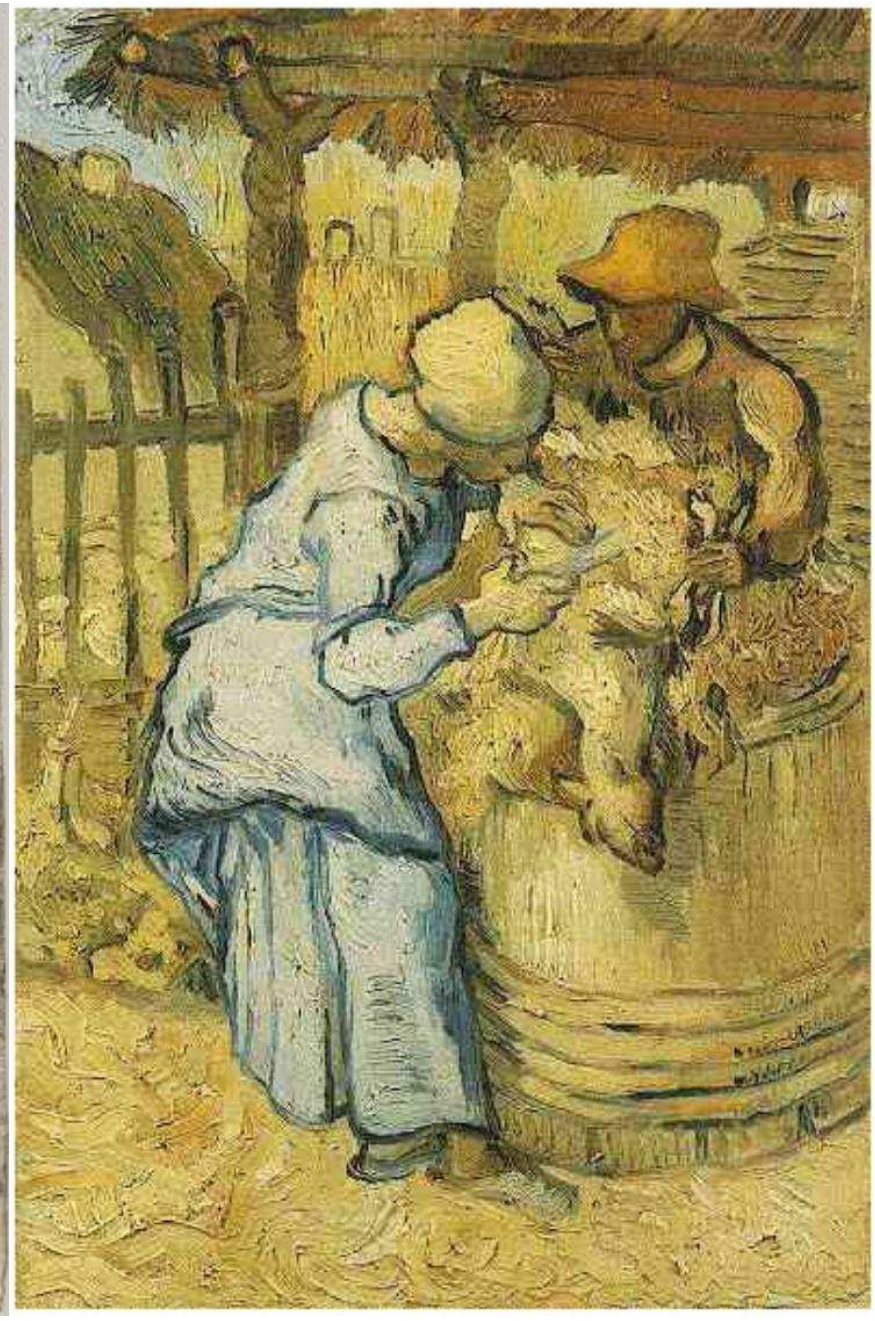


Overview of the 3rd International Competition on Plagiarism Detection

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

[1]Web Technology & Information Systems. Bauhaus-Universiät Weimar, Germany
[2]Natural Language Engineering Lab, ELiRF. Universidad Politécnica de Valencia, Spain

pan@webis.de http://pan.webis.de

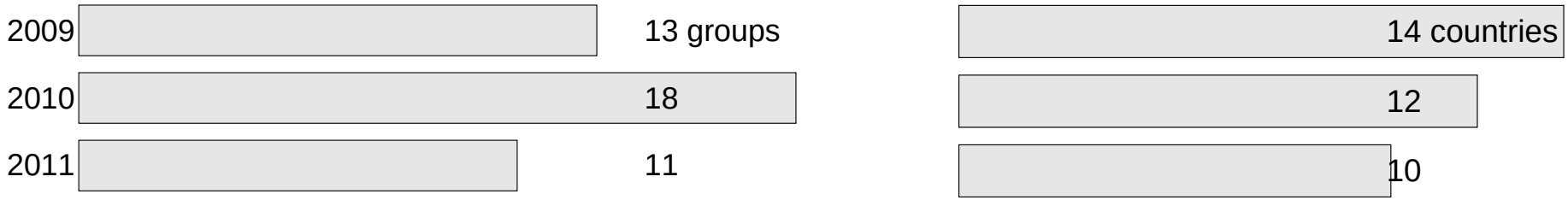
Introduction

Task:

