Uncovering Plagiarism, Authorship, and Social Software Misuse

PAN 2011 Results

[pan.webis.de]

The PAN Competition

Plagiarism Detection

The web is rife with text reuse: boilerplate, translations, paraphrases, summaries, and plagiarism.

The PAN Competition

Plagiarism Detection

The web is rife with text reuse: boilerplate, translations, paraphrases, summaries, and plagiarism.

Tasks:

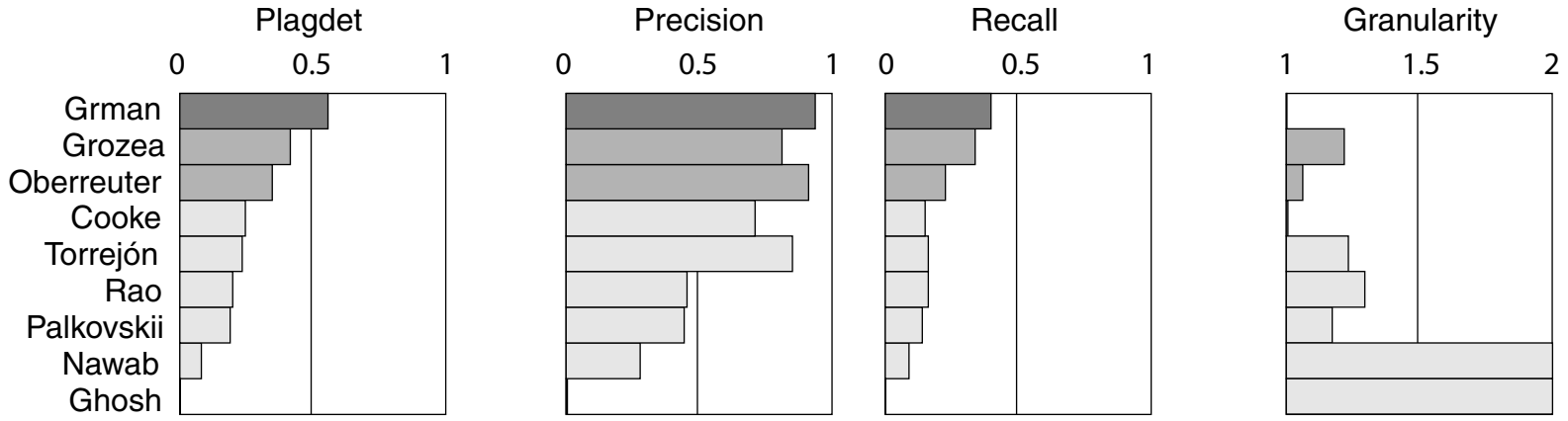
- *External Detection*. Given a suspicious document and a set of potential source documents, the task is to find all plagiarized passages in the suspicious document and their corresponding source passages in the source documents.
- *Intrinsic Detection*. Given a suspicious document, the task is to extract all plagiarized passages based on clues extracted from the document itself.

Corpus:

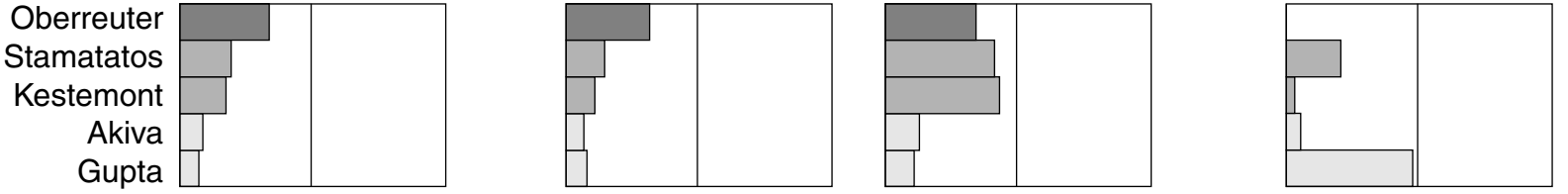
- PAN plagiarism corpus of 2010, 2011 [www.webis.de/research/corpora]
- 61 000 plagiarism cases hidden in about 27 000 documents
- 5 plagiarism-relevant parameters (length, language, task, obfuscation, fraction)

The PAN Competition

External plagiarism detection:



Intrinsic plagiarism detection:



- Plagdet combines the measures as F / log(granularity).
- Granularity measures the average number of times a plagiarism case is detected.

