Evaluation of plagiarism detection algorithms

Andreas Eiselt

Thesis defense
Bauhaus-Universität Weimar

1. Supervisor: Prof. Benno Stein
2. Supervisor: Junior-Prof. Dr. Hagen Höpfner
Tutor: Martin Potthast

23th of February, 2011

Table of Contents

1 Introduction

2 Plagiarism Corpus

3 Evaluation Measures

4 PAN-10

5 Conclusions

6 References

Figure: Dr. Karl-Theodor zu Guttenberg (Photo: Reuters)

Did Dr. Guttenberg commit plagiarism?

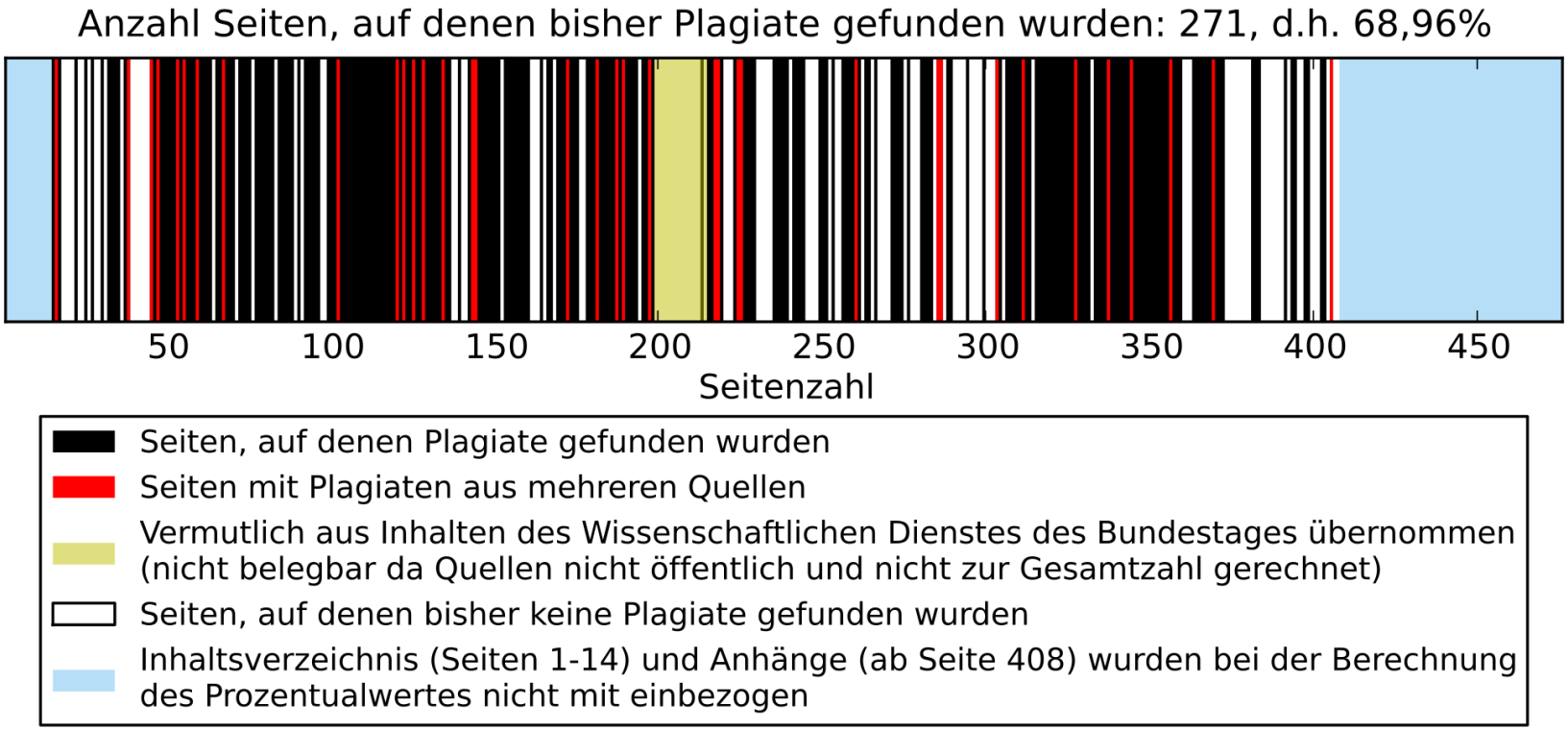


Figure: Plagiarism statistic of Dr. Guttenbergs dissertation (Source: GuttenPlag Wiki)

There is strong evidence against him.

Automatic Plagiarism Detection

Given a suspicious document...

