CoReMo 2.3 Plagiarism Detector Text Alignment Module
Notebook for PAN at CLEF 2014

Diego Antonio Rodríguez Torrejón, José Manuel Martín Ramos

Departamento de Tecnologías de la Información - Universidad de Huelva (Spain)
dartsystems@gmail.com, jmmartin@dti.uhu.com

Abstract. In this paper, the basics of the three tuning approaches of the evolving CoReMo Plagiarism Detector are shown, focused for the Text Alignment task. In the last PAN edition, it was observed that the different corpora could condition the necessary tuning, and the results using an overfitted tuning from a different corpus could be far from the expected ones. This year's goal has been to find the way to get the system could be self-tuned, looking for improving the performance of any fixed parameter tuning, and to be very closed to the over-fitted performance for any corpus. All of these tuning approaches have a high Plagdet performance for any corpus, but it's intended to show the different advances effect on each corpus and for all the years corpora. They include new features for parameters self-tuning, based on the size and the ratio of the compared documents. For the competition, our choice was based on the most constant detection quality tuning approach when any condition (called WideTuning).

1 Introduction

CoReMo (Contextual Reference Monotony) is an external Plagiarism Detector approach which has taken part in the PAN [1] competition since the 2010 edition, highlighting its efficacy (best ranked in the Text Alignment task in PAN 2013) and its efficiency, especially for large documents, as in PAN 2012, which proved to be at least 10 times faster than any rival approach. These amazing benefits were achieved through an innovative different type n-grams combining model, called Extended Contextual N-grams model (XCTnG) [2][3], and a self-tuning feature based on the analysis of the optimal tuning achieved when training by different corpora and sub-corpora, having rules only based on the n-grams matching ratio.

In a nutshell, the Extended Contextual N-Grams model used by CoReMo is the result of removing the stop-words and short-length words before a stems extraction. The stems are then combined to get 3-grams, by direct neighborhood or by skipping once or twice at most one stem. This gets four different n-grams beginning from each word stem. Finally, the inner stems combinations are alphabetically ordered. This process improves the matching probability when the plagiarism obfuscation happens.

To filter the chance matching, then the *Reference Monotony* is used, which rejects the matching n-gram when it happens far from any minimum length matching group. The minimum group length and maximum distance are rule based since 2.1 version.

The detection seeds are finally bound when their distance is below a fixed threshold, but currently, in 2.3 version this threshold is rule based too.

In the Overview of the 5th International Competition of Plagiarism Detection [4] PAN 2013, it led to the overall conclusion that almost all new 2013 algorithms versions had worse outcome than their 2012 version, or at least their behavior was not the expected for both years corpora. It was also felt, that the difficulty of detecting plagiarism in the 2012 corpus was higher than in the 2013 one, because in all the cases it was achieved an improved Plagdet than for 2012 corpus, as well as much lower required analysis time.

For CoReMo 2.1 (PAN2013) compared to CoReMo 1.9 (PAN2012), although the result was better in the corpus 2013, it was not the same case for the 2012 corpus. Although the Overview's authors believed an overfitting of the new system for the corpus of 2013 as the possible cause, the CoReMo authors considered to be the 1.9 version the really overfitted for the 2012 corpus, having an illogical but optimal granularity reduction binding distance of up to 80.000 characters, twenty times higher than the one used in version 2.1, being this last one large enough to achieve very similar but more accurate results thanks to the highest quality and number of its detection seeds.

However, the facts that the detection in the corpus PAN2012 could and have to be improved, and that its best fitting was far from the necessary one for the PAN2013 corpora, are leaving the door open to study the different features between the corpora, which may be the key to new self-tuning methods to achieve acceptable results for all the corpora together, or even to improve them for each case.

2 Analysis of the Different Corpora

To establish the way to set the automatic tuning of the granularity filter's binding distance (the greatest tuning parameter deviation), distinctive features were seek between 2012 and 2013 corpora that allowed how to decide it. The data used in the analysis can be found in Table 1.

Table 1. Data Analysis of file sizes in different corpora

PAN Corpus	Size	Suspicious	Sources	Pairs	Susp. Average	Src. Average	Susp/Src (avg)	Susp. Median	Src. Median	Susp/Src (median)
2012 test	656MB	3000	3500	3000	84499	115029	0,73	22867	12183	1,88
2012 train	1,35 GB	1804	4210	6000	165379	250102	0,66	55175	112334	0,49
2013 test	29,8 MB	1826	3169	5185	8762	4366	2,01	6269	2454	2,55
2013 train	30,4 MB	1827	3230	5185	8710	4487	1,94	6262	2382	2,63

3 Training Stage

By the analysis of the differences between the corpora used to train the system, possibly useful for tuning, it was essentially seen that the average size of the documents in PAN2013 was much lower than in PAN2012 and that the ratio between the sizes of suspicious documents and source also changed significantly for each corpus, so it was suspected its possible influence for the different optimal setting of CoReMo for each one.

For that purpose, the PAN12 and PAN13 competition and training corpora were analyzed by CoReMo 2.1, using different parameters for the chunk length and granularity filter binding distance to find the optimal settings for each one.

The CoReMo efficiency, ready for multi-core systems, allowed to arrange hundreds of runs with different configurations on these corpora, as it only takes 4.5 seconds to analyze any of the PAN13 corpora, and about 45 seconds for the PAN12 ones.

The first and obvious possible improvement lies in tuning the chunk length proportionally to the size of the documents analyzed instead of a fixed length. This adjustment improved the result with the corpus of PAN2012, though curiously it had slightly worse outcome in PAN2013.

