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

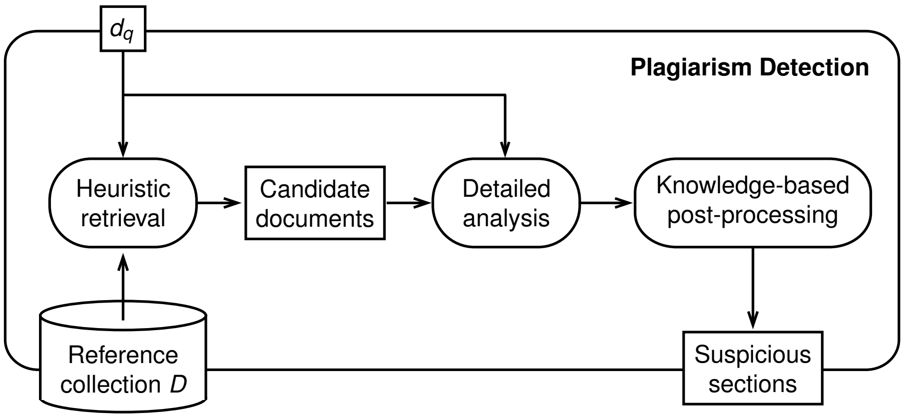
Introduction

- 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.



[Fig.] Benno Stein, Sven Meyer zu Eissen, and Martin Potthast. Strategies for Retrieving Plagiarized Documents. In Clarke, Fuhr, Kando, Kraaij, and de Vries, editors, 30th Annual International ACM SIGIR Conference, pages 825-826, July 2007. ACM. ISBN 987-1-59593-597-7.

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:

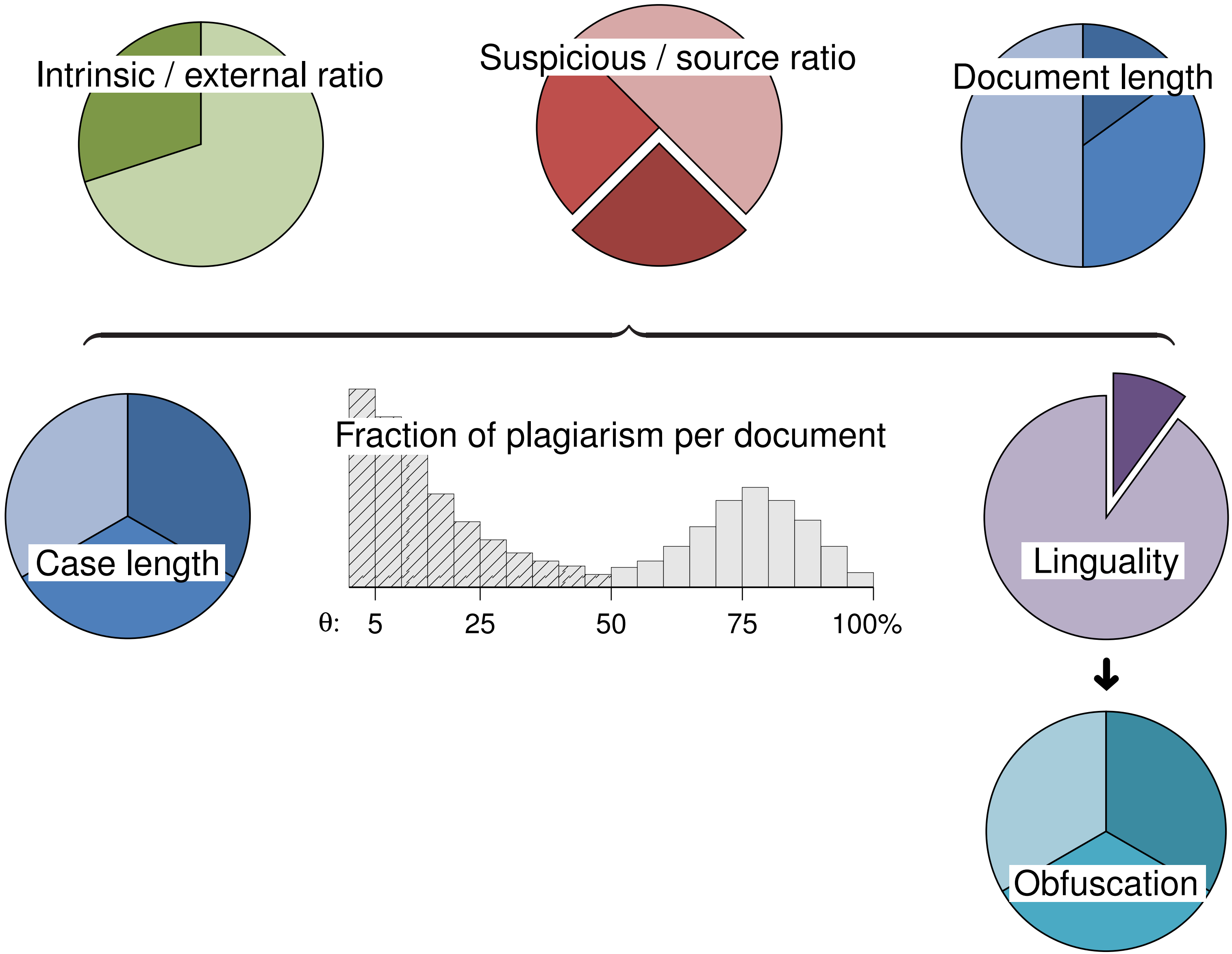
- 41 223 text documents (obtained from 22 874 books from the Project Gutenberg [2])
- 94 202 plagiarism cases
- 70% is dedicated to external plagiarism detection,
  30% is dedicated to intrinsic plagiarism detection

- Types of cases: monolingual with and without obfuscation, and cross-lingual
- Authenticity of cases: real, emulated, and artificial

[1] Webis at Bauhaus-Universität Weimar and NLEL at Universidad Politécnica de Valencia. PAN Plagiarism Corpus PAN-PC-09. `http://www.uni-weimar.de/medien/webis/research/corpora`, 2009. M. Potthast, A. Eiselt, B. Stein, A. Barrón-Cedeño, and P. Rosso (editors).

