Text Alignment Module in CoReMo 2.1 Plagiarism Detector

Diego A. Rodríguez-Torrejón[1,2]
José Manuel Martín-Ramos[1]

[1] Universidad de Huelva
jmmartin@dti.uhu.es

[2] I.E.S. José Caballero
dartsystems@gmail.com

Contextual Reference Monotony
CoReMo
The Best & Quickest Plagiarism Detection Comparator

PAN 2013
Best & Quickest Plagiarism Detection Comparator

http://coremodetector.com

The attendance of Diego A. Rodríguez is **Penalized by Junta de Andalucía Educational Admistration** :-(

Index

- **Introduction**
- Model Used in Tests
- Context Influence & Surrounding Context N-grams
- Tests Framework
- Test Results
- Conclusions

Introduction

Comparison from PAN Analysis since '10 to '12 editions shows **the mainly common limits** to any competitor proposals:

- **Short plagiarism cases** (more frequents into PAN-PC-11) are hardest to detect.

- The former effect is more accused when **crosslingual** cases happens.

- **Simulated, low and high paraphrasing** cases are much more difficult to detect.

Introduction

Hardest cases uses **methods** as words removal / replacement / inclusion, sentence reordering, similar appearance character changes…

N-gram based plagiarism detection methods are **the most commonly** used.

Synonym normalization by WordNet got best results in PAN'11, but it's **not enough**.

… **We need new ways** to solve the hardest obfuscation conditions...

Index

- Introduction
- **Model Used in Tests**
- Context Influence & Surrounding Context N-grams
- Tests Framework
- Test Results
- Conclusions

Model Used in Tests
Crosslingual CoReMo

CoReMo System has competed since PAN'10 to PAN'13 achieving the **current best Plagdet** performance.

The most significant features are the **high speed** detection and no external translation system dependence , both **ideal for intensive tests**.

For our first tests, we used our own External PDS: **Crosslingual CoReMo 1.7**, **improved by** new Surrounding Context N-grams (**SCnG**) method. However, SCnG are extensible to any N-gram based PDS (and other IR / NLP tasks).

Model Used in Tests
Crosslingual CoReMo

CoReMo Basics:

- ***Extended Contextual N-grams*** (*xCTnG*)
- ***HAIRS*** High Accuracy Inf. Retrieval System only based on n-grams **idf** for local corpora.
- ***Reference Monotony Pruning*** (*RMP*)
- ***Self-Adaptive Alignment*** parameters settings
- ***Fast Local Translation*** dictionary based
- ***External Translation possibility*** by scripting
- ***Speed Optimized*** C/C++ parallel programming

Model Used in Tests
Crosslingual CoReMo

Contextual N-grams* (**CTnG**) a way to get wide recall and lower index size in sentence order changed environment (translations, active to passive forms ...) got by:

- *Case Folding* characters normalization
- *Stopwords* and short length words removal
- *Stemming* by Porter's Stemmer Algorithm
- *N-grams Inner Sort* (after stems selection*)

* **Extended mode** includes stems skipping

Context Influence and **Extended Contextual N-Grams**

Humans can **guess a word by near context**. In 1977 [16] determined the easiest way: using surrounding context words (a group former and just later).

Usual n-grams belong to closed near context.

Surrounding Context N-grmas (SCnG) were new concept in '2012 extending CTnG by including new others made from words surrounding a discarded word.

This year **OddEven N-grams (OEnG)** are also included in the model: skip n-grams obtained from odd-only or even-only stems.

Context Influence and **Extended Context N-Grams**

Let's see the classic text example (starts from *quick*):

“~~*The*~~ ***quick brown fox jumps*** ~~*over the*~~ ***lazy dog***”

To get **direct** type xCT3G (CT3G):

1_2_3 → quick brown fox → **brown_fox_quick**

Left-hand and Right-hand Context types (SC3G):

1_2_4 → quick brown jump → **brown_jump_quick**

1_3_4 → quick fox jump → **fox_jump_quick**

Odd n-gram type (OEnG):

1_3_5 → quick fox laz → **laz_fox_quick**

Context Influence and **Extended Contextual N-Grams**

All these n-grams are indexed or compared together. No matter if matching different xCT3G types. This way gets **4 times more n-grams than words** from the same document, **increasing the matching opportunities**, but **most selectively** than using CT2G: acting as a **magnifier effect for the matching context**

Let's see matching possibilities when changes happen:

A) **Changed** word by synonym or any other cause:

*“The **quick dark fox** is **jump**ing where the dog is”*

B) **Text enriching** with new word:

“*The **quick dark brown fox**y jumps where the dog is*”

Context Influence and **Surrounding Context N-Grams**

C) **Deleted words (summary)**:

*“The **brown** one **jump**s over the **dog**”*

D) **Translation Errors, writing faults, incorrect term disambiguation**: will match as in A case.

The biggest matching quantity enables **lowest chunk length** to **tackle shortest plagiarism cases**, without granularity sacrifice or using thesaurus.

xCT3G will get almost the “good” matching opportunities of CT2G, and almost the exceptional precision of CT3G, but improved reliability by its biggest amount, **almost without chance noisy matches**.

