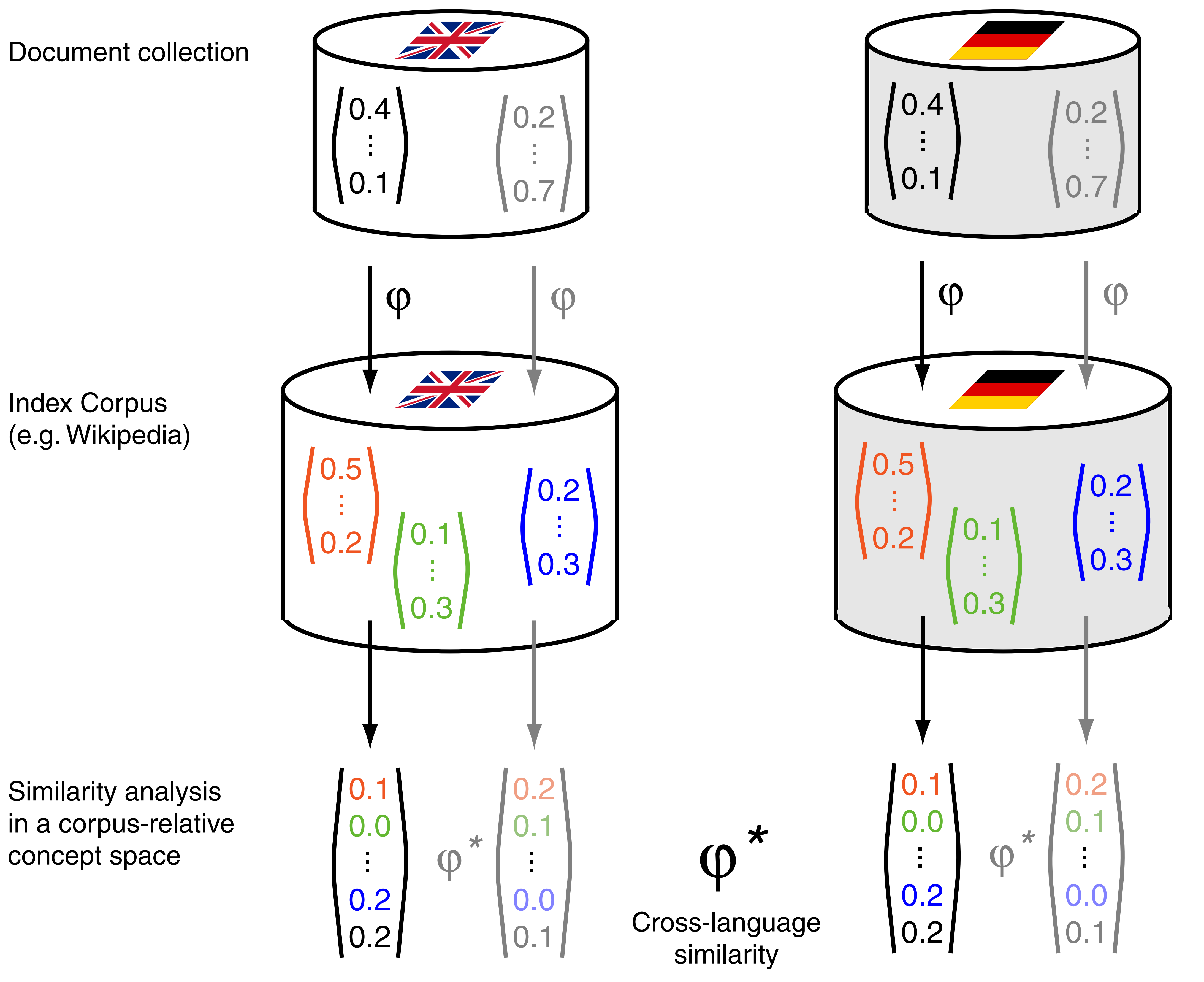
A Wikipedia-based Multilingual Retrieval Model

