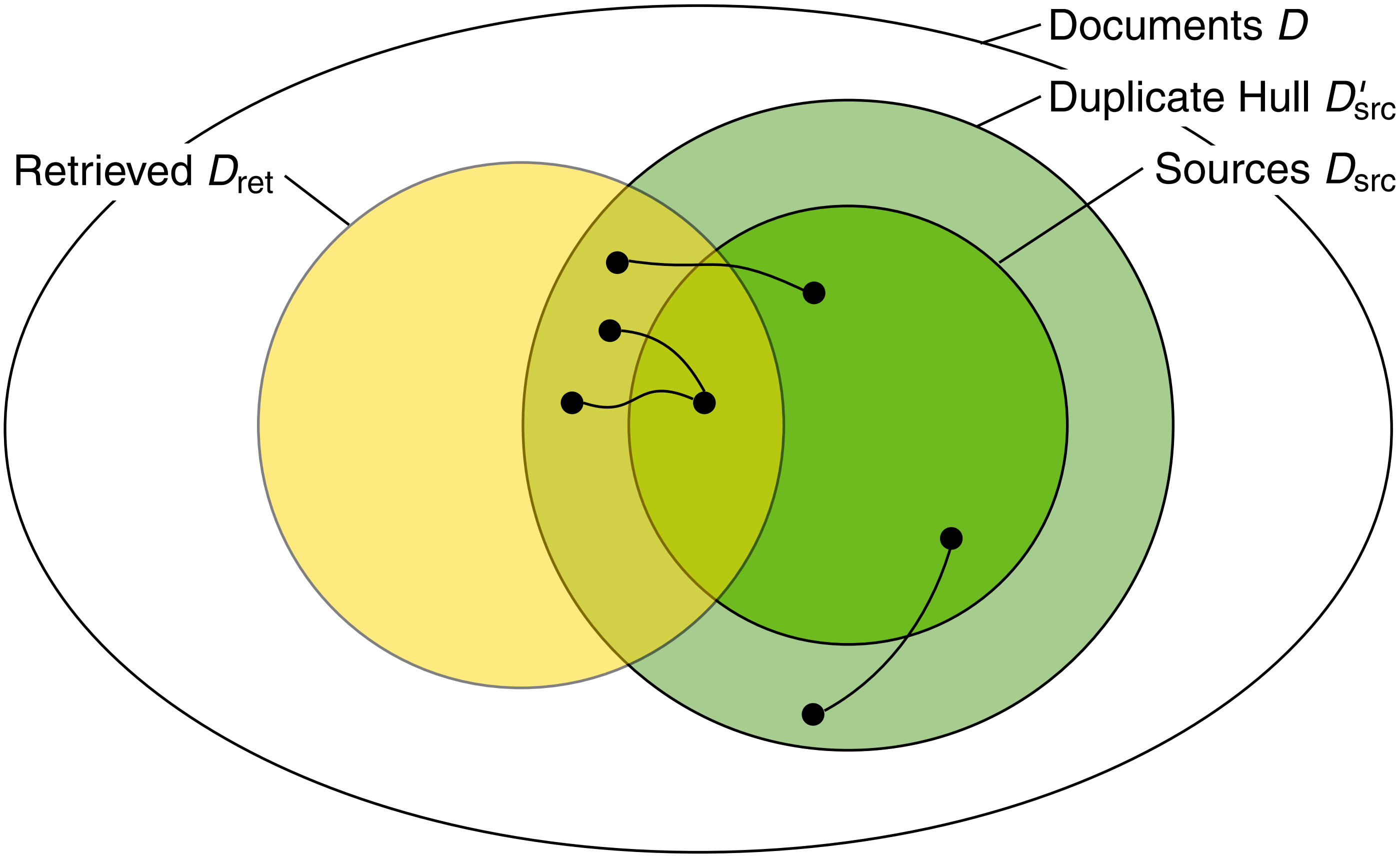
- Standard information retrieval situation:

$$\text{precision} = \frac{|D_{\text{ret}} \cap D_{\text{src}}|}{|D_{\text{ret}}|}, \qquad \text{recall} = \frac{|D_{\text{ret}} \cap D_{\text{src}}|}{|D_{\text{src}}|}$$

