

A Pairwise Document Analysis Approach for Monolingual Plagiarism Detection



Nava Ehsan, Azadeh Shakery

School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Iran

Introduction

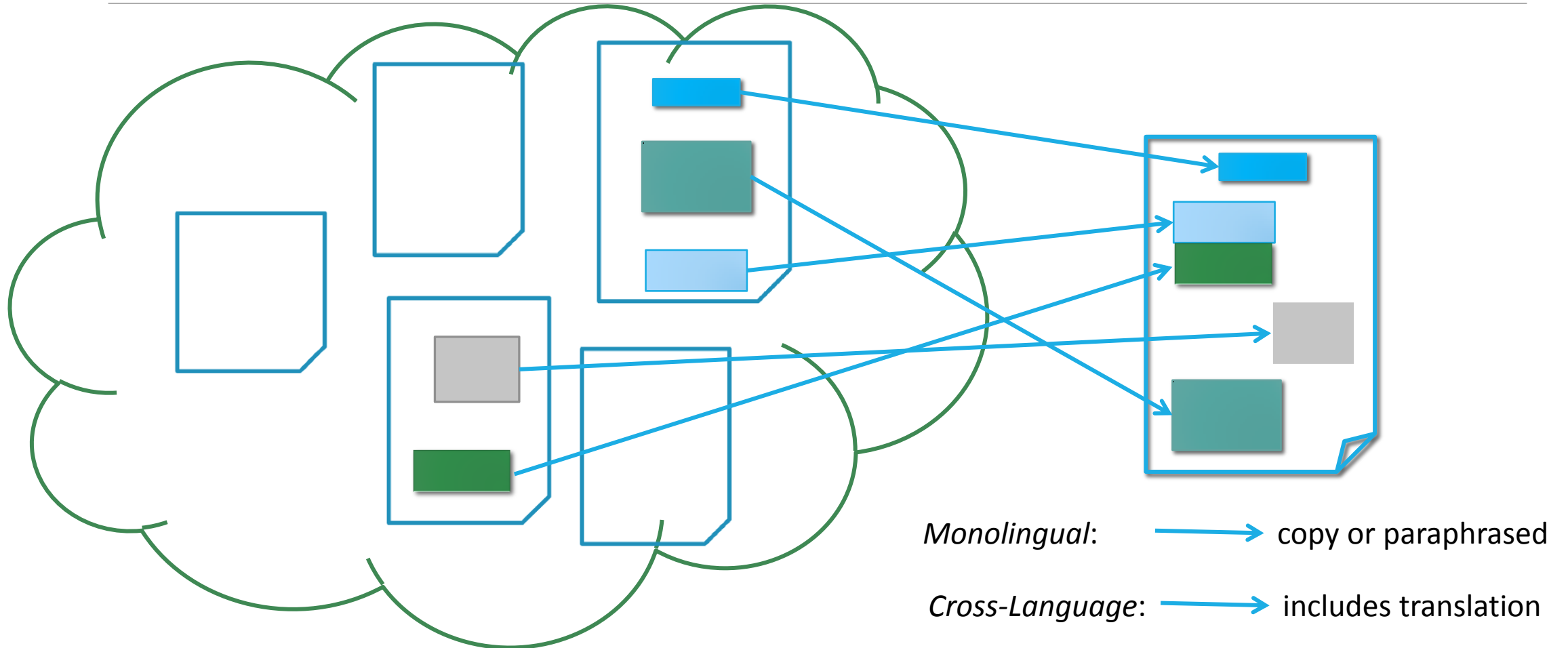
Plagiarism:

- Unauthorized use of **Text**, code, idea,

Plagiarism detection research area has received increasing attention

- The rapid growth of documents in different languages
- Increased accessibility of electronic documents

Prototypical Plagiarism



Problem definition has two steps

Candidate document retrieval

- D : set of source documents
- d' : suspicious document with fragments $d'_{f'}$

$$\text{Candidate documents } (D, d') = \{ d \in D \mid d_f \subseteq d, d'_{f'} \subseteq d', \text{Sim}(d_f, d'_{f'}) > \alpha \}$$