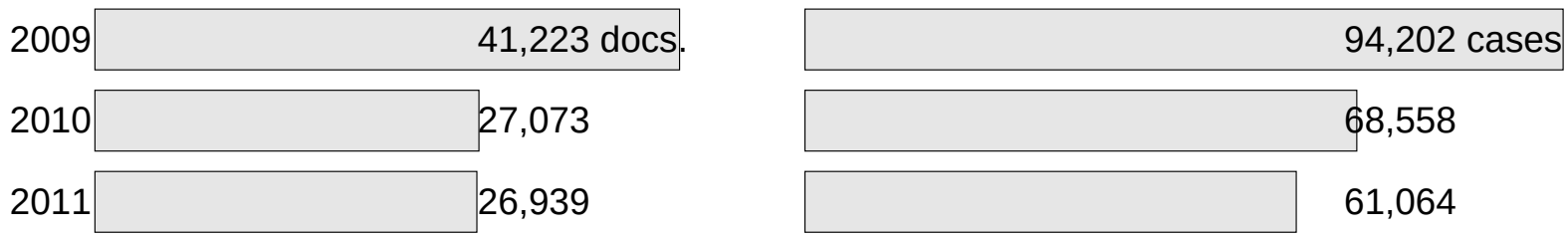
- Given a set of suspicious documents and a set of source documents, find all plagiarized sections in the suspicious documents and, if available, the corresponding source sections.

Introduction: Facts

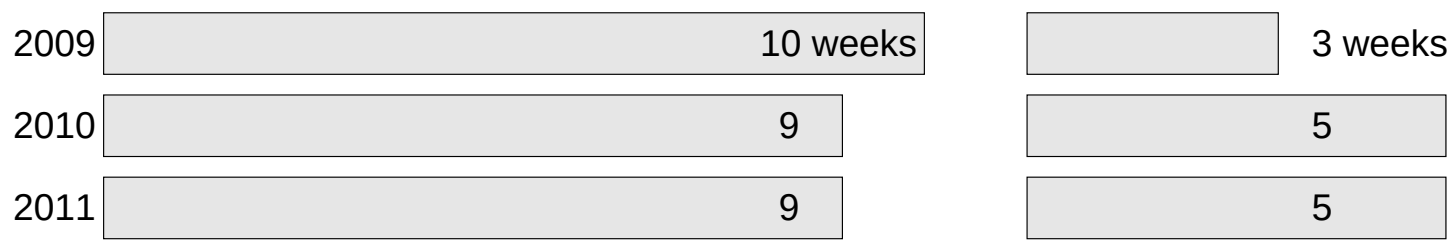
Participation



Corpus size



Competition phases: training / test



The PAN Competition 2011: Corpus PAN-PC-11

The PAN Competition 2011: Corpus PAN-PC-11

Document length

The PAN Competition 2011: Corpus PAN-PC-11

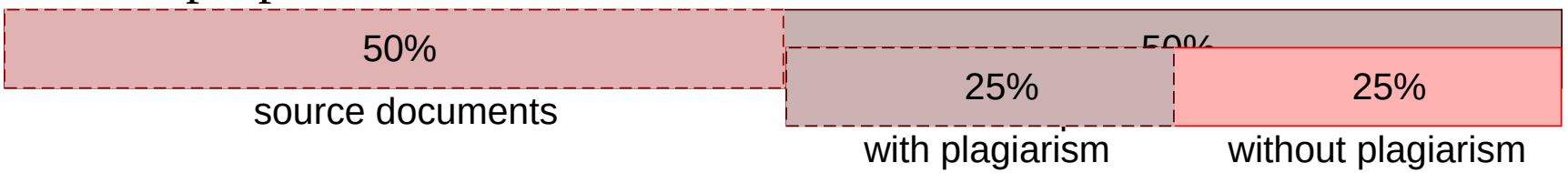
Document length

Document purpose

The PAN Competition 2011: Corpus PAN-PC-11

Document length

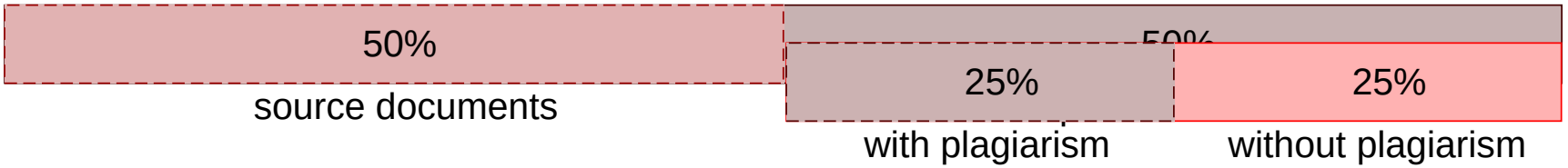
Document purpose



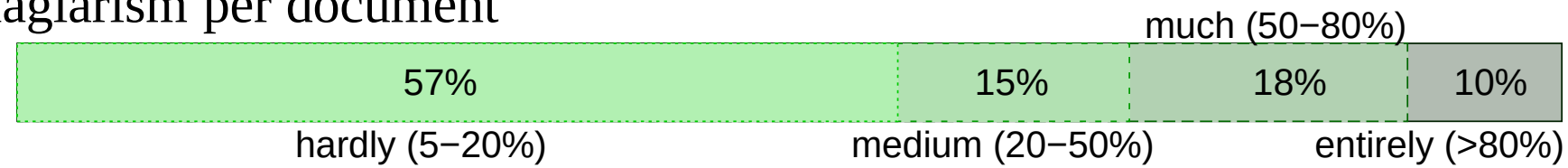
The PAN Competition 2011: Corpus PAN-PC-11