The PAN Competition

Authorship Identification

Many texts on the web are of uncertain authorship.

The PAN Competition

Authorship Identification

Many texts on the web are of uncertain authorship.

Tasks:

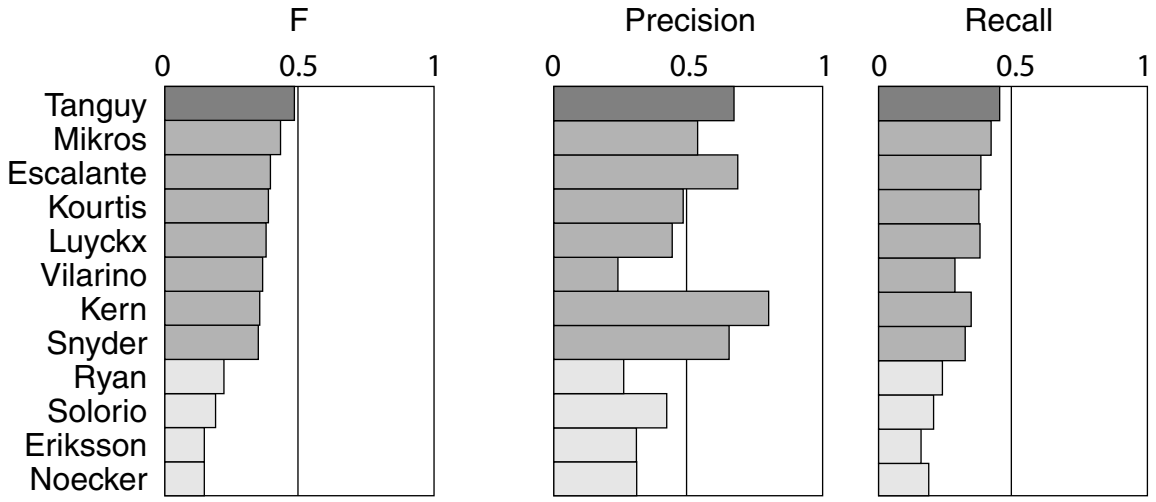
- *Authorship Attribution.* Given a document of uncertain authorship and documents from a set of candidate authors, the task is to map the document onto its true authors among the candidates.
- *Authorship Verification.* Given a document of uncertain authorship and a document from a specific author, the task is to determine whether the given text has been written by that author.

Corpus:

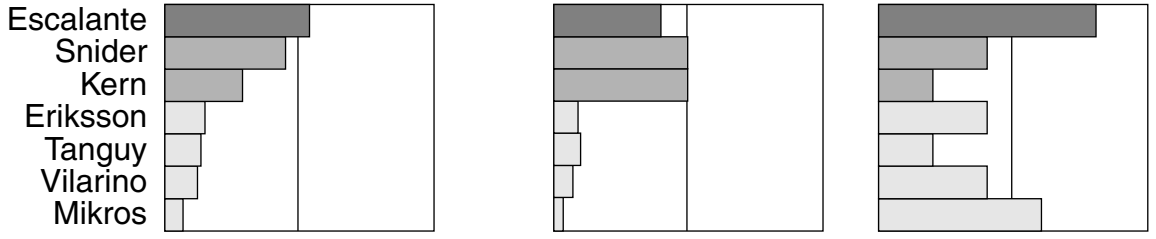
- Subset of the Enron Email Dataset [www.cs.cmu.edu/~enron]
- More than 12 000 documents written by 118 authors.
- 3 relevant parameters (task, canidate set size, closed vs. open canidate set)

The PAN Competition

Authorship attribution:



Authorship verification:



The PAN Competition

Wikipedia Vandalism Detection

Every edit on Wikipedia has to be double-checked for integrity.

The PAN Competition

Wikipedia Vandalism Detection

Every edit on Wikipedia has to be double-checked for integrity.

