Corpus and Evaluation Measures for Automatic Plagiarism Detection

Alberto Barrón-Cedeño[1], Martin Potthast[2], Paolo Rosso[1], Benno Stein[2], Andreas Eiselt[2]

[1]NLE Lab, Universidad Politécnica de Valencia, Spain
{lbarron, prosso}@dsic.upv.es

[2]Webis, Bauhaus-Universität Weimar, Germany
{martin.potthast, benno.stein, andreas.eiselt}@uni-weimar.de

LREC 2010
May, 2010

Outline

Introduction

Text reuse

- The reuse (even after modification) of text.

(from [Clough et al., 2002], [IEEE, 2008], and [Bierce, 1911])

Introduction

Text reuse

- The reuse (even after modification) of text.

Plagiarism

- the reuse of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source

(from [Clough et al., 2002], [IEEE, 2008], and [Bierce, 1911])

Introduction

Text reuse

- The reuse (even after modification) of text.

Plagiarism

- the reuse of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source
- to take the thought or style of another writer whom one has never, never read

(from [Clough et al., 2002], [IEEE, 2008], and [Bierce, 1911])

Introduction: Relevance

1986 In a survey over 380 students, 30% admitted cheating on their assignments [Haines et al., 1986]

Introduction: Relevance

1986 In a survey over 380 students, 30% admitted cheating on their assignments [Haines et al., 1986]

2000 With the advent of the Web, plagiarism is on the rise, it is even named cyberplagiarism [Baty, 2000, Anderson, 1999]

Introduction: Relevance

1986 In a survey over 380 students, 30% admitted cheating on their assignments [Haines et al., 1986]

2000 With the advent of the Web, plagiarism is on the rise, it is even named cyberplagiarism
[Baty, 2000, Anderson, 1999]

2007 Copy-paste syndrome
[Weber, 2007, Kulathuramaiyer and Maurer, 2007]

Introduction: Relevance

1986 In a survey over 380 students, 30% admitted cheating on their assignments [Haines et al., 1986]

2000 With the advent of the Web, plagiarism is on the rise, it is even named cyberplagiarism
[Baty, 2000, Anderson, 1999]

2007 Copy-paste syndrome
[Weber, 2007, Kulathuramaiyer and Maurer, 2007]

2008 Some professors estimate that around 28% of their pupils reports include plagiarism
[Association of Teachers and Lecturers, 2008]

Introduction: Relevance

1986 In a survey over 380 students, 30% admitted cheating on their assignments [Haines et al., 1986]

2000 With the advent of the Web, plagiarism is on the rise, it is even named cyberplagiarism [Baty, 2000, Anderson, 1999]

2007 Copy-paste syndrome [Weber, 2007, Kulathuramaiyer and Maurer, 2007]

2008 Some professors estimate that around 28% of their pupils reports include plagiarism [Association of Teachers and Lecturers, 2008]

2009 Wikipedia is considered a preferred source for plagiarists [Martínez, 2009]

Introduction: Automatic Plagiarism Detection

Goal Identifying the plagiarized sections in a suspicious document d_q.

Introduction: Automatic Plagiarism Detection

Goal Identifying the plagiarized sections in a suspicious document d_q.

Objective Providing experts with evidence to decide whether a case of plagiarism is at hand.

Introduction: Automatic Plagiarism Detection

Goal Identifying the plagiarized sections in a suspicious document d_q.

Objective Providing experts with evidence to decide whether a case of plagiarism is at hand.

Approaches

- *intrinsic*
- *external*

Introduction: Intrinsic Plagiarism Detection

An expert is often able to detect plagiarism by reading a document

Insertion of text from a different author into d_q causes **style** and **complexity** irregularities

[Meyer zu Eißen and Stein, 2006], [Stamatatos, 2009]

Introduction: Intrinsic Plagiarism Detection

An expert is often able to detect plagiarism by reading a document

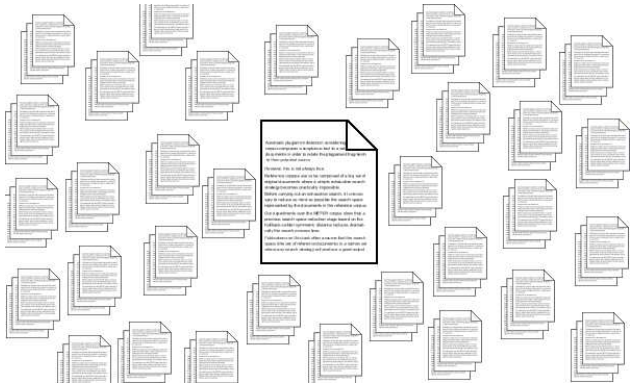
Insertion of text from a different author into d_q causes **style** and **complexity** irregularities