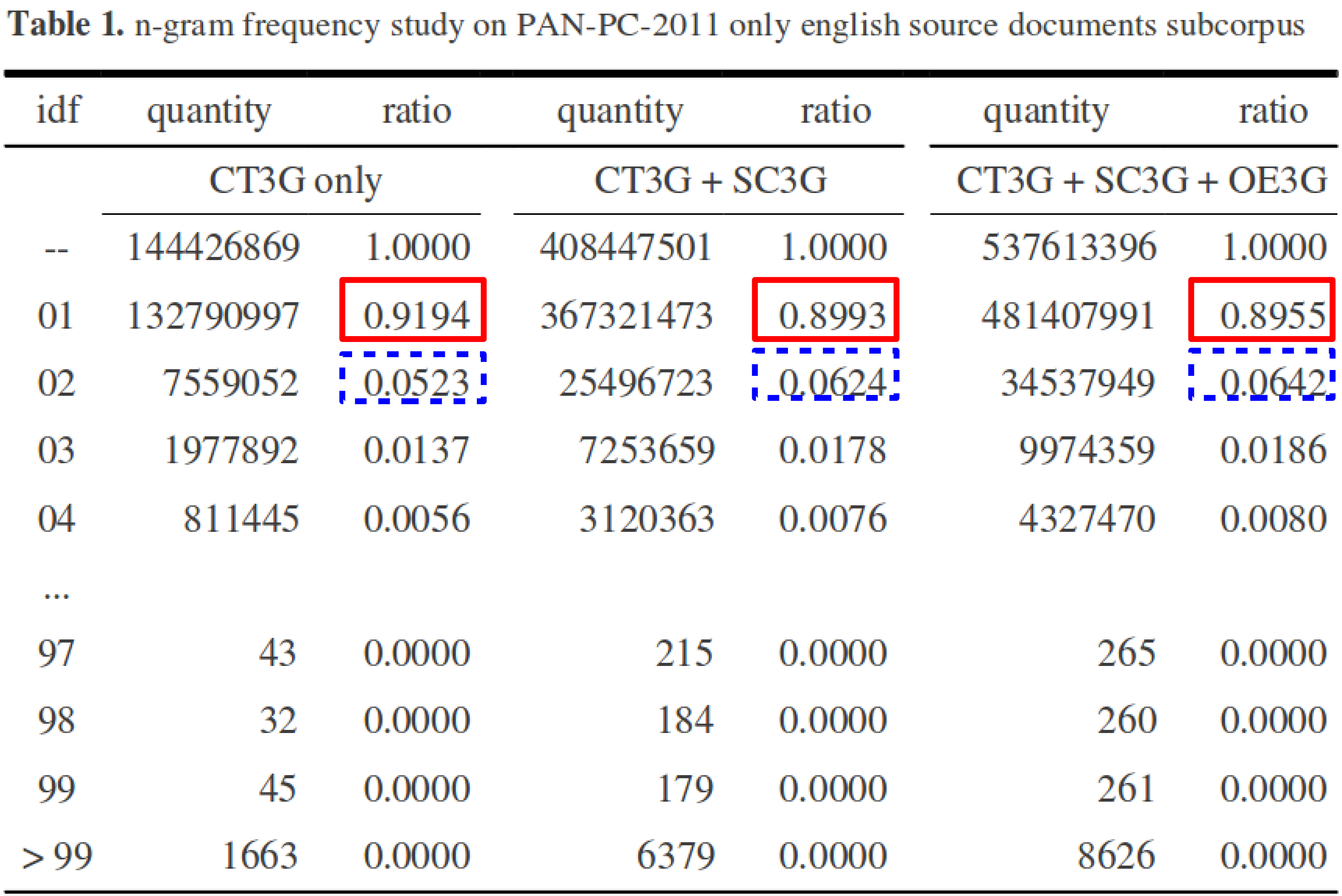


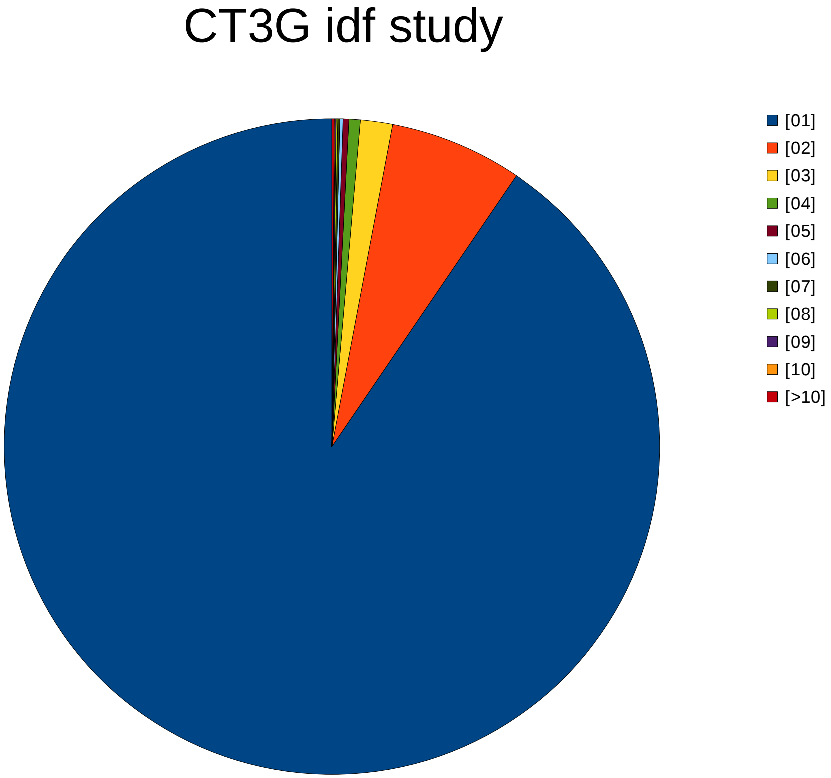
Table 1. n-gram frequency study on PAN-PC-2011 only english source documents subcorpus

