INSTITUTO POLITÉCNICO NACIONAL

Centro de Investigación en Computación

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

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

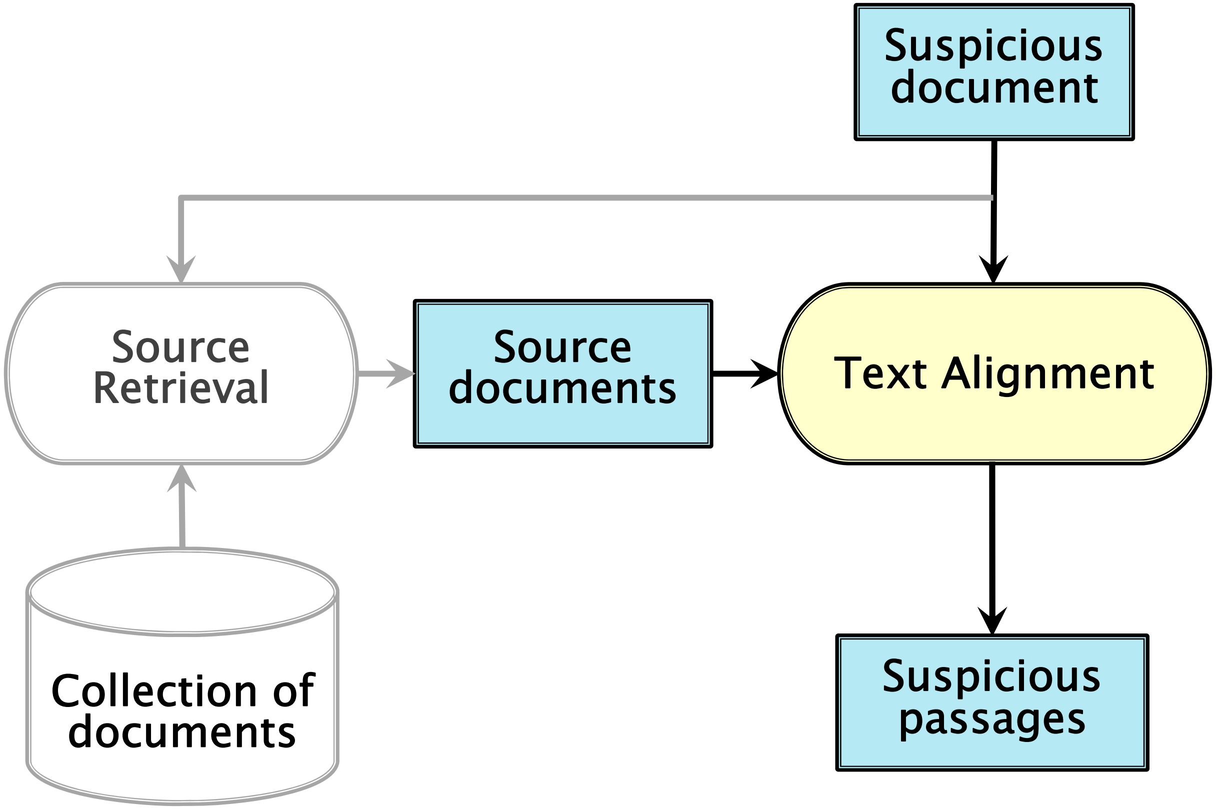
Tuesday, 16 September 2014

Content

1. Task
2. Methodology
3. Adaptative behavior
4. Results
5. Conclusions
6. Future Work

Task

Text Alignment: Given a **pair of documents**, the task is to identify all **contiguous maximal-length** passages of **reused text** between them.



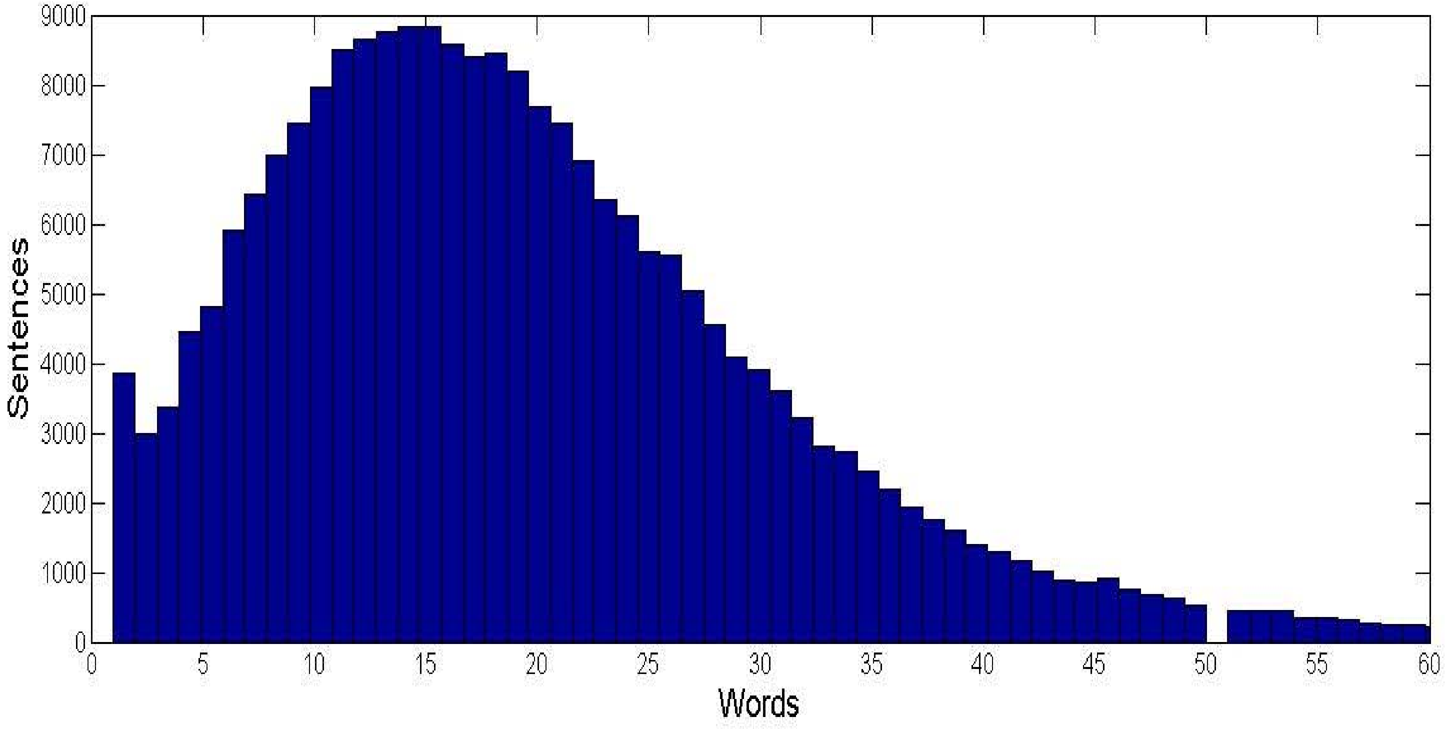
Methodology

- Preprocessing
- Seeding
- Extension
- Filtering

Preprocessing

- Sentence splitting (Kiss pretrained punkt model)
- Tokenizing (Treebank word tokenizer)
- Keeping tokens starting with a letter or digit
- Reducing to lowercase
- Stemming (Porter algorithm)
- Joining small sentences (1–2 words) with the next one

Preprocessing



PAN 2014 training corpus
Sentences length histogram (words)