Martin Potthast, Benno Stein, and Maik Anderka

Contribution

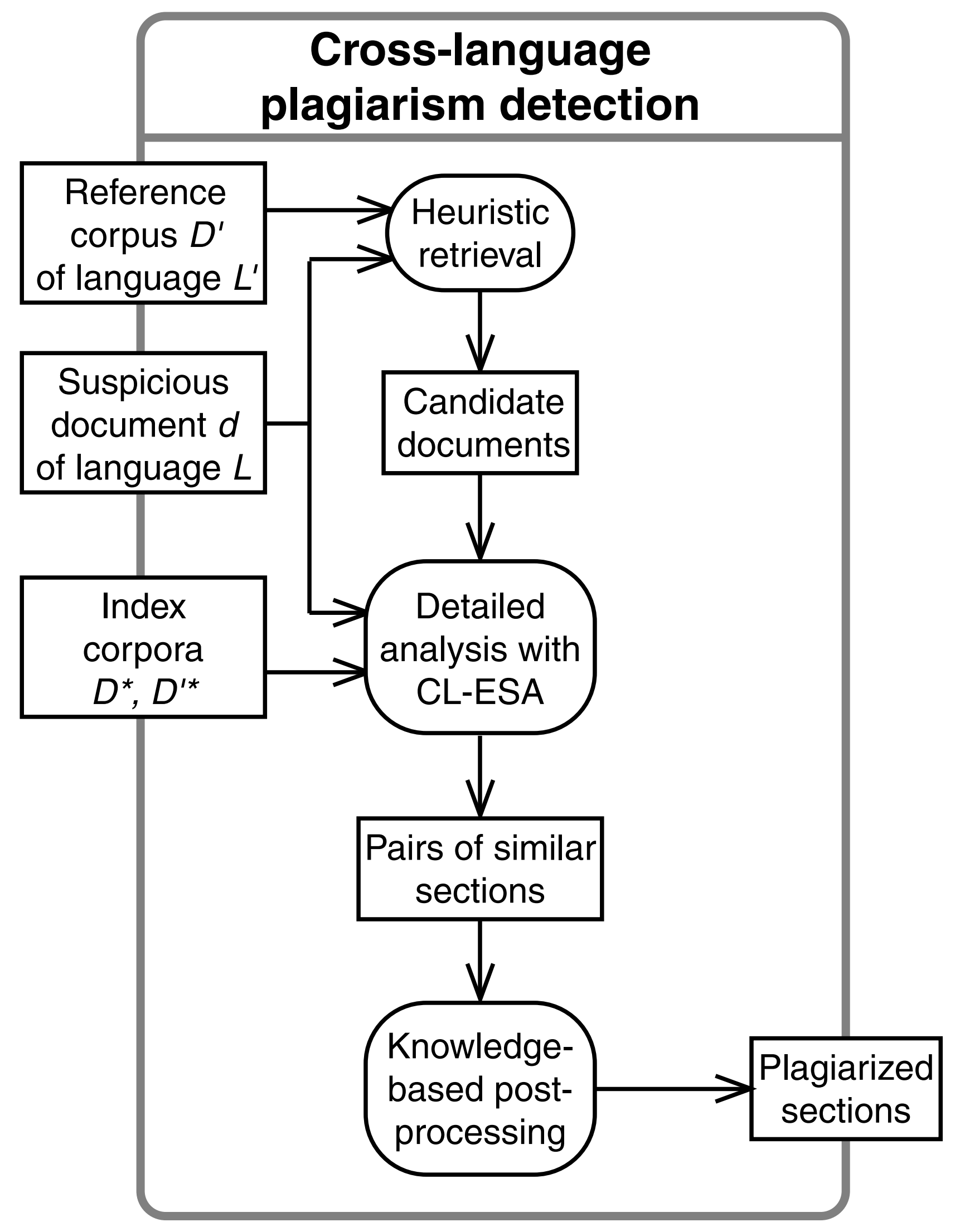

Document collection

Index Corpus (e.g. Wikipedia)