Quantification can be made by measuring…

Text readability	Gunning Fog, Flesch–Kincaid
Vocabulary richness	types/tokens ratio
Basic statistics	avg. sentence length, avg. word length
n-grams profiles	character level statistics

[Meyer zu Eißen and Stein, 2006], [Stamatatos, 2009]

Introduction: External Plagiarism Detection

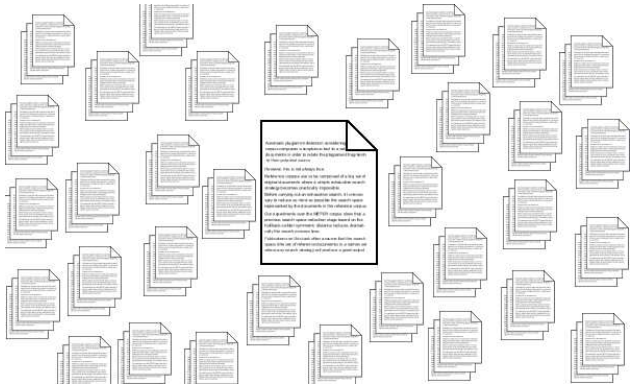


Better evidence than style irregularities is if the source of a plagiarism case can be provided

It is closer to information retrieval

[Potthast et al., 2009]

Introduction: External Plagiarism Detection



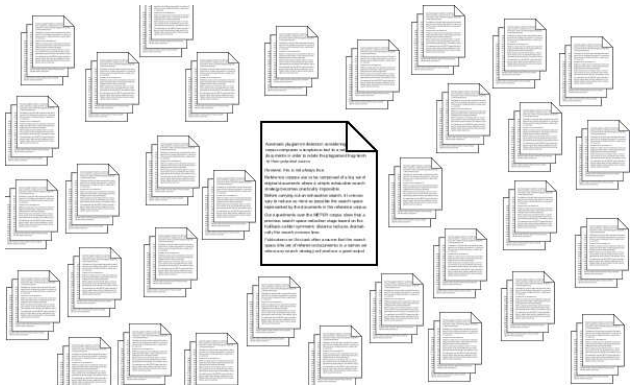
Better evidence than style irregularities is if the source of a plagiarism case can be provided

It is closer to information retrieval

d_q and a collection of potential source documents D are given. The task is to identify the plagiarized sections in d_q (if there are any), and their respective source sections in D

[Potthast et al., 2009]

Introduction: External Plagiarism Detection



Better evidence than style irregularities is if the source of a plagiarism case can be provided

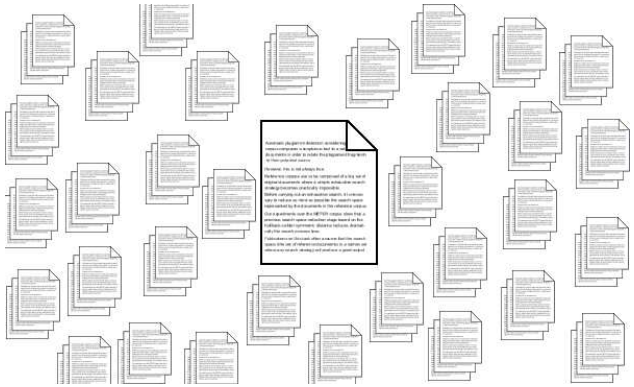
It is closer to information retrieval

Issues that render this task difficult

- Number of potential source documents, $|D|$;
- Plagiarizing a text often implies paraphrasing, summarizing, and even translation.

[Potthast et al., 2009]

Introduction: External Plagiarism Detection



Better evidence than style irregularities is if the source of a plagiarism case can be provided

It is closer to information retrieval

Models

Vector Space Models	[Broder, 1997], [Maurer et al., 2006]
Fingerprinting techniques	SPEX [Bernstein and Zobel, 2004], Winnowing [Schleimer et al., 2003]

[Potthast et al., 2009]

Introduction: Drawbacks

- Plagiarism implies an ethical issue
- Nobody would like to be included in a corpus of plagiarism!
- Properly anonymizing actual cases of plagiarism is a hard task
- No standard evaluation measures have been previously defined

Introduction: Drawbacks

- Plagiarism implies an ethical issue
- Nobody would like to be included in a corpus of plagiarism!
- Properly anonymizing actual cases of plagiarism is a hard task
- No standard evaluation measures have been previously defined
- Evaluations use to be incomparable and often not even reproducible.