[2] `http://www.gutenberg.org`

Plagiarism Corpus



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:

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

$s_q =$ "The quick brown dodger leaps over the lazy canine."
$s_q =$ "The quick brown canine jumps over the lazy canine."
$s_q =$ "The quick brown vixen leaps over the lazy puppy."

Plagiarism Corpus

Plagiarism Obfuscation Synthesis

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:

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

POS $=$ “DT JJ JJ NN VBZ IN DT JJ NN .”

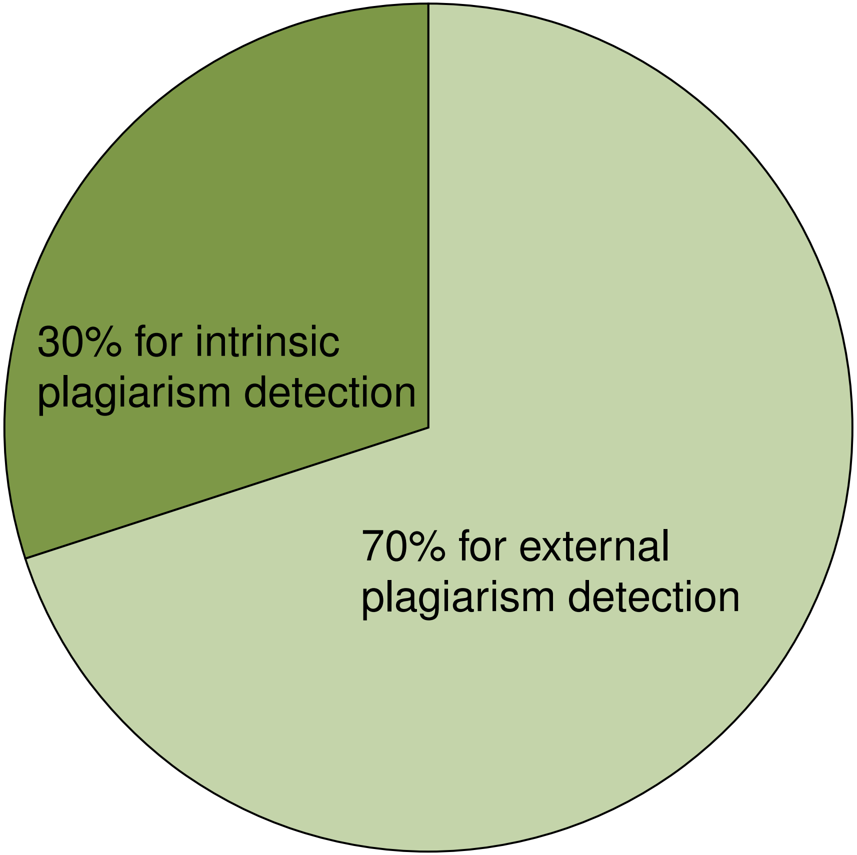
$s_q =$ “The brown lazy fox jumps over the quick dog.”
$s_q =$ “The lazy quick dog jumps over the brown fox.”
$s_q =$ “The brown lazy dog jumps over the quick fox.”