Document length

Document purpose



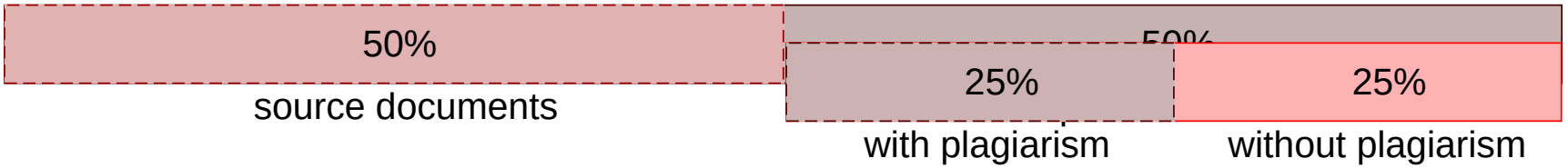
Plagiarism per document



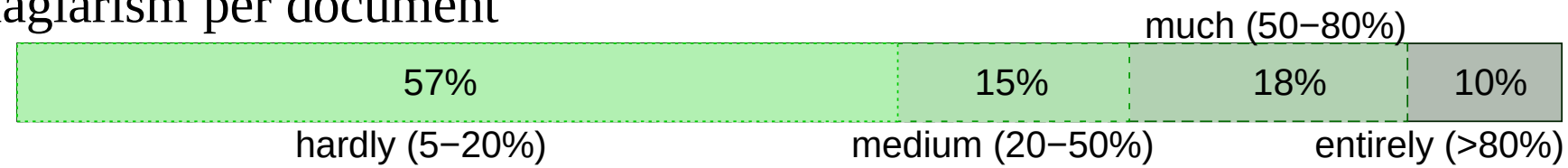
The PAN Competition 2011: Corpus PAN-PC-11

Document length

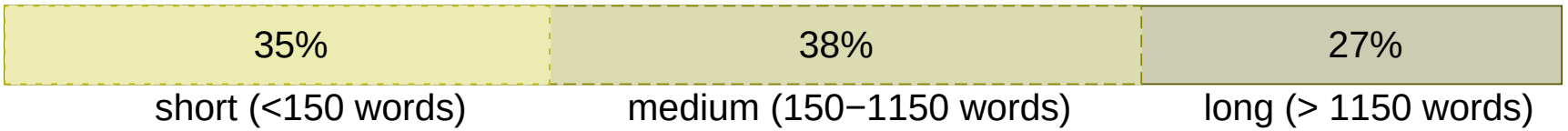
Document purpose



Plagiarism per document



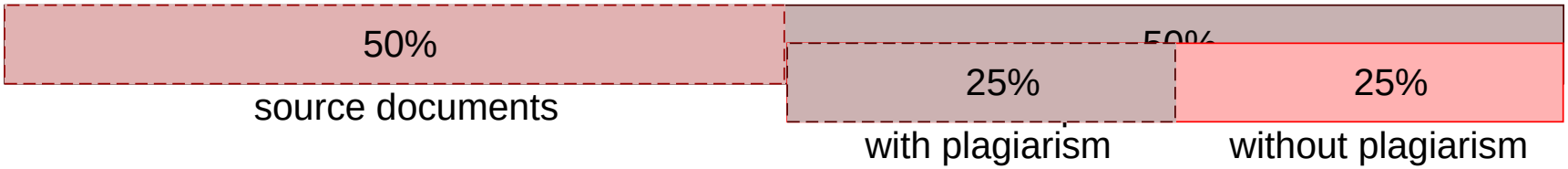
Case length



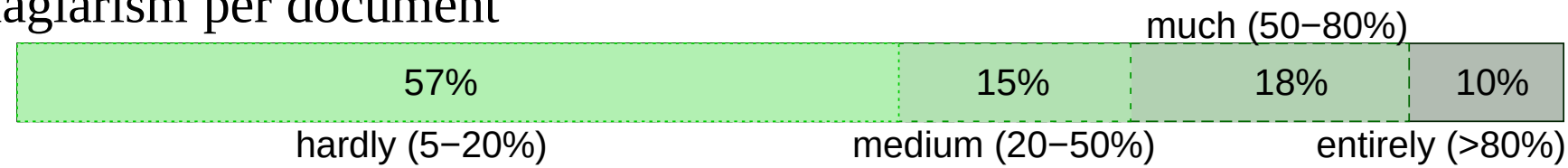
The PAN Competition 2011: Corpus PAN-PC-11