Outline

PAN-PC-09

“A newly developed large-scale corpus of *artificial* plagiarism”

- 41 223 documents
- 94 202 artificial plagiarism cases
- It includes cases for intrinsic and external detection methods

http://www.webis.de/research/corpora

PAN-PC-09: Corpus Parameters

Document Length

- 50% short: 1-10 pages
- 35% medium: 10-100 pages
- 15% large: 100-1000 pages

PAN-PC-09: Corpus Parameters

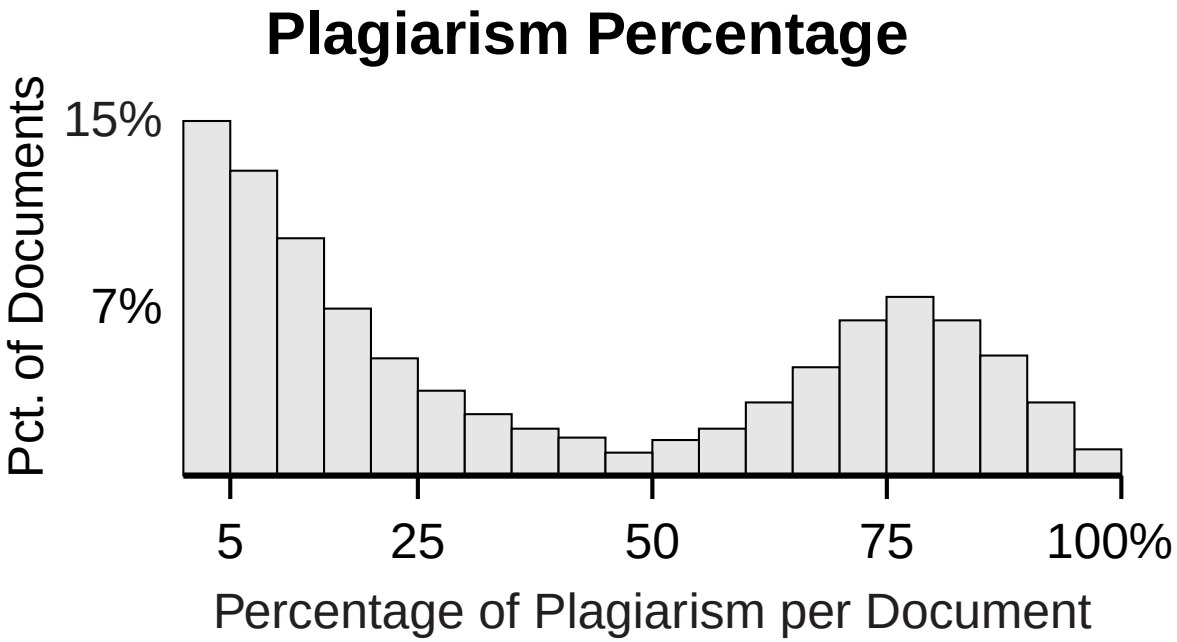
Document Length

- 50% short: 1-10 pages
- 35% medium: 10-100 pages
- 15% large: 100-1000 pages

Suspicious-to-Source Ratio

- 50% are designated as suspicious documents D_q
- 50% are designated as source documents D

PAN-PC-09: Corpus Parameters



- 50% of D_q contain no plagiarism at all

PAN-PC-09: Corpus Parameters

Cases Length

- 250–750 chars; ∼50–150 words
- 1500–5000 chars; ∼300–1000 words
- 15000–25000 chars; ∼3000-5000 words

PAN-PC-09: Corpus Parameters

Cases Length

- 250–750 chars; ∼50–150 words
- 1500–5000 chars; ∼300–1000 words
- 15000–25000 chars; ∼3000-5000 words

Plagiarism Languages

- 90% are monolingual English plagiarism
- 10% are cross-language plagiarism (German or Spanish into English)