Plagiarism Corpus

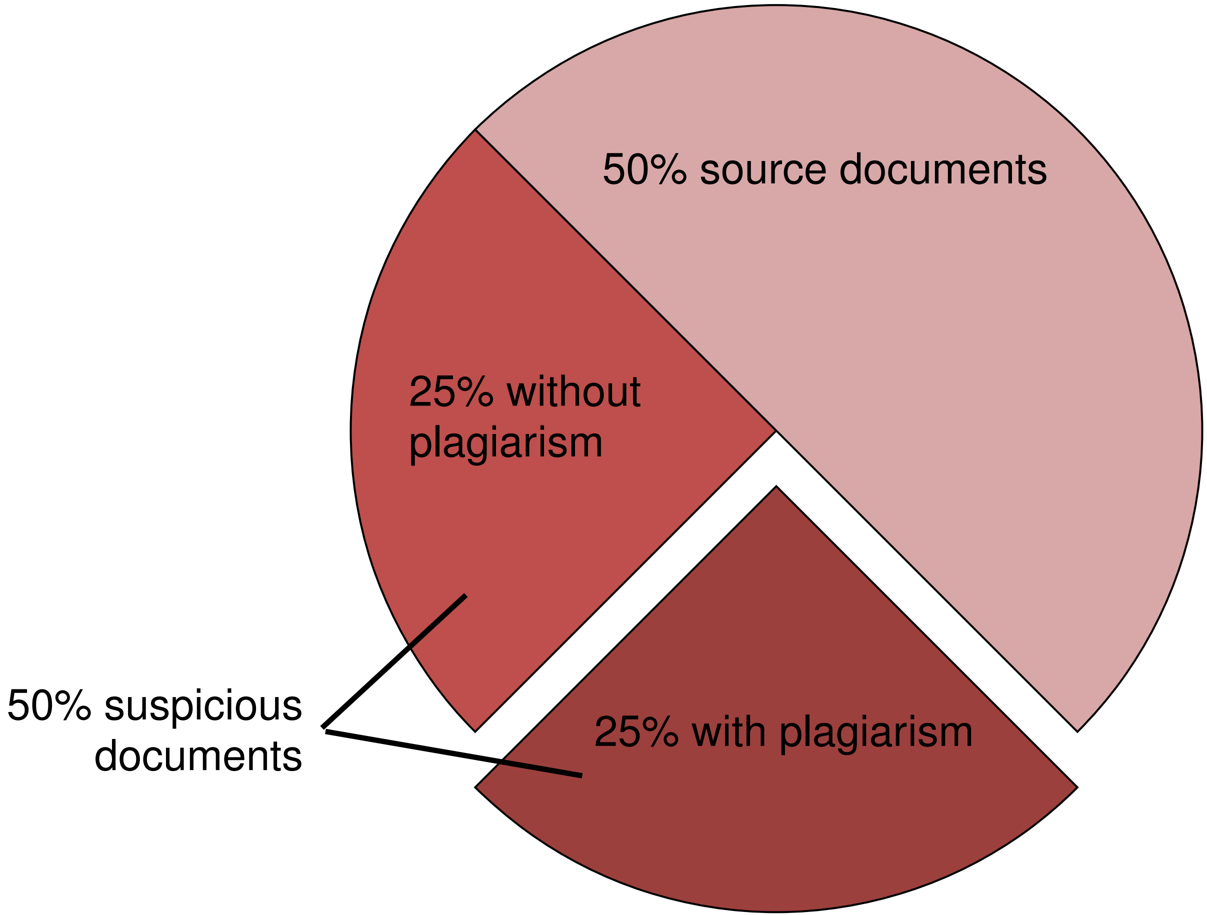
Critical Remarks

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

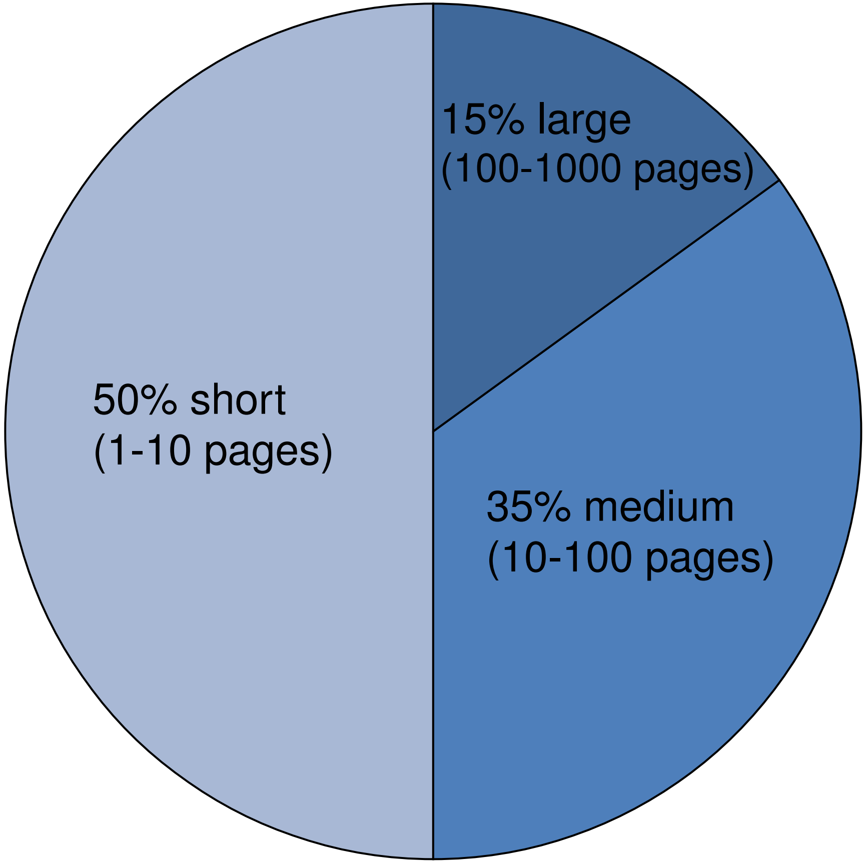