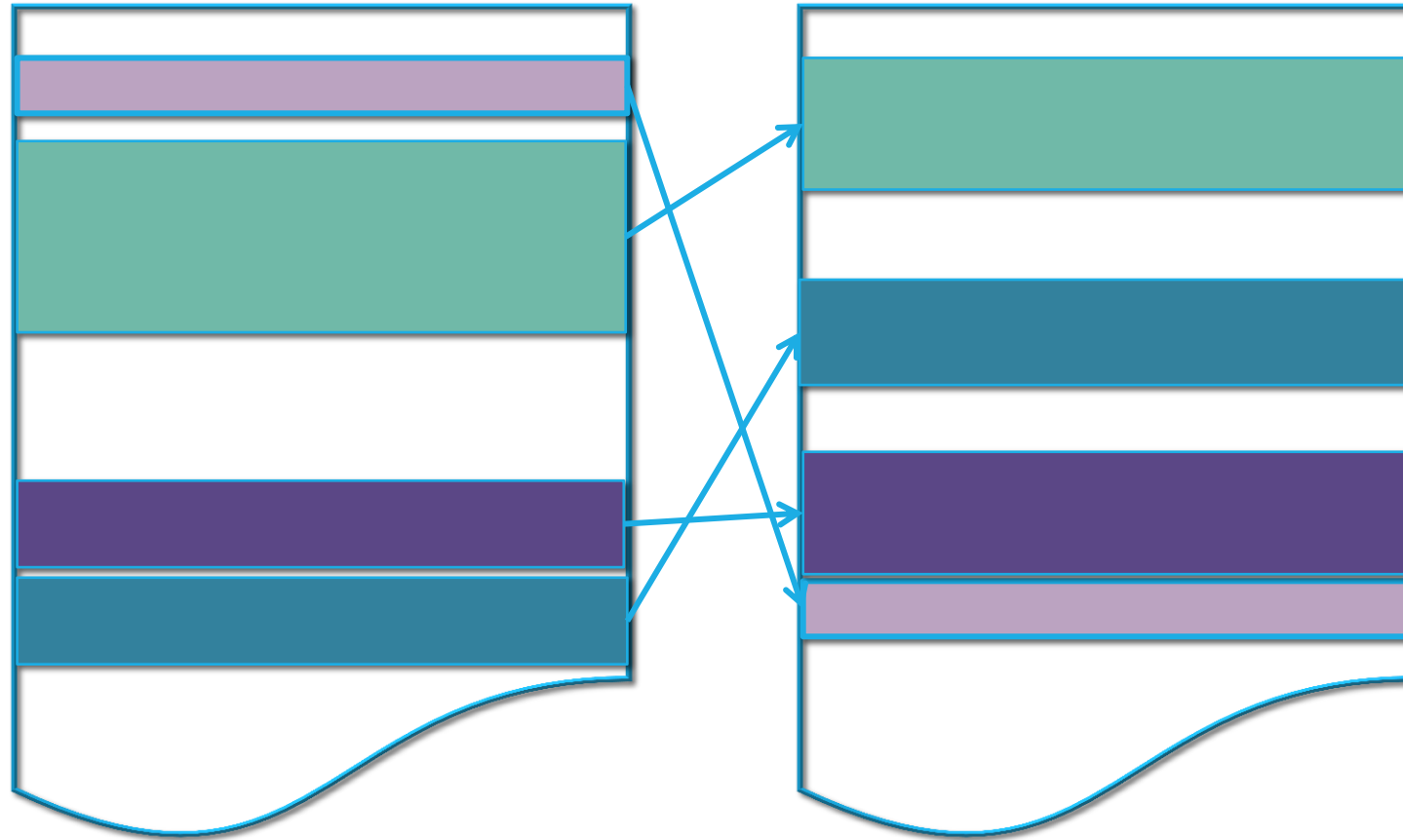


Pairwise document similarity

- d : source document with fragments d_f
- d' : suspicious document with fragments $d'_{f'}$

$$\text{Copied pairs}(d, d') = \{ \langle d_f, d'_{f'} \rangle \mid d_f \subseteq d, d'_{f'} \subseteq d', \text{Sim}(d_f, d'_{f'}) > \beta \}$$

Detailed analysis in a pair of documents



Possible **errors** in detecting plagiarism:

- Text that is not plagiarized might be erroneously reported
- Part or whole of plagiarized source or target text might be unreported
- Parts of one plagiarism case might be reported as separate cases