PAN-PC-09: Corpus Parameters

Cases Obfuscation

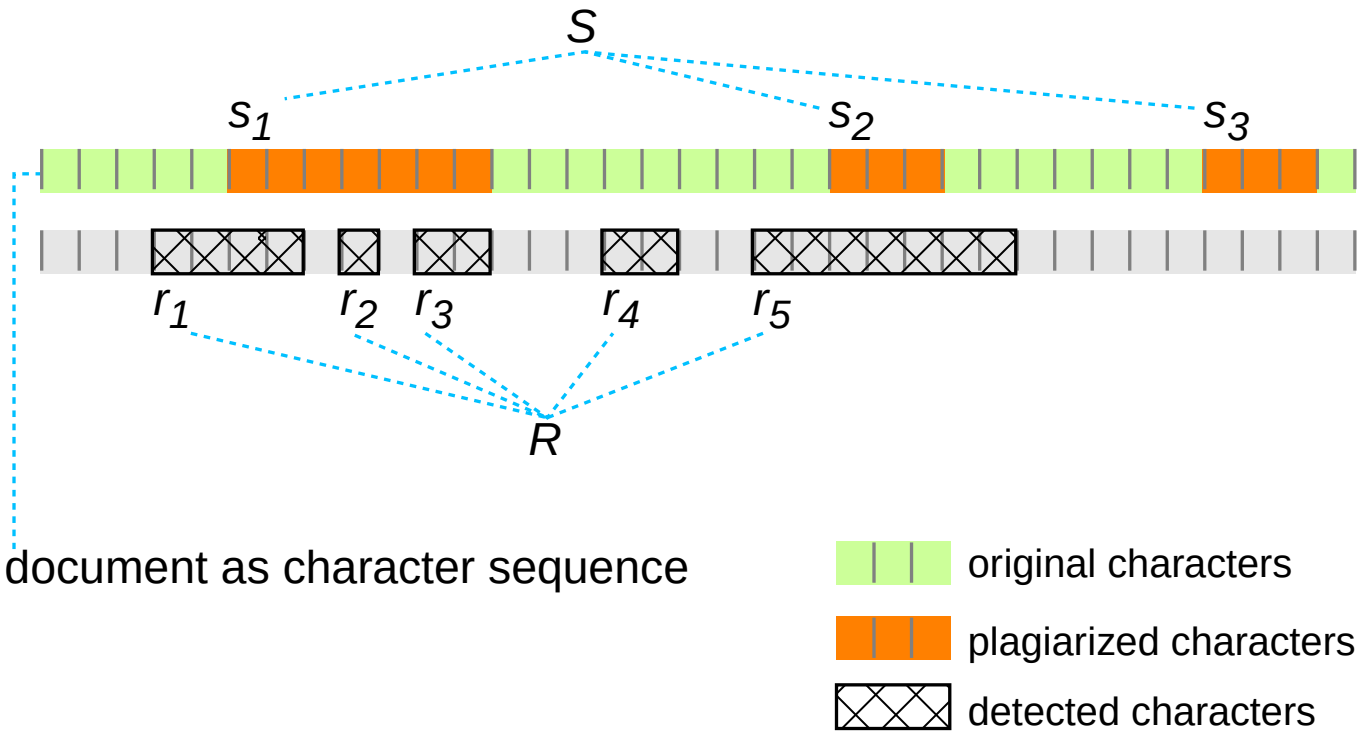
- small
- medium
- high

Paraphrasing, summarization, etc. is simulated by…