©www.webis.de 2013

Source Retrieval

Retrieval Performance

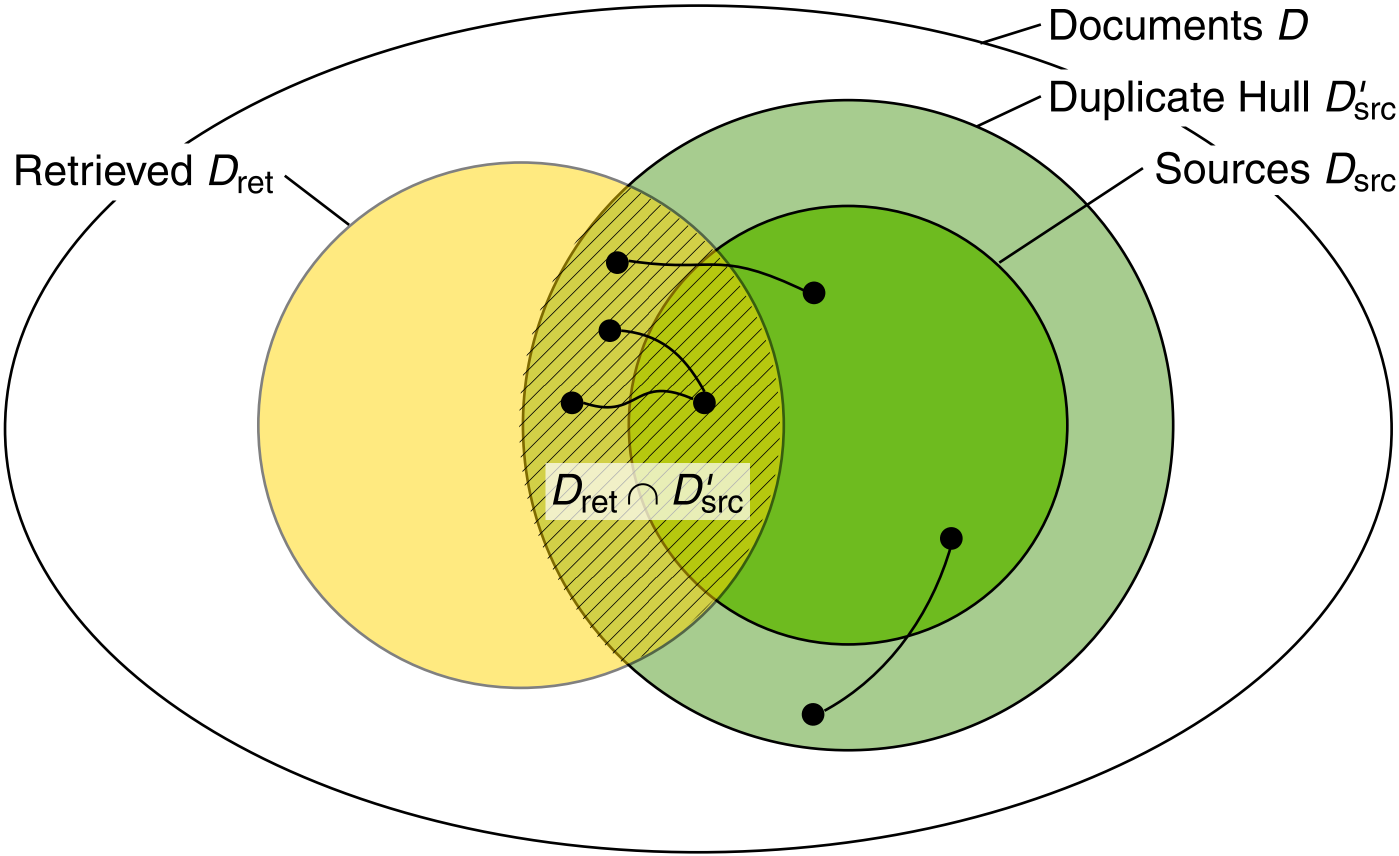


- How to deal with near-duplicates of source documents?
- Detect them by measuring equality, similarity, and containment [details]