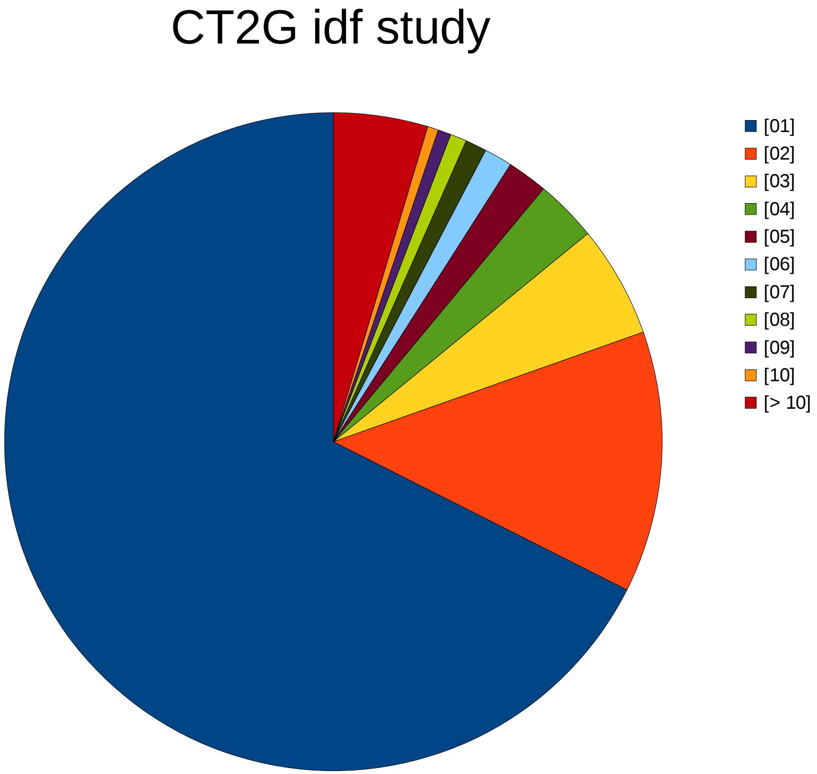
idf	quantity	ratio	quantity	ratio	quantity	ratio
	CT3G only		CT3G + SC3G		CT3G + SC3G + OE3G	
--	144426869	1.0000	408447501	1.0000	537613396	1.0000
01	132790997	0.9194	367321473	0.8993	481407991	0.8955
02	7559052	0.0523	25496723	0.0624	34537949	0.0642
03	1977892	0.0137	7253659	0.0178	9974359	0.0186
04	811445	0.0056	3120363	0.0076	4327470	0.0080
...						
97	43	0.0000	215	0.0000	265	0.0000
98	32	0.0000	184	0.0000	260	0.0000
99	45	0.0000	179	0.0000	261	0.0000
> 99	1663	0.0000	6379	0.0000	8626	0.0000

About 12.000 docs (1.5 Gbytes plain text)

Model Used in Tests
Crosslingual CoReMo

HAIRS is based in Inverse Document Frequency
CTnG study. The best results are got by CT3G





Model Used in Tests
Crosslingual CoReMo

Reference Monotony Prune strategy: **discard matching if not happening monotonously**.