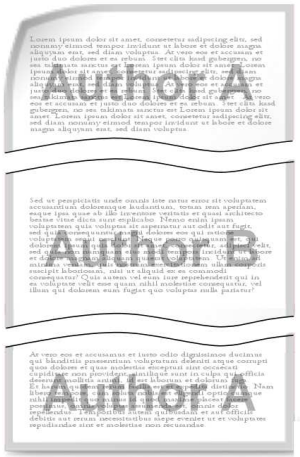
Task

- Find all plagiarized passages
- Provide the corresponding sources, if available

Approaches

- Intrinsic plagiarism detection
- External plagiarism detection

Automatic Plagiarism Detection - Approaches

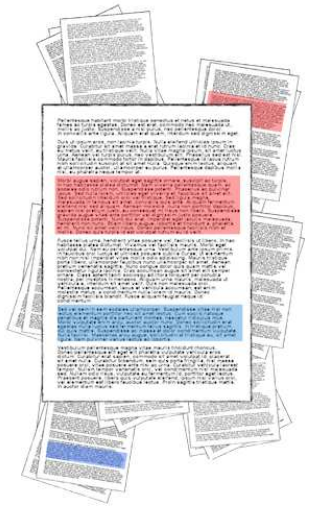


Insertion of text from author **B** into a text from author **A** causes style irregularities.

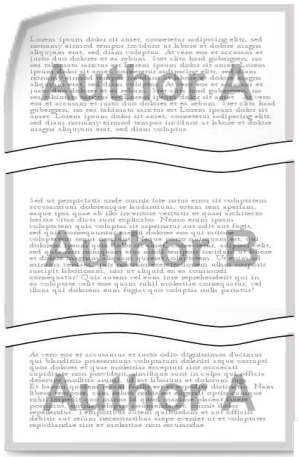
→ can be detected by **intrinsic** plagiarism detection algorithms

Better evidence than style irregularities is the source of a plagiarism case.

→ can be detected by **external** plagiarism detection algorithms



Automatic Plagiarism Detection - Approaches

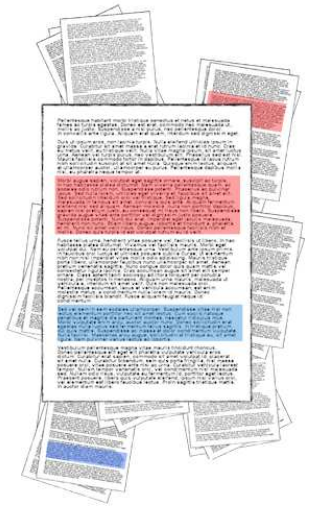


Insertion of text from author **B** into a text from author **A** causes style irregularities.

→ can be detected by **intrinsic** plagiarism detection algorithms

Better evidence than style irregularities is the source of a plagiarism case.

→ can be detected by **external** plagiarism detection algorithms



Research Question

A lot of plagiarism detection algorithms:

Research Question

But how to compare them?

Contributions

Implementation of an Evaluation Framework for Plagiarism Detection Algorithms

1 Corpus of Plagiarism Cases

- Different types of plagiarism:
  - copy & paste plagiarism
  - paraphrased plagiarism
  - cross-lingual plagiarism
- According plagiarism cases for intrinsic and external plagiarism detection

2 Evaluation Measures

- Two measures to quantify *Precision* and *Recall*
- *Granularity* to quantify whether the contiguity between plagiarized text passages is properly recognized