©www.webis.de 2013

Source Retrieval

Retrieval Performance



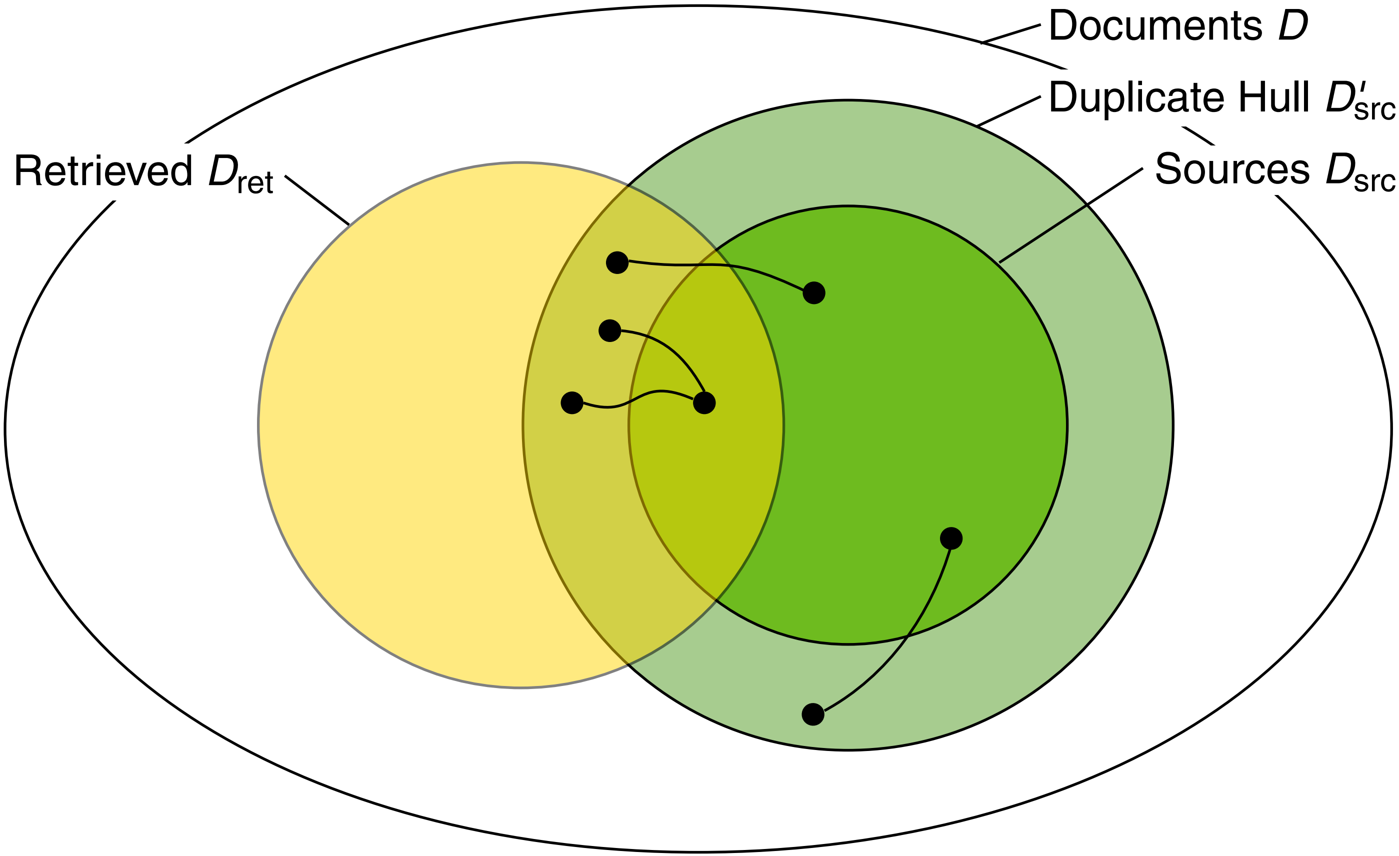
- Detecting near-duplicates shall not decrease precision:

$$\text{precision} = \frac{|D_{ret} \cap D'_{src}|}{|D_{ret}|}$$

©www.webis.de 2013

Source Retrieval

Retrieval Performance

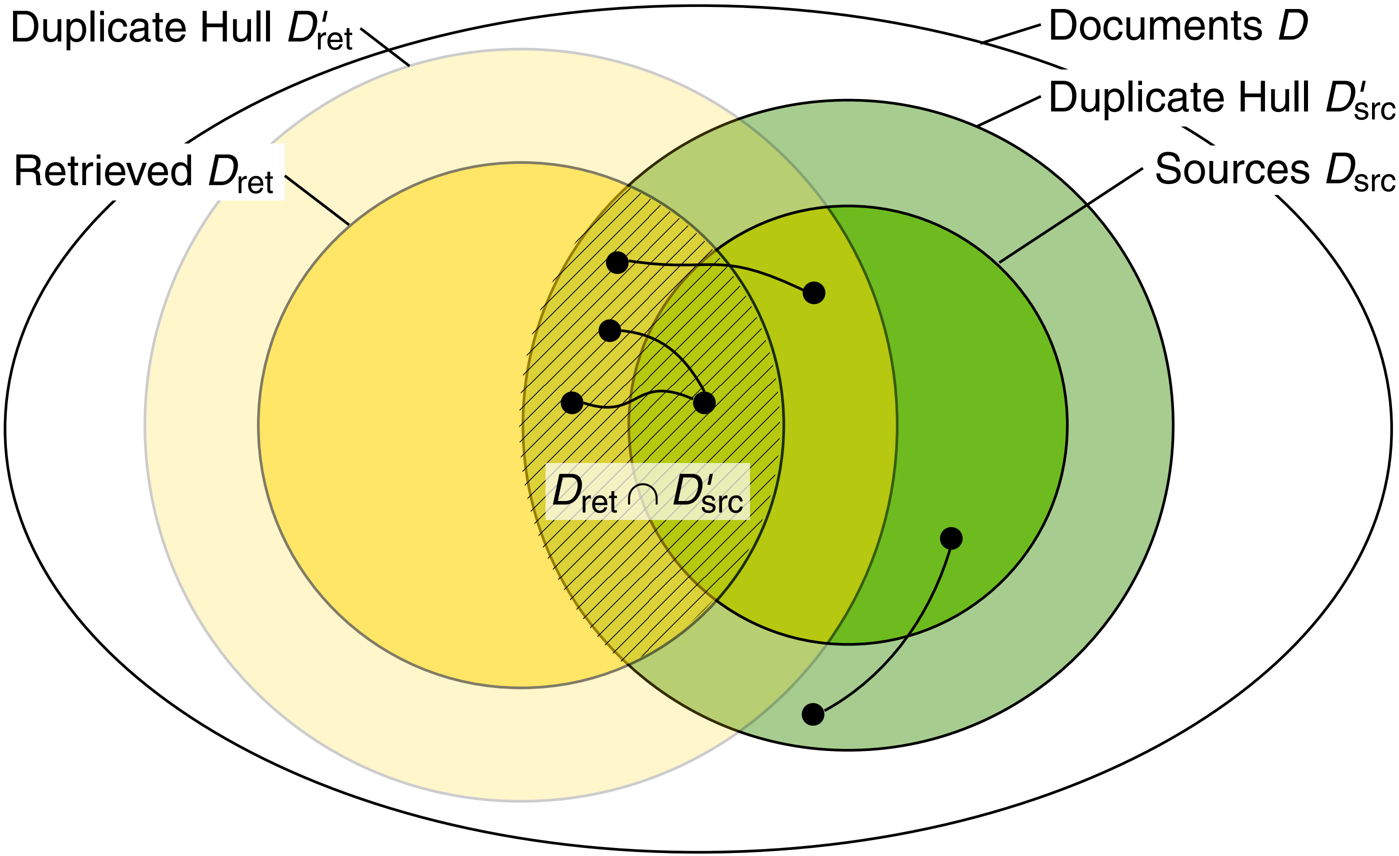


- Considering recall, the reference set is uncertain
- Strategies: include all duplicates, retrieved duplicates, or no duplicates

©www.webis.de 2013

Source Retrieval

Retrieval Performance

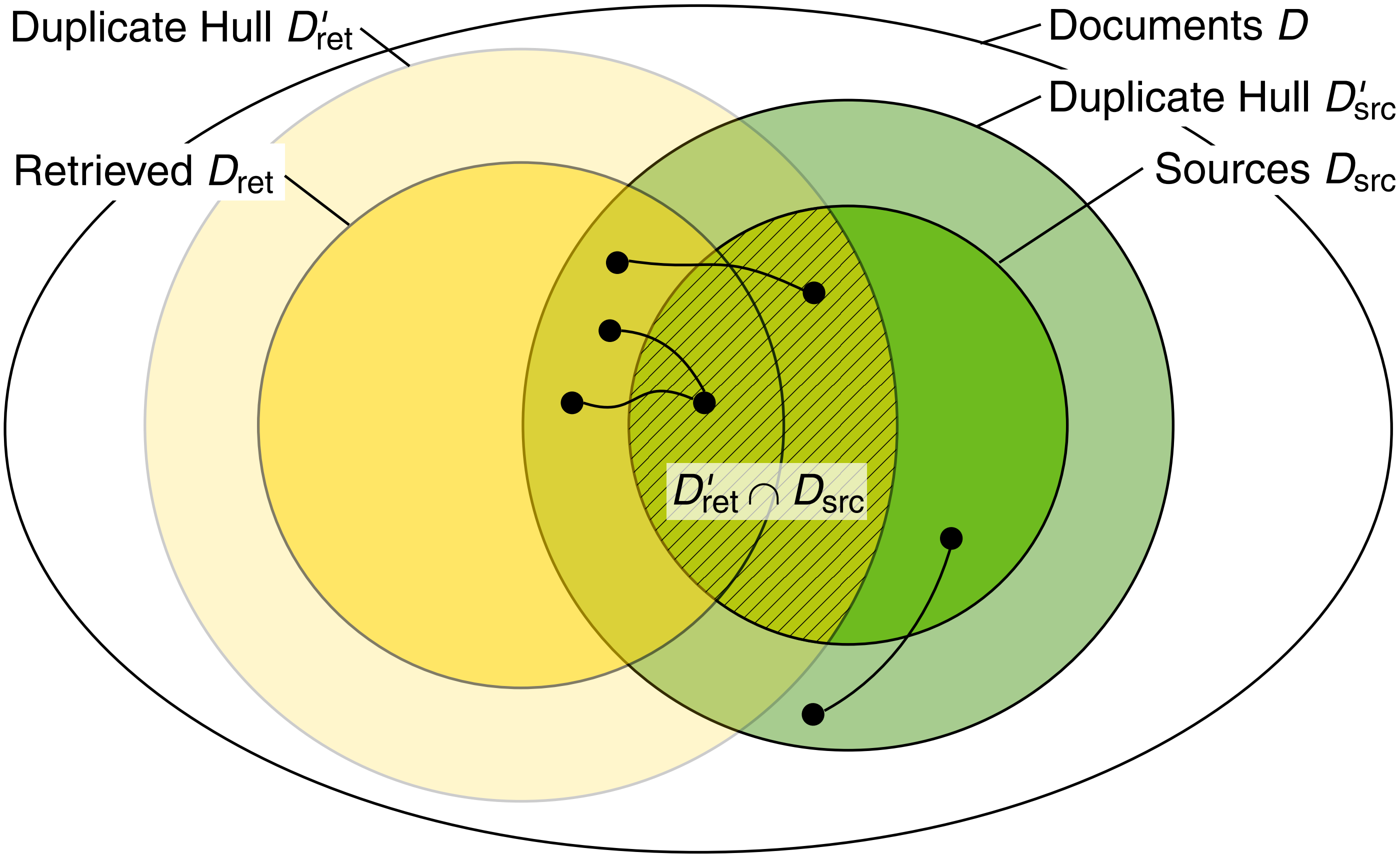


- Considering recall, the reference set is uncertain
- Strategies: include all duplicates, retrieved duplicates, or no duplicates

©www.webis.de 2013

Source Retrieval

Retrieval Performance



- Detecting near-duplicates shall not increase recall:

$$\text{recall} = \frac{|D'_{ret} \cap D_{src}|}{|D_{src}|}$$

©www.webis.de 2013

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.