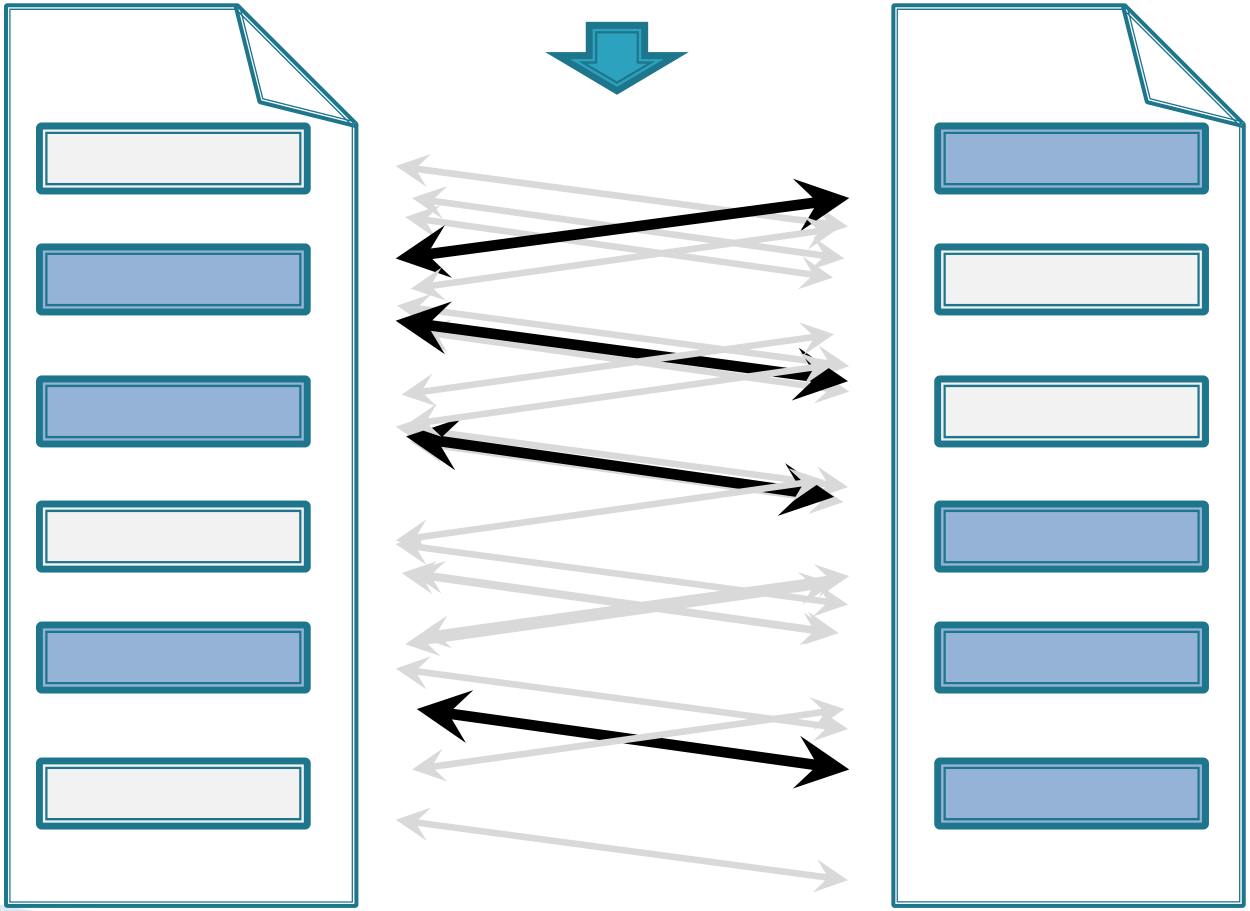
Seeding

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

Vector similarity:
Cosine similarity $\geq$ threshold *th1*
AND Dice similarity $\geq$ threshold *th2*