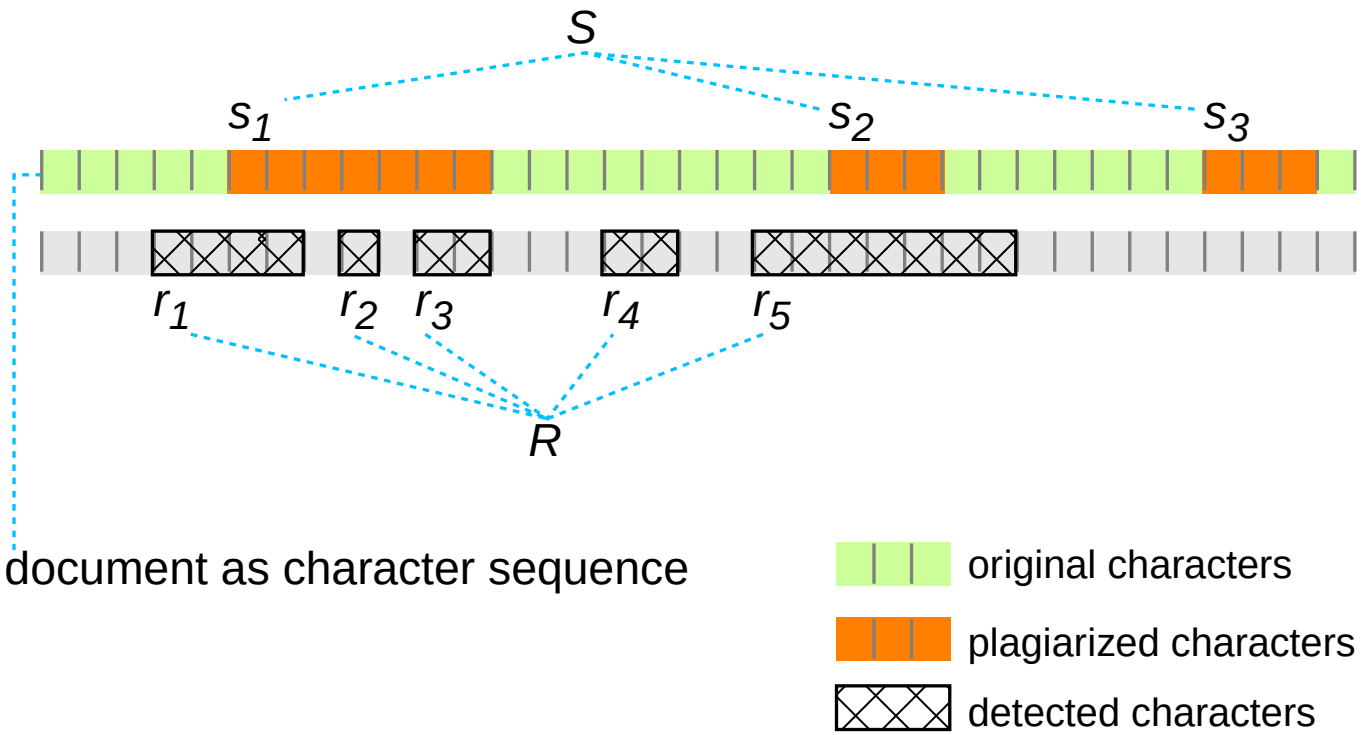
- shuffling, removing, inserting short phrases
- replacing semantically related words
- POS preserving shuffling