Document length

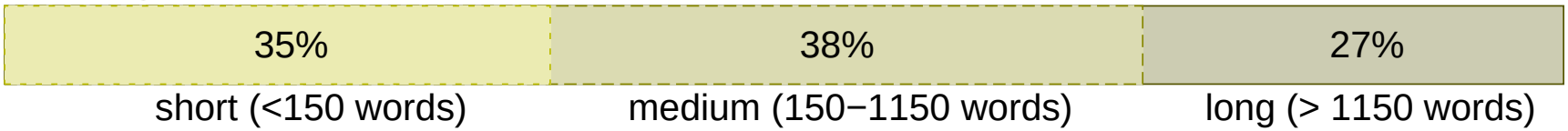
Document purpose



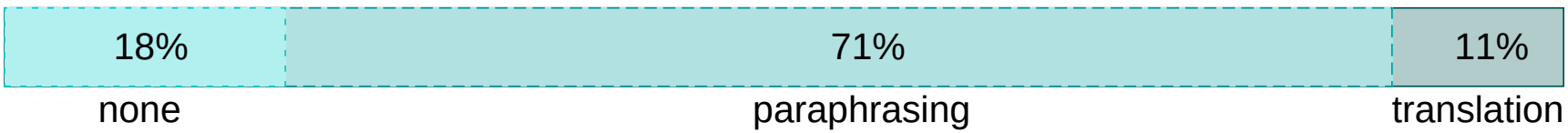
Plagiarism per document



Case length



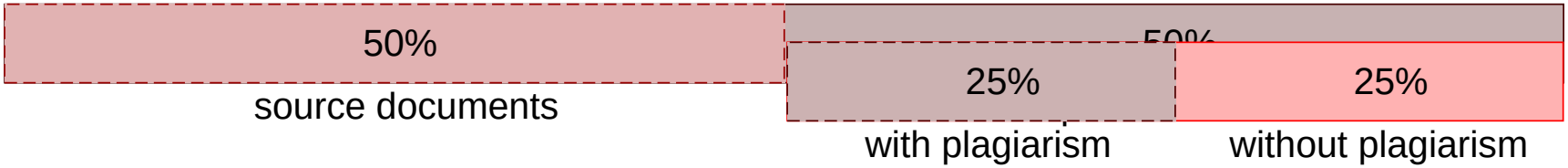
Obfuscation



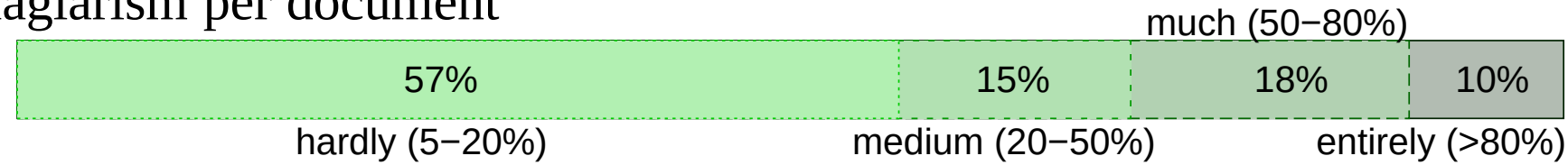
The PAN Competition 2011: Corpus PAN-PC-11

Document length

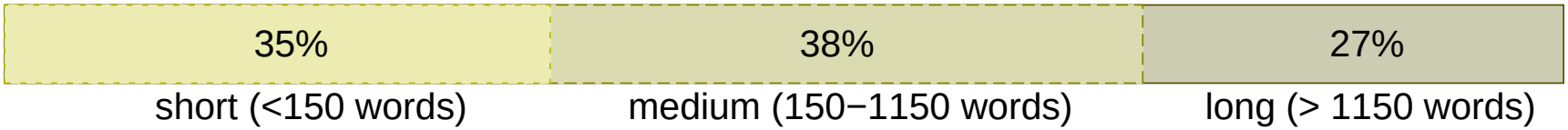
Document purpose



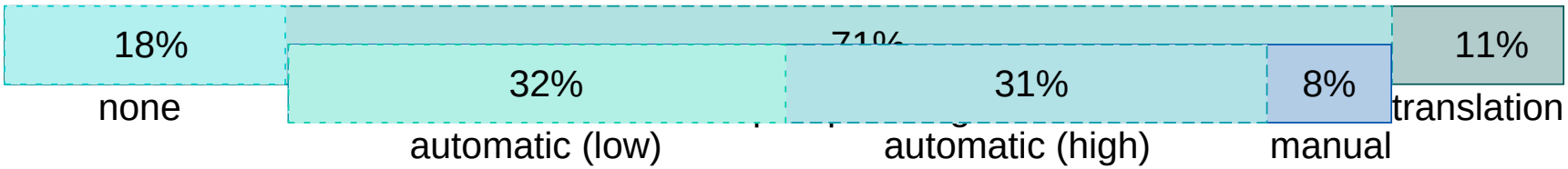
Plagiarism per document



Case length



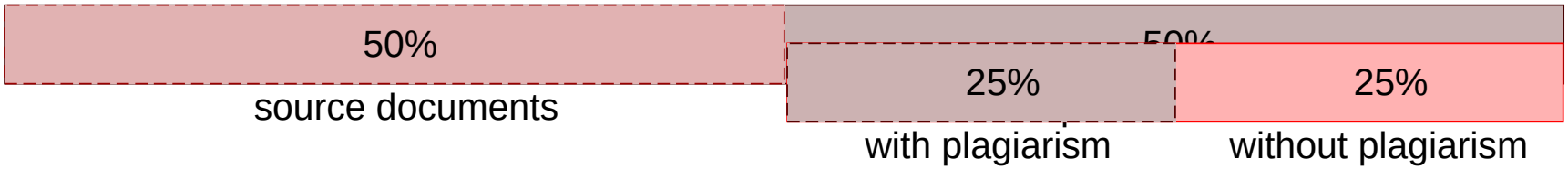
Obfuscation



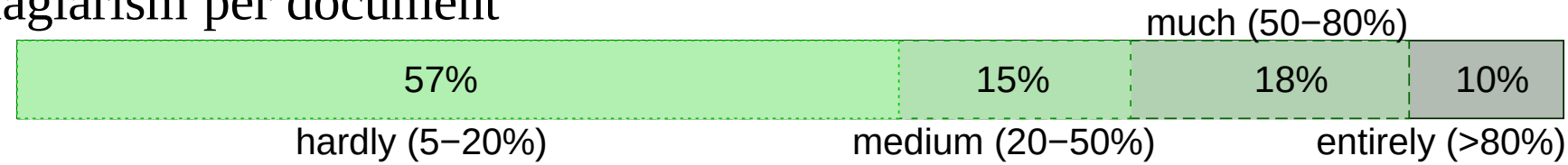
The PAN Competition 2011: Corpus PAN-PC-11