©www.webis.de 2013

Source Retrieval

Evaluation Results

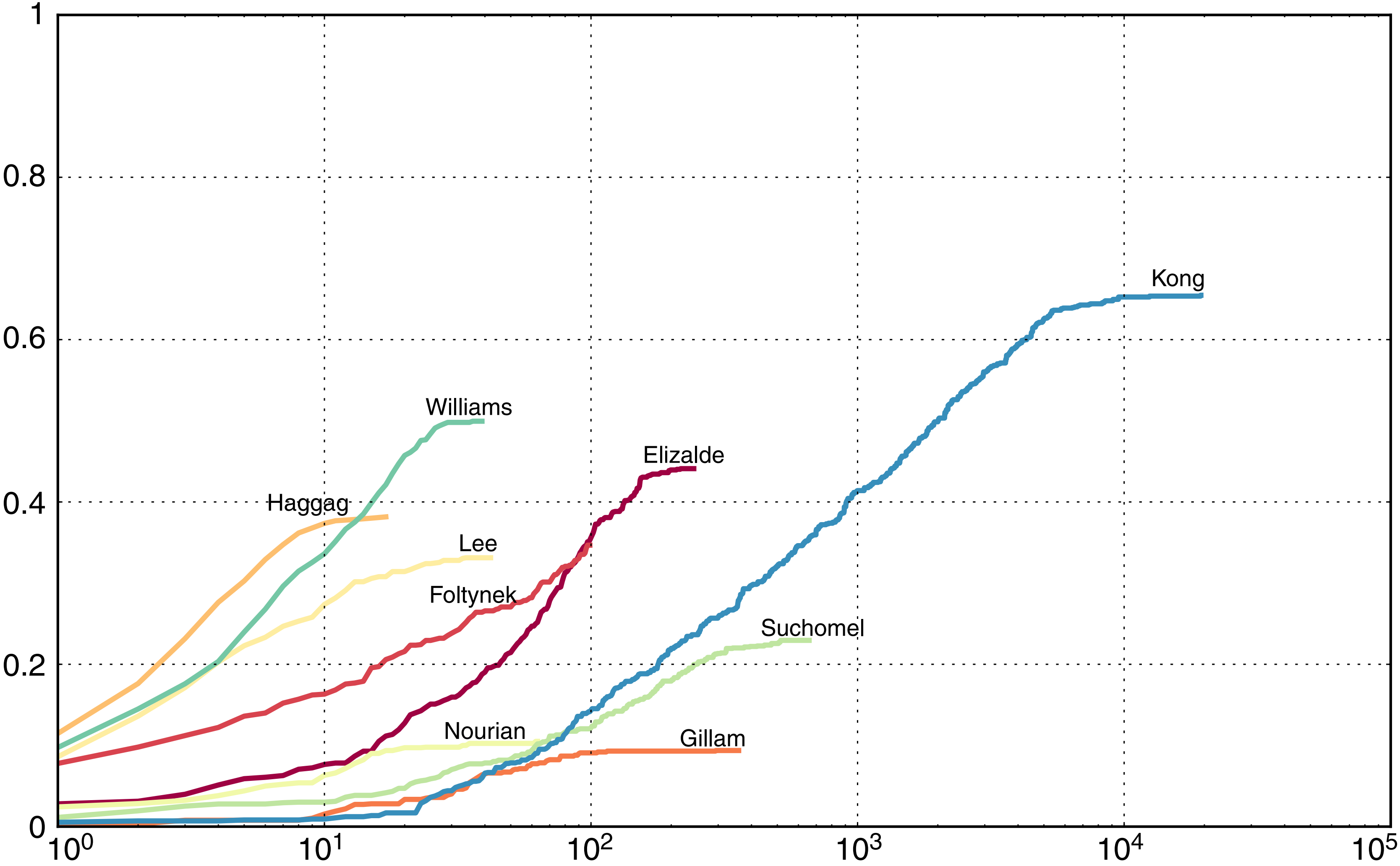
Team (alphabetical order)	Downloaded Sources			Total Workload		Workload to 1st Detection		No Detection	Runtime
	F_1	Precision	Recall	Queries	Downloads	Queries	Downloads		
Elizalde	0.17	0.12	0.44	44.50	107.22	16.85	15.28	5	241.7 m
Gillam	0.04	0.02	0.10	16.10	33.02	18.80	21.70	38	**15.1 m**
Haggag	0.44	**0.63**	0.38	32.04	**5.93**	8.92	**1.47**	9	152.7 m
Kong	0.01	0.01	**0.65**	48.50	5691.47	2.46	285.66	**3**	4098.0 m
Lee	0.35	0.50	0.33	44.04	11.16	7.74	1.72	15	310.5 m
Nourian	0.10	0.15	0.10	**4.91**	13.54	**2.16**	5.61	27	25.3 m
Suchomel	0.06	0.04	0.23	12.38	261.95	2.44	74.79	10	1637.9 m
Veselý	0.15	0.11	0.35	161.21	81.03	184.00	5.07	16	655.3 m
Williams	**0.47**	0.55	0.50	116.40	14.05	17.59	2.45	5	1163.0 m