Outline

Evaluation Measures



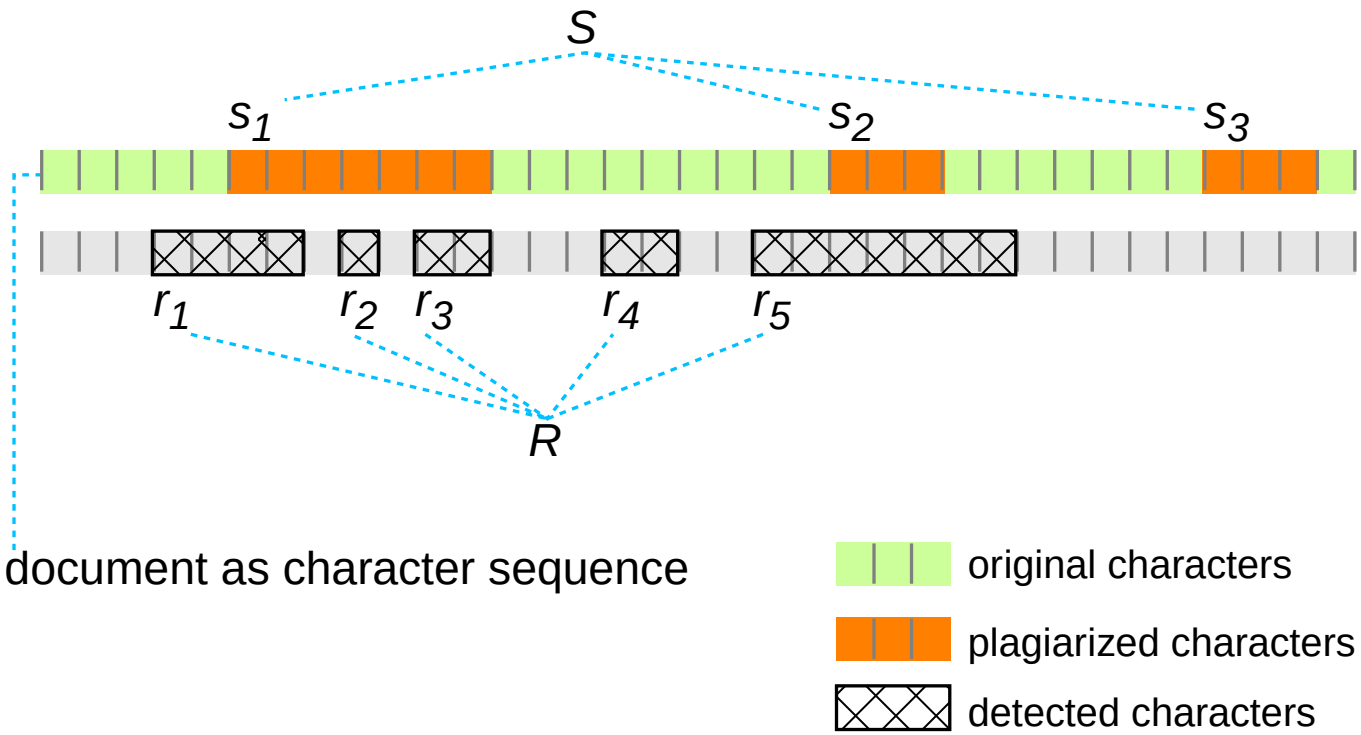
Evaluation Measures



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

($\sqcap$ computes the positionally overlapping characters)

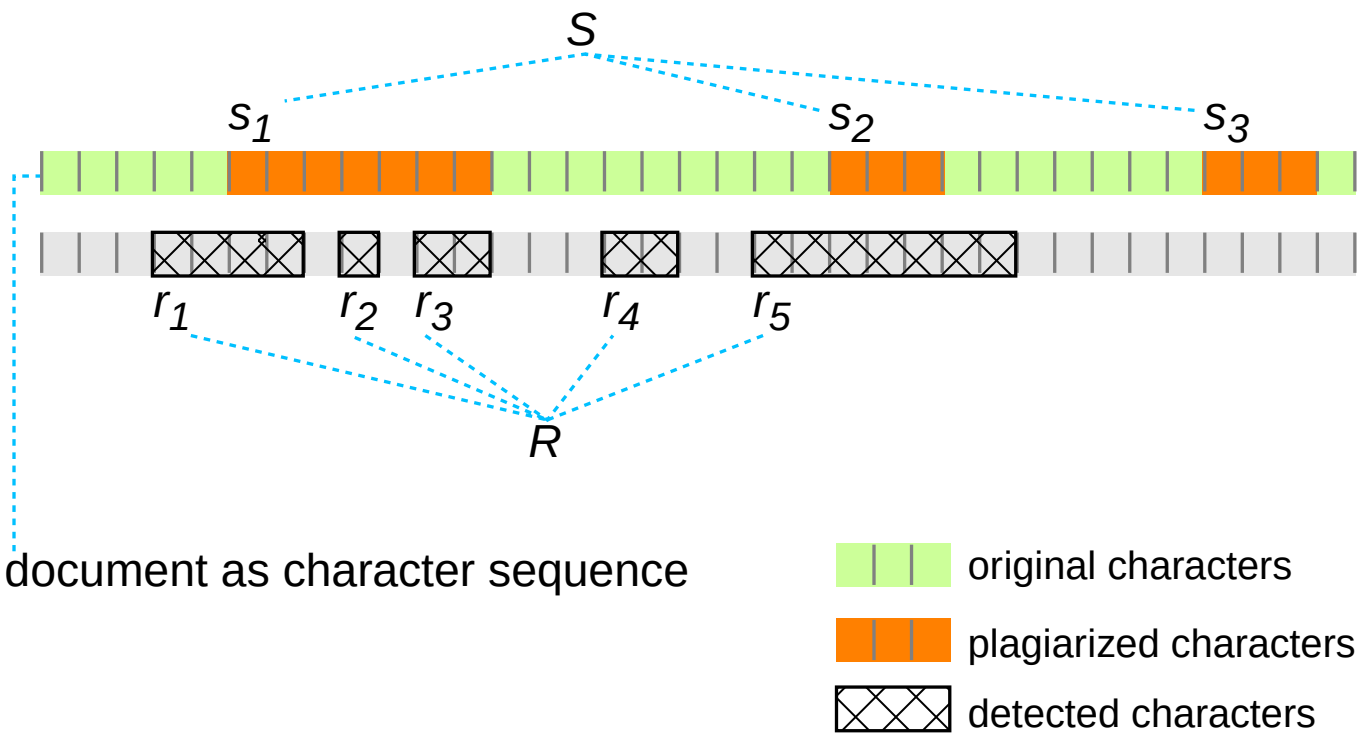
Evaluation Measures



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

($\sqcap$ computes the positionally overlapping characters)

