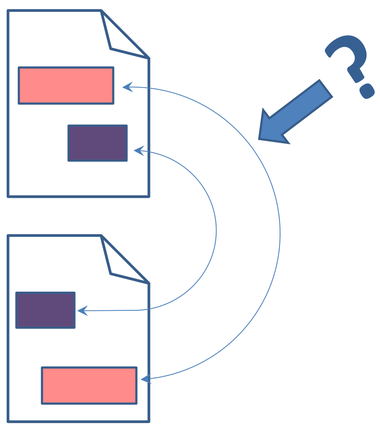
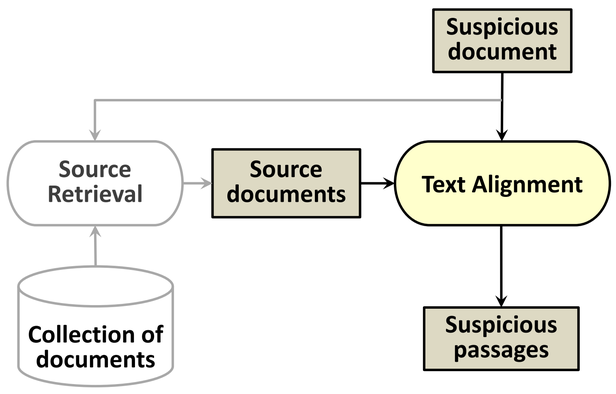
A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014

Miguel A. Sanchez-Perez, Grigori Sidorov, Alexander Gelbukh

Centro de Investigación en Computación
Instituto Politécnico Nacional

1. Task





2. Methodology

2.1. Pre-processing

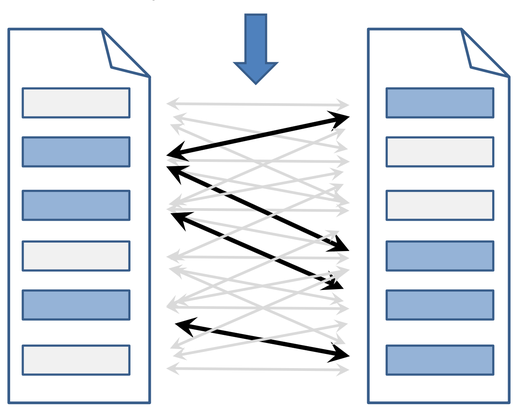
Sentence splitting, tokenizing, removal of tokens that do not start from a letter or digit, reducing to lowercase, stemming, joining small sentences (1-2 words) with the next one.

2.1. Seeding

Vector representation of sentences:
TF-IDF, where **sentences** are "documents,"
thus called TF-I**S**F: inverse **sentence** freq.
"Documents": union of sentences of both docs

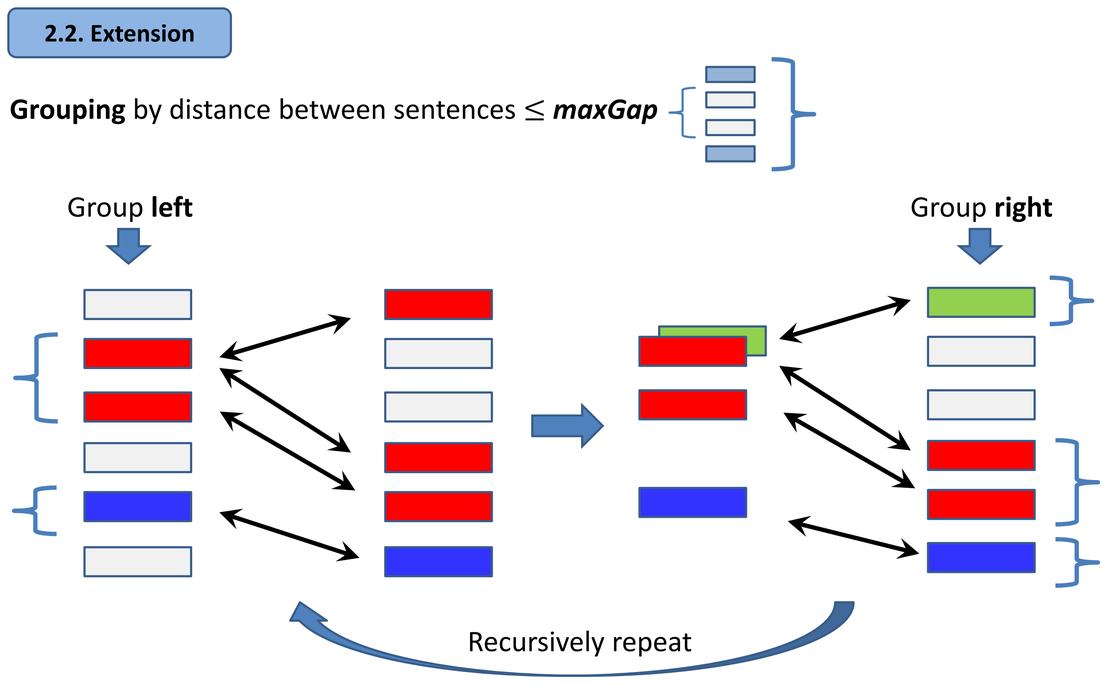
Vector similarity:
Cosine similarity ≥ threshold *th1*
AND Dice similarity ≥ threshold *th2*