Plagiarism Corpus

1 Introduction

2 Plagiarism Corpus

3 Evaluation Measures

4 PAN-10

5 Conclusions

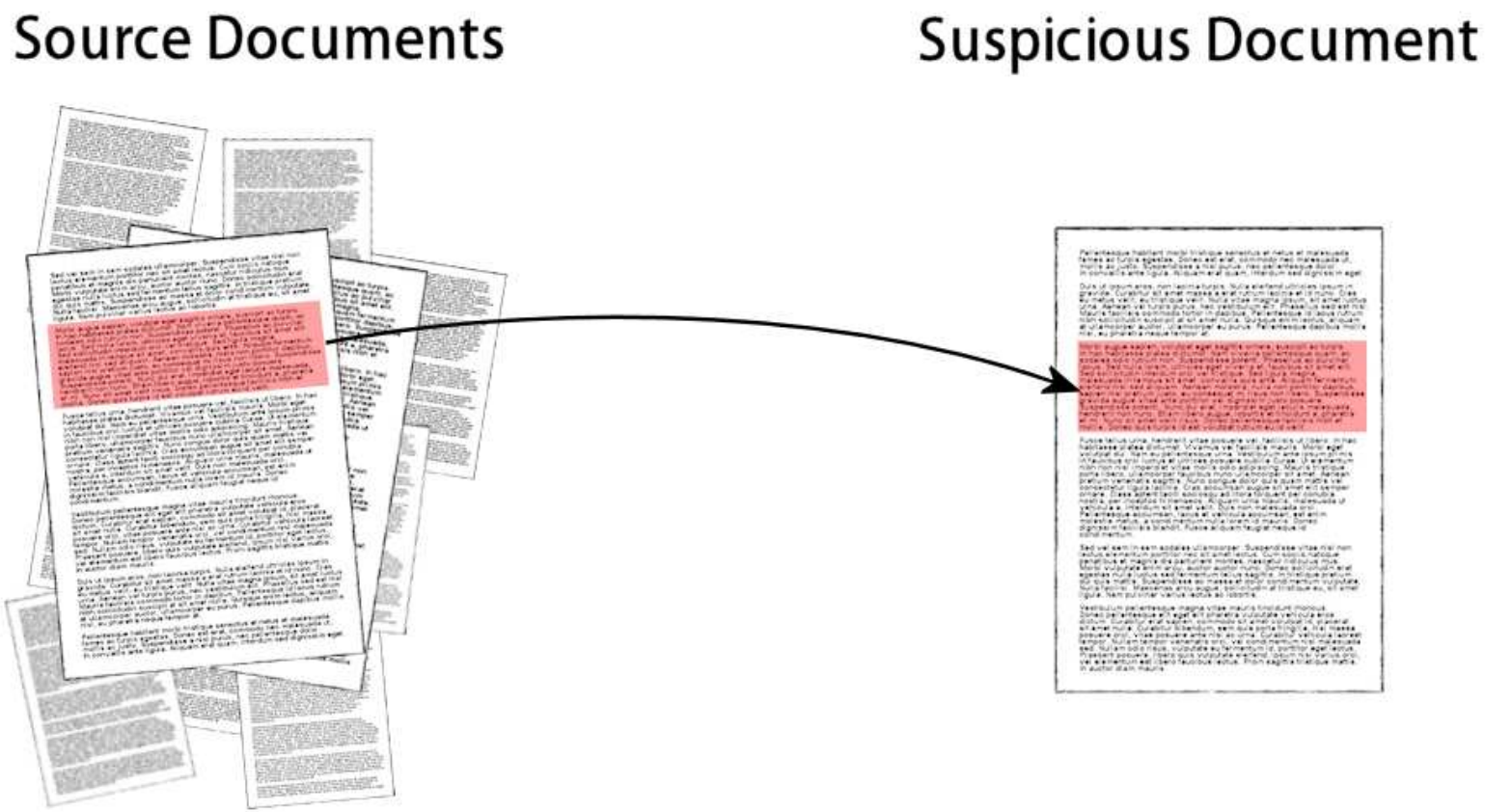
6 References

Plagiarism Process

Source Documents

Suspicious Document

Plagiarism Process

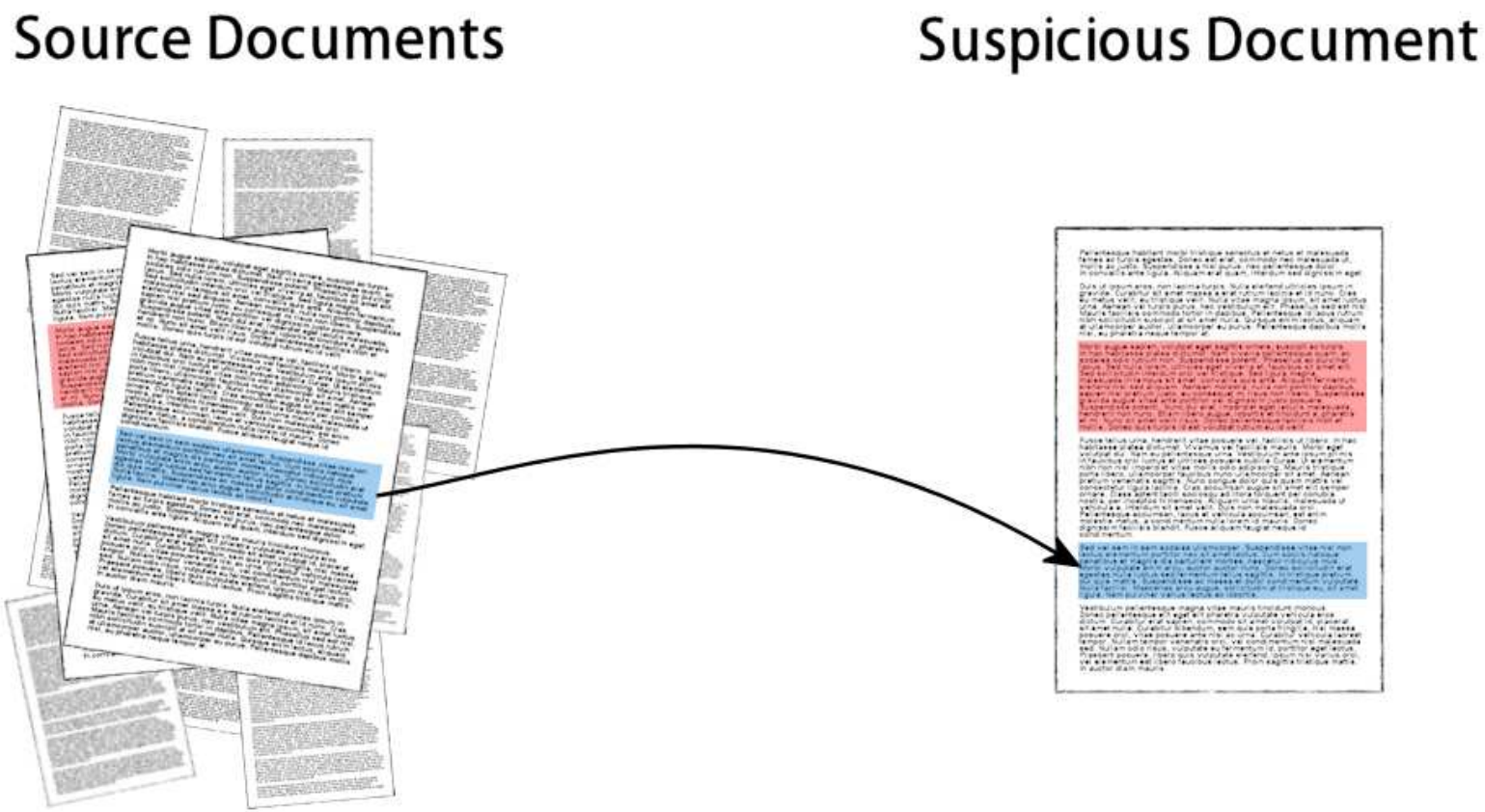


Plagiarism Process

Source Documents

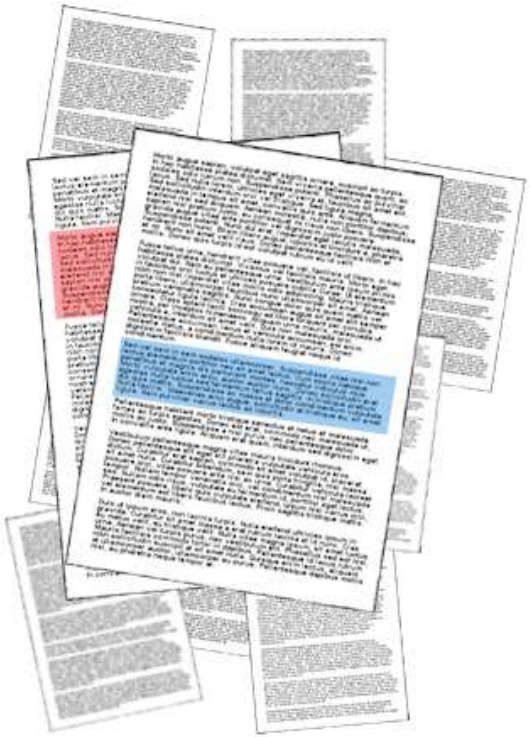