Evaluation Metrics

- S : set of true plagiarism cases , R : set of detections reported

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

Fraction of reported detections (at character level) that are truly plagiarized

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

Fraction of plagiarism cases (at character level) that are detected

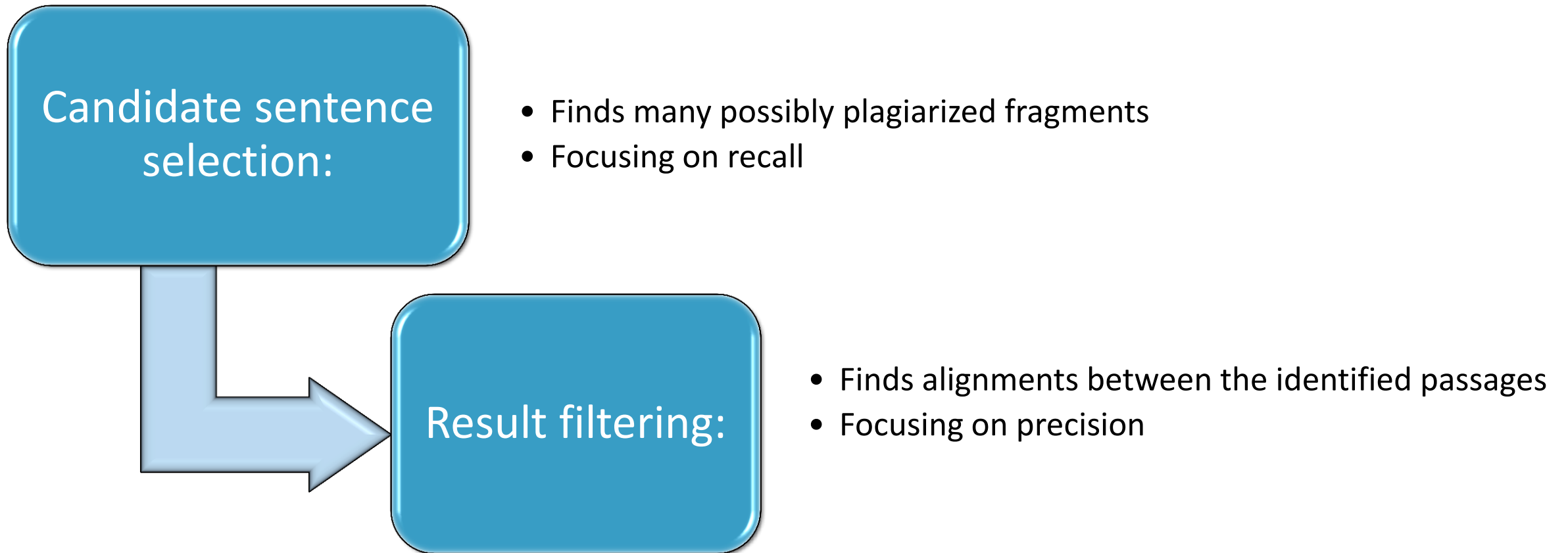
$$Granularity(R, S) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|$$