Task:

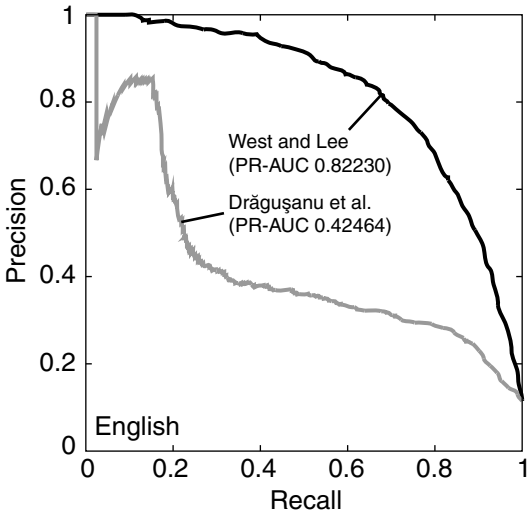
- Given a set of edits on Wikipedia articles, separate the ill-intentioned edits from the well-intentioned edits.

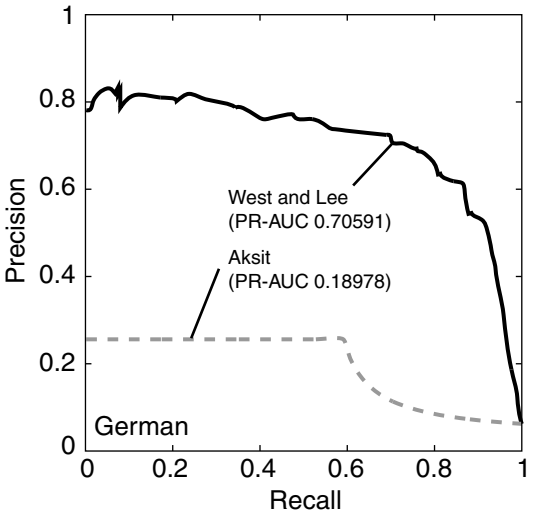
Corpus:

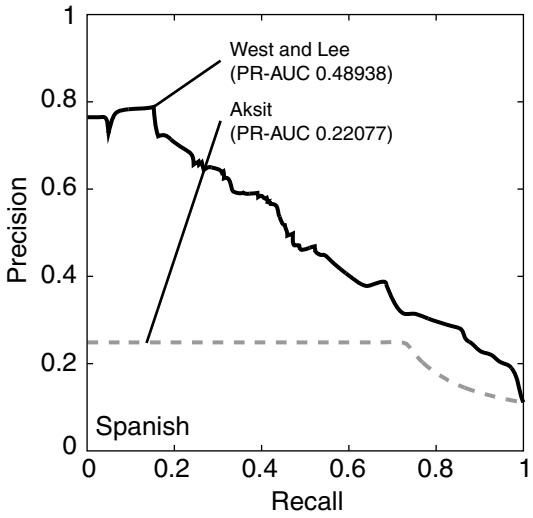
- PAN Wikipedia vandalism corpus of 2010, 2011 [www.webis.de/research/corpora]
- About 2 800 vandalism cases among about 30 000 edits
- 3 languages with corpus annotations obtained from Mechanical Turk.