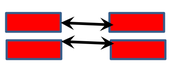
Seeds: pairs of **similar** sentences



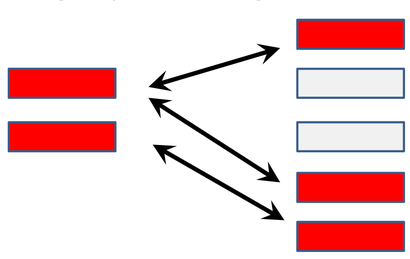
2.2. Extension

Grouping by distance between sentences ≤ ***maxGap***

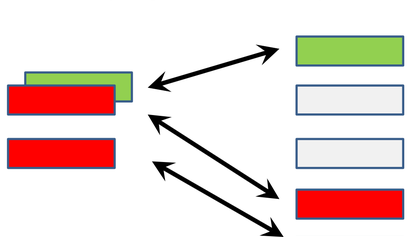


For **each** group (like this ):
If **cosine similarity** between **left** and **right** sides of the group ≤ threshold *th3*
then form groups **again** with ***maxGap - 1***

Example:
After group left and right with ***maxGap = 2***



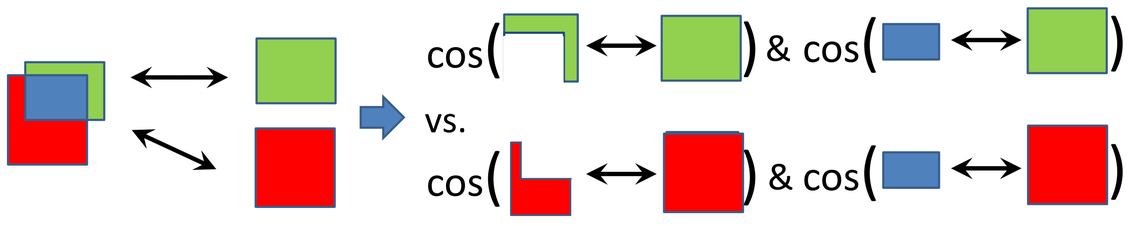
Grouping with ***maxGap = 1***



Resulting **groups** are considered **plagiarism cases**

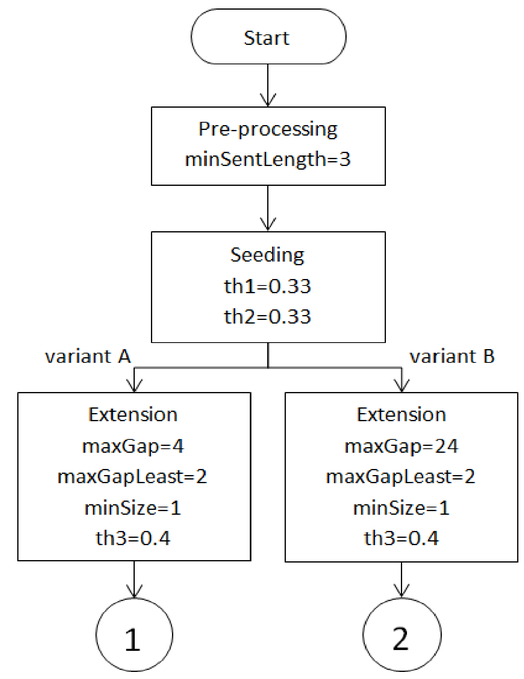
2.3. Filtering

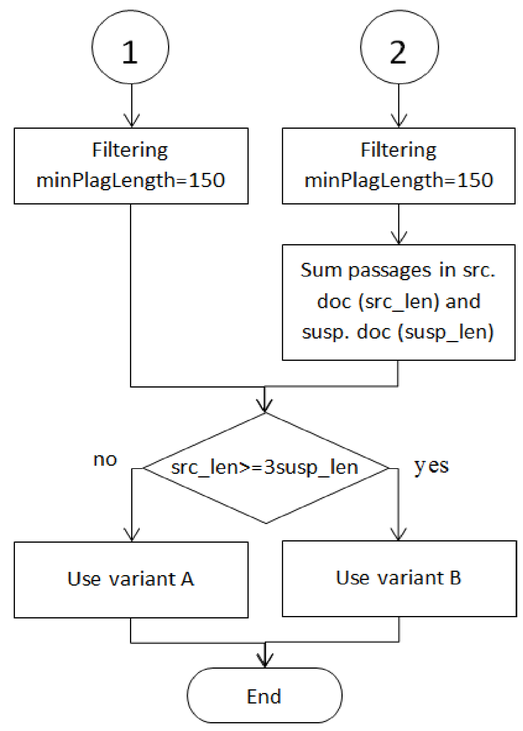
Resolving **overlapping**



Plagiarism cases shorter than ***minSentLength*** characters are **removed**

3. Adaptive behavior





4. Results

Training: PAN 2014 = PAN 2013 training corpus. **Evaluation**: PAN 2014, PAN 2013.

Obfus-cation	2014=2013 training corpus				PAN **2013** test corpus			
	Plagdet	Recall	Prec	Granul	Plagdet	Recall	Prec	Granul
None	0.893	0.978	0.822	1.000	0.900	0.978	0.833	1.000
Random	0.888	0.858	0.921	1.000	0.884	0.860	0.910	1.000
Translation	0.883	0.890	0.877	1.000	0.886	0.889	0.884	1.000
Summary	0.577	0.424	0.994	1.043	0.560	0.412	0.999	1.058
Entire	0.877	0.879	0.877	1.002	0.878	0.879	0.881	1.003

Team	Year	None	Random	Translation	Summary	Entire corpus