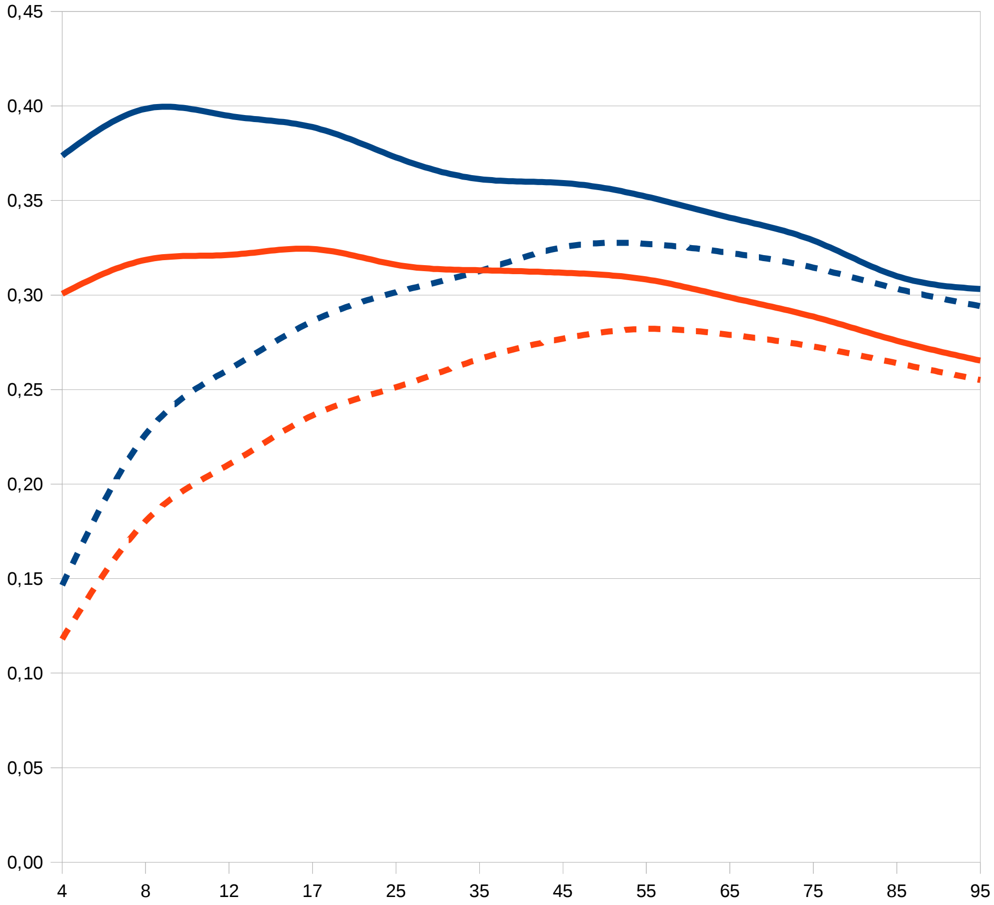
Used in several steps to gets fastest runtime, by discarding noisy matching, reducing documents pairs, or complete document comparison even.

- i.e.: Suspicious documents are divided in equal N-gram length chunks. ***HAIRS*** will get one only document for every chunk