Intrinsic / external ratio



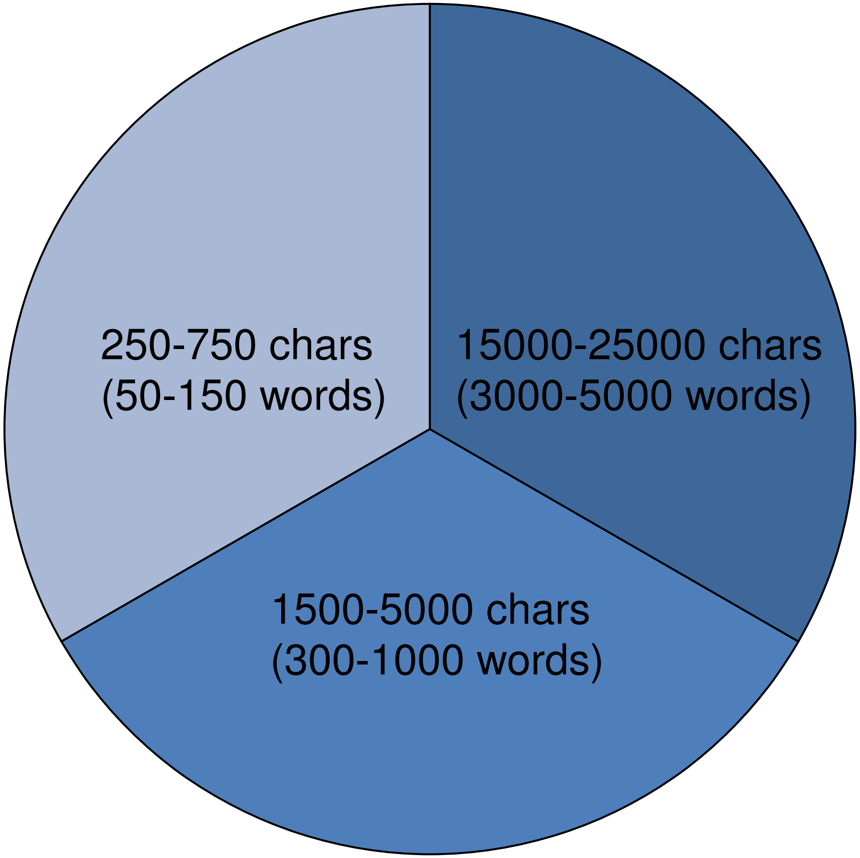
Suspicious / source ratio



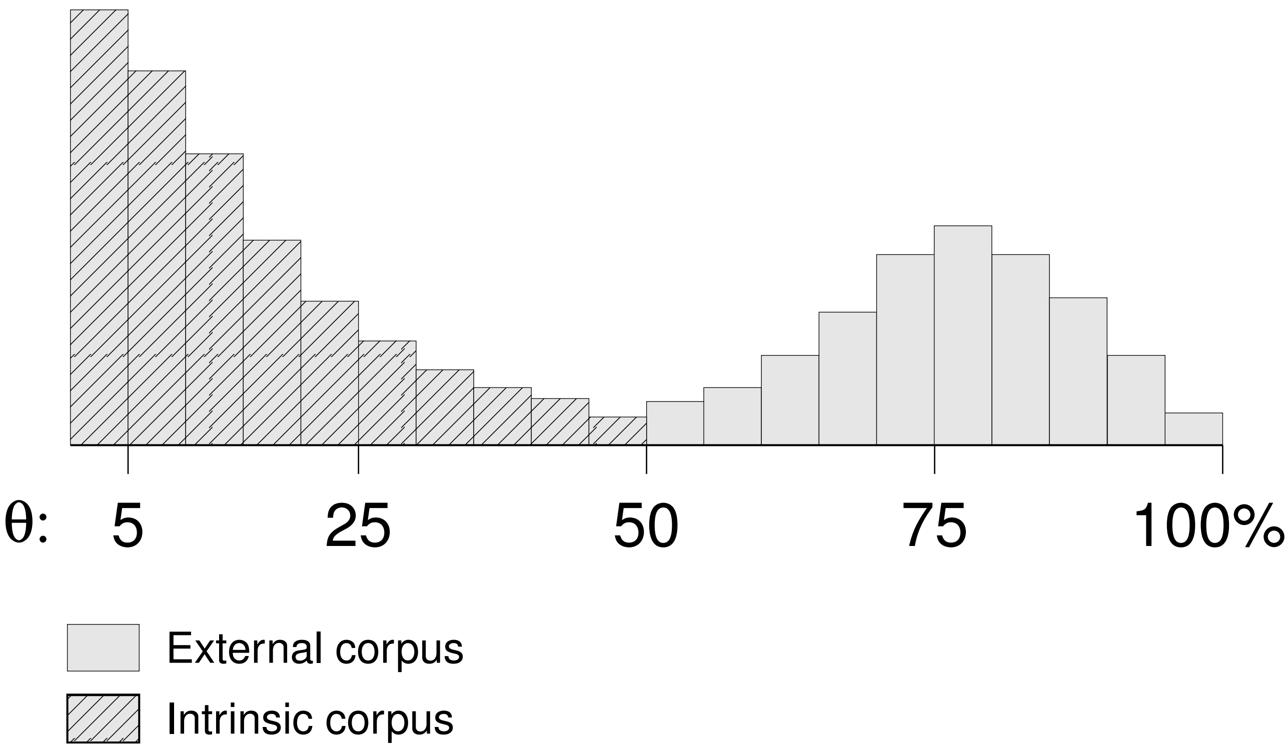
Document length