Average number of reported detections per detected plagiarism case

$$Plagdet(R, S) = \frac{F_1(R, S)}{\log_2(1 + Granularity(R, S))}$$

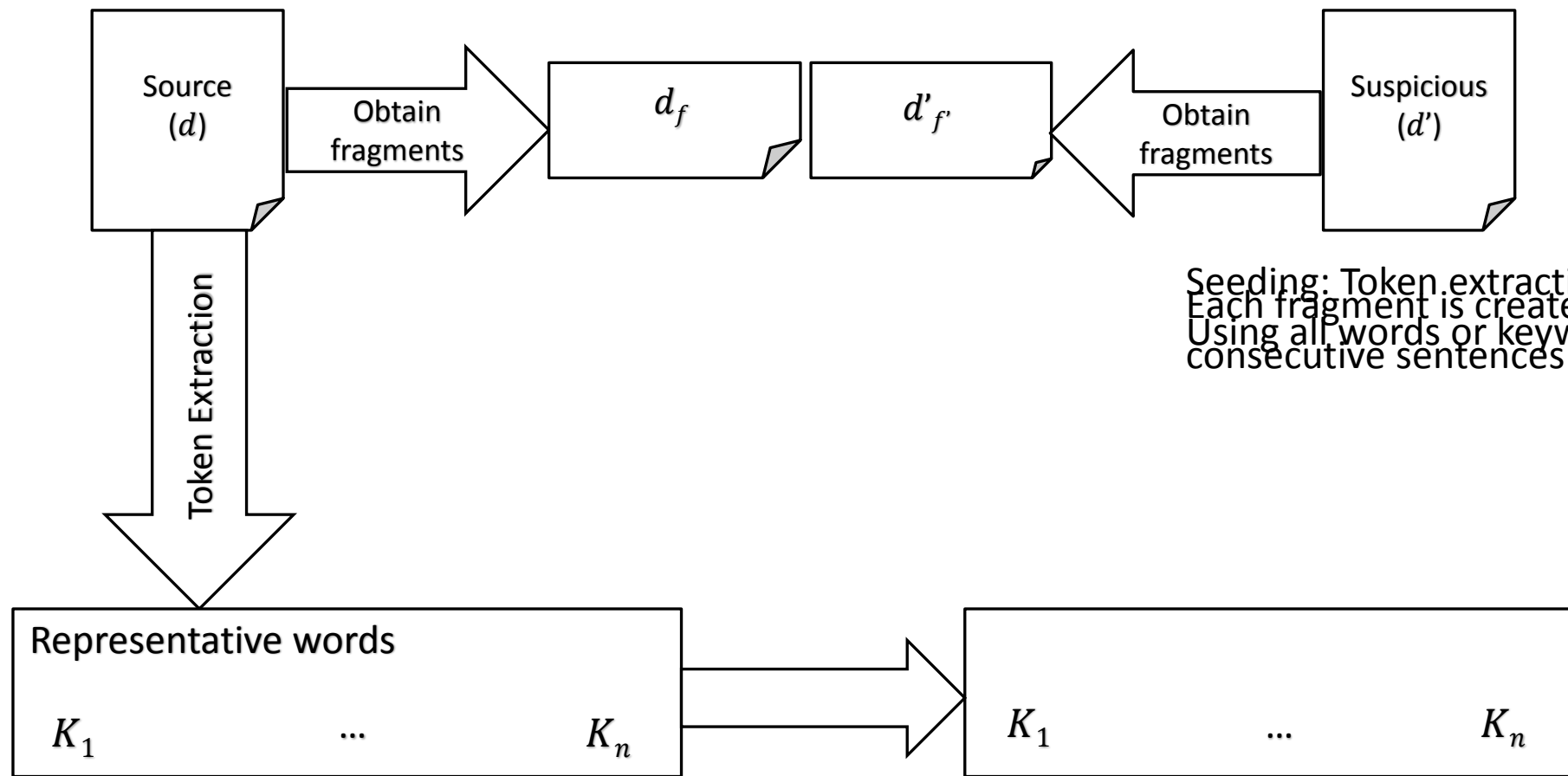
Combined metric

Two phase algorithm for identifying plagiarized text fragments

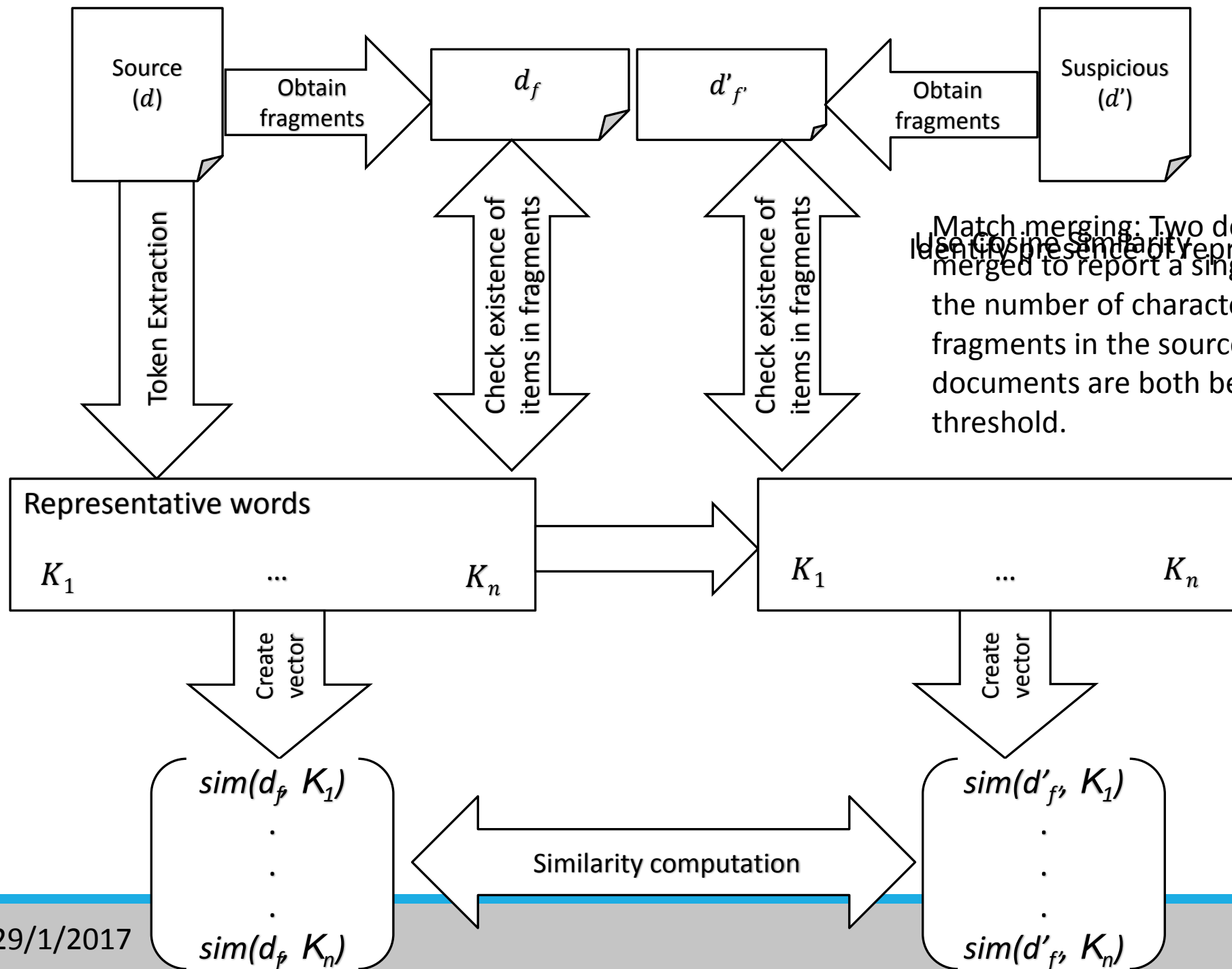


Step 1:

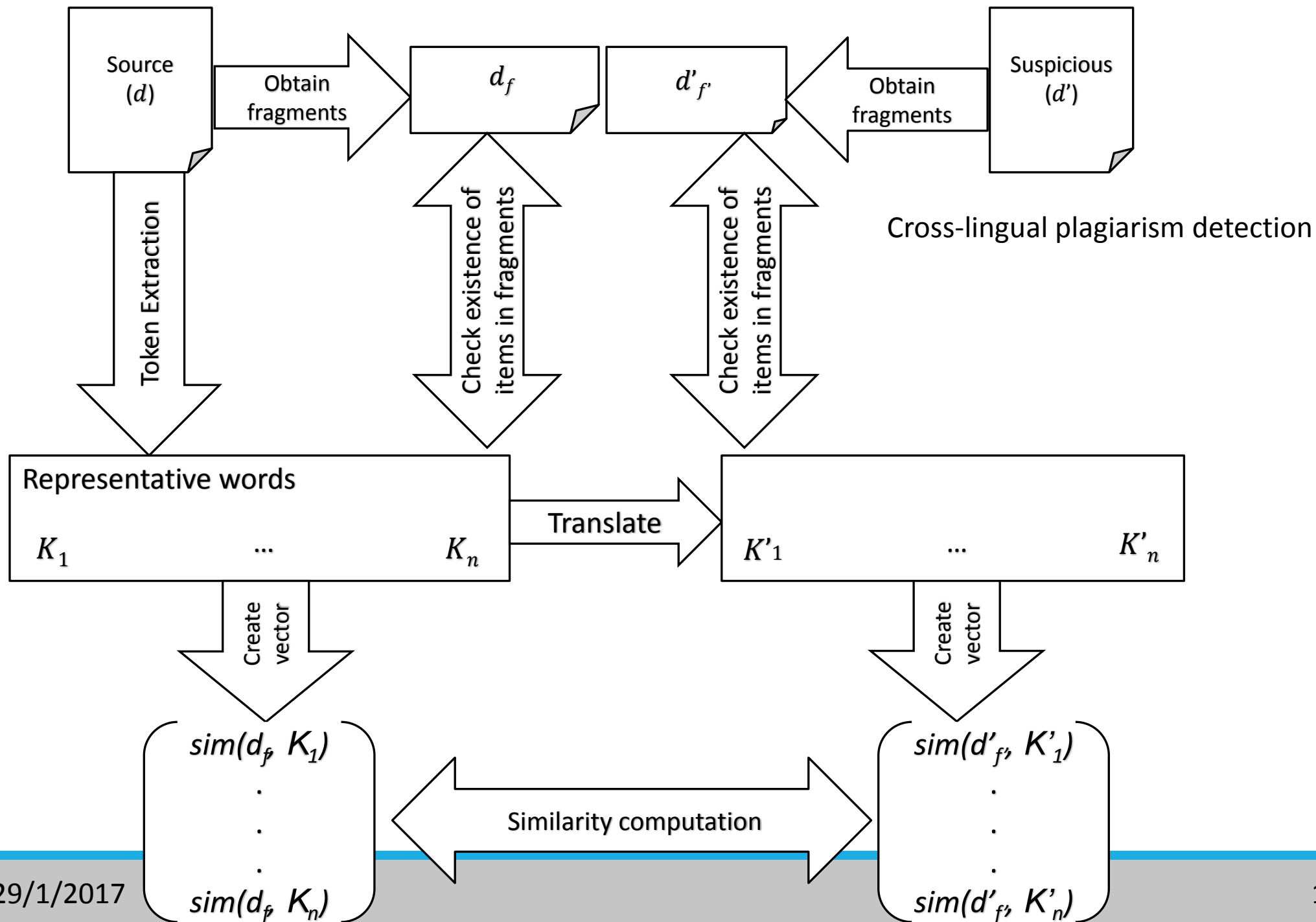
Candidate Sentence Selection



Seeding: Token extraction.
Each fragment is created from a sequence of k
consecutive sentences using a sliding window.
Using all words or keywords



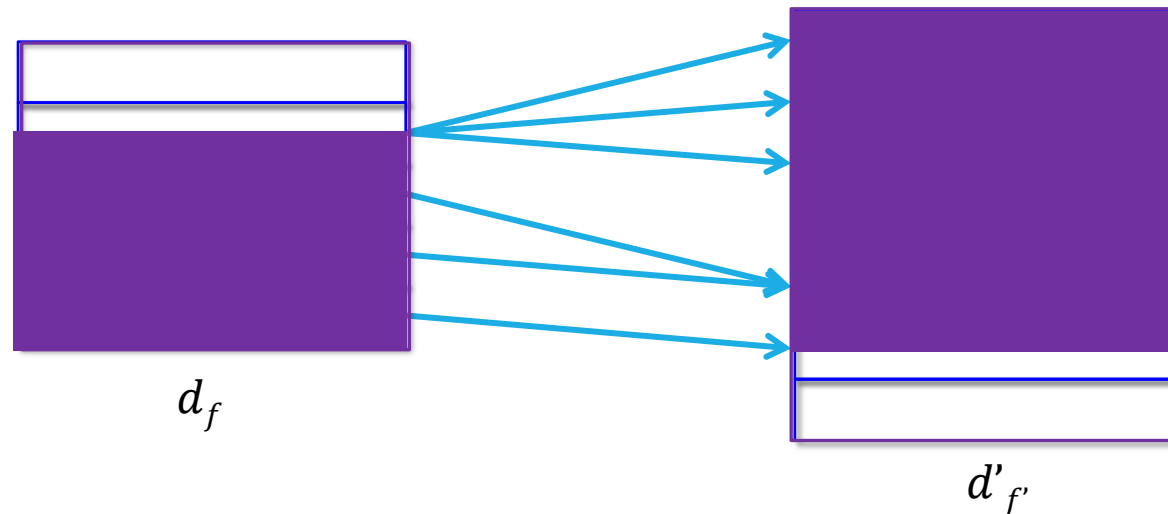
Match merging: Two detected fragments are identified as similar if representative terms merged to report a single plagiarism case if the number of characters between those fragments in the source and suspicious documents are both below a proximity threshold.