Sanchez-Perez	-	0.90032	0.88417	0.88659	0.56070	0.87818
Torrejón	2013	0.92586	0.74711	0.85113	0.34131	0.8222
Kong	2013	0.8274	0.82281	0.85181	0.43399	0.81896
Suchomel	2013	0.81761	0.75276	0.67544	0.61011	0.74482
Saremi	2013	0.84963	0.65668	0.70903	0.11116	0.69913
Shrestha	2013	0.89369	0.66714	0.62719	0.1186	0.69551
Palkovskii	2013	0.82431	0.49959	0.60694	0.09943	0.61523
Nourian	2013	0.90136	0.35076	0.43864	0.11535	0.57716
Baseline	2013	0.93404	0.07123	0.1063	0.04462	0.42191
Gillam	2013	0.85884	0.04191	0.01224	0.00218	0.40059
Jayapal	2013	0.3878	0.18148	0.18181	0.0594	0.27081

Final result of PAN **2014** Text Alignment

Plagdet	Team
0.87818	Miguel A. Sanchez-Perez, Grigori Sidorov, and Alexander Gelbukh Instituto Politécnico Nacional, Mexico
0.86933	Gabriel Oberreuter and Andreas Eiselt

5. Conclusions

Text alignment task: best result of all 11 participating systems, thanks to:

1. TF-ISF (inverse *sentence* frequency) measure for "soft" removal of stopwords
2. Recursive extension algorithm: dynamic adjustment of tolerance to gaps
3. Novel algorithm for resolution of overlapping cases by comparison of competing cases
4. Dynamic adjustment of parameters by type of case (summary vs. other types)

Contact

Miguel A. Sanchez-Perez
masp1988@hotmail.com
Centro de Investigación en Computación,
Instituto Politécnico Nacional, http://www.cic.ipn.mx