Seeding

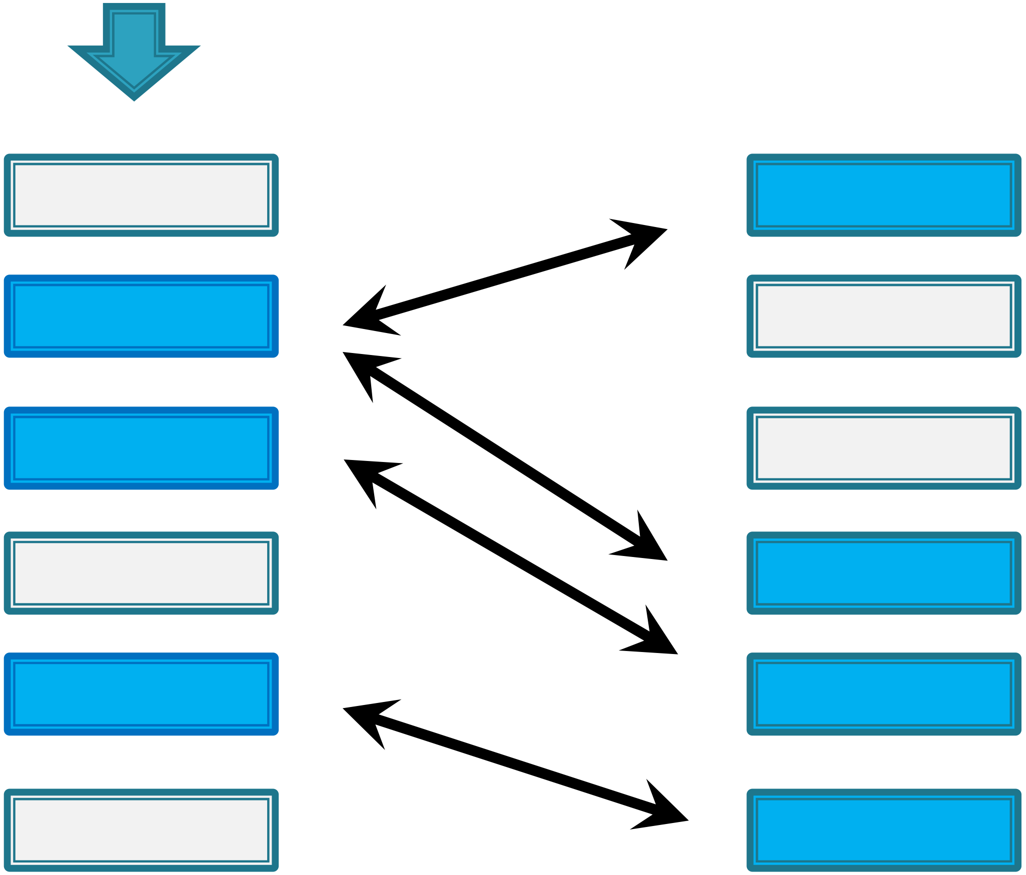
Seeds: pairs of **similar** sentences



Extension

Grouping

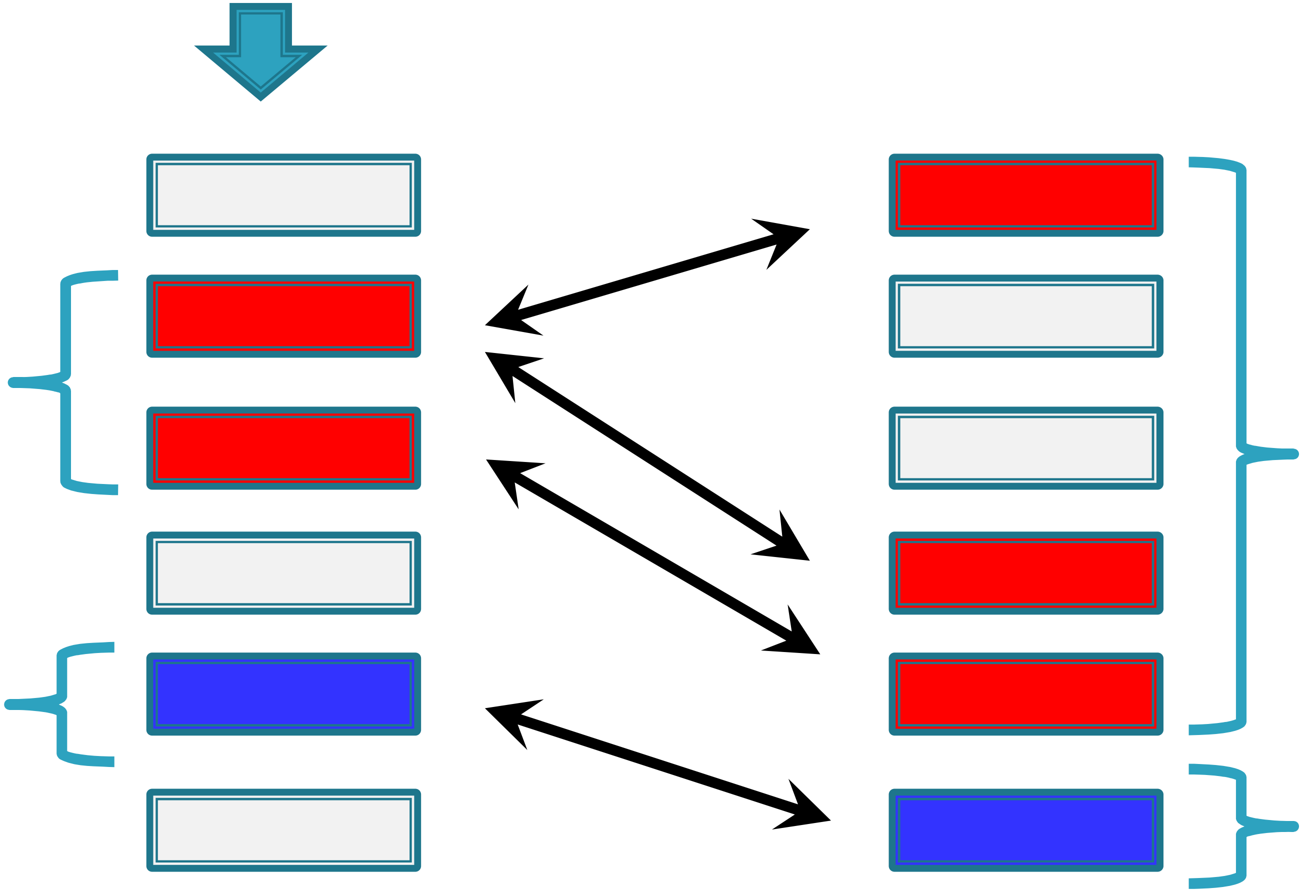
Group left



Extension

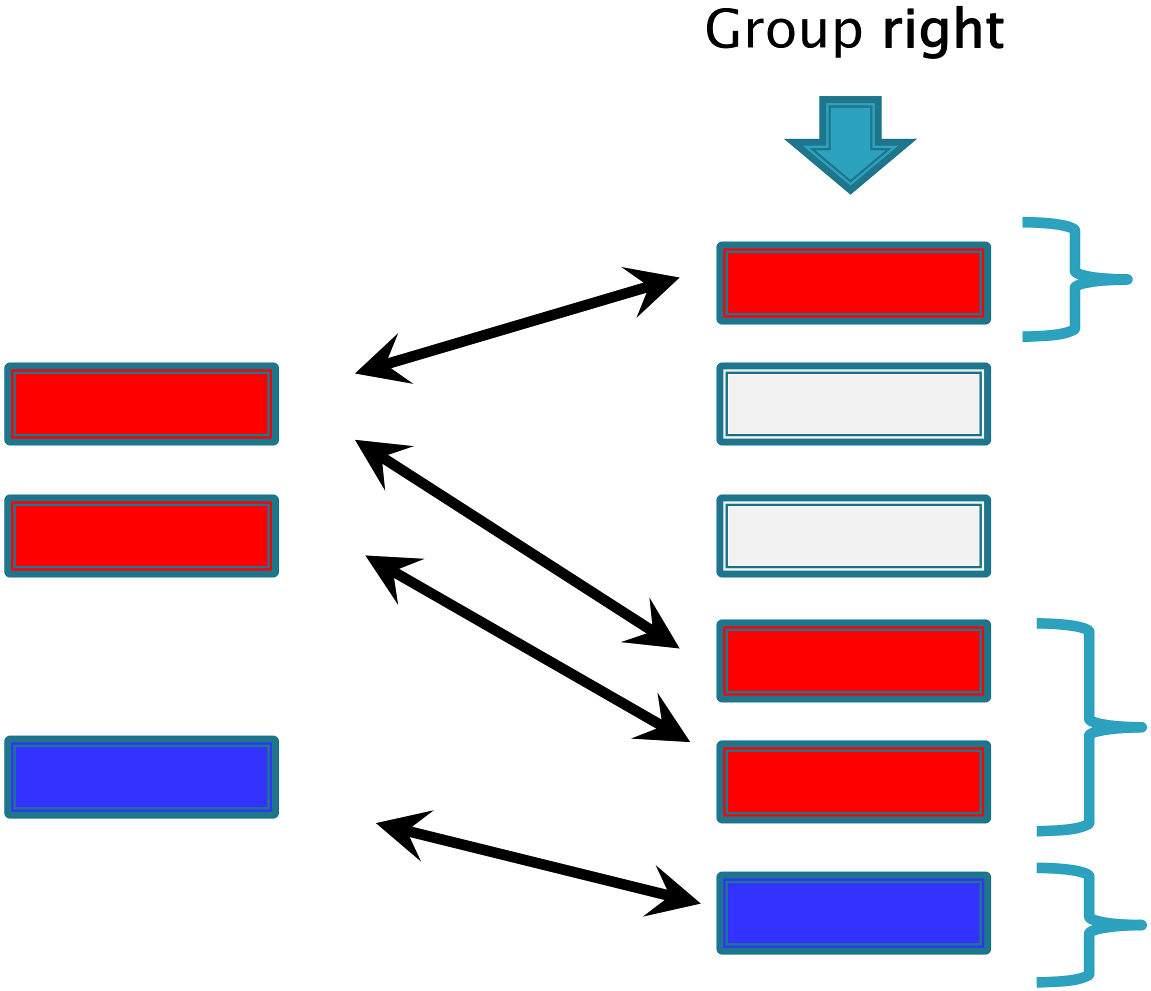
Grouping

Group left



Extension

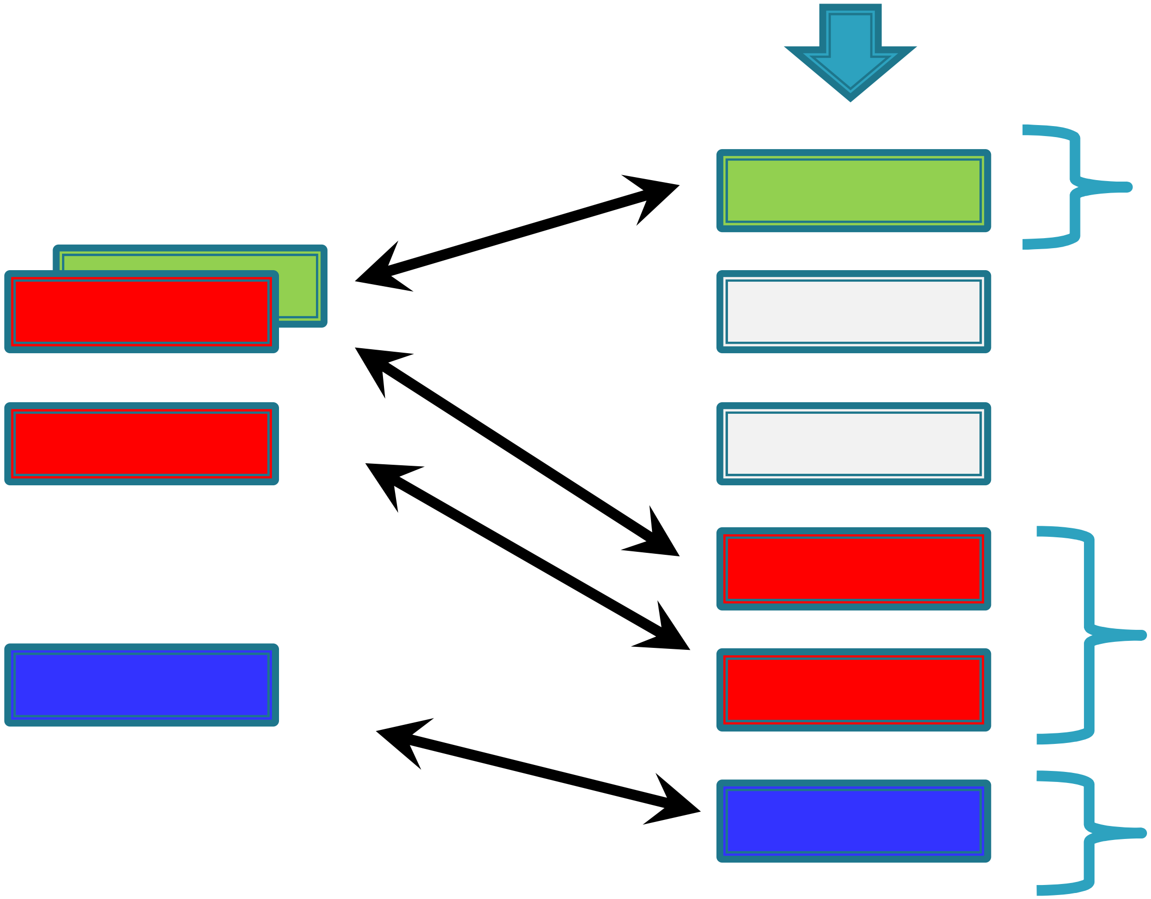
Grouping



