

An Evaluation Framework for Plagiarism Detection

Martin Potthast

Benno Stein

Alberto Barrón-Cedeño

Paolo Rosso

Motivation

Observations from a survey of 205 evaluations of plagiarism detectors:
(101 on text, 104 on code)

Evaluation Aspect	Text	Code
<i>Performance Measures</i>		
precision, recall	43%	18%
manual, other	57%	82%
<i>Corpus Acquisition</i>		
existing corpus	20%	18%
homemade corpora	80%	82%
<i>Comparative Evaluation</i>		
no	46%	51%
yes	54%	49%



Measuring detection performance is not well-understood.



There is no standardized evaluation corpus.



Half of the evaluations don't compare different approaches.

We introduce the first standardized evaluation framework for plagiarism detection.

Performance Measures

Let $s \in S$ denote plagiarism cases.

Let $r \in R$ denote plagiarism detections.

The formulas below measure the detection performance of R with regard to S .

The well-known precision and recall:

$$\text{prec}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup_{s \in S} (s \cap r)|}{|r|}$$

$$\text{rec}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup_{r \in R} (s \cap r)|}{|s|}$$

The granularity measures the average number of detections of all detected cases:

$$\text{gran}(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|$$

The domain of the granularity is $[1, |R|]$.

Combining the three concepts:

$$\text{plagdet}(S, R) = \frac{F_1}{\log_2(1 + \text{gran}(S, R))}$$

Evaluation Corpus

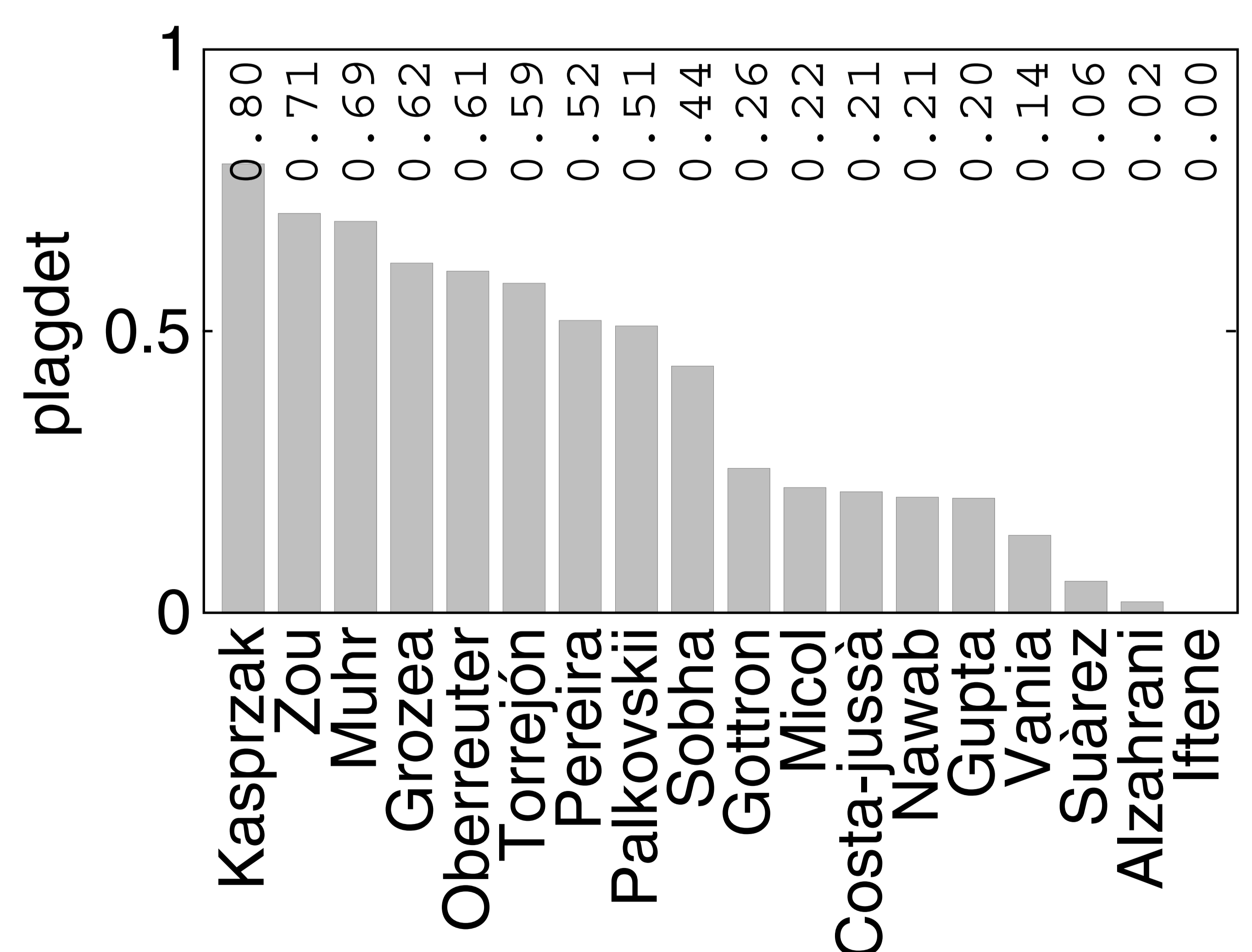
The PAN plagiarism corpus 2010 (PAN-PC-10):

27 073 documents in which
68 558 plagiarism cases have been inserted.

4 000 cases were created manually using
Amazon's Mechanical Turk, the rest artificially.

A high diversity of cases was achieved by
varying 7 different parameters.

PAN at CLEF'10



See the framework in action at <http://pan.webis.de>