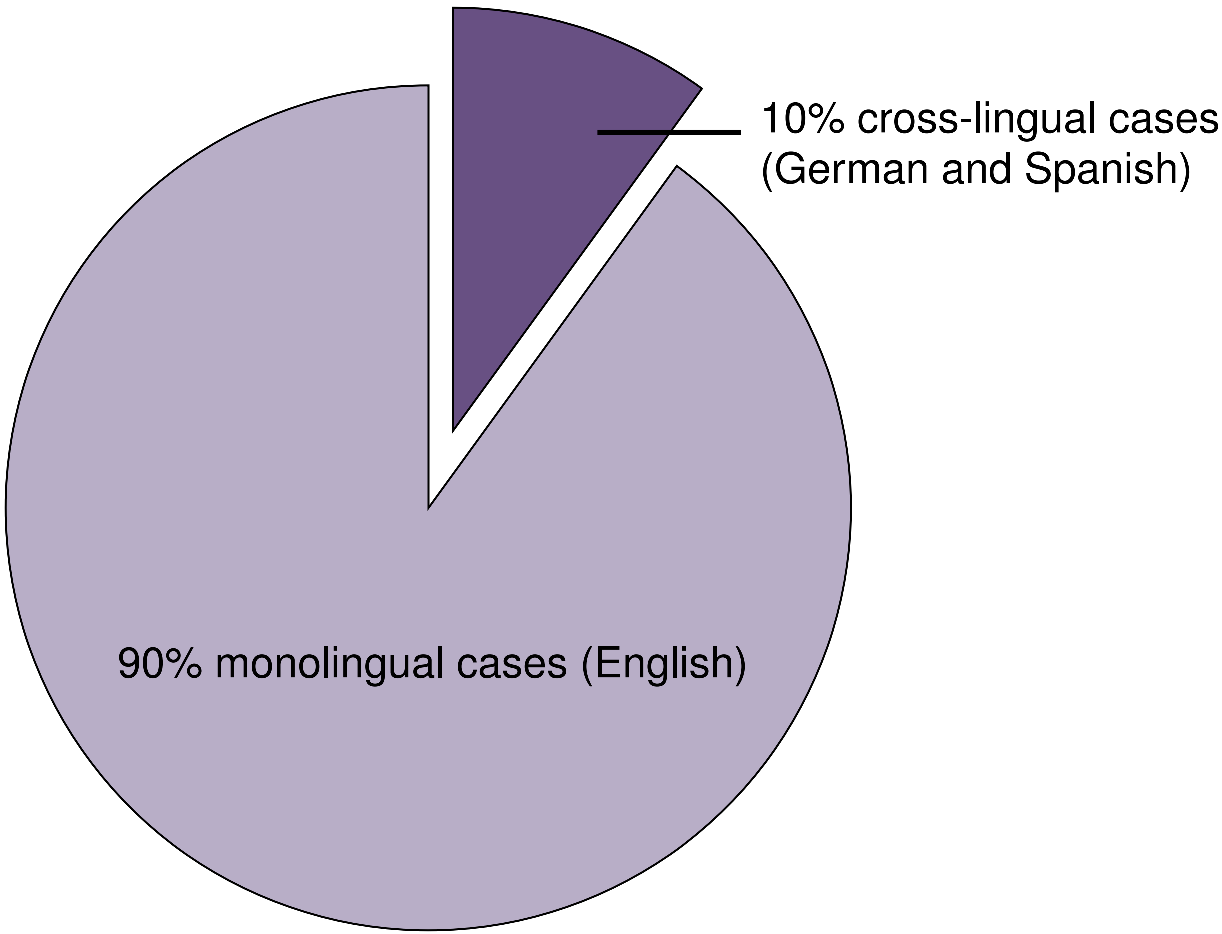
Case length


250-750 chars
(50-150 words)
15000-25000 chars
(3000-5000 words)
1500-5000 chars
(300-1000 words)