Suspicious Document

Plagiarism Process



Plagiarism Process

Source Documents



Suspicious Documents

Document Source

Project Gutenberg

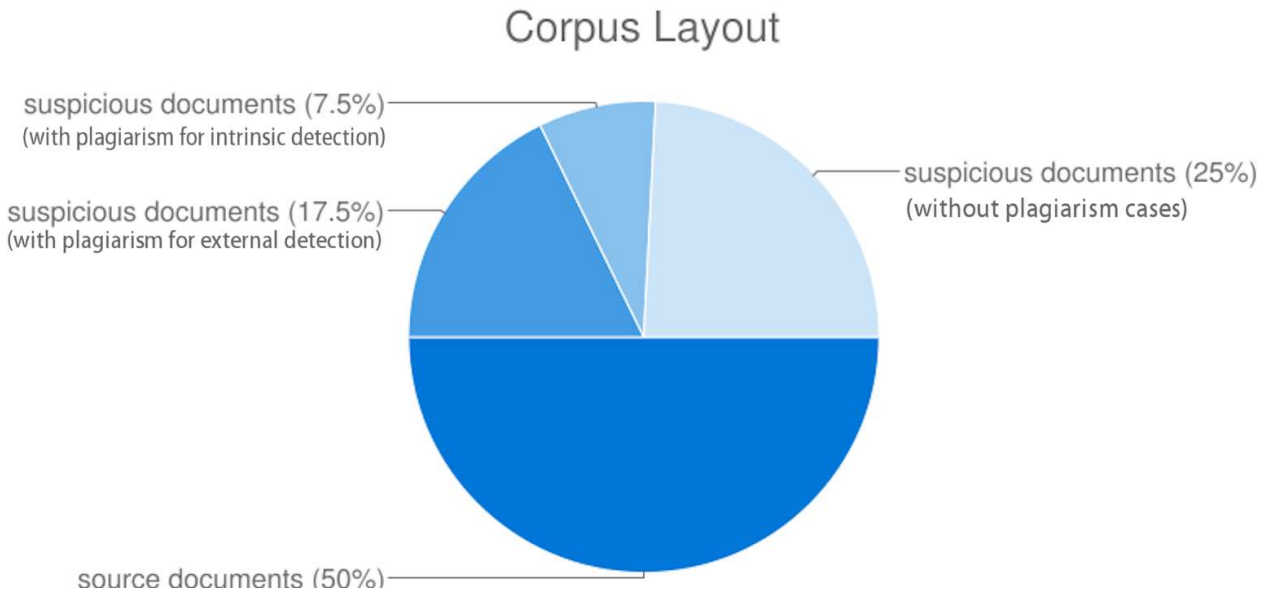
- $>$ 22,000 documents (en,de,es)
- $>$ 80 thematic categories (philosophy, architecture, children's book, etc.)
- Documents are *Public Domain*