73	-1	6	49	11	-1	31	91	91	91	91	91	6	92	5	7	98	91	-1	-1

Plagdet / chunk length

CoReMo 1.6 version only

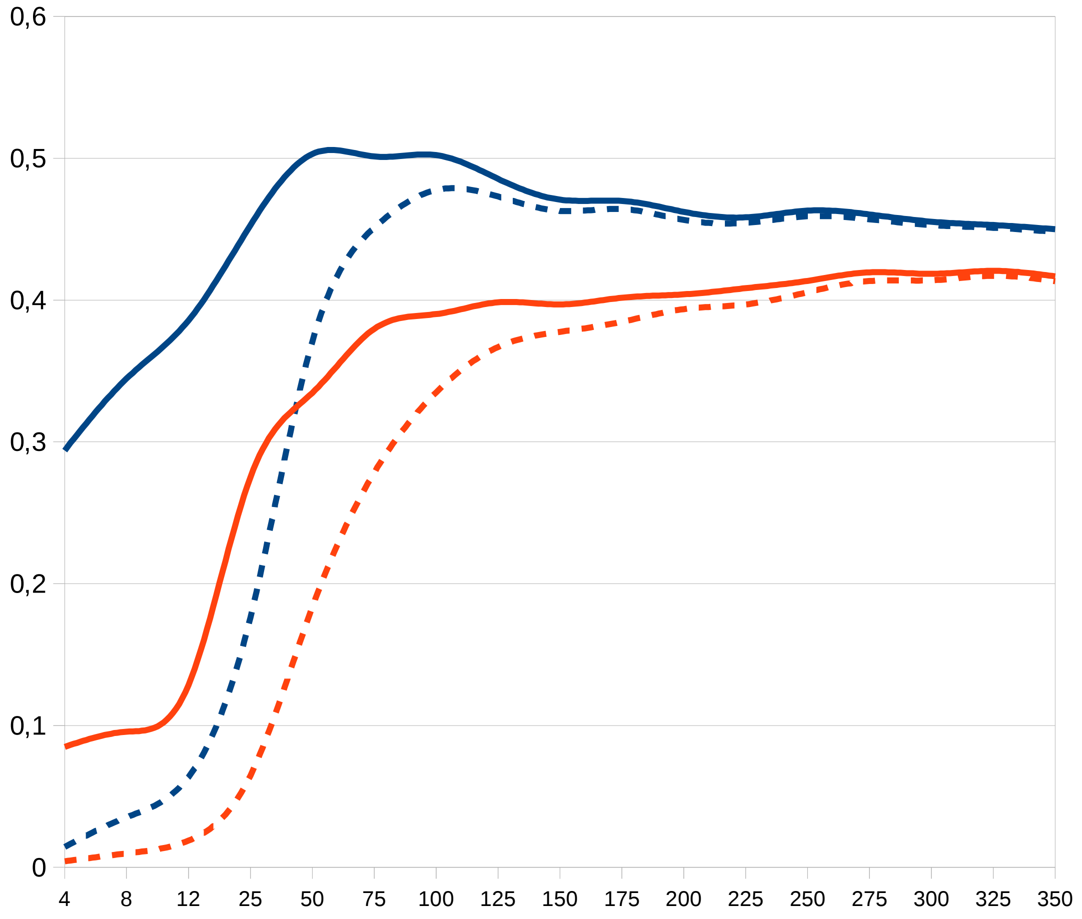


PAN-PC-2011

monolingual analysis only

Plagdet / chunk length

CoReMo 1.6 version only

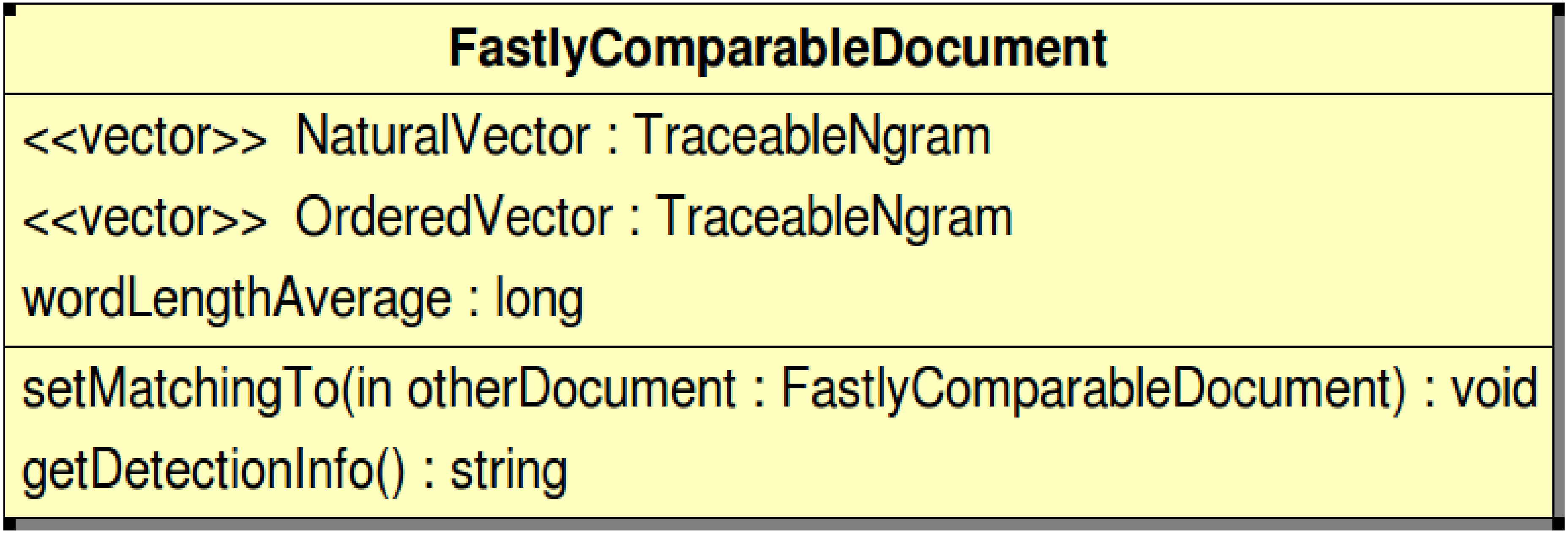


PAN-PC-2011

Translated cases only

Text Alignment Module

- Every document is modelled having two xCTnG reference lists: naturally ordered and alfabetically ordered ones.