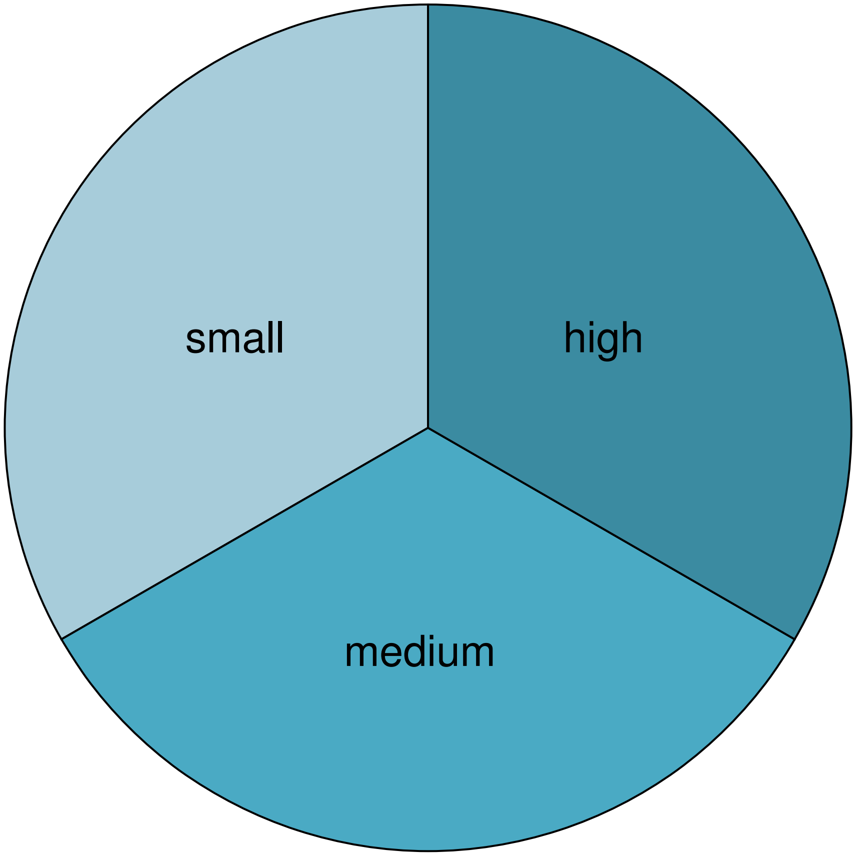
Fraction of plagiarism per document



Linguality

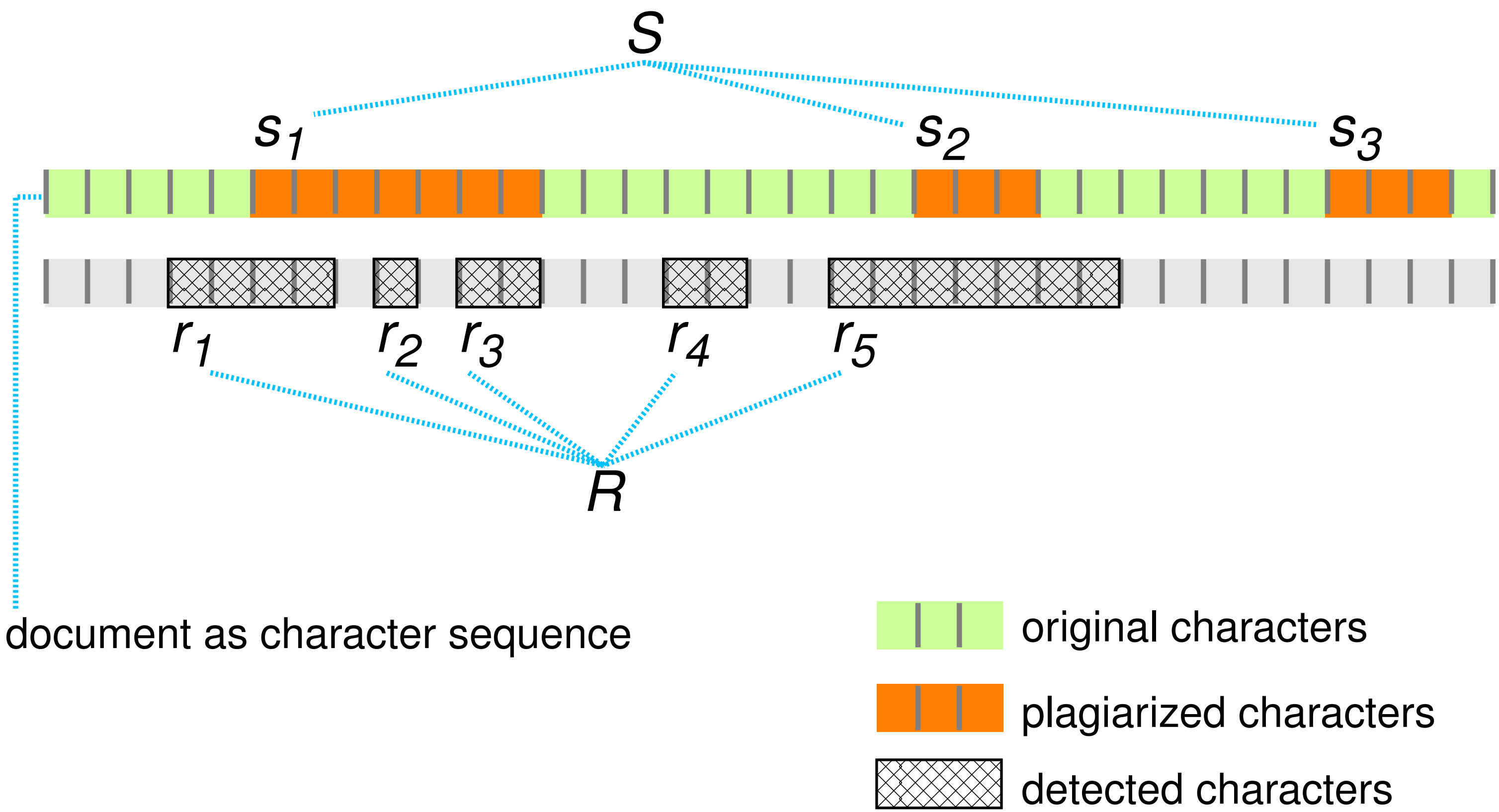

10% cross-lingual cases
(German and Spanish)
90% monolingual cases (English)

Obfuscation



Detection Performance Measures

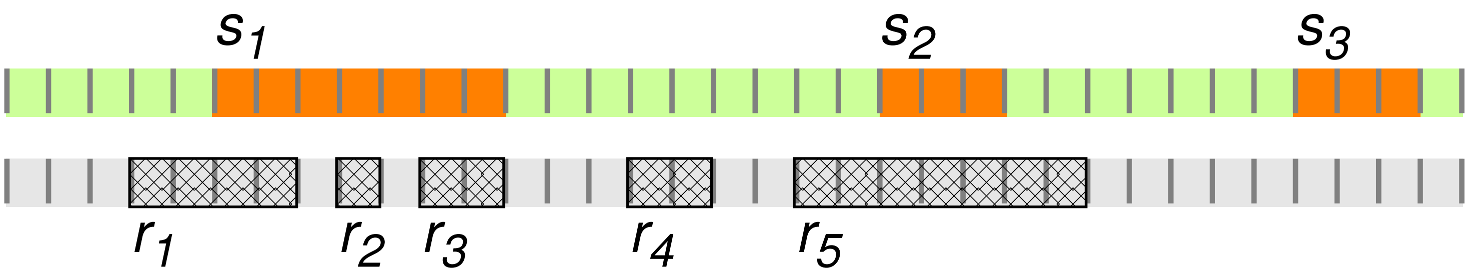
Terminology



- $s_i \in S$ 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:

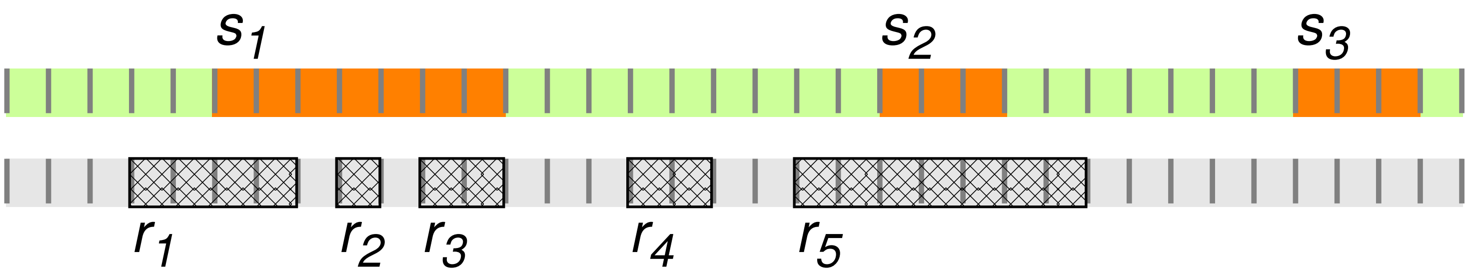
$$rec_{PDA} = \frac{8}{13} \qquad prec_{PDA} = \frac{8}{16}$$

\+ Simple to understand and simple to compute by counting char overlaps.

– Rewards the detection of long sections which are typically easier to detect.

Detection Performance Measures

Macro-averaged Recall and Precision



- Macro-averaged recall computes straightforward:

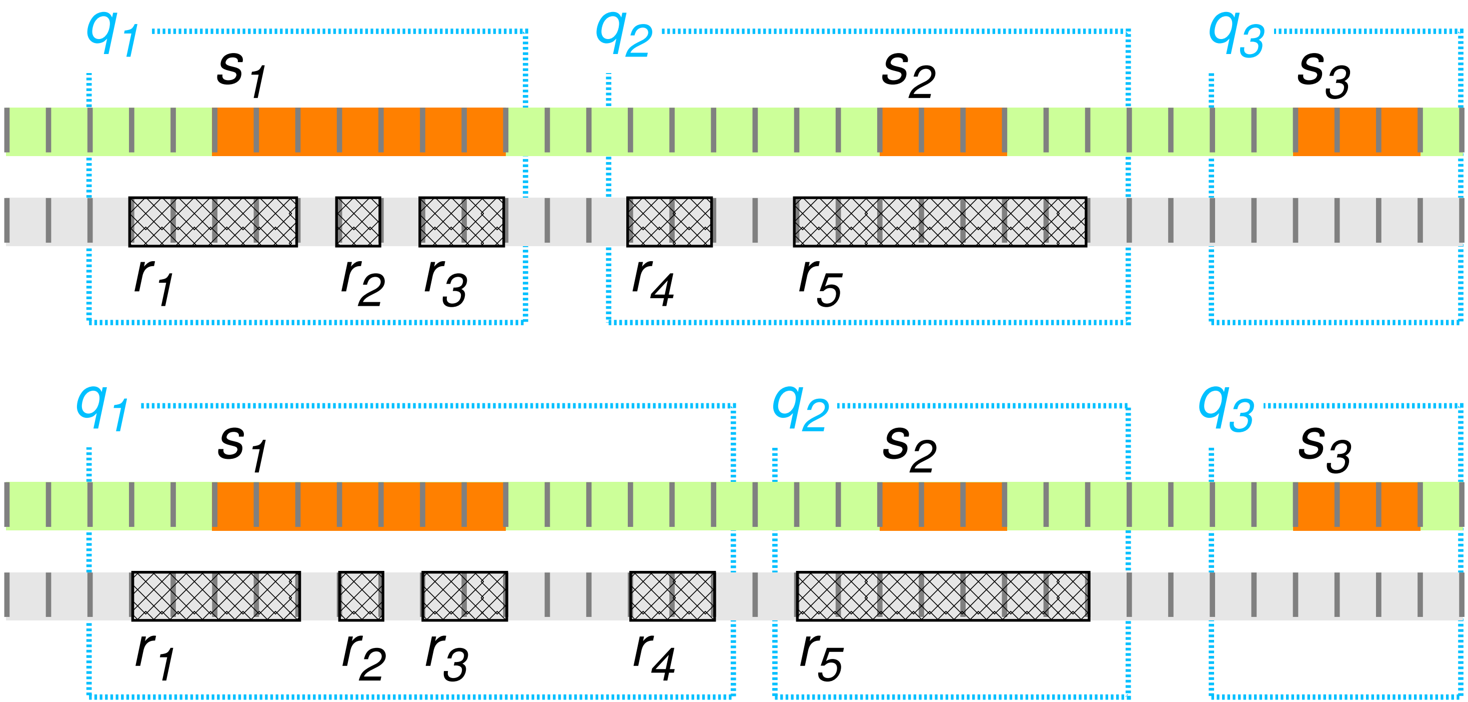
$$rec_{PDA}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|s \sqcap \bigcup_{r \in R} r|}{|s|},$$

  where $\sqcap$ computes the positionally overlapping characters.

- But macro-averaged precision is undefined!