- Kong achieves best recall, Haggag best precision, Williams best tradeoff
- Results indicate paradigmatically different approaches
- Ensemble recall: 0.82

©www.webis.de 2013

Source Retrieval

Evaluation Results (continued)



©www.webis.de 2013

Text Alignment

©www.webis.de 2013

Text Alignment

PAN Plagiarism Corpus 2013

Source documents

- 145 topics
- 10 630 web documents manually retrieved from the ClueWeb09
- Between 1 and 270 documents per topic

➜ Topic-homogeneity of source documents per suspicious document

Suspicious documents

- Generated from passages drawn from the source documents
- Document length, plagiarism length, number of sources drawn at random

Obfuscation

- None
- Random obfuscation
- Cyclic translation
- Summarization

©www.webis.de 2013

Text Alignment

Obfuscation

Obfuscation is an author's attempt to hide text reuse from being identified by means of paraphrasing, summarization, or translation.

Random

- Random shuffling, adding, deleting, and replacing words and short phrases

Cyclic translation

- English ➜ IL_1 ➜ IL_2 ➜ IL_3 ➜ English (IL = intermediate language)
- IL_i one of $\{$fr, de, es, se, ar, cn, he, hi, ja$\}$
- Usage of Google Translate, Microsoft Translate, and MyMemory

Summarization

- Manual summaries taken from DUC 2001 text summarization corpus
- Inserted in documents of similar genre from another DUC 2006 corpus
- Named entities replaced to foreclose easy detection

©www.webis.de 2013

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.

©www.webis.de 2013

Text Alignment

Evaluation Results

Team	PlagDet	Recall	Precision	Granularity	Runtime
R. Torrejón	0.82	0.76	0.90	1.00	1.2 m
Kong	0.82	0.81	0.83	1.00	6.1 m
Suchomel	0.75	0.77	0.73	1.00	28.0 m
Saremi	0.70	0.77	0.87	1.25	446.0 m
Shrestha	0.70	0.79	0.88	1.22	684.5 m
Palkovskii	0.62	0.54	0.82	1.07	6.5 m
Nourian	0.58	0.43	0.95	1.04	40.1 m
Baseline	0.42	0.34	0.93	1.28	30.5 m
Gillam	0.40	0.26	0.89	1.00	21.3 m
Jayapal	0.27	0.38	0.88	2.91	4.8 m

- R. Torrejón and Kong perform best
- Granularity is mostly under control
- Runtime varies from minutes to hours

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

©www.webis.de 2013

Text Alignment

Evaluation Results (continued)

Team	PlagDet per Obfuscation Strategy			
	None	Random	Cyclic transl.	Summary
R. Torrejón	0.93	0.75	0.85	0.34
Kong	0.83	0.82	0.85	0.43
Suchomel	0.82	0.75	0.68	0.61
Saremi	0.85	0.66	0.71	0.11
Shrestha	0.89	0.67	0.63	0.12
Palkovskii	0.82	0.50	0.61	0.10
Nourian	0.90	0.35	0.44	0.16
Baseline	0.93	0.07	0.11	0.04
Gillam	0.86	0.04	0.01	0.00
Jayapal	0.39	0.18	0.18	0.06

- Unobfuscated plagiarism not a problem; very competitive baseline
- Kong performs best on random plagiarism
- Cyclic translations pose no bigger problem than random plagiarism
- Summaries are extremely difficult; outstanding performance of Suchomel