Document length

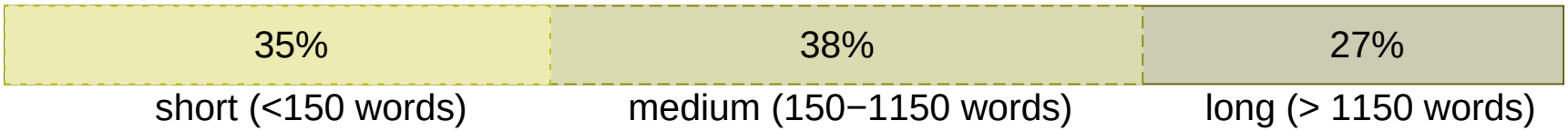
Document purpose



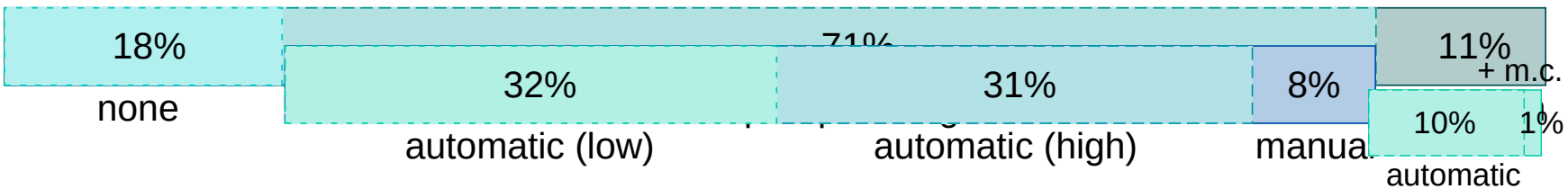
Plagiarism per document



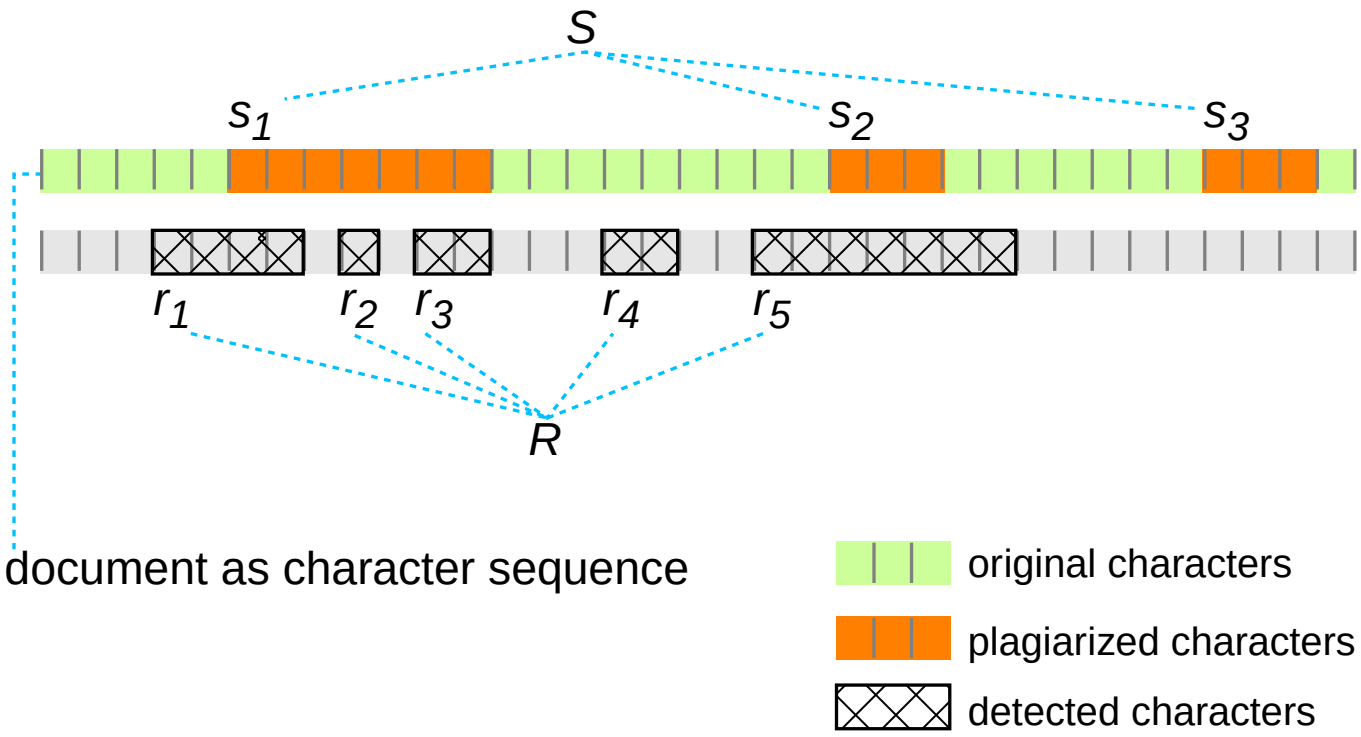
Case length



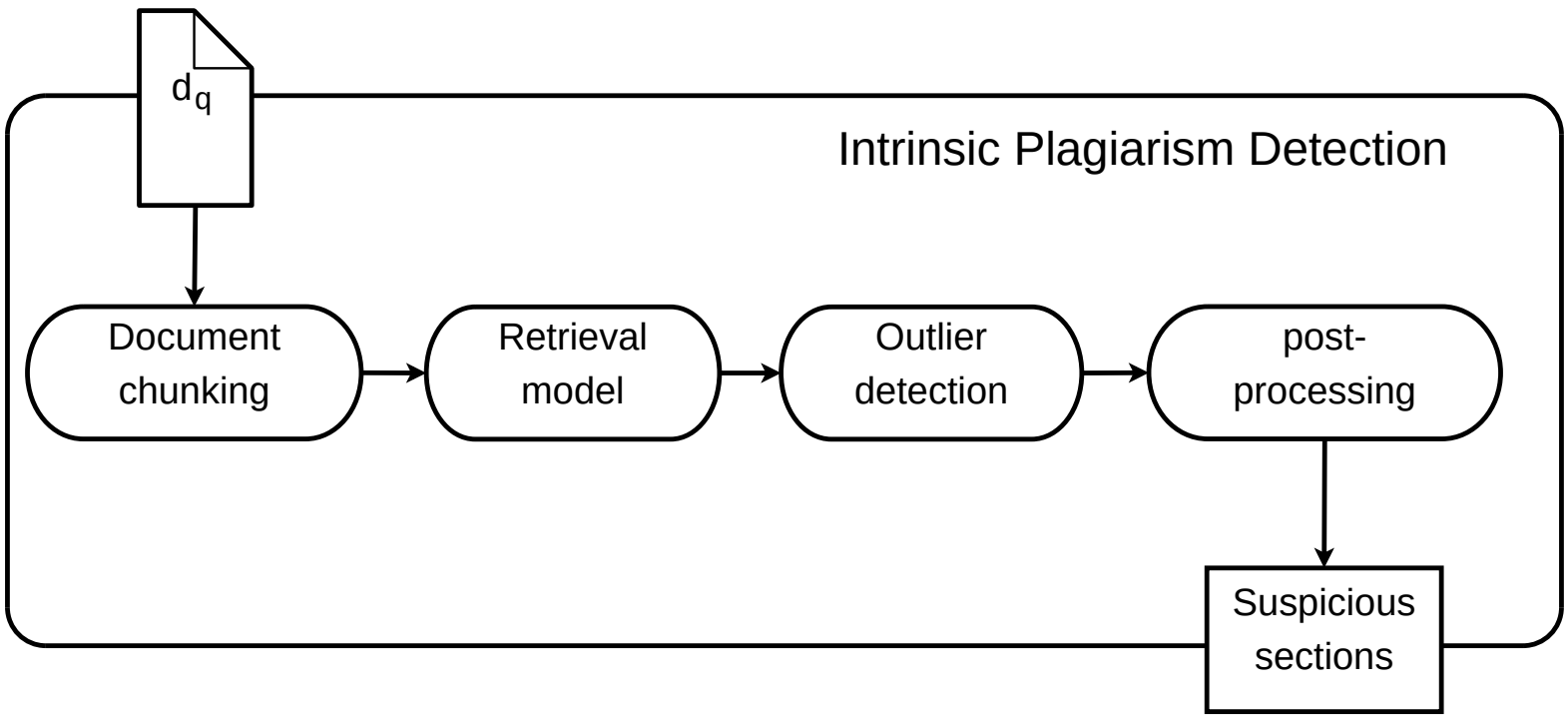
Obfuscation



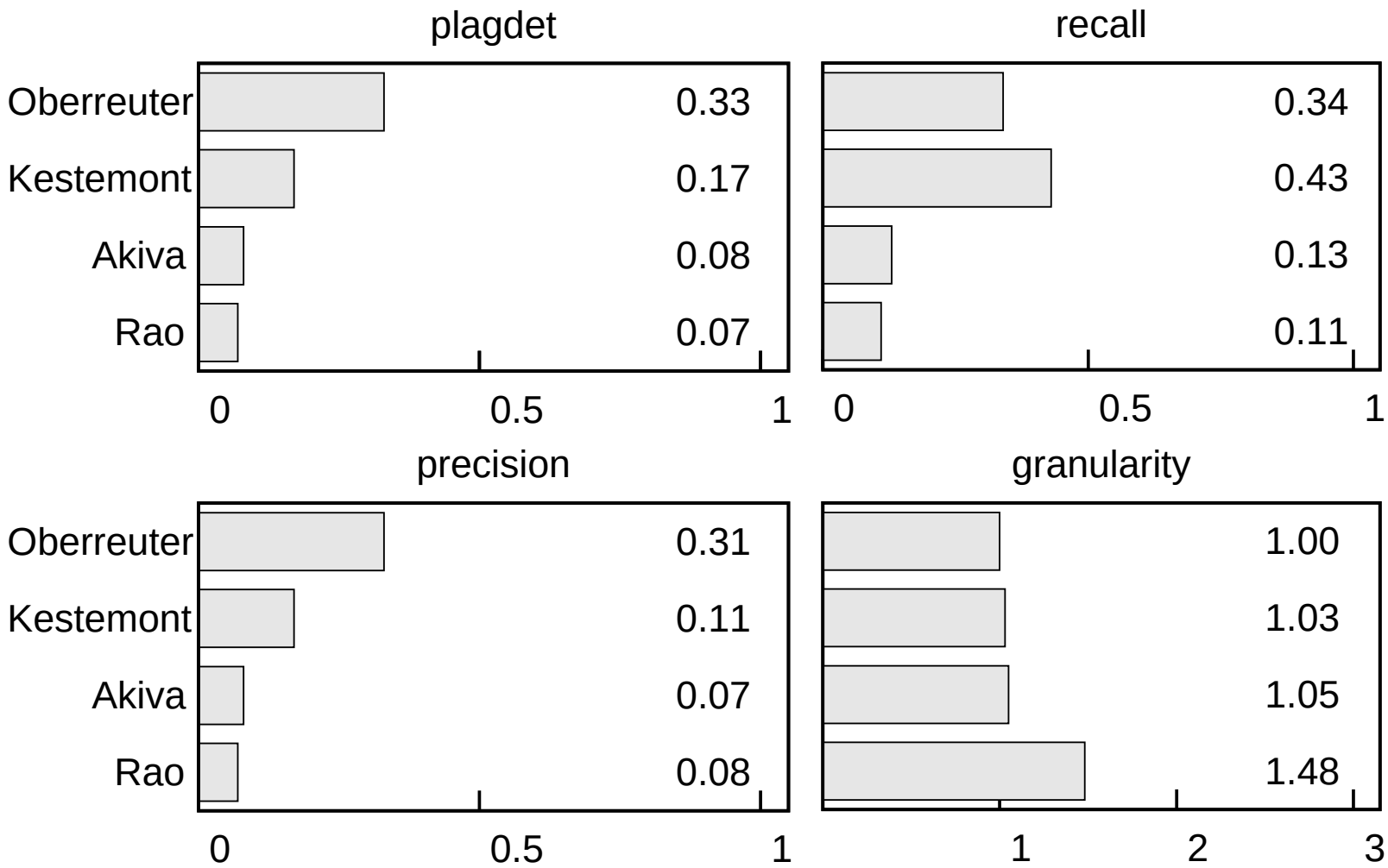
The PAN Competition 2011: Evaluation



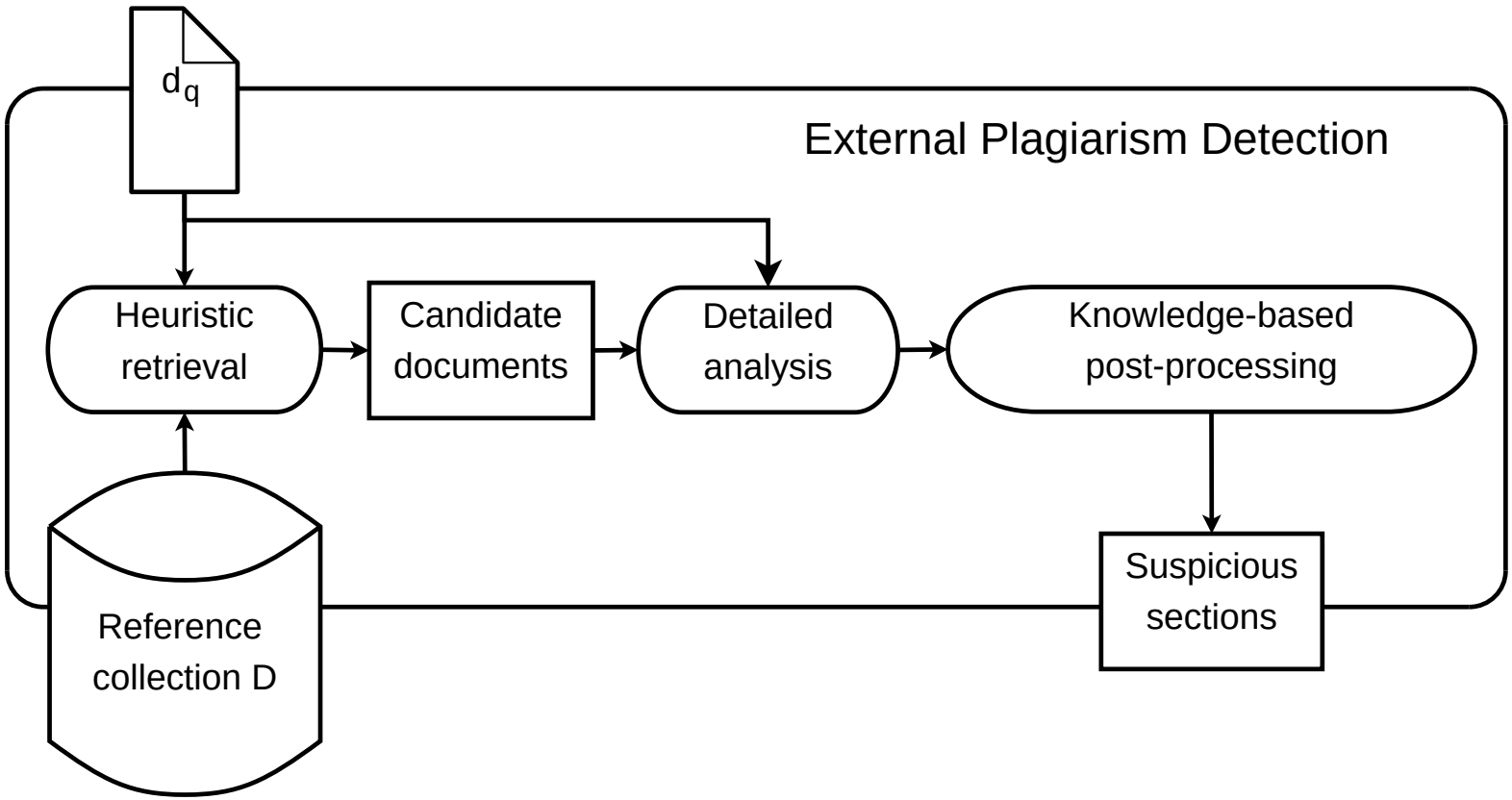
Intrinsic Detection



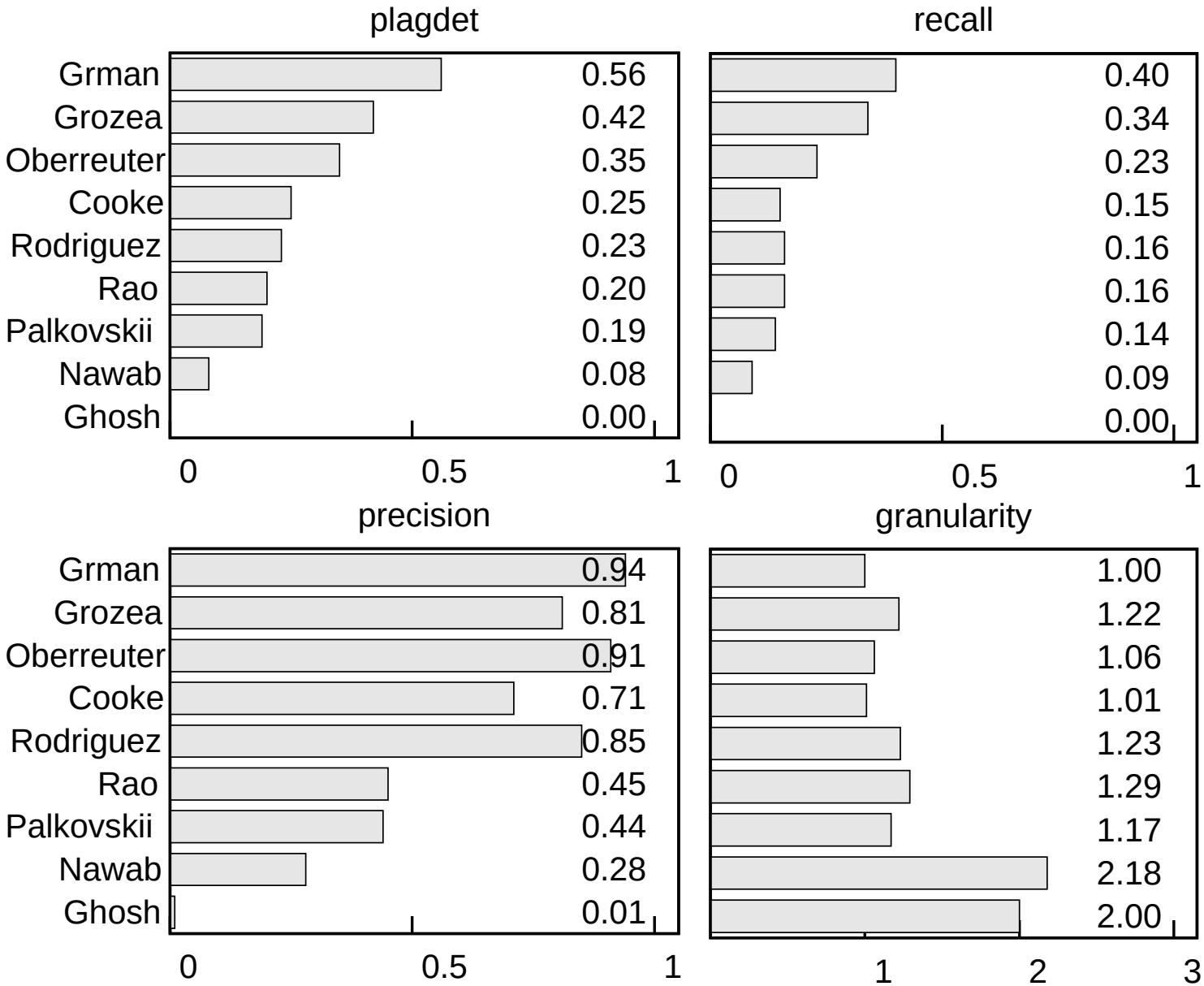
Intrinsic Detection