Similarity analysis in a corpus-relative concept space

Cross-language similarity

Current Work


Cross-language plagiarism detection

Reference corpus D′ of language L′

Suspicious document d of language L

Index corpora D*, D′*

Heuristic retrieval

Candidate documents

Detailed analysis with CL-ESA

Pairs of similar sections

Knowledge-based post-processing

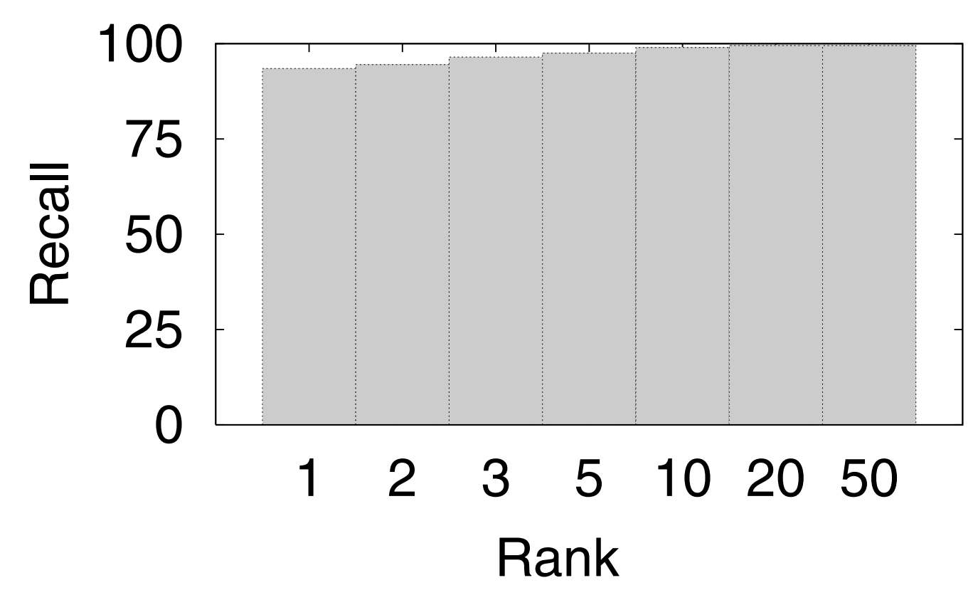
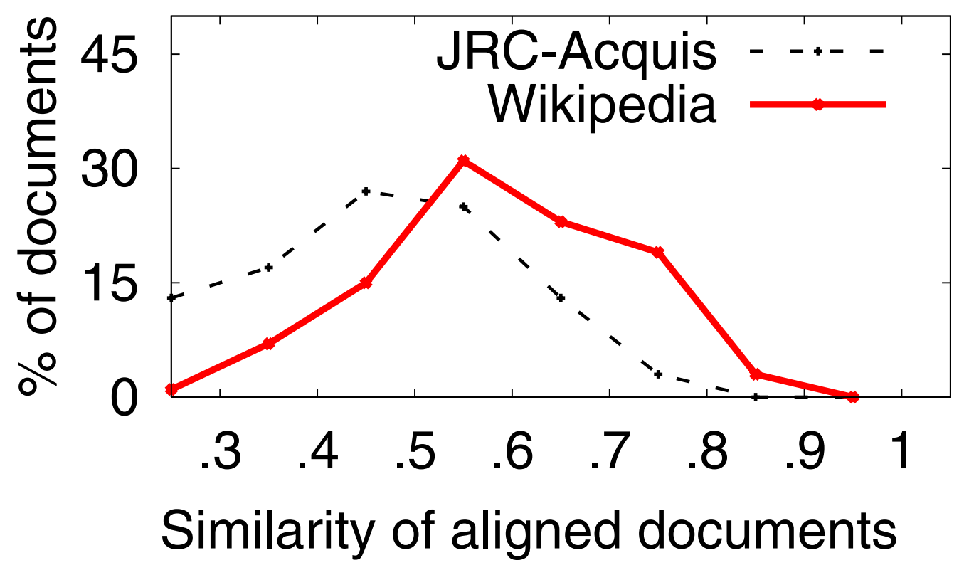
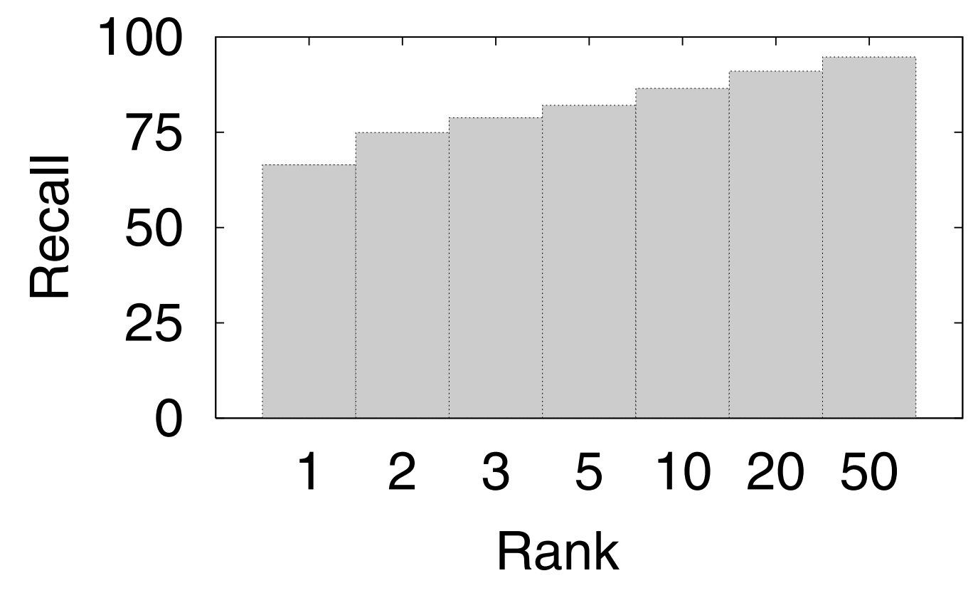