Problem

- Not all documents are suitable
- Documents already contain "plagiarism"
- Documents contain unwanted meta information

$\rightarrow$ extensive preprocessing (about 16,000 documents left)

Layout



Document Statistics (27,073 documents)

Document Length			*Plagiarism per Document*		
short	(1-10 pp.)	50%	hardly	(5%-20%)	45%
medium	(10-100 pp.)	35%	medium	(20%-50%)	15%
long	(100-1000 pp.)	15%	much	(50%-80%)	25%
			entirely	(>80%)	15%

Plagiarism generation

Cross-lingual

- Machine Translation (Google Translation)

Paraphrased

- **Manual:** using Crowdsourcing (low cost, big community)
- **Automatic:**
  - Semantic word variation
  - POS-preserving word shuffling
  - Random text operations
  - Sentence/Phrase shuffling

Obfuscation Examples

- **Original Text**
  The quick brown fox jumps over the lazy dog.
- **Manual Obfuscation (by a human)**
  Dogs are lazy which is why brown foxes quickly jump over them.
- **Semantic Word Variation**
  The quick brown dodger leaps over the lazy canine.
- **POS-preserving Word Shuffling**
  The brown lazy fox jumps over the quick dog.
- **Random Text Operations**
  over The. the dog quick lazy human jumps brown fox
- **Phrase Shuffling**
  The lazy dog jumps over the quick brown fox.

Obfuscation Examples

- **Original Text**
  The quick brown fox jumps over the lazy dog.
- **Manual Obfuscation (by a human)**
  Dogs are lazy which is why brown foxes quickly jump over them.
- **Semantic Word Variation**
  The quick brown dodger leaps over the lazy canine.
- **POS-preserving Word Shuffling**
  The brown lazy fox jumps over the quick dog.
- **Random Text Operations**
  over The. the dog quick lazy human jumps brown fox
- **Phrase Shuffling**
  The lazy dog jumps over the quick brown fox.

Obfuscation Examples

- **Original Text**
  The quick brown fox jumps over the lazy dog.
- **Manual Obfuscation (by a human)**
  Dogs are lazy which is why brown foxes quickly jump over them.
- **Semantic Word Variation**
  The quick brown dodger leaps over the lazy canine.
- **POS-preserving Word Shuffling**
  The brown lazy fox jumps over the quick dog.
- **Random Text Operations**
  over The. the dog quick lazy human jumps brown fox
- **Phrase Shuffling**
  The lazy dog jumps over the quick brown fox.

Obfuscation Examples

- **Original Text**
  The quick brown fox jumps over the lazy dog.
- **Manual Obfuscation (by a human)**
  Dogs are lazy which is why brown foxes quickly jump over them.
- **Semantic Word Variation**
  The quick brown dodger leaps over the lazy canine.
- **POS-preserving Word Shuffling**
  The brown lazy fox jumps over the quick dog.
- **Random Text Operations**
  over The. the dog quick lazy human jumps brown fox
- **Phrase Shuffling**
  The lazy dog jumps over the quick brown fox.

Obfuscation Examples

- **Original Text**
  The quick brown fox jumps over the lazy dog.
- **Manual Obfuscation (by a human)**
  Dogs are lazy which is why brown foxes quickly jump over them.
- **Semantic Word Variation**
  The quick brown dodger leaps over the lazy canine.
- **POS-preserving Word Shuffling**
  The brown lazy fox jumps over the quick dog.
- **Random Text Operations**
  over The. the dog quick lazy human jumps brown fox
- **Phrase Shuffling**
  The lazy dog jumps over the quick brown fox.

Obfuscation Examples

- **Original Text**
  The quick brown fox jumps over the lazy dog.
- **Manual Obfuscation (by a human)**
  Dogs are lazy which is why brown foxes quickly jump over them.
- **Semantic Word Variation**
  The quick brown dodger leaps over the lazy canine.
- **POS-preserving Word Shuffling**
  The brown lazy fox jumps over the quick dog.
- **Random Text Operations**
  That the quick brown human fox jumps over the dog.
- **Phrase Shuffling**
  The lazy dog jumps over the quick brown fox.

Obfuscation Examples

- **Original Text**
  The quick brown fox jumps over the lazy dog.
- **Manual Obfuscation (by a human)**
  Dogs are lazy which is why brown foxes quickly jump over them.
- **Semantic Word Variation**
  The quick brown dodger leaps over the lazy canine.
- **POS-preserving Word Shuffling**
  The brown lazy fox jumps over the quick dog.
- **Random Text Operations**
  That the quick brown human fox jumps over the dog.
- **Phrase Shuffling**
  The lazy dog jumps over the quick brown fox.