Extension

Grouping

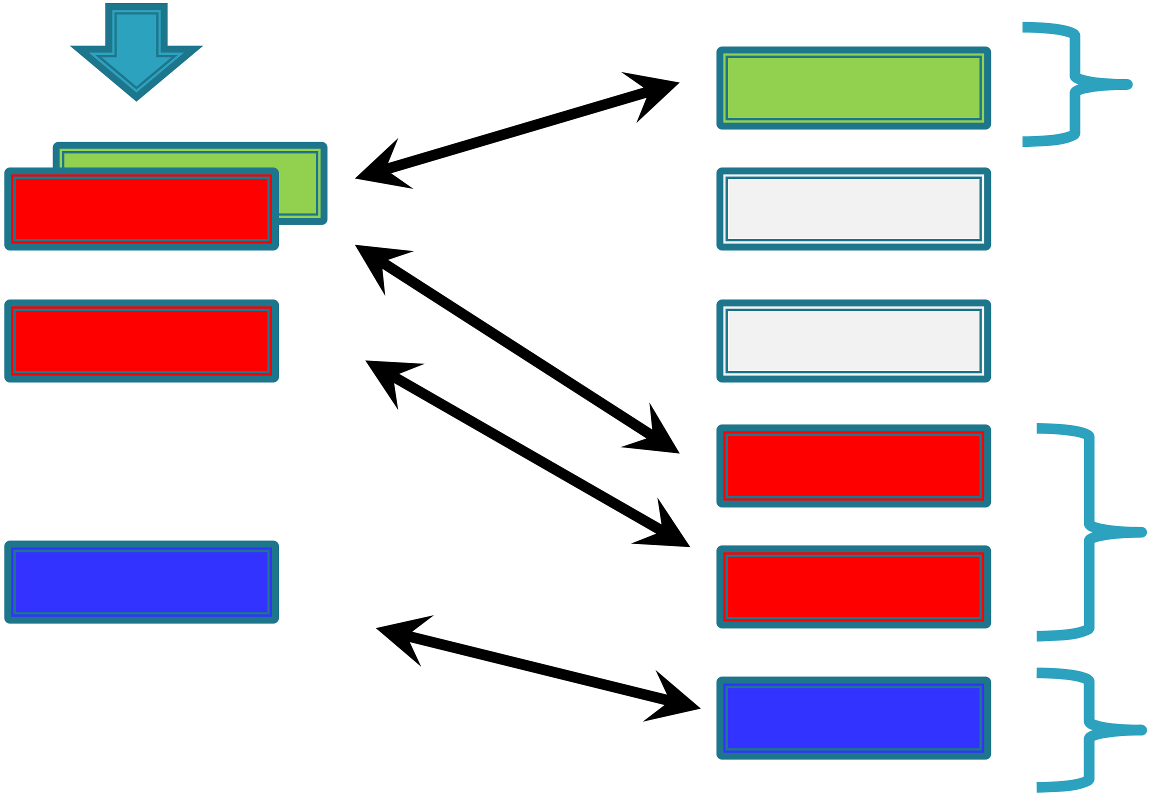
Group right



Extension

Grouping

Group left

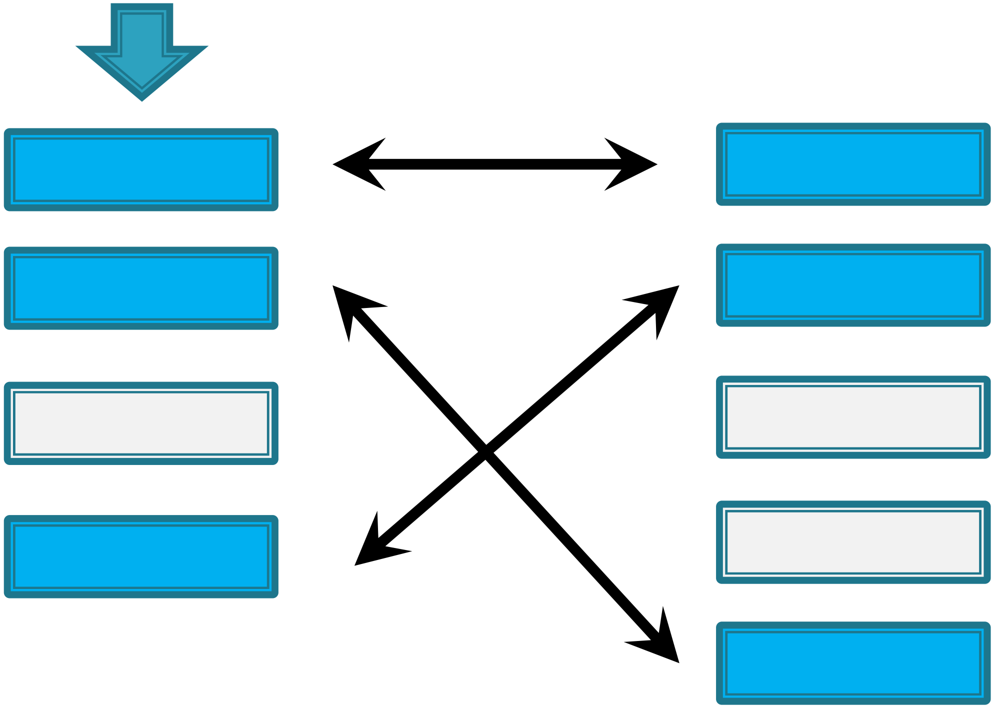


Extension

Grouping

Example:
maxGap = 1

Group **left**

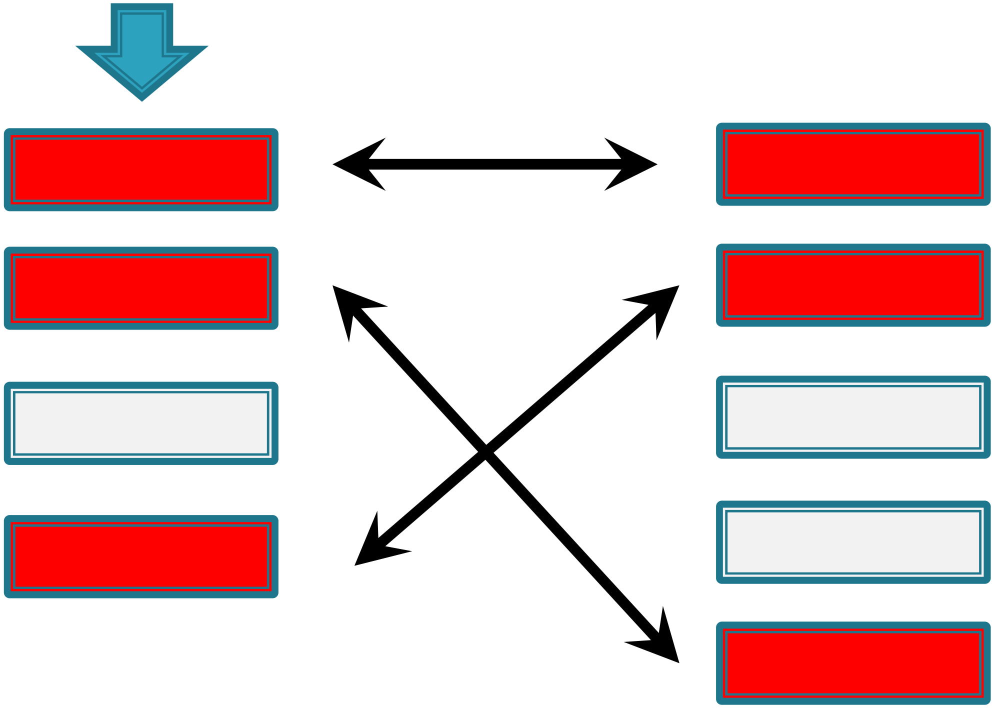


Extension

Grouping

Example:
maxGap = 1