External Detection



External Detection



Summary

Overview paper

- This year's best practices for intrinsic and external detection.
- Detection results with regard to every corpus parameter.
- Comparison to PAN 2009 and PAN 2010.

Lessons & frontiers

- Detection performances decreased by the increased detection difficulty
- Intrinsic detection results may be biased due to the corpus nature
- Both approaches are important (also to win the competition)
- Short plagiarism cases remain being the hardest to detect
- Manual translation shows to be much harder to detect than automatic (result less biased)

CL!TR: Cross-Language !ndian Text Reuse

- Task on cross-language text re-use detection
- Potential source texts in English, suspicious texts in Hindi
- Document level task (no specific fragments are expected to be identified)

http://users.dsic.upv.es/grupos/nle/fire-workshop-clitr.html

Jean-François Millet (1854)
Sheep Shearing Beneath a Tree

Jean-François Millet (1854)
Sheep Shearing Beneath a Tree

Vincent van Gogh (1889)
The Sheep Shearers

Jean-François Millet (1854)
Sheep Shearing Beneath a Tree

Vincent van Gogh (1889)
The Sheep Shearers (after Millet)

Jean-François Millet (1854)
Sheep Shearing Beneath a Tree

Vincent van Gogh (1889)
The Sheep Shearers (after Millet)

"[I am] translating the black and white impressions into another language –that of colour"