Plagiarism Statistics

Plagiarism Case Statistics (68,558 plagiarism cases)

Obfuscation		*Case Length*		
none	40%	short	(50-150 words)	34%
automatic		medium	(300-500 words)	33%
– low obfuscation	20%	long	(3000-5000 words)	33%
– high obfuscation	20%			
manual	6%			
translated ({de,es} to en)	14%			

Evaluation Measures

1 Introduction

2 Plagiarism Corpus

3 Evaluation Measures

4 PAN-10

5 Conclusions

6 References

Evaluation Measures

Initial situation

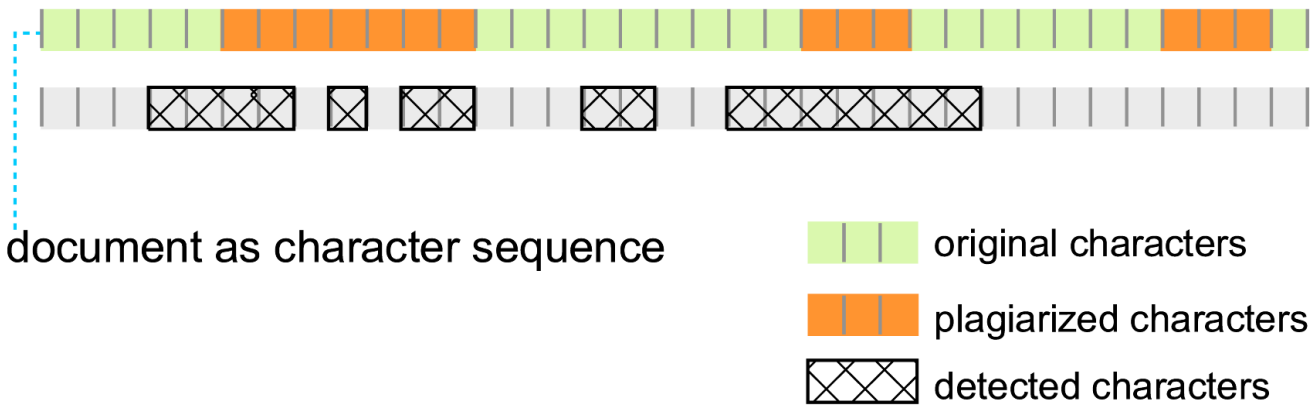
- No standard evaluation measures have been previously defined
- Different evaluations are hard to compare

Approach
[Potthast et al., 2010b]

Four metrics to quantify the performance of a plagiarism detection algorithm:

- *Precision*
- *Recall*
- *Granularity*
- *Plagdet* (overall score)

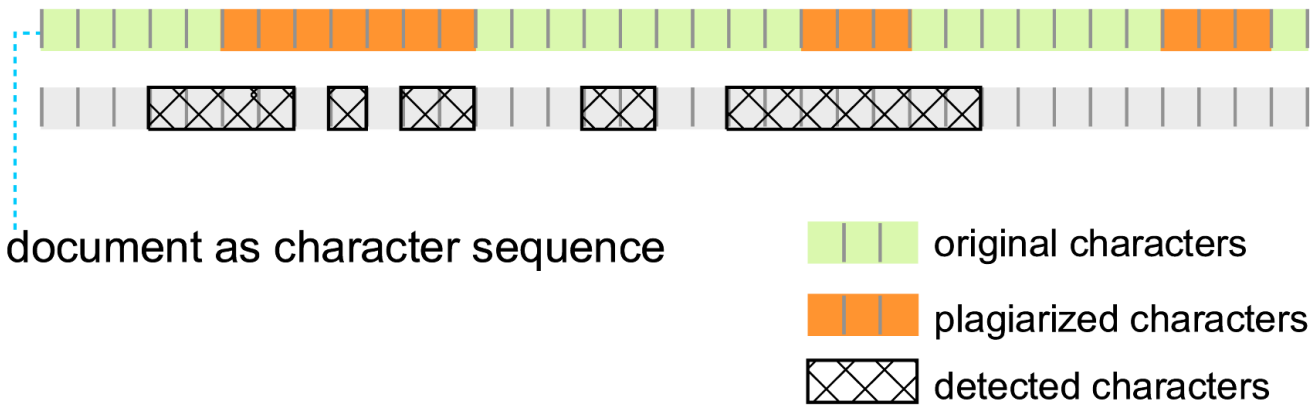
Evaluation Measures



Recall:

$$recall = \frac{|\{plagiarized\ \ characters\} \cap \{detected\ \ characters\}|}{|\{plagiarized\ \ characters\}|}$$