Group **left**

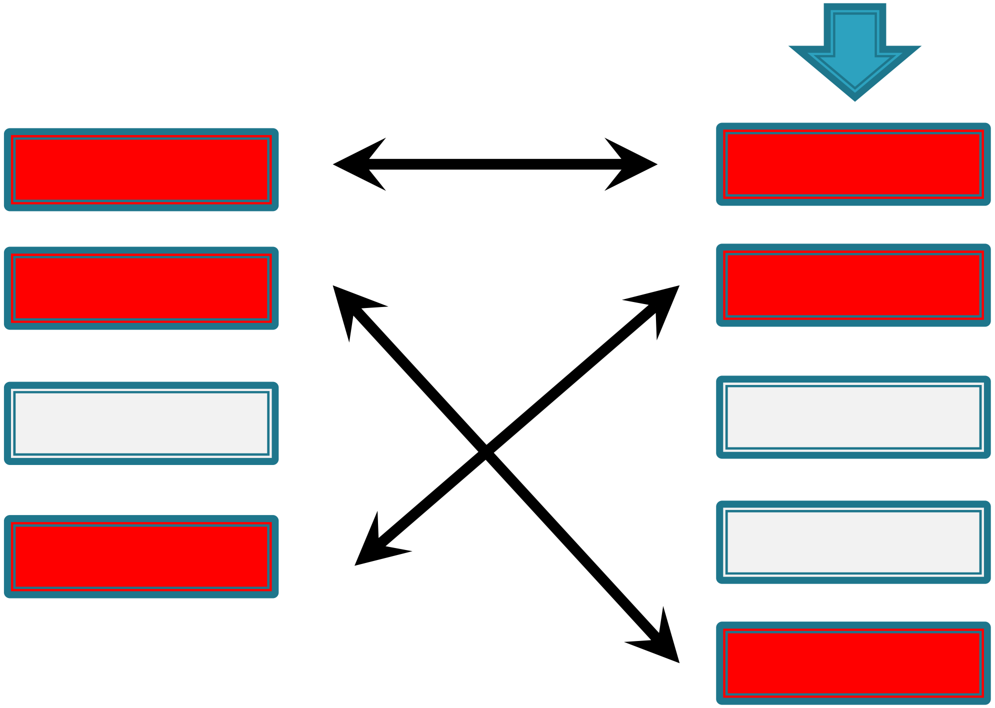


Extension

Grouping

Example:
maxGap = 1

Group **right**

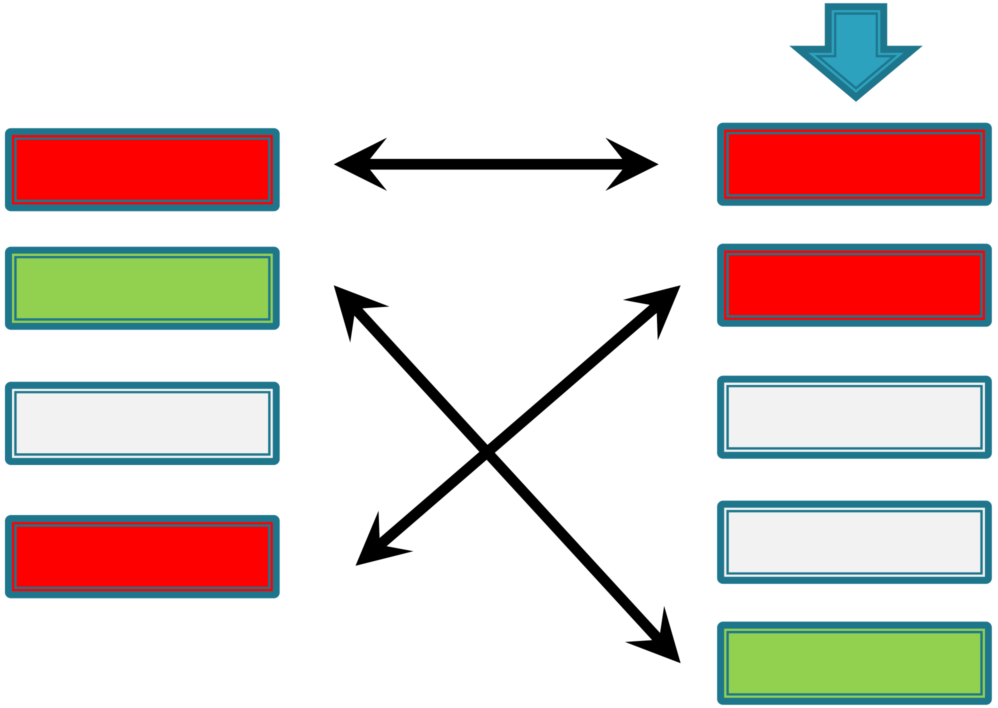


Extension

Grouping

Example:
maxGap = 1

Group **right**

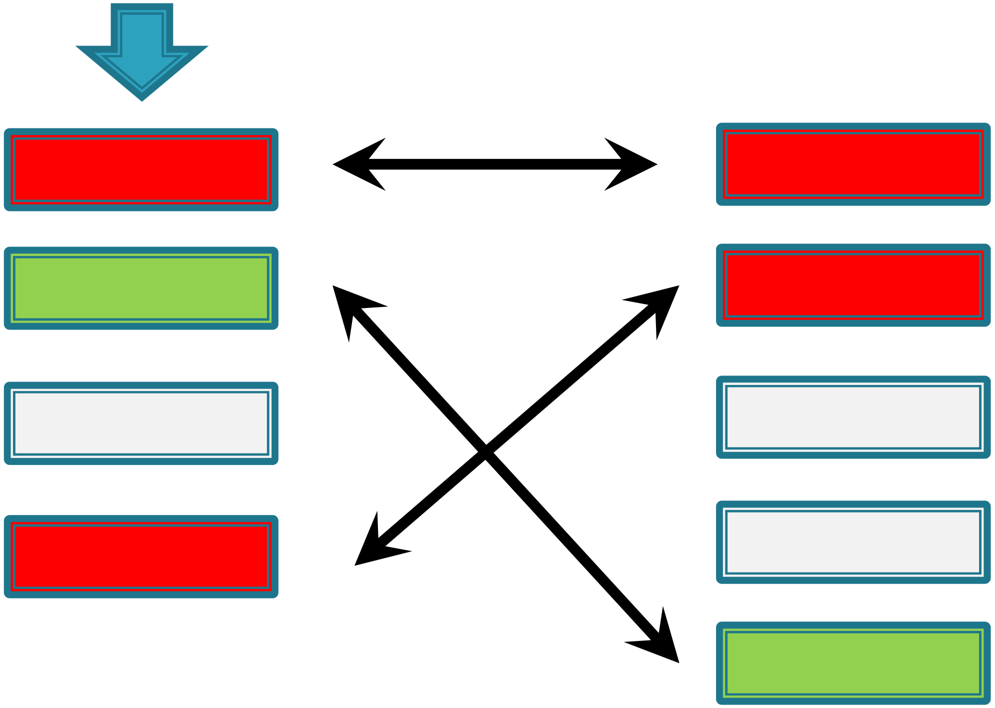


Extension

Grouping

Example:
maxGap = 1