FastlyComparableDocument

<<vector>> NaturalVector : TraceableNgram

<<vector>> OrderedVector : TraceableNgram

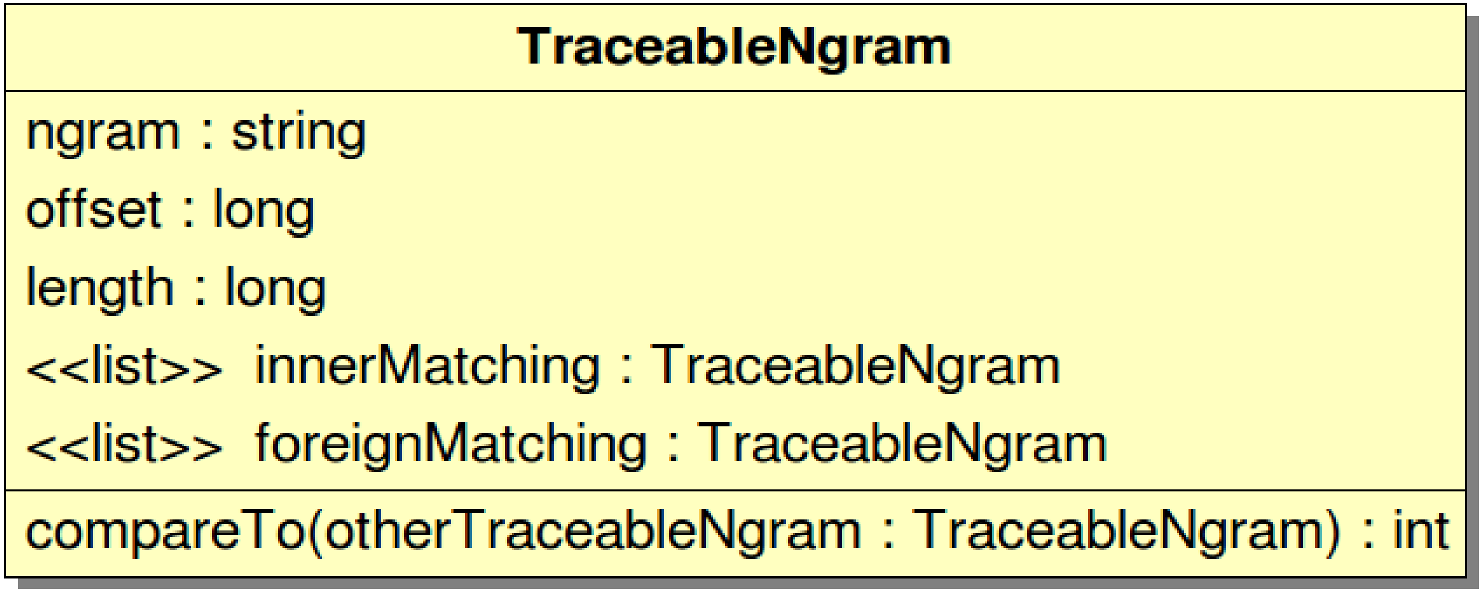
wordLengthAverage : long

setMatchingTo(in otherDocument : FastlyComparableDocument) : void

getDetectionInfo() : string

Text Alignment

- When internall order is arranged, internal matching is registered for each xCTnG as a references list.
- The document’s matching cases are got from the ordered lists by a merge-sort modified algorithm, interchanging the *references* information when matching happens.



TraceableNgram

ngram : string

offset : long

length : long

<<list>> innerMatching : TraceableNgram

<<list>> foreignMatching : TraceableNgram

compareTo(otherTraceableNgram : TraceableNgram) : int

Text Alignment

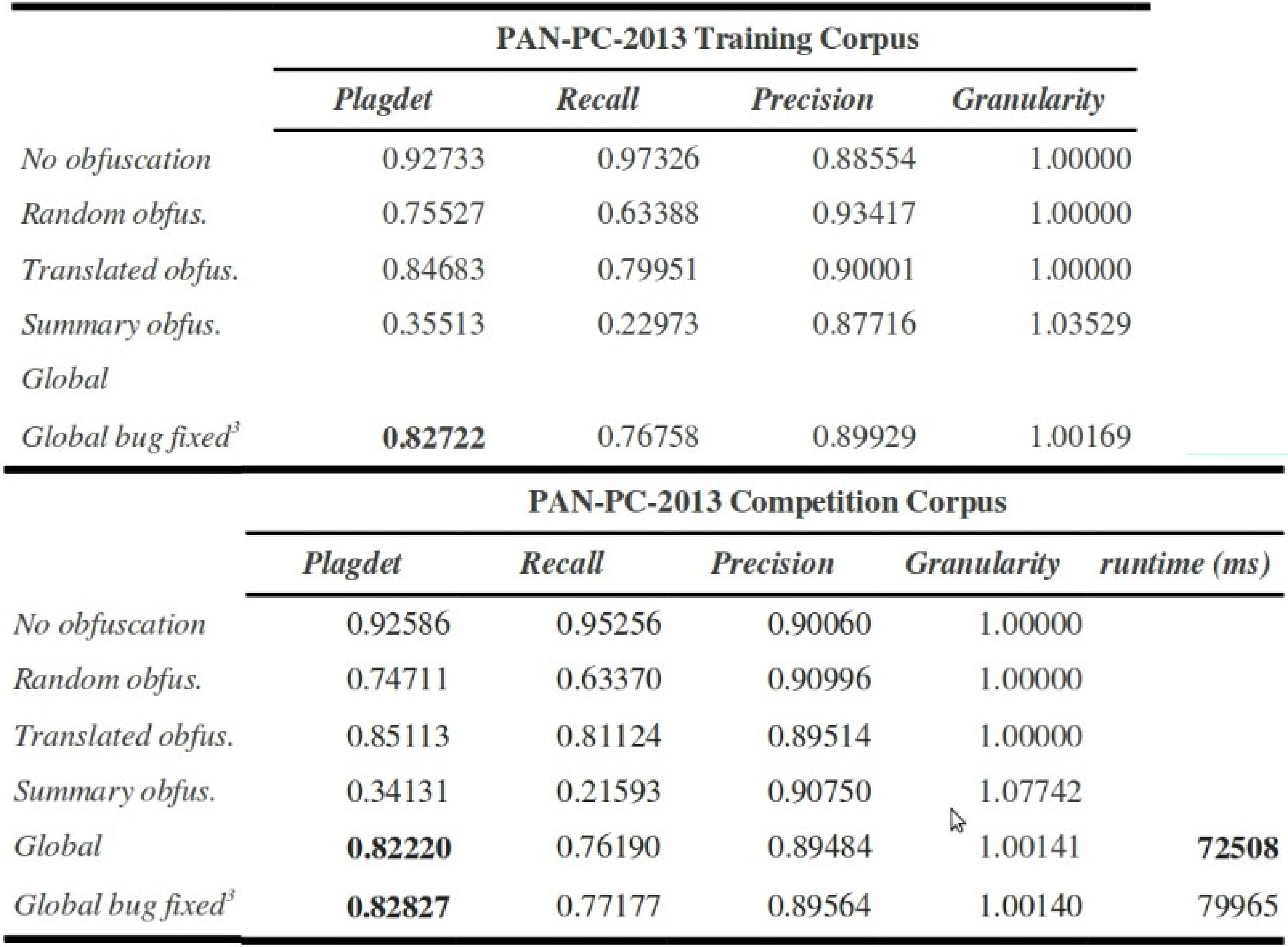
- Reliable matching are those with **foreign dtf = 1** and **positionally closed to another** reliable one in both suspicious and source documents.
- When the distance from last reliable match is over the chunk length, the fragment detection finishes, but only will be **registered if it's larger than a chunk** between the first and the last matches
- The direct detections (**seeds**) are good, but a bit fragmented. The granularity filter process will **join overlapped or closed detections** in both documents. We used “only” **4.000 characters distance** for this step.
- **Distances** are taken **in n-grams** for suspicious fragments **and in characters** for source ones.