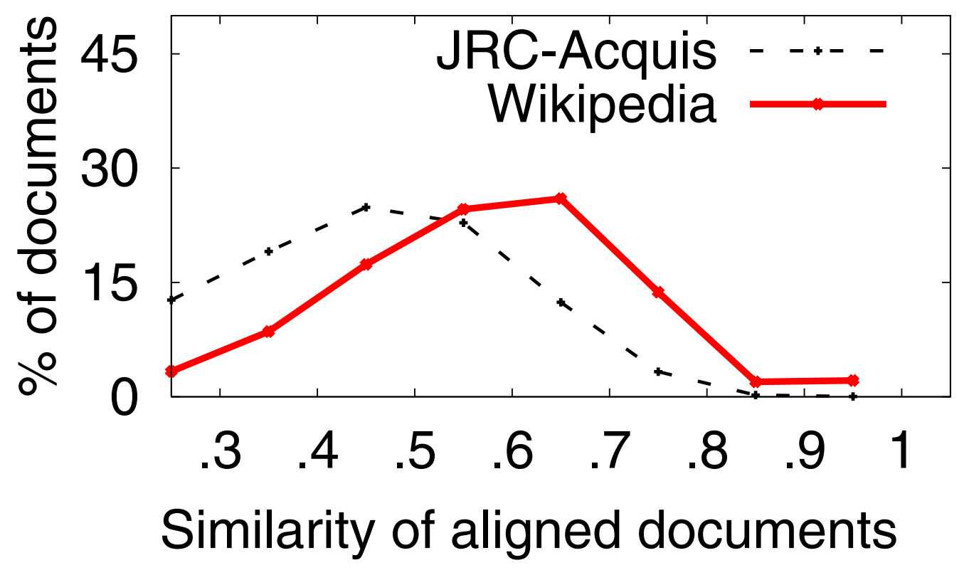
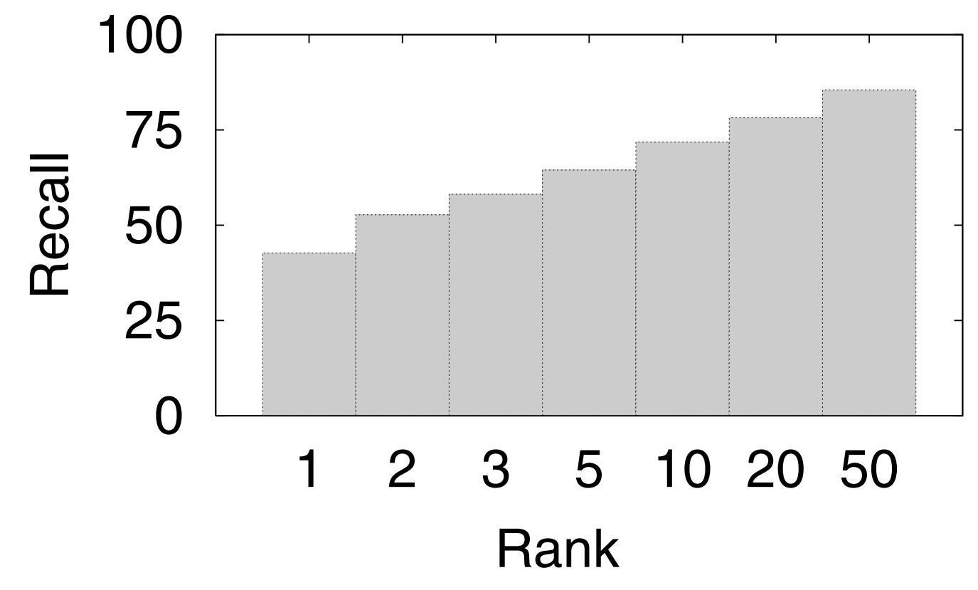
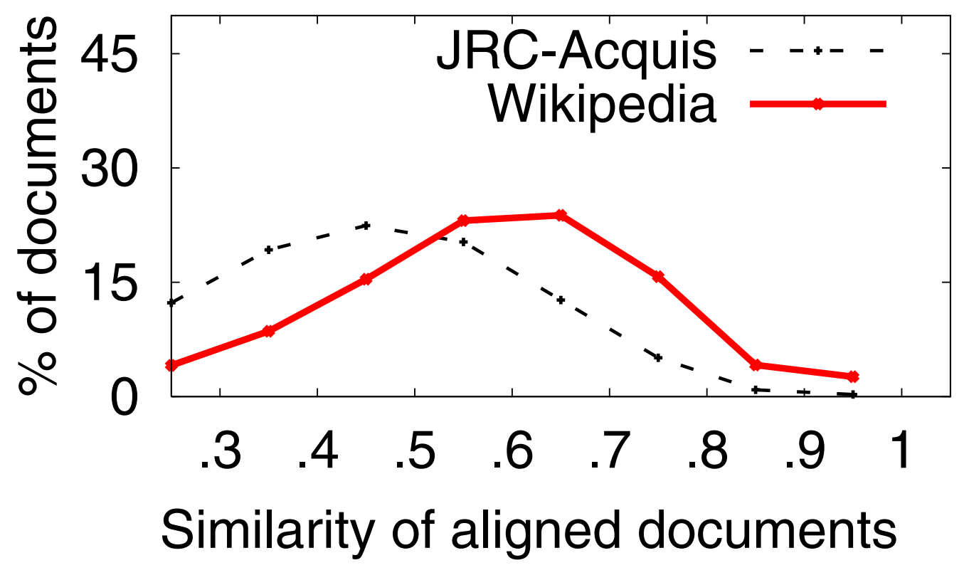
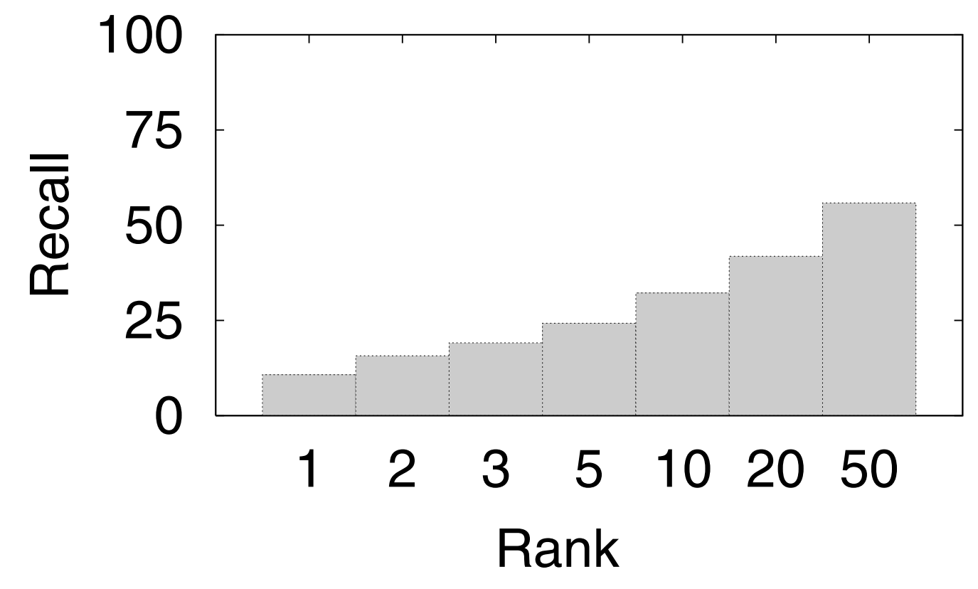
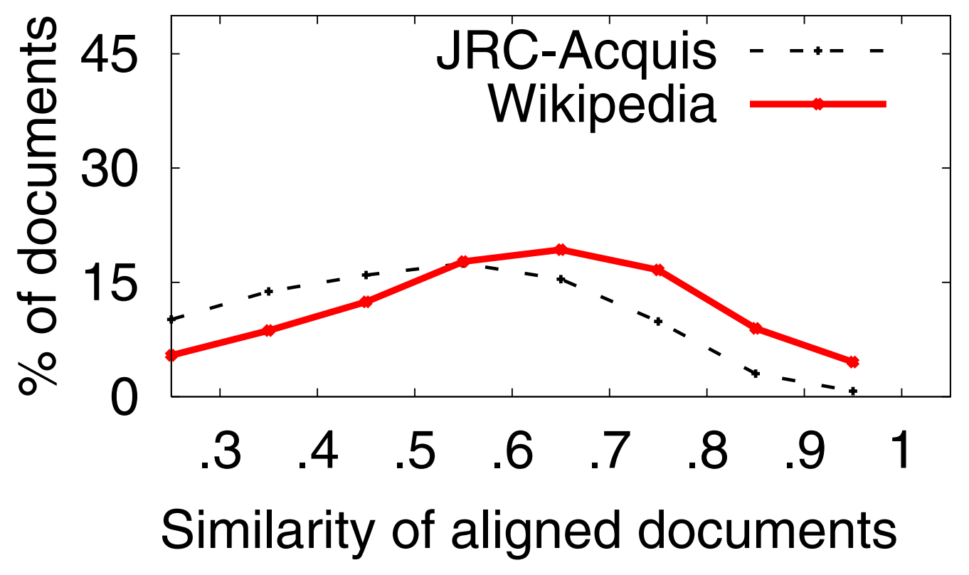
Plagiarized sections