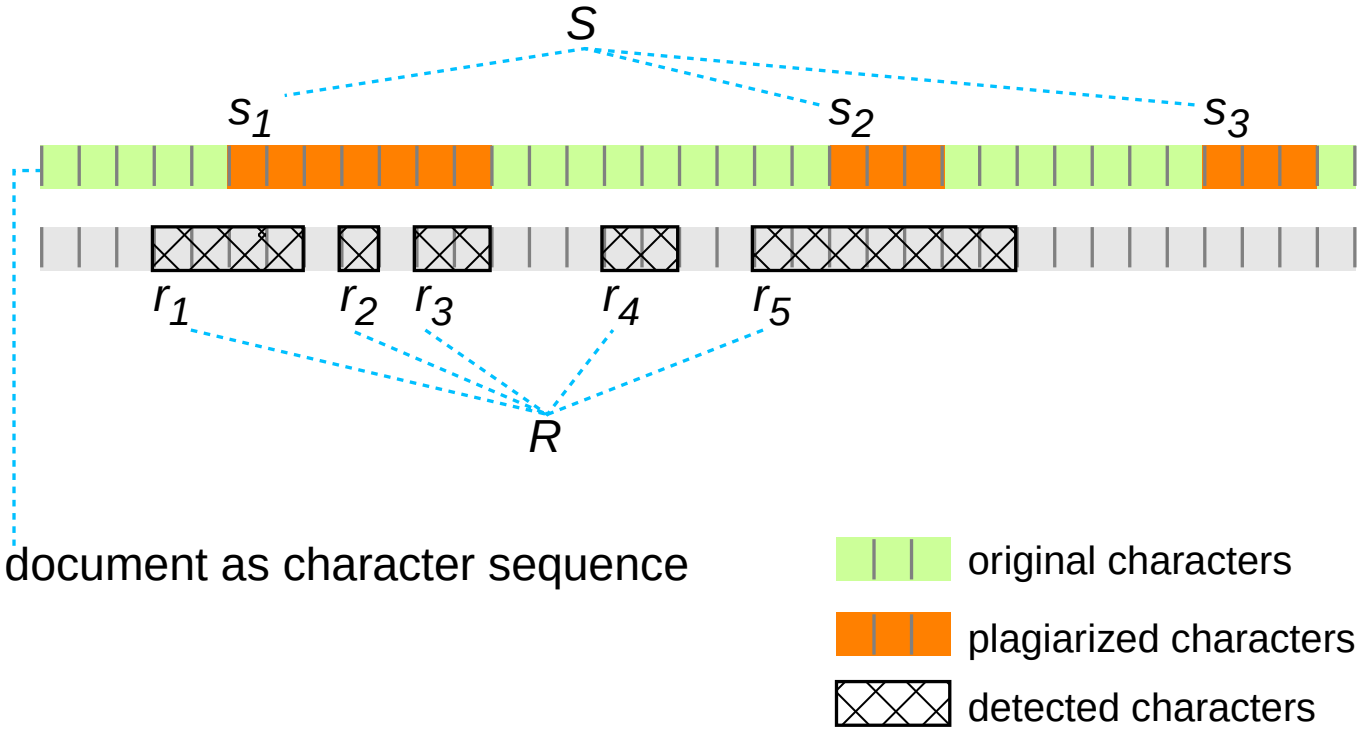
Evaluation Measures



$$gran_{PDA}(S,R) = \frac{1}{|S_R|} \sum_{s \in S_R} |C_s| \quad \in [1, |R|]$$

$$C_s = \{r \mid r \in R \ \wedge \ s \cap r \neq \emptyset\}$$
$$S_R = \{s \mid s \in S \ \wedge \ \exists r \in R: \ s \cap r \neq \emptyset\}$$

Evaluation Measures



$$overall_{PDA}(S,R) = \frac{F}{\log_2(1 + gran_{PDA})}$$

Outline

1st Intl. Competition on Plagiarism Detection

http://www.webis.de/research/workshopseries/pan-09/competition.html

http://ceur-ws.org/Vol-502

1st Intl. Competition on Plagiarism Detection

13 research teams

Europe (8) America (3)
Asia (1.5) Africa (0.5)

Intrinsic Approaches (4 teams)

Participant	Analyzed features
Stamatatos	character n-grams
Zechner, Muhr, Kern, Granitzer	word freq. class + text frequencies
Seaward, Matwin	Kolmogorov complexity measures

http://www.webis.de/research/workshopseries/pan-09/competition.html

http://ceur-ws.org/Vol-502

1st Intl. Competition on Plagiarism Detection

13 research teams

Europe (8) America (3)
Asia (1.5) Africa (0.5)

External Approaches (10 teams)