Step 2: Result Filtering

Aligning segments within fragment pairs

- Fragment pair from the first step retrieved
- Fragments split into smaller segments
- Segments aligned using a dynamic programming algorithm
 - allowing 1:0, 0:1, 1:1, 2:1, 1:2, 3:1 and 1:3 alignments
 - exclude sentences at start or end of fragment with >50% content in 1:0 or 0:1 alignments



Alignment details

$$S(i, j) = \max \begin{cases} S(i-1, j) \\ S(i, j-1) \\ S(i-1, j-1) + \text{sim}(\text{susp}_i, \text{src}_j) \\ S(i-1, j-2) + \text{sim}(\text{susp}_i, \text{src}_{j-1:j}) \\ S(i-1, j-3) + \text{sim}(\text{susp}_i, \text{src}_{j-2:j}) \\ S(i-2, j-1) + \text{sim}(\text{susp}_{i-1:i}, \text{src}_j) \\ S(i-3, j-1) + \text{sim}(\text{susp}_{i-2:i}, \text{src}_j) \end{cases}$$

where $S(i, j)$ represents the score of the optimal alignment from the beginning of the fragment to the i^{th} suspicious segment and the j^{th} source segment

- To penalize 1-0 and 0-1 alignments and also to make all scores comparable, we keep track of the number of alignments obtained so far, and the score in each step is normalized by the number of alignments.

Granularity level of alignment

Sentence level:

- Using sentences as the granularity level of alignment

n -gram level:

- A plagiarized fragment may omit pieces from the source, but it is likely that at least some of the smallest units are preserved
- n is the expected number of terms in each segment

Results

Result of detailed analysis sub-task using PersinaPlagdet2016 training corpus

t = Similarity threshold, n=Number of sentences

	Precision	Recall	Granularity	Plagdet
(t = 0.2, n = 5)	0.4004	0.8151	1	0.5370
(t = 0.3, n = 5)	0.7630	0.7486	1	0.7557
(t = 0.4, n = 5)	0.8532	0.5357	1	0.6582
(t = 0.3, n = 3)	0.7867	0.8304	1	0.8080
(t = 0.4, n = 3)	0.8604	0.6567	1	0.7449

Result of detailed analysis sub-task using PersinaPlagdet2016 test corpus

	Precision	Recall	Granularity	Plagdet	Runtime
(t = 0.3, n = 5)	0.7496	0.7050	1	0.7266	00:24:08

Results

Evaluation of the second phase, result filtering step:

$t = 0.3, n = 3$

	Precision	Recall	Granularity	Plagdet
Without result filtering	0.6029	0.8602	1	0.7087
After result filtering	0.7867	0.8304	1	0.8080

Results

Evaluation of the seeding phase, using keywords:

	Precision	Recall	Granularity	Plagdet
(t = 0.3, n = 3)	0.5118	0.8858	1	0.6487
(t = 0.4, n = 3)	0.6431	0.8928	1	0.7476
(t = 0.5, n = 3)	0.7475	0.8862	1	0.8110
(t = 0.6, n = 3)	0.8117	0.8459	1	0.8282
(t = 0.7, n = 3)	0.8522	0.7531	1	0.7996


Cross-lingual detailed analysis for plagiarism detection

Ehsan, N., Tompa, F.W., Shakery, A.: Using a dictionary and n-gram alignment to improve fine-grained cross-language plagiarism detection. In: *Proceedings of the 2016 ACM Symposium on Document Engineering*. pp. 59–68. ACM (2016).

	Precision	Recall	Granularity	Plagdet
Using PAN2012 English-German dataset	0.9301	0.8193	1	0.8712

Summary

- The proposed method is a two phase approach for identifying plagiarized fragments
- The first phase tries to find possibly plagiarized fragments
- The second phase tries to improve the precision metric
- The framework is applicable in any language
- The approach could be adapted for cross language domain

An illustration featuring a large, light brown thought bubble. Inside the bubble is a glowing yellow lightbulb with radiating lines. A stylized figure of a person in a blue suit is shown from the back, reaching out with their right arm towards the lightbulb. The text "Thanks for your attention" is centered within the thought bubble.

Thanks for your
attention