Detection Performance Measures

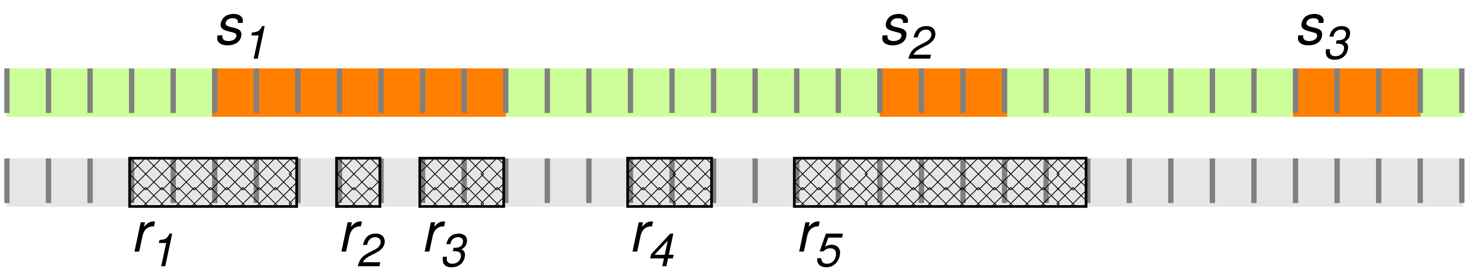
Macro-averaged Recall and Precision



- 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.

Detection Performance Measures

Macro-averaged Recall and Precision



- Therefore we define precision in an new way:

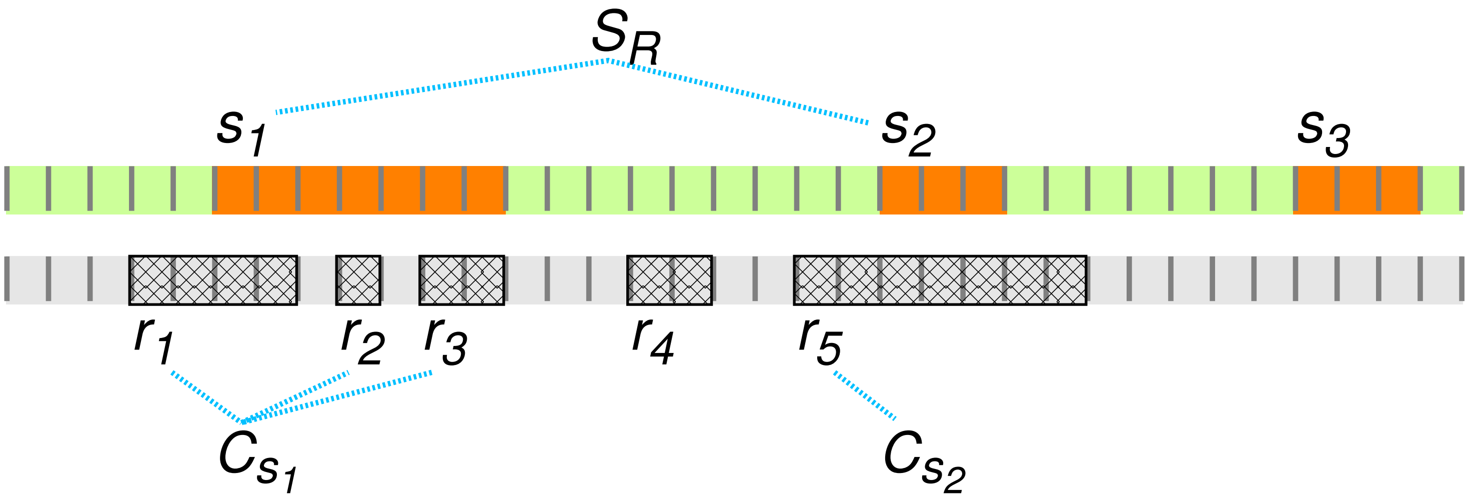
$$prec_{PDA}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|r \sqcap \bigcup_{s \in S} s|}{|r|},$$

where $\sqcap$ 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.

Detection Performance Measures

Detection Granularity



- PDAs often report the same s_i with multiple detections.

- We therefore define the granularity of a PDA as follows:

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

where

- $S_R = \{s \mid s \in S \ \wedge \ \exists r \in R: \ s \cap r \neq \emptyset\}$ denotes the detected subset of S, and
- $C_s = \{r \mid r \in R \ \wedge \ s \cap r \neq \emptyset\}$ 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:

$$\mathit{overall}_{PDA}(S, R) = \frac{F}{\log_2(1 + \mathit{gran}_{PDA})},$$

  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

Heuristic Retrieval	Detailed Analysis	Participant
Retrieval Model Character-16-gram VSM (frequency weights, cosine similarity) *Comparison of D_q and D* Exhaustive *Candidates $D_x \subset D$ for a d_q* The 51 documents most similar to d_q.	*Exact Matches of d_q and $d_x \in D_x$* Character-16-grams *Match Merging Heuristic to get (s_q, s_x)* Computation of the distances of adjacent matches. Joining of the matches based on a Monte Carlo optimization. Refinement of the obtained section pairs, e.g., by discarding too small sections.	Grozea et al.
Retrieval Model Word-5-gram VSM (boolean weights, Jaccard similarity) *Comparison of D_q and D* Exhaustive *Candidates $D_x \subset D$ for a d_q* Documents which share at least 20 n-grams with d_q.	*Exact Matches of d_q and $d_x \in D_x$* Word-5-grams *Match Merging Heuristic to get (s_q, s_x)* 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.	Kasprzak et al.

• • •

Competition on Plagiarism Detection

Detection Performance in the External Plagiarism Detection Task

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

Competition on Plagiarism Detection

Detection Performance in the Intrinsic Plagiarism Detection Task

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

* 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

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