©www.webis.de 2013

Text Alignment

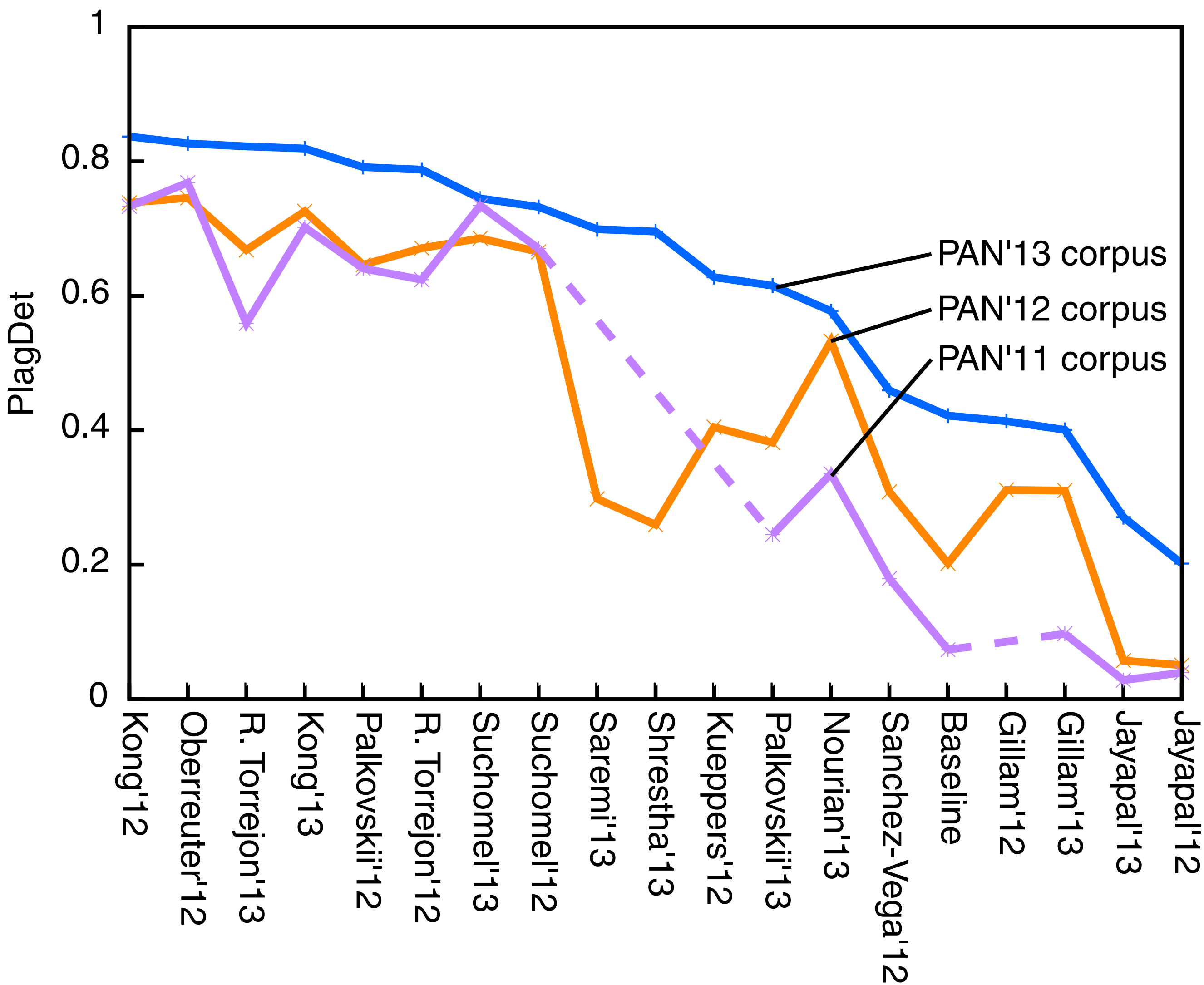
Cross-year Evaluation 2011-2013

Software Submission		PlagDet on PAN Plagiarism Corpus		
Team	Year	2013	2012	2011
Kong	**2012**	0.84	0.74	0.73
Oberreuter	**2012**	0.83	0.75	0.77
R. Torrejón	2013	0.82	0.67	0.56
Kong	2013	0.82	0.73	0.70
Palkovskii	**2012**	0.79	0.65	0.64
R. Torrejón	**2012**	0.79	0.67	0.62
Suchomel	2013	0.74	0.69	0.73
Suchomel	**2012**	0.73	0.67	0.67
Saremi	2013	0.70		
Shrestha	2013	0.70		
Kueppers	**2012**	0.63	0.40	
Palkovskii	2013	0.62	0.38	0.25
Nourian	2013	0.58	0.53	0.34
Sánchez-Vega	**2012**	0.46	0.31	0.18
Baseline		0.42	0.20	0.07
Gillam	**2012**	0.41	0.31	0.10
Gillam	2013	0.40	0.31	0.10
Jayapal	2013	0.27	0.06	0.03
Jayapal	**2012**	0.20	0.05	0.04