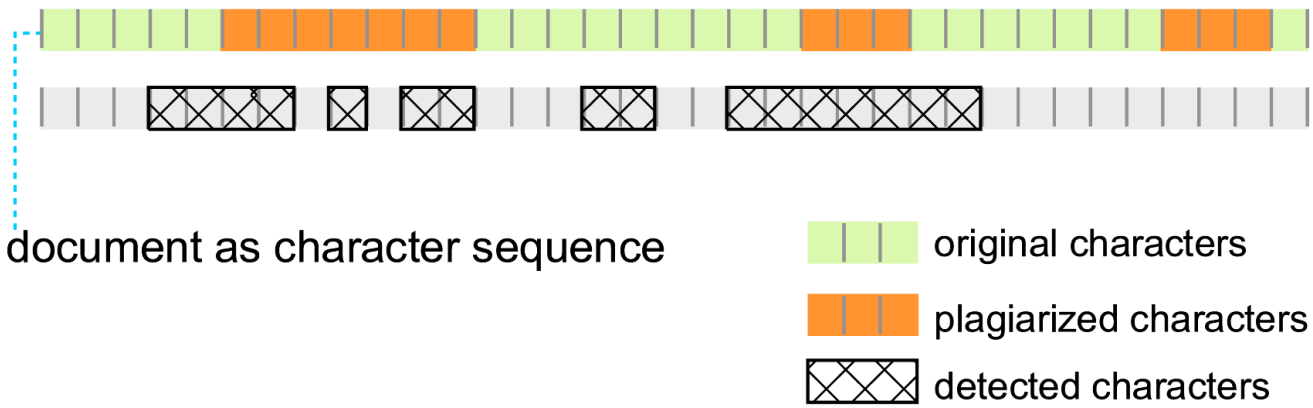
Evaluation Measures



Precision:

$$precision = \frac{|\{plagiarized\ \ characters\} \cap \{detected\ \ characters\}|}{|\{detected\ \ characters\}|}$$

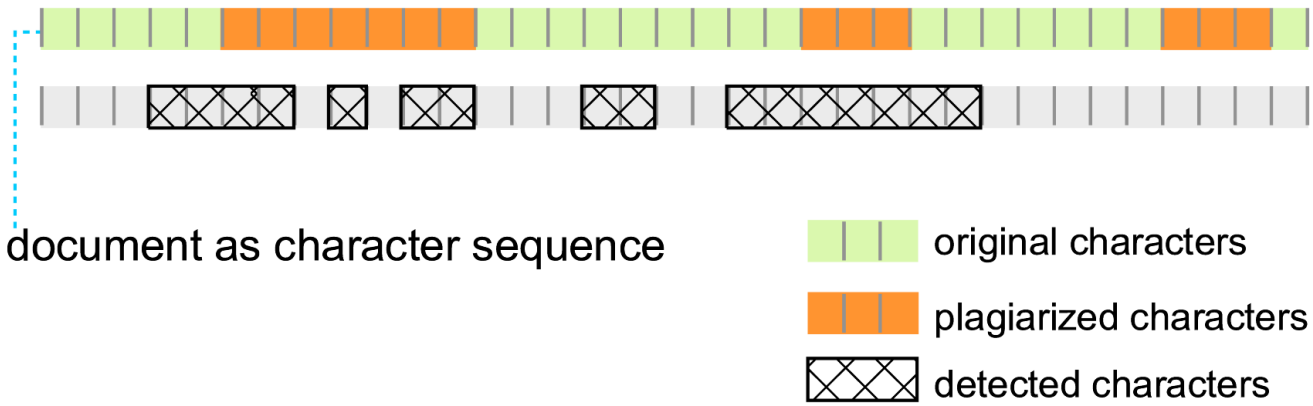
Evaluation Measures



F-Measure:

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

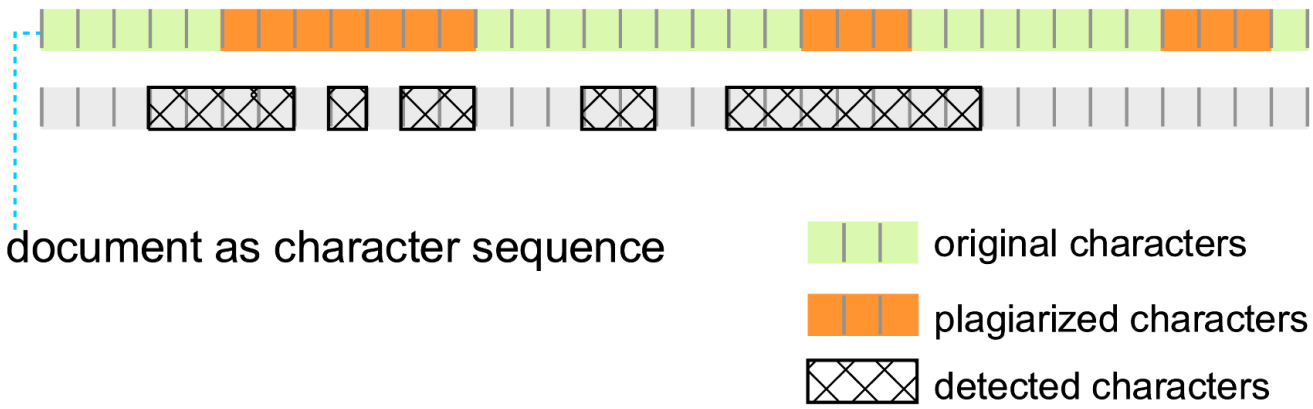
Evaluation Measures



Granularity:

In average, how many plagiarism cases are reported per plagiarism case?