Group **left**

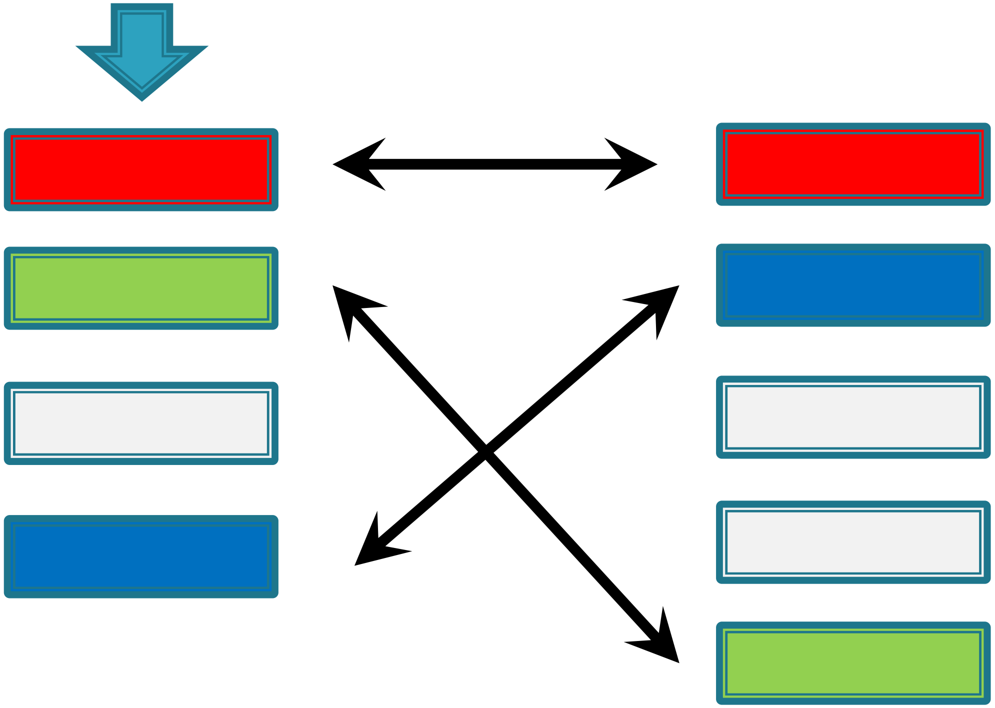


Extension

Grouping

Example:
maxGap = 1

Group **left**



Extension

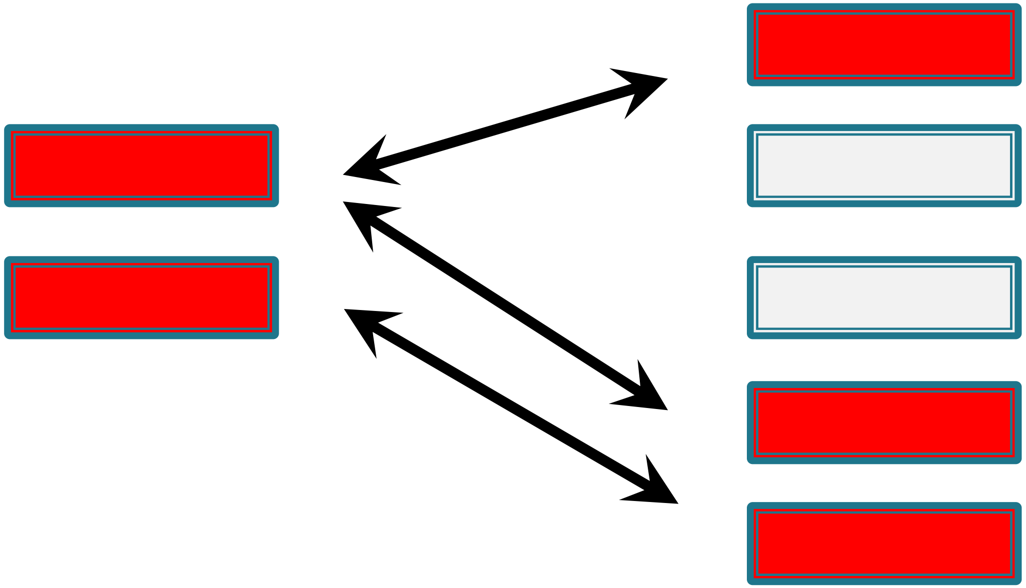
Grouping

Iteration	No plagiarism	None	Random	Translation	Summary
1	674	6803	6436	7637	3074
2	3	278	180	246	294
3	0	7	7	3	3
4	0	1	0	0	0