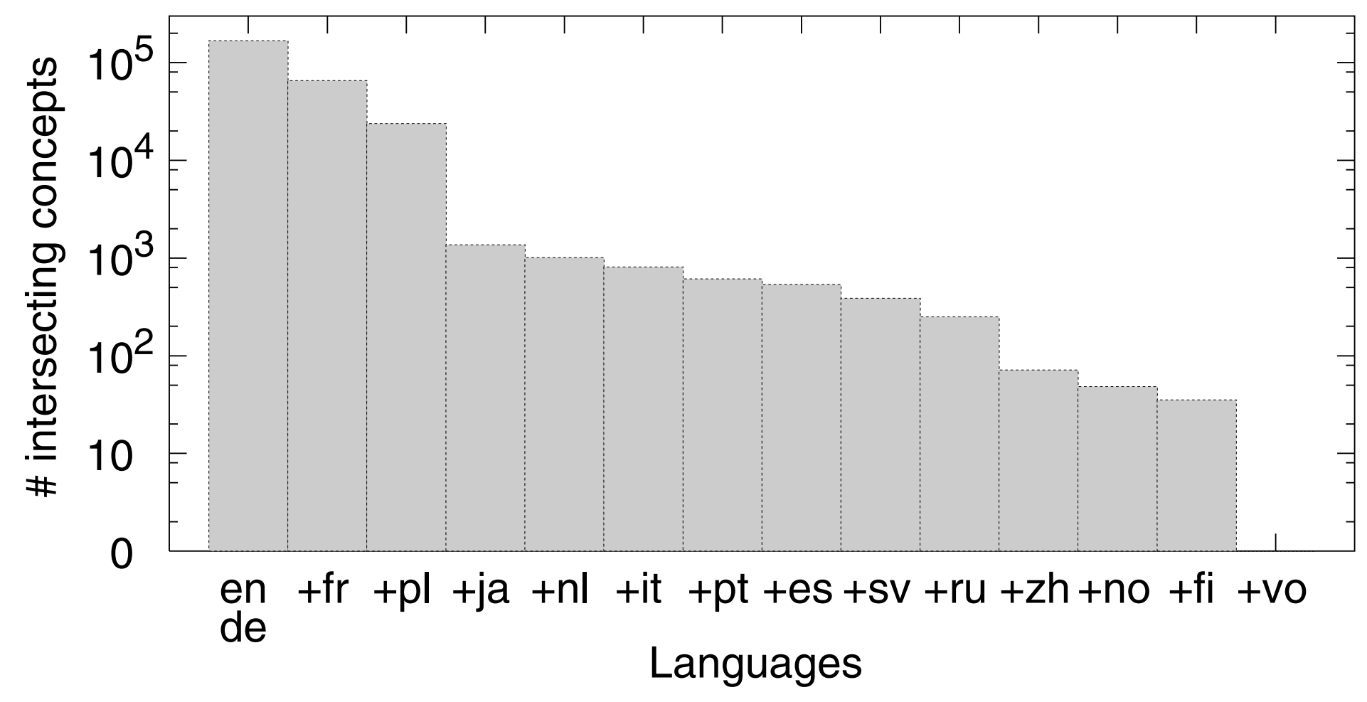
Heuristic retrieval: A subset of *D′* is retrieved which contains candidate documents that are likely to be sources for plagiarism with respect to the content of *d*.