Evaluation Measures



Plagdet (overall score):

$$plagdet = \frac{F}{\log_2(1+granularity)}$$

PAN-10 Plagiarism Detection Competition

1 Introduction

2 Plagiarism Corpus

3 Evaluation Measures

4 PAN-10

5 Conclusions

6 References

PAN-10 - 2nd Competition on Plagiarism Detection

Participants

- 48 researchers in 18 groups
- from 15 countries in Europe, Asia and South America
- 3 months development time

PAN-10 - Overall Results

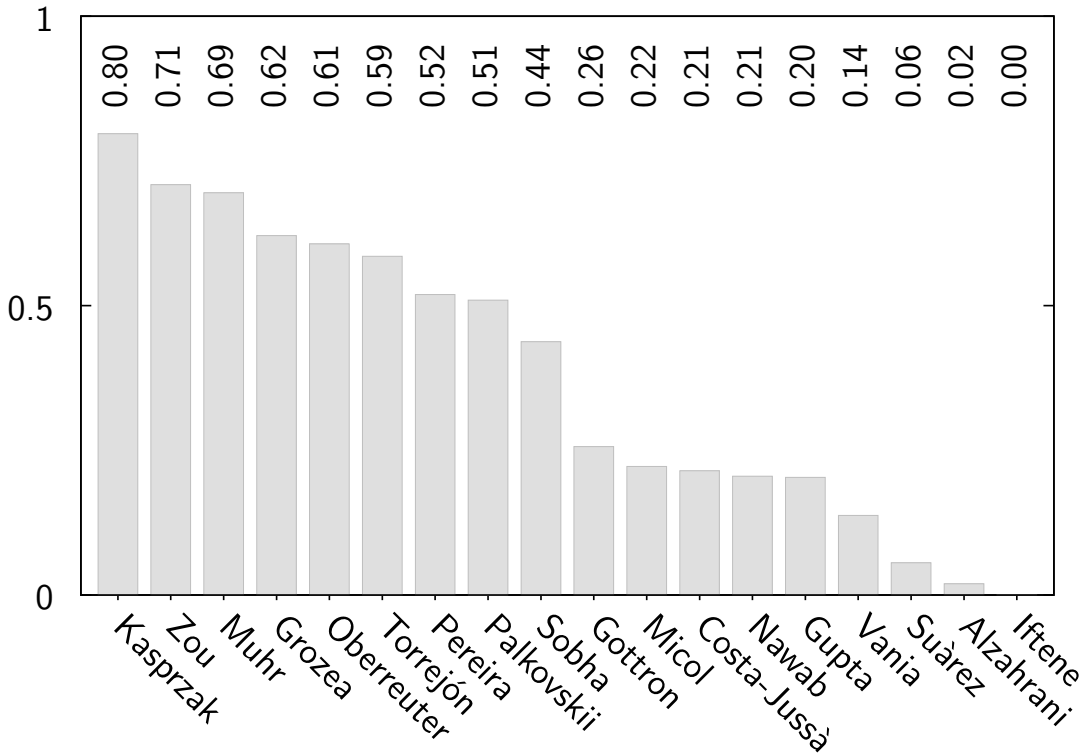


Figure: plagdet score (y-axis) for all participants at PAN-10 plagiarism competition.

Conclusions

1 Introduction

2 Plagiarism Corpus

3 Evaluation Measures

4 PAN-10

5 Conclusions

6 References

Summary

- Large-scale corpus and tailored performance measures for plagiarism detection for the controlled evaluation of detection algorithms
- Corpus has proven its suitability in practice (PAN09 & PAN10)
- Corpus already features various kinds of plagiarism cases
- In relation to previous corpora our corpus reveals a high degree of maturity [Potthast et al., 2010b]
- 31 plagiarism detectors have been compared using our evaluation framework

Discussion

- Corpus is an approximation of the world of plagiarism
+ Corpus provides a wide spectrum of possible plagiarism characteristics
- Evaluation framework currently doesn't consider the document retrieval performance

→ Projected for PAN-11/PAN-12

- Currently no differentiation between correctly cited text and plagiarism

→ Projected for PAN-11/PAN-12

References I


Barrón-Cedeño, A., Potthast, M., Rosso, P., Stein, B., and Eiselt, A. (2010).
Corpus and Evaluation Measures for Automatic Plagiarism Detection.
In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 10)*. European Language Resources Association (ELRA).


Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., and Rosso, P. (2010a).
Overview of the 2nd International Competition on Plagiarism Detection.
In Braschler, M. and Harman, D., editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*.


Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. (2010b).
An Evaluation Framework for Plagiarism Detection.
In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China. Association for Computational Linguistics.


Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., and Rosso, P. (2009).
Overview of the 1st International Competition on Plagiarism Detection.
In Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Agirre, E., editors, *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pages 1–9. CEUR-WS.org.

Note of Thanks

Thank you for your ongoing support
Martin Potthast
Prof. Benno Stein
Alberto Barrón-Cedeño
Prof. Paolo Rosso