Extension

Validation

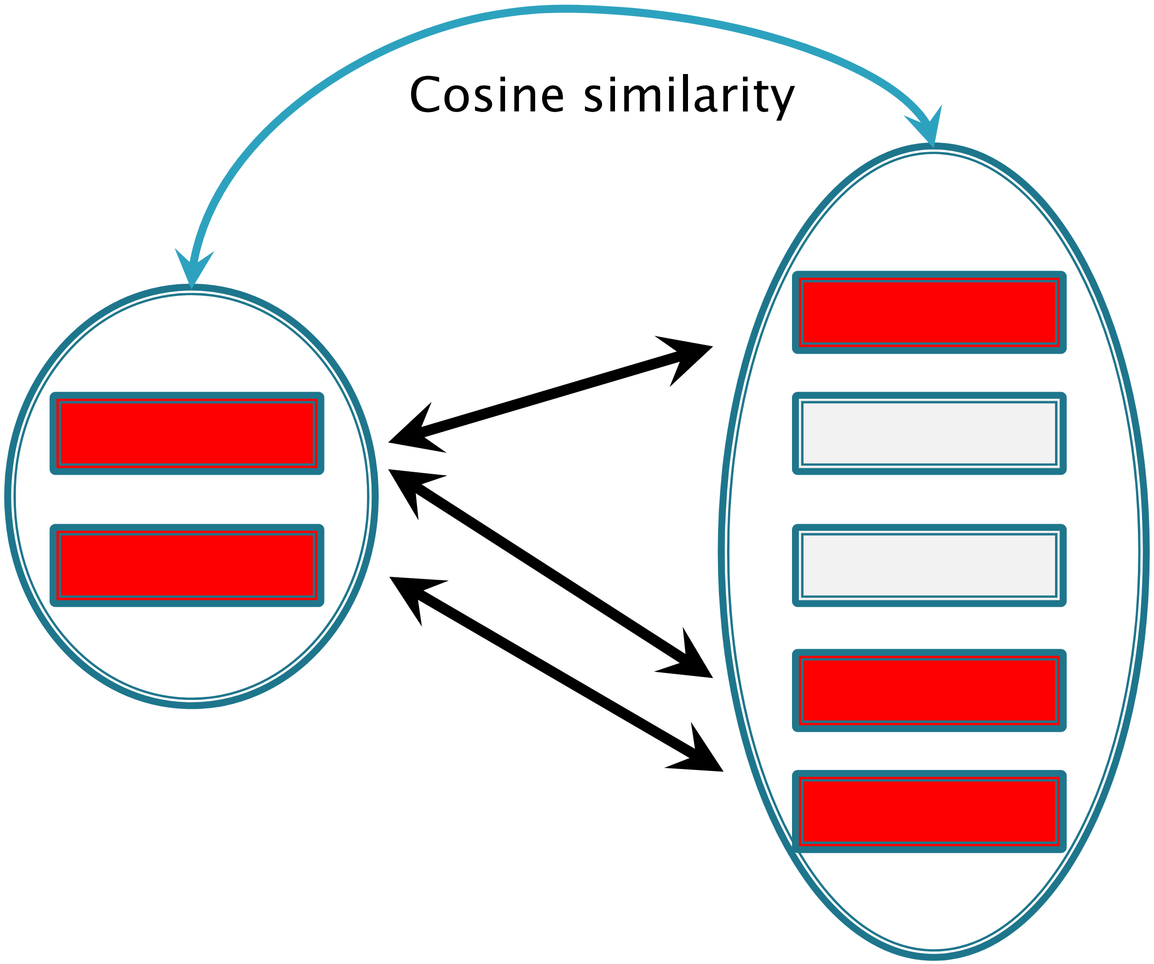
Example:
maxGap = 2



Extension

Validation

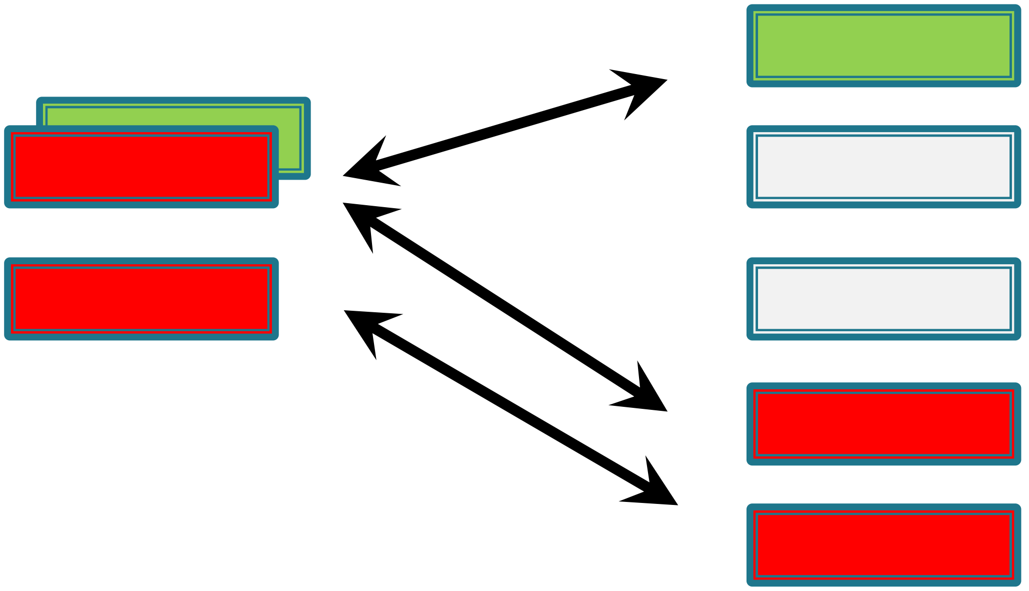
Example:
maxGap = 2



If cosine similarity < *th3*
Regroup with *maxGap – 1*

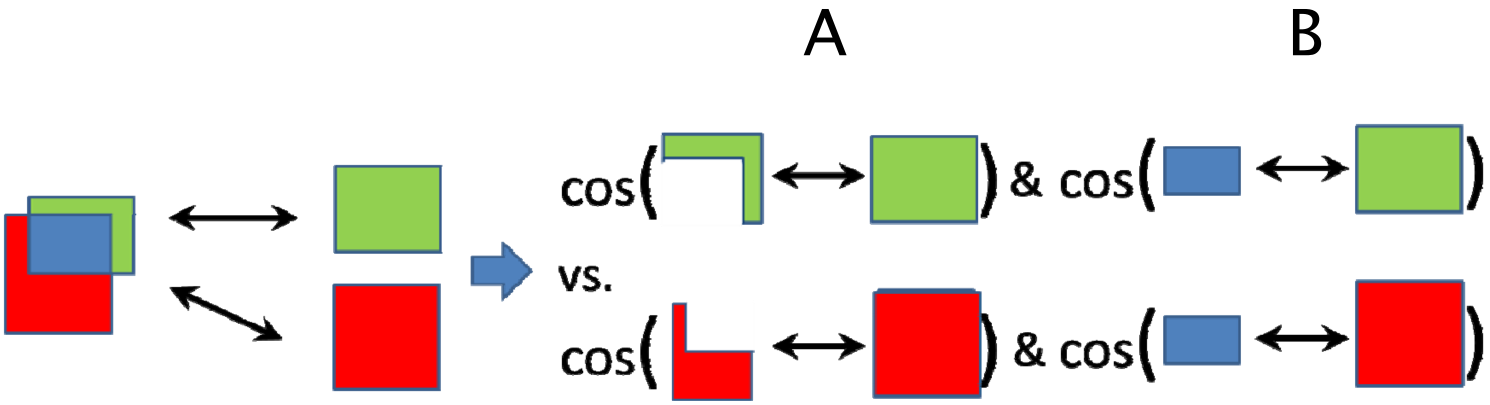
Extension

Validation



Filtering

1. Resolving overlapping



$$score = B + (1 - B) \times A,$$

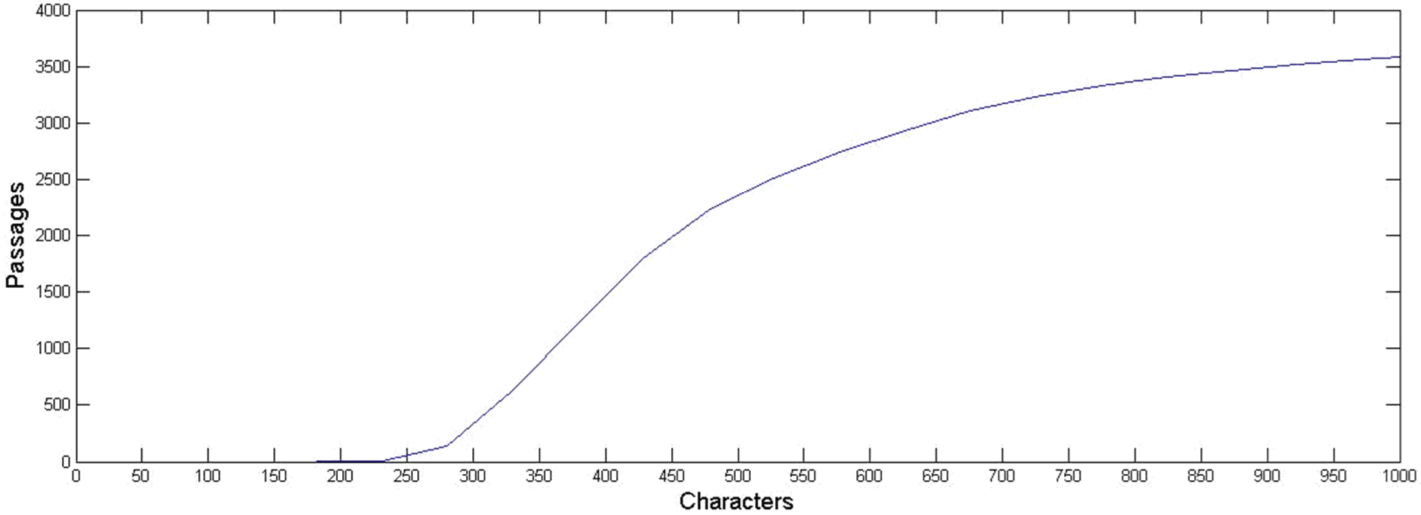
2. Removing small cases

If **n° characters** in left side **OR** rigth side < ***minPlagLength*** then the case is removed