Detailed analysis: The candidate documents are compared section-wise to d using CL-ESA for each pair of sections.

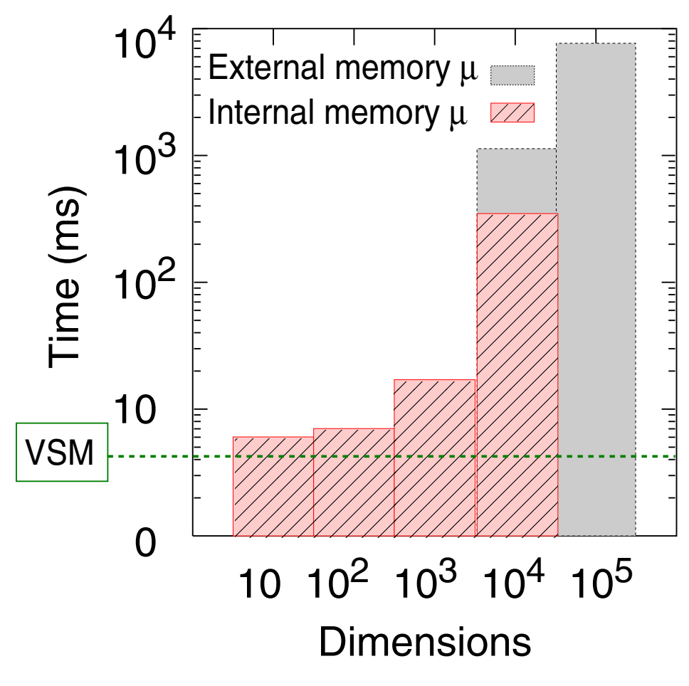
Knowledge-based postprocessing: Sections from the candidate documents that are similar to a section in d are processed in detail, for instance to filter cases of proper citation.

Evaluation

Experiment 1 Cross-language Ranking	Experiment 2 Bilingual Rank Correlation		Experiment 3 CL Similarity Distribution	Experiment 4 Dimension
	JRC-Aquis	Wikipedia		
	0.81	0.72		100 000
	0.46	0.61		10 000
	0.20	0.44		1 000
	0.09	0.22		100



Experiment 5: Starting with the two most prominent languages in Wikipedia, English and German, we study how many concepts are described in both languages, and how many are in the intersection set if more languages are considered.



Test documents: Two test document collections, *D* and *D′*, comprising 3 000 documents each were chosen (1 000 translation-algined, 1 000 concept-algined, 1 000 not aligned).

Experiment 1: Given an aligned document *d* ∈ *D*, all documents in *D′* are ranked according to their cross-language similarity to *d*. Let *d′* ∈ *D′* be the aligned document of *d* ∈ *D*, then the retrieval rank of *d′* is recorded. Ideally, *d′* should be on the first or at least on one of the top ranks. The experiment was repeated for all of the aligned documents in *D*.

Experiment 2: For a pair of aligned documents *d* ∈ *D* and *d′* ∈ *D′* the documents from D are ranked twice: *(i)* with respect to their cross-language similarity to *d* using a cross-language retrieval model, and, *(ii)* with respectto their monolingual similarity to *d* using the vector space model. The top 100 ranks of the two rankings are compared using a rank correlation coefficient, e. g. Spearma's ρ, which measures their (dis-)agreement as a value between -1 and 1 respectively.

Experiment 3: This experiment contrasts the distribution of pairwise similarities of translation-aligned and concept-aligned documents.

Experiment 4: Both retrieval quality and runtime depend on the concept space dimension of CL-ESA, which in turn corresponds the size of a language's index corpus.

Experiment 6: The time to index a document is between 10 to 100 milliseconds, which is comparable to the time to compute a vector space representation. Employed hardware: Intel Core 2 Duo processor at 2 GHz and with 1 GB RAM.

Web Technology and Information Systems

Bauhaus University Weimar