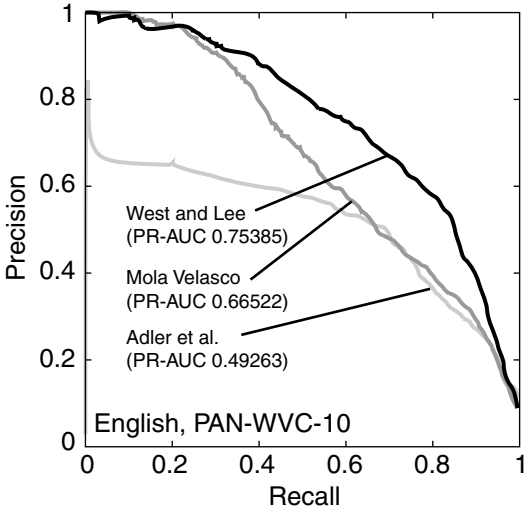
The PAN Competition

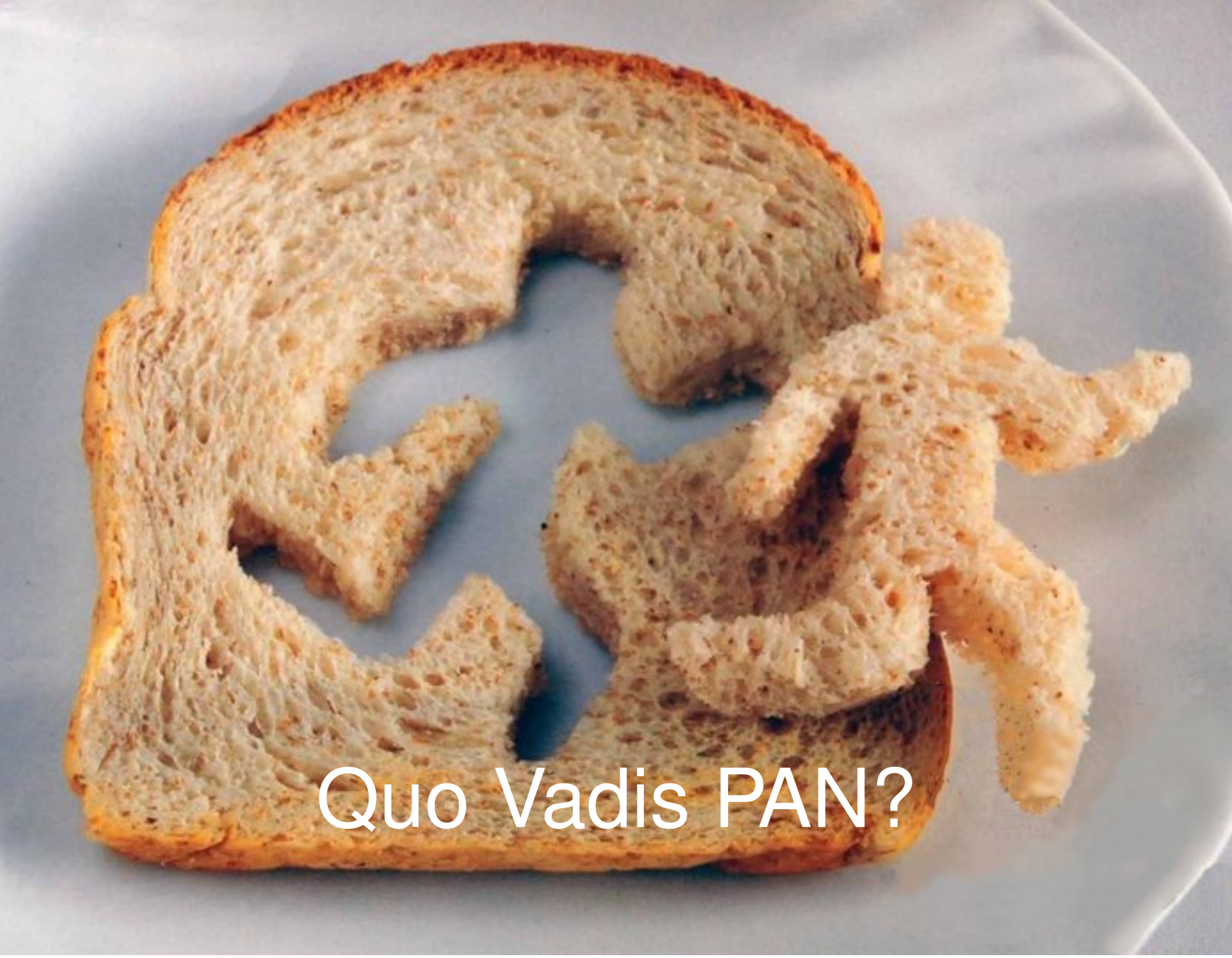
Wikipedia Vandalism Detection










Quo Vadis PAN?

Quo Vadis PAN?

Lessons Learned and Outlook

- Focus & Simplicity
  - ➜ Focus on specific aspects of the tasks.
  - ➜ Reduced number of task variants.
  - ➜ Reduced number of parameters and limited ranges.
- Realism & Scale
  - ➜ New corpora for plagiarism detection and authorship identification.
  - ➜ Scale up where necessary, scale down otherwise.
- Contributions & Challenges
  - ➜ Inclusion of real plagiarism and real cases of disputed authorship.
  - ➜ Distinguishing text reuse and plagiarism.
  - ➜ Considering human performance.

Thank you!

Visit us at pan.webis.de.

Mail us at pan@webis.de.