Text Alignment

- These **distances** are got **from** the **chunk-legth** parameter, and also **combined** with word **length average** obtained from the source document.
- In order to optimize the tuning for the best performance in the most difficult plagiarism types (summarized) **avoiding false positives when no plagiarism cases** happens, the **chunk length** (*cl*) to different regions **depends of the foreign matching rate** (*emr*) for both documents:

base case: *cl* = 8 * multiplicty factor (4)
emr1 > 4% & *emr2* < 15% → *cl* = 3 *cl* / 7
emr1 > 30% & *emr2* >= 15% → *cl* = 2 *cl* / 3

Test Results



PAN-PC-2013 Training Corpus

	Plagdet	*Recall*	*Precision*	*Granularity*
No obfuscation	0.92733	0.97326	0.88554	1.00000
Random obfus.	0.75527	0.63388	0.93417	1.00000
Translated obfus.	0.84683	0.79951	0.90001	1.00000
Summary obfus.	0.35513	0.22973	0.87716	1.03529
Global				
Global bug fixed[3]	**0.82722**	0.76758	0.89929	1.00169

PAN-PC-2013 Competition Corpus

	Plagdet	*Recall*	*Precision*	*Granularity*	*runtime (ms)*
No obfuscation	0.92586	0.95256	0.90060	1.00000	
Random obfus.	0.74711	0.63370	0.90996	1.00000	
Translated obfus.	0.85113	0.81124	0.89514	1.00000	
Summary obfus.	0.34131	0.21593	0.90750	1.07742	
Global	**0.82220**	0.76190	0.89484	1.00141	**72508**
Global bug fixed[3]	**0.82827**	0.77177	0.89564	1.00140	79965

Test Results

- Most significant improvement are due to SCnG
- Including OEnG and self-tuning improves **seeds for precision and Recall, enabling shorter GF**.
- Granularity Filter distance is now 1/20th than '12
- A **late corrected bug**, achieves a even best score:

PlagDet,	*Recall,*	*Precision,*	*Granularity,*	*Runtime*
0.82827	*0.77177*	*0.89564*	*1.00140*	*79965ms*

- Single core VMs **Runtime** don't shows **real** analysis power: CoReMo is now **multicore optimized**, and we can get same analysis in **only 4,5 seconds** using 8 cores AMD FX8120 / 4GHz + SSD drive.

Conclusions

- **xCTnG** gets **improved detection** when **harder obfuscation or crosslingual** conditions, getting also lower length plagiarism detection.
- **xCTnG** mode gets hoped **CT2G *Recall*** and practical CT3G ***Precision. More and Most Reliable matching Seeds.***
- **Defragmentation** filter gets **improved scores** at lower detection chunk length. Must be used **cautiously** however.
- **xCTnG** possibilities open to **other IR/NLP** tasks.

Future Jobs

- **Improving self-tunig** by studing matching rates distributions, but for chunk length and filter distance also.
- **Improving filtering** by using information of unconnected matches previously discarded.
- **Testing** the possible positive influence of using **Wordnet synsets** reductions, as proposed in PAN'10 and successfully exploded in PAN'11 by J. Grman and R. Ravas.

Acknowledges

- Thanks to the PAN group and all the teams for keeping so interesting challenge every year.
- None entity has supported the Diego Rodríguez job or attendance. It's company (Andalusian Educational Administration) will cut off its salary for the days attending to CLEF2013 **: (**
- To my family, who has enforced me to be here, but its economy (and stability) can not support “Vicious” Research: it has been my …

  ... last-time **: (???**

End … or Begining?

… But CoReMo will have an opportunity to go on improving only if demonstrates self-financial capability as non-free web services, hoped to start next month and get fully operational about 2014 mid January.

http://www.coremodetector.com

THANKS FOR YOUR ATTENTION

We can improve this slide-show

diego@dartsystems.es

dartsystems@gmail.com

jmmartin@dti.uhu.es

info @ coremodetector.com

References (1)

1. Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. Overview of the 1st International Competition on Plagiarism Detection. In [16]
2. Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. Overview of the 2nd International Competition on Plagiarism Detection. In [24]
3. Jan Kasprzak and Michal Brandejs. Improving the Reliability of the Plagiarism Detection System: Lab Report for *PAN* at CLEF 2010. In Braschler et al. [24]
4. Du Zou, Wei-Jiang Long, and Ling Zhang. A Cluster-Based Plagiarism Detection Method: Lab Report for *PAN* at CLEF 2010. In Braschler et al. [24]
5. Markus Muhr, Roman Kern, Mario Zechner, and Michael Granitzer. External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System: Lab Report for *PAN* at CLEF 2010. In Braschler et al. [24].
6. Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. Overview of the 3rd International Competition on Plagiarism Detection. In [25]
7. Ján Grman and Rudolf Ravas. Improved Implementation for Finding Text Similarities in Large Collections of Data: Notebook for *PAN* at CLEF 2011. In [25]
8. Cristian Grozea and Marius Popescu. The Encoplot Similarity Measure for Automatic Detection of Plagiarism: Notebook for *PAN* at CLEF 2011. In [25]
9. Gabriel Oberreuter, Gaston L’Huillier, Sebastián A. Ríos, and Juan D. Velásquez. Approaches for Intrinsic and External Plagiarism Detection: Notebook for *PAN* at CLEF 2011. In [25]

References (2)

10. Steven Burrows, Martin Potthast, and Benno Stein. Paraphrase Acquisition via Crowdsourcing and Machine Learning. Transactions on Intelligent Systems and Technology (ACM TIST) (to appear), 2012.

11. Diego Antonio Rodríguez Torrejón and José Manuel Martín Ramos. *CoReMo* System (Contextual Reference Monotony) A Fast, Low Cost and High Performance Plagiarism Analyzer System: Lab Report for *PAN* at CLEF 2010. In Braschler et al. [24]

12. Diego A. Rodríguez Torrejón and José Manuel Martín Ramos. Crosslingual *CoReMo* System: Notebook for *PAN* at CLEF 2011. In [25]

13. Palkovskii, Yurii Anatol'yevich, Alexei Vitalievich Belov, and Irina Alexandrovna Muzika. "Counter plagiarism detection software" and "Counter counter plagiarism detection" methods - 2009. Submission to the 1st International Competition on Plagiarism Detection. From the Zhytomyr State University, Ukraine.

14. Diego A. Rodríguez Torrejon y José Manuel Martín Ramos. (2010b). Detección de plagio en documentos. Sistema externo monolingüe de altas prestaciones basado en n-gramas contextuales. Procesamiento del Lenguaje Natural, 45:49–57.

15. Rodríguez-Torrejón D.A.: Detección de plagio en documentos. Propuesta de sistema externo monolingüe de altas prestaciones basada en n-gramas. Master Dissertation – Universidad de Huelva (2009)

16. David C. Rubin. The effectiveness of context before, after and around a missing word. In Perceptions & Phychophysics 1976, 19(2), 214-216.

17. Alberto Barrón-Cedeño, Paolo Rosso. On Automatic Plagiarism Detection based on n-grams Comparison. In: Boughanem et al. (Eds.) ECIR 2009, LNCS 5478, pp. 696-700, Springer-Verlag Berlin Heidelberg (2009)

References (3)

18. Alberto Barrón-Cedeño, Martin Potthast, Paolo Rosso, Benno Stein, and Andreas Eiselt. Corpus and Evaluation Measures for Automatic Plagiarism Detection. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias, editors, 7th Conference on International Language Resources and Evaluation (LREC 10), May 10. European Language Resources Association (ELRA).

19. Meyer zu Eissen, Sven and Benno Stein. 2006. Intrinsic plagiarism detection. In Mounia Lalmas, Andy MacFarlane, Stefan M. Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, editors, Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006),London, Volume 3936 of Lecture Notes in Computer Science, pages 565–569. Springer.

20. Benno Stein, Nedim Lipka, and Peter Prettenhofer. Intrinsic Plagiarism Analysis. Language Resources and Evaluation (LRE), 45 (1): 63-82, 11.

21. Jan Kasprzak and Michal Brandejs. Improving the Reliaility of the Plagiarism DetectionSystem: Lab Report for *PAN* at CLEF 10. In Braschler et al. [13].

22. Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. An Evaluation Framework for Plagiarism Detection. In 23rd International Conference on Computational Linguistics (COLING 10), August 10. Association for Computational Linguistic

23. Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, Paolo Rosso. Cross-Language Plagiarism Detection. Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis, vol. 45, num. 1. DOI: 10.1007/s10579-009-9114-z, 11

24. Benno Stein, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors, SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (*PAN* 09), pages 1-9, September 2009. CEUR-WS.org. ISSN 1613-0073

Universidad de Huelva

Seeds Comparison

JUNTA DE ANDALUCIA

I.E.S. José Caballero

Torrejon13/PAN13 training

Seeds
Plagdet Score 0.77915100343
Recall 0.750258541782
Precision 0.923206830702
Granularity 1.08845070423

Torrejon12/PAN13 training

Seeds
Plagdet Score 0.656719889391
Recall 0.670569425935
Precision 0.922594444295
Granularity 1.26988085342

Torrejon13 /PAN12 Compet.
(locally translated)

seeds
Plagdet Score 0.408856888467
Recall 0.441193683693
Precision 0.856176743299
Granularity 1.6837565884

Torrejon12 / PAN12 Compet.
(locally translated)

seeds
Plagdet Score 0.346070995453
Recall 0.419077935863
Precision 0.844858063703
Granularity 2.07139364303