Overview of the 2nd International Competition on Plagiarism Detection

Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt,
Benno Stein, Paolo Rosso

Bauhaus-Universität Weimar & Universidad Politécnica de Valencia
http://pan.webis.de

The PAN Competition

The PAN Competition

2nd International Competition on Plagiarism Detection, PAN 2010

These days, plagiarism and text reuse is rife on the Web.

Task:

> Given a set of suspicious documents and a set of source documents, find all plagiarized sections in the suspicious documents and, if available, the corresponding source sections.

The PAN Competition

2nd International Competition on Plagiarism Detection, PAN 2010

These days, plagiarism and text reuse is rife on the Web.

Task:

> Given a set of suspicious documents and a set of source documents, find all plagiarized sections in the suspicious documents and, if available, the corresponding source sections.

Facts:

- 18 groups from 12 countries participated
- 15 weeks of training and testing (March – June)
- training corpus was the PAN-PC-09
- test corpus was the PAN-PC-10, a new version of last year's corpus.
- performance was measured by precision, recall, and granularity

The PAN Competition

Plagiarism Corpus PAN-PC-10[1]

Large-scale resource for the controlled evaluation of detection algorithms:

- 27 073 documents (obtained from 22 874 books from the Project Gutenberg[2])
- 68 558 plagiarism cases (about 0-10 cases per document)

[1] www.webis.de/research/corpora/pan-pc-10
[2] www.gutenberg.org

The PAN Competition

Plagiarism Corpus PAN-PC-10[1]

Large-scale resource for the controlled evaluation of detection algorithms:

- 27 073 documents (obtained from 22 874 books from the Project Gutenberg[2])
- 68 558 plagiarism cases (about 0-10 cases per document)

[1] www.webis.de/research/corpora/pan-pc-10
[2] www.gutenberg.org

PAN-PC-10 addresses a broad range of plagiarism situations by varying reasonably within the following parameters:

1. document length
2. document language
3. detection task
4. plagiarism case length
5. plagiarism case obfuscation
6. plagiarism case topic alignment

The PAN Competition

PAN-PC-10 Document Statistics

100% 27 073 documents

The PAN Competition

PAN-PC-10 Document Statistics

100% 27 073 documents

Document length:

50% short (1-10 pages)	35% medium (10-100 pages)	15% long (100-1 000 pp.)

The PAN Competition

PAN-PC-10 Document Statistics

100% 27 073 documents

Document length:

50% short (1-10 pages)	35% medium (10-100 pages)	15% long (100-1 000 pp.)

Document language:

80% English	10% de	10% es

The PAN Competition

PAN-PC-10 Document Statistics

100% 27 073 documents

Document length:

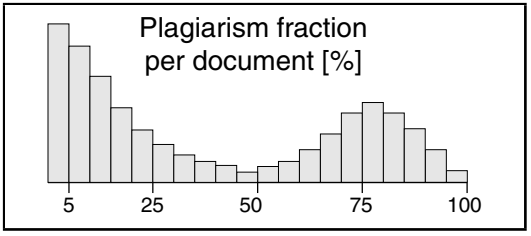
50% short (1-10 pages)	35% medium (10-100 pages)	15% long (100-1 000 pp.)

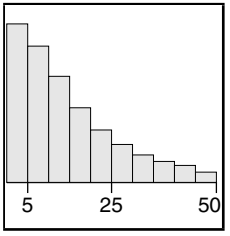
Document language:

80% English	10% de	10% es

Detection task:

70% external analysis		30% intrinsic analysis	
plagiarized	unmodified (plagiarism source)	plagiarized	unmodified





The PAN Competition

PAN-PC-10 Plagiarism Case Statistics

100% 68 558 plagiarism cases

The PAN Competition

PAN-PC-10 Plagiarism Case Statistics

100% 68 558 plagiarism cases

Plagiarism case length:

34% short (50-150 words)	33% medium (300-500 words)	33% long (3 000-5 000 words)

The PAN Competition

PAN-PC-10 Plagiarism Case Statistics

100% 68 558 plagiarism cases

Plagiarism case length:

34% short (50-150 words)	33% medium (300-500 words)	33% long (3 000-5 000 words)

Plagiarism case obfuscation:

40% none	40% artificial[3]		6%[4]	14%[5]	
	low obfuscation	high obfuscation	AMT	de	es

[3] Artificial plagiarism: algorithmic obfuscation.
[4] Simulated plagiarism: obfuscation via Amazon Mechanical Turk.
[5] Cross-language plagiarism: obfuscation due to machine translation de→en and es→en.