©www.webis.de 2013

Text Alignment

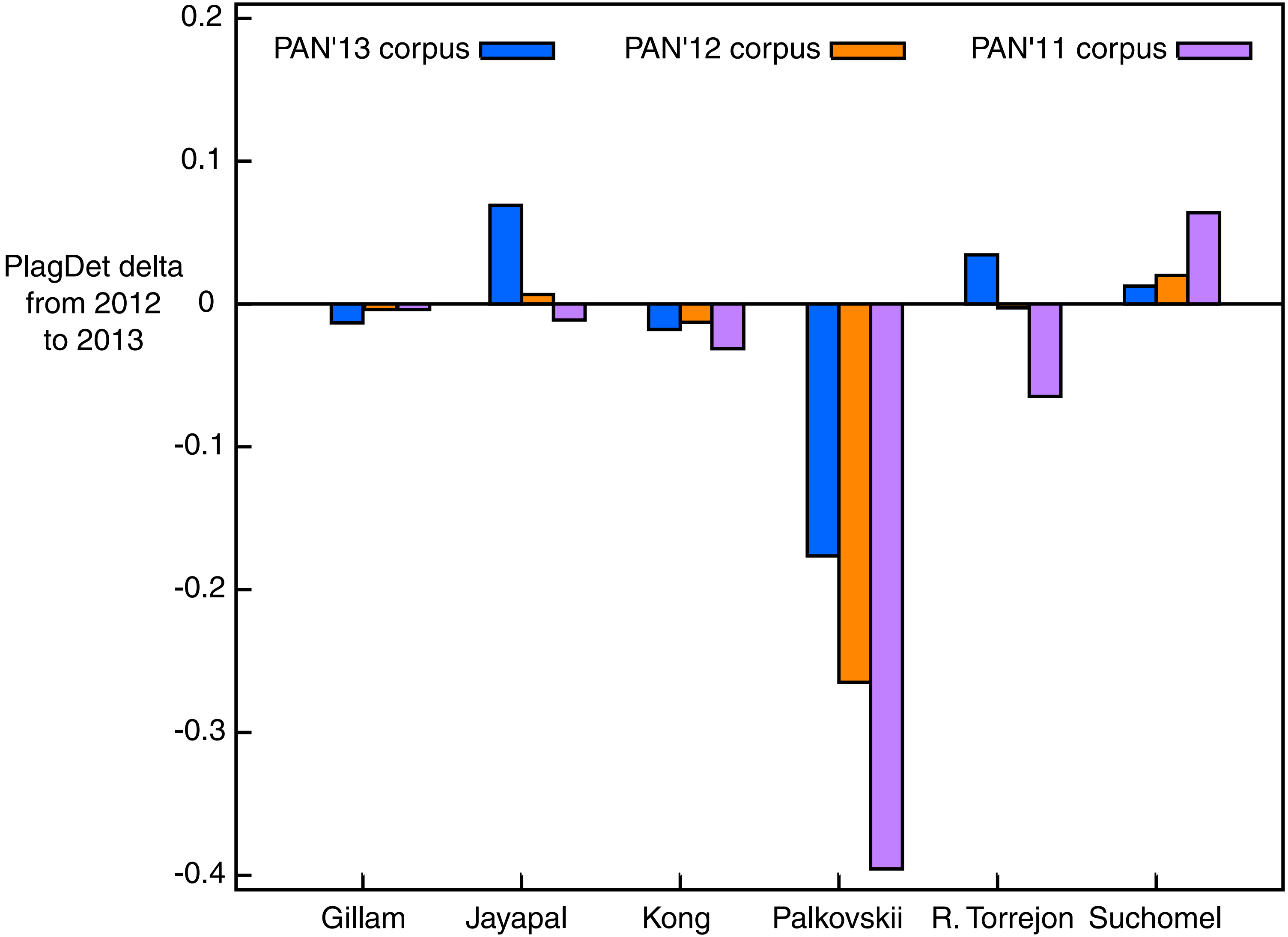
Cross-year Evaluation 2011-2013 (continued)



©www.webis.de 2013

Text Alignment

Cross-year Evaluation 2011-2013 (continued)



©www.webis.de 2013

Summary

PAN 2013

- Emergence of new source retrieval paradigms
- Source oracle to separate source retrieval from text alignment
- Consolidation and many small errors fixed
- New text alignment corpus
- New kinds of obfuscation (summaries and cyclic translation)
- First time cross-year evaluation
- Corpus difficulty analysis
- Performance difference across versions

PAN 2014 and beyond

- All-time ranking for text alignment and source retrieval
- Automation toward self-service evaluation
- New tools for error analysis

©www.webis.de 2013

Summary

PAN 2013

- Emergence of new source retrieval paradigms
- Source oracle to separate source retrieval from text alignment
- Consolidation and many small errors fixed
- New text alignment corpus
- New kinds of obfuscation (summaries and cyclic translation)
- First time cross-year evaluation
- Corpus difficulty analysis
- Performance difference across versions

PAN 2014 and beyond

- All-time ranking for text alignment and source retrieval
- Automation toward self-service evaluation
- New tools for error analysis

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

©www.webis.de 2013

©www.webis.de 2013

Near-duplicate Detection

We say that d' is a near-duplicate of d if one of the following conditions holds:

1. *Equality.* $d' = d$.
2. *Similarity.* Under n-gram Jaccard similarity φ_1,
   $\varphi_1(d', d) > 0.8$ for $n = 3$,
   $\varphi_1(d', d) > 0.5$ for $n = 5$, and
   $\varphi_1(d', d) > 0$ for $n = 8$
3. *Containment.* Under asymmetrical n-gram overlap φ_2 of d' toward d,
   $\varphi_2(d', d) > 0.8$ for $n = 3$,
   $\varphi_2(d', d) > 0.5$ for $n = 5$, and
   $\varphi_2(d', d) > 0$ for $n = 8$

For source retrieval, we employ partial containment instead of containment:

- let d be a source document of a plagiarized document d_{plg}
- let d' be retrieved by a source retrieval algorithm analyzing d_{plg}
- then we consider d' partially contained in d iff the passages of d that are reused in d_{plg} are contained in d' as defined above

©www.webis.de 2013