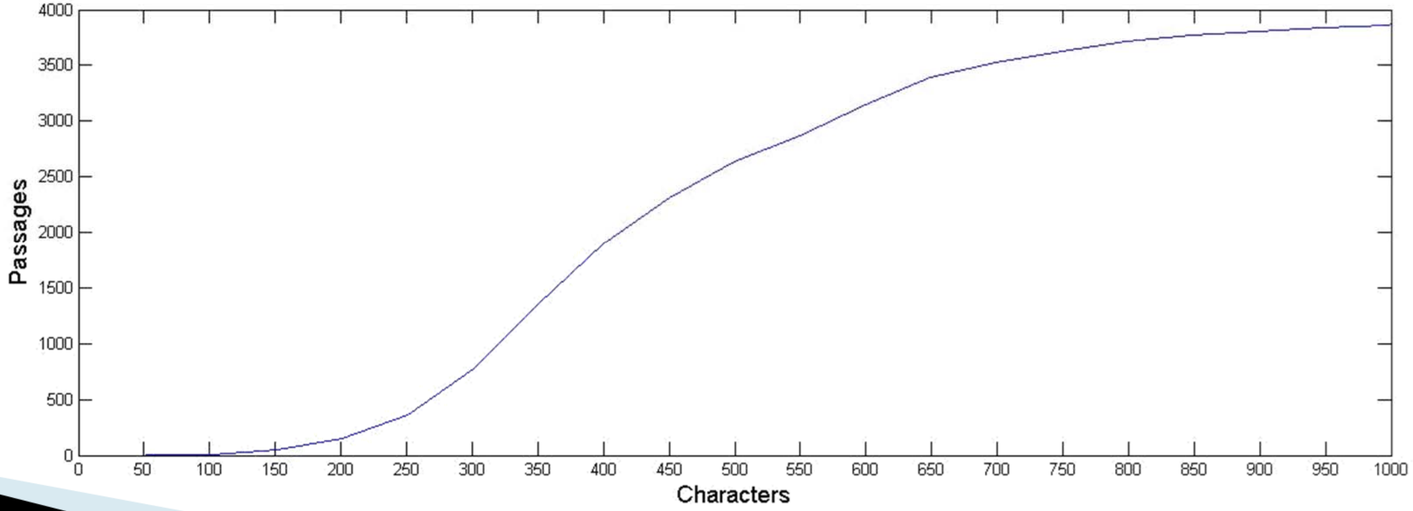
Filtering

Cumulative histogram of plagiarism cases passages

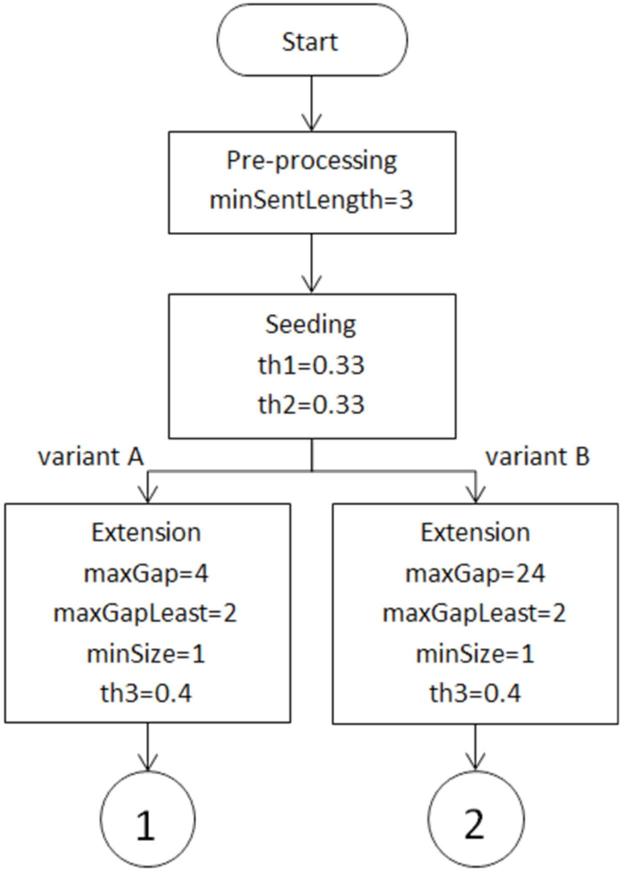
Source documents

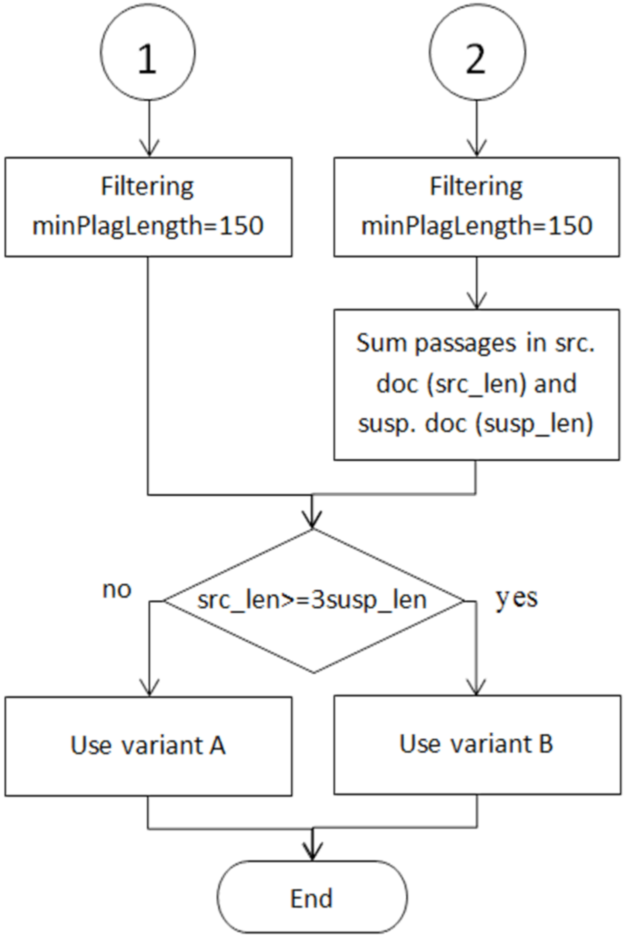


Suspicious documents



Adaptative behavior

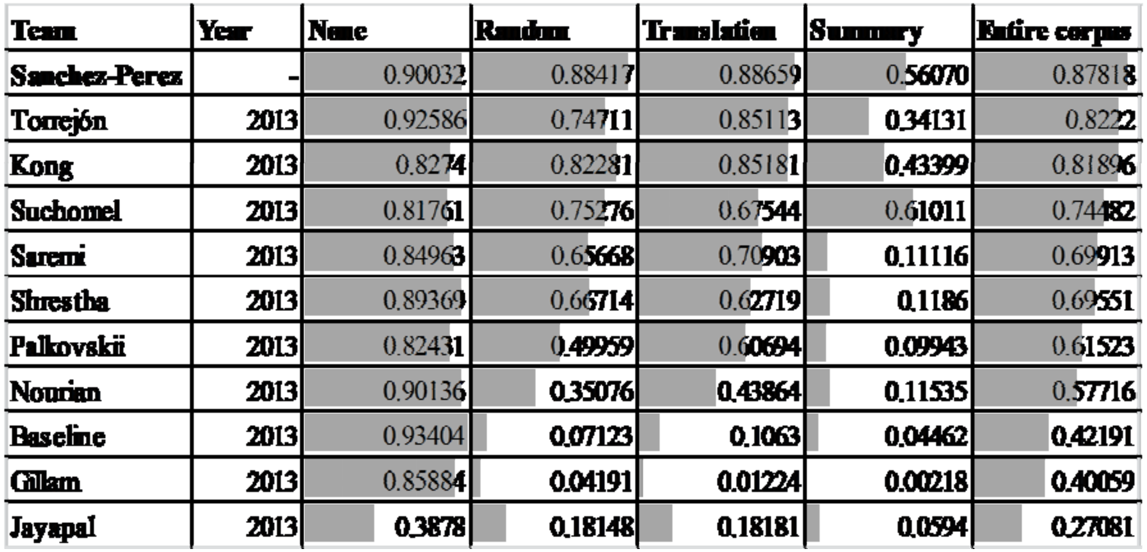




Results

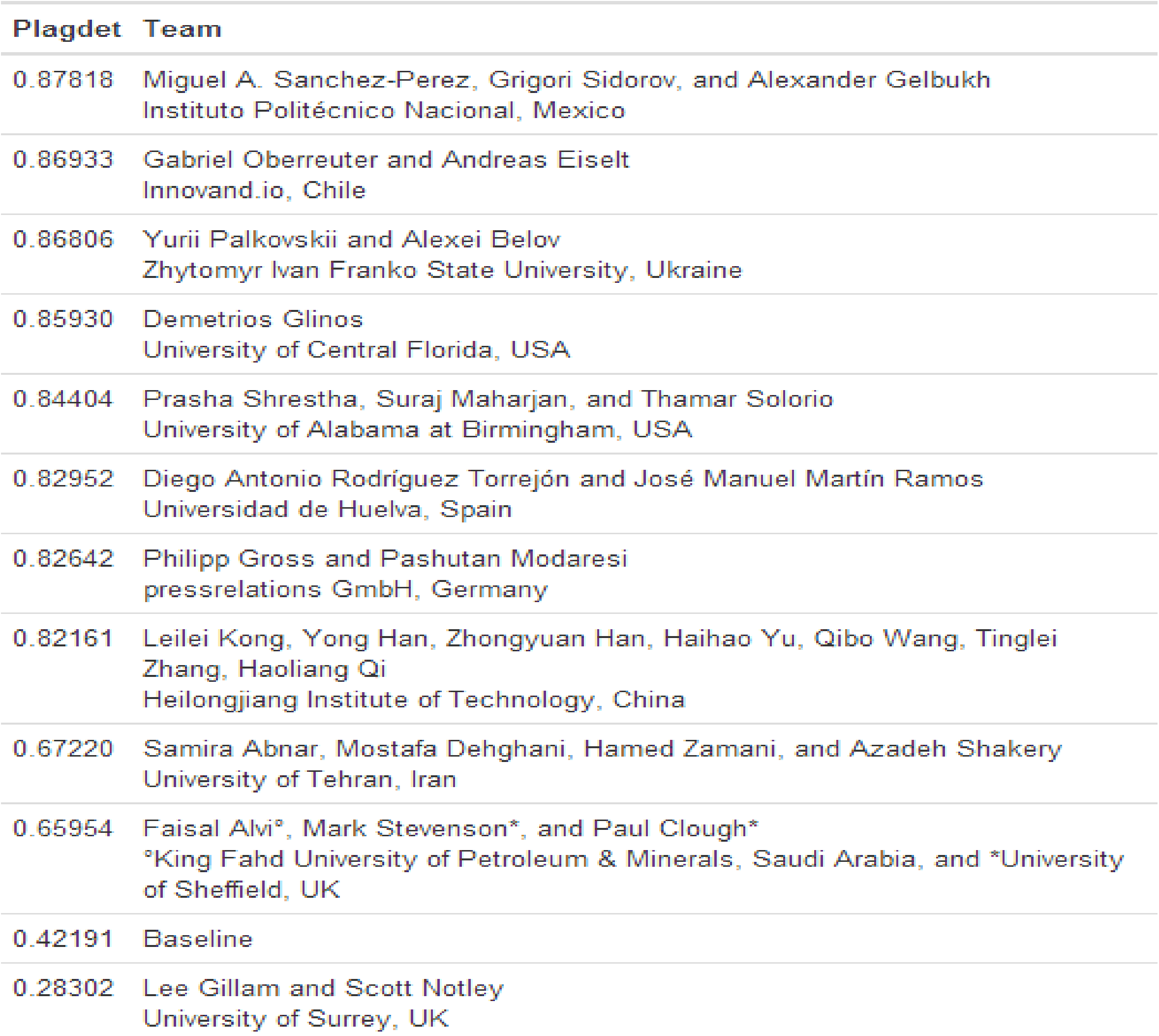
Training: PAN 2014 = PAN 2013 training corpus. **Evaluation**: PAN 2014, PAN 2013.

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

Results



Plagdet	Team
0.87818	Miguel A. Sanchez-Perez, Grigori Sidorov, and Alexander Gelbukh Instituto Politécnico Nacional, Mexico
0.86933	Gabriel Oberreuter and Andreas Eiselt Innovand.io, Chile
0.86806	Yurii Palkovskii and Alexei Belov Zhytomyr Ivan Franko State University, Ukraine
0.85930	Demetrios Glinos University of Central Florida, USA
0.84404	Prasha Shrestha, Suraj Maharjan, and Thamar Solorio University of Alabama at Birmingham, USA
0.82952	Diego Antonio Rodríguez Torrejón and José Manuel Martín Ramos Universidad de Huelva, Spain
0.82642	Philipp Gross and Pashutan Modaresi pressrelations GmbH, Germany
0.82161	Leilei Kong, Yong Han, Zhongyuan Han, Haihao Yu, Qibo Wang, Tinglei Zhang, Haoliang Qi Heilongjiang Institute of Technology, China
0.67220	Samira Abnar, Mostafa Dehghani, Hamed Zamani, and Azadeh Shakery University of Tehran, Iran
0.65954	Faisal Alvi°, Mark Stevenson*, and Paul Clough* °King Fahd University of Petroleum & Minerals, Saudi Arabia, and *University of Sheffield, UK
0.42191	Baseline
0.28302	Lee Gillam and Scott Notley University of Surrey, UK

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. Algorithm for resolution of overlapping cases by comparison of competing cases
4. Dynamic adjustment of parameters by type of obfuscation (summary vs. other types)

Future work

- Text reuse focused on paraphrase
- Soft cosine to measure similarity between features
- New strategy to resolve overlapping

Thanks!

http://www.gelbukh.com/plagiarism-detection/PAN-2014