The PAN Competition

PAN-PC-10 Plagiarism Case Statistics

100% 68 558 plagiarism cases

Plagiarism case length:

34% short (50-150 words)	33% medium (300-500 words)	33% long (3 000-5 000 words)

Plagiarism case obfuscation:

40% none	40% artificial[3]		6%[4]	14%[5]	
	low obfuscation	high obfuscation	AMT	de	es

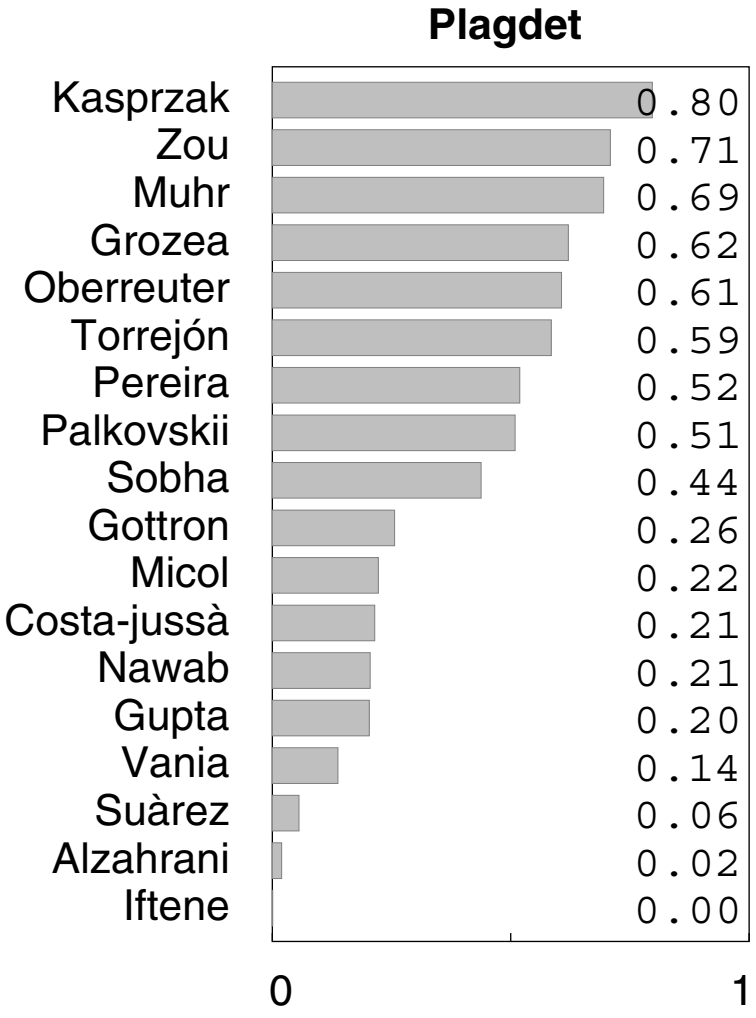
[3] Artificial plagiarism: algorithmic obfuscation.
[4] Simulated plagiarism: obfuscation via Amazon Mechanical Turk.
[5] Cross-language plagiarism: obfuscation due to machine translation de→en and es→en.

Plagiarism case topic alignment:

50% intra-topic	50% inter-topic

The PAN Competition

Plagiarism Detection Results

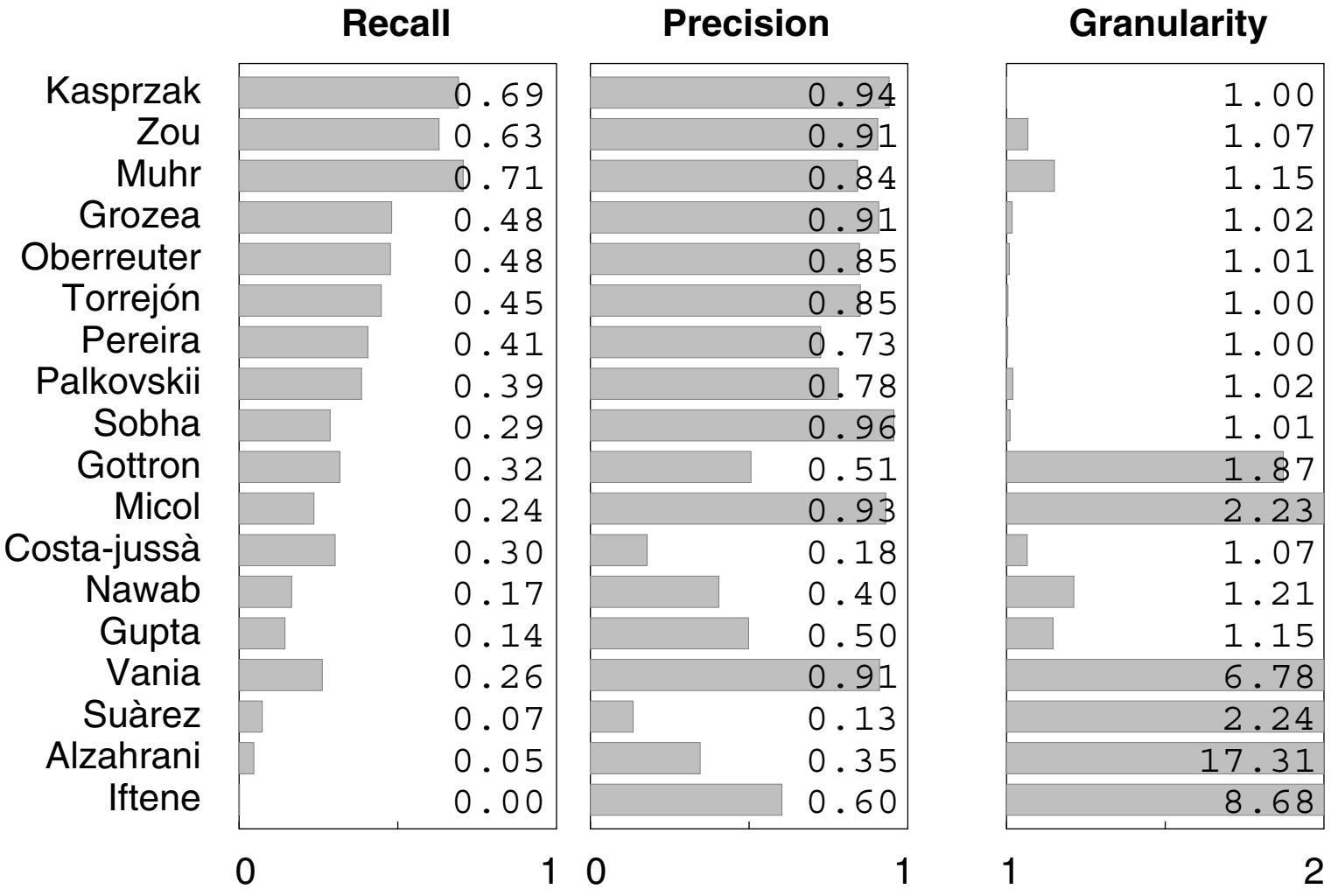


- Plagdet combines precision, recall, and granularity.
- Precision and recall are well-known, yet not often used in plagiarism detection.
- Granularity measures the number of times a single plagiarism case has been detected.

[Potthast et al., COLING 2010]

The PAN Competition

Plagiarism Detection Results



Summary

Summary

- More in the overview paper
  - This year's best practices for external detection.
  - Detection results with regard to every corpus parameter.
  - Comparison to PAN 2009.
- Lesson's learned & frontiers
  - Too much focus on local comparison instead of Web retrieval.
  - Intrinsic detection needs more attention.
  - Machine translated obfuscation is easily defeated in the current setting.
  - Short plagiarism cases and simulated plagiarism cases are difficult to detect.

Obfuscation

Real plagiarists modify their plagiarism to prevent detection,
i.e., to *obfuscate* their plagiarism.

Our task:

> Given a section s_{src}, create a section s_{plg} that has a high content similarity to s_{src} under some retrieval model but a different wording.

Obfuscation

Real plagiarists modify their plagiarism to prevent detection,
i.e., to *obfuscate* their plagiarism.

Our task:

> Given a section s_{src}, create a section s_{plg} that has a high content similarity to s_{src} under some retrieval model but a different wording.

Obfuscation strategies:

1. simulated: human writers
2. artificial: random text operations
3. artificial: semantic word variation
4. artificial: POS-preserving word shuffling
5. artificial: machine translation

Obfuscation Strategy: Human Writers

s_{plg} is created by manually rewriting s_{src}.

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

Examples:

- $s_{\text{plg}} =$ “Over the dog, which is lazy, quickly jumps the fox which is brown.”
- $s_{\text{plg}} =$ “Dogs are lazy which is why brown foxes quickly jump over them.”
- $s_{\text{plg}} =$ “A fast bay-colored vulpine hops over an idle canine.”

Reasonable scales can be achieved with this strategy via payed crowdsourcing, e.g., on Amazon’s Mechanical Turk.

Obfuscation Strategy: Random Text Operations

s_{plg} is created from s_{src} by shuffling, removing, inserting, or replacing words or short phrases at random.

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

Examples:

- $s_{\text{plg}} =$ “over The. the quick lazy dog context jumps brown fox”
- $s_{\text{plg}} =$ “over jumps quick brown fox The lazy. the”
- $s_{\text{plg}} =$ “brown jumps the. quick dog The lazy fox over”

Obfuscation Strategy: Semantic Word Variation

s_{plg} is created from s_{src} by replacing each word by one of its synonyms, antonyms, hyponyms, or hypernyms, chosen at random.

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

Examples:

- $s_{\text{plg}} =$ “The quick brown dodger leaps over the lazy canine.”
- $s_{\text{plg}} =$ “The quick brown canine jumps over the lazy canine.”
- $s_{\text{plg}} =$ “The quick brown vixen leaps over the lazy puppy.”

Obfuscation Strategy: POS-preserving Word Shuffling

Given the part of speech sequence of s_{src}, s_{plg} is created by shuffling words at random while retaining the original POS sequence.

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

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

Examples:

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

Obfuscation Strategy: Machine Translation

s_{plg} is created from s_{src} by translating it using machine translation (services).

$s_{\text{src}} =$ “Der flinke braune Fuchs hüpft über den faulen Hund.”

Examples:

- $s_{\text{plg}} =$ “The quick brown fox jumps over the lazy dog.”
- $s_{\text{plg}} =$ “The speedy brown fox hops over the lazy dog.”