Participant	Comparison units
Grozea, Gehl, Popescu	character n-grams
Kasprzak, Brandejs, Kripac	word n-grams
Basile, Benedetto, Caglioti, Degli Esposti	length n-grams

http://www.webis.de/research/workshopseries/pan-09/competition.html

http://ceur-ws.org/Vol-502

2nd Intl. Competition on Plagiarism Detection

17 teams registered

Europe (9) Asia (5) America (3)

http://pan.webis.de

2nd Intl. Competition on Plagiarism Detection

PAN 2010 LAB
Uncovering Plagiarism, Authorship, and Social Software Misuse

sponsored by YAHOO! RESEARCH

held in conjunction with CLEF2010 Padua

17 teams registered

Europe (9) Asia (5) America (3)

- PAN-PC-09 corpus → PAN 2010 training corpus
- PAN 2010 test corpus composed of around 40,000 documents

http://pan.webis.de

Outline

Final Remarks

- First standardized corpus dedicated to the evaluation of automatic plagiarism detection

Final Remarks

- First standardized corpus dedicated to the evaluation of automatic plagiarism detection
- New performance measures to evaluate plagiarism detection have been proposed

Final Remarks

- First standardized corpus dedicated to the evaluation of automatic plagiarism detection
- New performance measures to evaluate plagiarism detection have been proposed
- Two weeks to submit detections for PAN 2010's competition!

Thank you!

http://pan.webis.de

Alberto Barrón-Cedeño	lbarron@dsic.upv.es
Martin Potthast	martin.potthast@uni-weimar.de
Paolo Rosso	prosso@dsic.upv.es
Benno Stein	benno.stein@uni-weimar.de
Andreas Eiselt	andreas.eiselt@uni-weimar.de

References I

Anderson, G. (1999).
Cyberplagiarism. a look at the web term paper sites.
College & Research Libraries News, 60(5):371–373.

Association of Teachers and Lecturers (2008).
School Work Plagued by Plagiarism - ATL Survey.
Technical report, Association of Teachers and Lecturers, London, UK.
Press release.

Baty, P. (2000).
Copycats roam in era of the net.
Times Higher Education.

Bernstein, Y. and Zobel, J. (2004).
A Scalable System for Identifying Co-Derivative Documents.
In *Proceedings of the Symposium on String Processing and Information Retrieval*, pages 55–67. Springer.

Bierce, A. (1911).
The Devil's Dictionary.
Doubleday, Page & Company.

Broder, A. (1997).
On the Resemblance and Containment of Documents.
In *Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29. IEEE Computer Society.

Clough, P., Gaizauskas, R., Piao, S., and Wilks, Y. (2002).
Measuring Text Reuse.
In *Proceedings of Association for Computational Linguistics (ACL2002)*, pages 152–159, Philadelphia, PA.

References II

Haines, V., Diekhoff, G., LaBeff, G., and Clarck, R. (1986).
College Cheating: Inmaturity, Lack of Commitment, and the Neutralizing Attitude.
Research in Higher Education, 25(4):342–354.

IEEE (2008).
A plagiarism FAQ.
http://www.ieee.org/web/publications/rights/plagiarism_FAQ.htm.
[Online; accessed 3-March-2010].

Kulathuramaiyer, N. and Maurer, H. (2007).
Coping With the Copy-Paste-Syndrome.
In *E-Learn 2007*, pages 1072—1079, Quebec, CA.

Martínez, I. (2009).
Wikipedia usage by Mexican students. The constant usage of copy and paste.
In *Wikimania 2009*, Buenos Aires, Argentina.

Maurer, H., Kappe, F., and Zaka, B. (2006).
Plagiarism - A Survey.
Journal of Universal Computer Science, 12(8):1050–1084.

Meyer zu Eißen, S. and Stein, B. (2006).
Intrinsic plagiarism detection.
Advances in Information Retrieval: Proceedings of the 28th European Conference on IR Research (ECIR 2006), LNCS (3936):565–569.

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., and Rosso, P. (2009).
Overview of the 1st International Competition on Plagiarism Detection.
In [Stein et al., 2009], pages 1–9.

References III

Schleimer, S., Wilkerson, D., and Aiken, A. (2003).
Winnowing: Local Algorithms for Document Fingerprinting.
In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, New York, NY. ACM.

Stamatatos, E. (2009).
Intrinsic Plagiarism Detection Using Character n-gram Profiles.
In [Stein et al., 2009], pages 38–46.

Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Agirre, E., editors (2009).
SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), San Sebastian, Spain. CEUS-WS.org.

Weber, S. (2007).
Das Google-Copy-Paste-Syndrom. Wie Netzplagiate Ausbildung und Wissen gefahrden.
Telepolis.