The second improvement was to propose different granularity filter binding distances for the suspicious and source documents. This step actually achieved improvements for all the corpora, although the optimum settings for the 2012 and 2013 corpora were quite different. By this way the best optimal setting was achieved for the PAN2013 competition corpus, achieving a 0.8349 Plagdet score when the distance filter is 8% of the suspicious document size, without affecting the distance between the source detections (we used the 100%). In the 2012 corpus, the optimum is achieved by binding distances of 30% of the the suspicious document size and 50% of the source document size.

The third improvement was to apply rules based on the ratio between suspicious and source document sizes, to decide when applying the above optimal settings, which improved the performance achieved for PAN2012 significantly (0.6905) at the expense of a slight decrease in the performance for PAN2013 to 0.8319.

Finally, noting that despite having very similar characteristics in terms of documents and corpus size, the optimal chunk size settings were slightly different for each PAN corpus, similar relationship rules between document sizes were arranged to tune the chunk size, improving the results obtained for all the corpus (i.e. 0.6930 in PAN2012) but for the PAN2013 competition one, which had shown again another slight decrease, until 0.8995.

The Table 2 describes the rules to define these self-tuning. The authors think that these rules can still be more optimized by any kind of interpolation.

Table 2. Self-adjusting Rules

parameter	susp/sour < 1.6	1.6 > = susp/sour < 3.0	3.0 < = susp /sour
Susp. filter distance	susp file length 8%	30%suspfile length	
Sour. filter distance	100% source file length	50%source file length	
chunk_length	chunk_length - -	nominal chunk_length	chunk_length + +

In Fig. 1, it can be seen the effect of each analyzed possible optimization on its own corpus and the other ones. The left hand side is for PAN2013 test corpus optimizations, the center is for the rule based tuning (focused to any corpora) and the right hand side is for PAN 2012 test corpus optimizations.

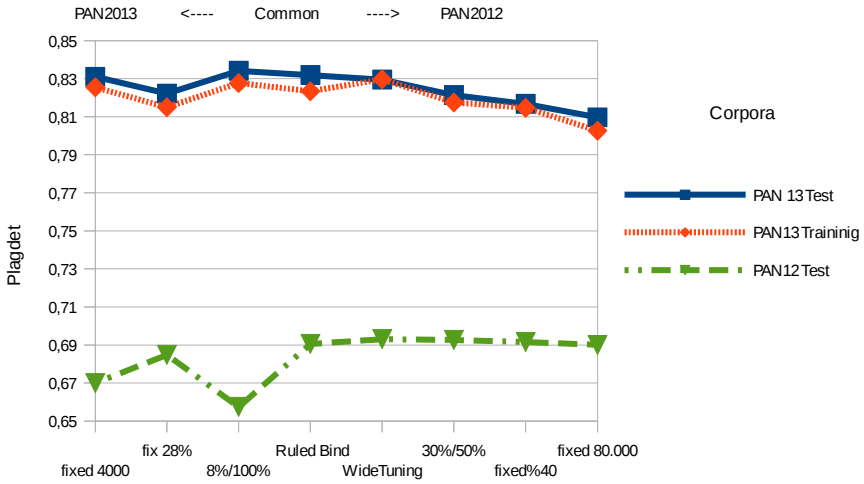

Fig. 1. Optimal settings for each optimization applied to different corpora

Concluding the analysis of the tests, it can be stated that the auto setting will not get optimal performance for any corpus, but a very closed one to it and good enough to meet the expectations for all of them, despite their considerable differences.

4 New Features in CoReMo 2.3

In its continuous evolution, CoReMo 2.3 includes the following new features since the 2.1 (PAN2013) version:

- Support for Hindi language, since CoReMo 2.2
- Granularity Filter dumping distance tunable as fixed one or as proportional to document size.
- Independent Granularity Filter settings for source and suspicious documents.
- Rules for auto-tuning Granularity Filter by the ratio suspicious/source sizes.
- New auto-tuning chunk size rule based on the suspicious/source sizes ratio.

5 Tuning Approaches Presented

Due to the new TIRA environment [5] features, for this issue, there different CoReMo tuning approaches have been provided:

1. *CoReMo 2.1 Settings*: the same as proposed last year, without the explained improvements, but without the existing bug when the 2013 competition, which left the last detection unregistered when it was near the end of the document. Though curiously it was the best performing of all the three at the Early-Bird evaluation, this fact doesn't match with the training experiments by 2013 and 2012 corpora or by the final competition corpus.
2. *Optimum Settings for PAN2013*: it only includes the granularity filer distance proportional to the documents size, different for the source document (100% of its size) and the suspicious document (8% of its size). It achieved the best performance when training, but for PAN2012 corpora, which was almost a 5% below the *WideTuning* version. It also got the best PAN2014 performance of the three tuning approaches, as expected.
3. *Wide Tuning Version*: It exploits the four new features above explained, with self-tuning for both granularity filter distances and chunk length, by rules based on the ratio between documents size. It Achieves the best Plagdet average for all the corpora, being significantly better than the others for the PAN2012 corpora, although slightly lower for 2013 ones.

Table 3 shows the Plagdet score achieved by the three tuning approaches for the different corpora used in training and the PAN2014 competition corpus, including the Early-Bird evaluation, where curiously, the previous year's tuning approach (CoReMo 2.1) achieved the best results.

Table 3. Plagdet Score achieved on different corpora

Tuning approach	PAN14 test	PAN14 Early-Bird	PAN13 test	PAN13 training	PAN12 test	Average 12~13	Average 12~14
CoReMo 2.3 2013 optimum (Soft. 2)	***0.85006***	0.84490	***0.83417***	***0.82766***	0.65747	0.77310	0.79234
CoReMo 2.3 WideTuning (Soft. 3)	0.84823	0.84870	0.82952	0.82478	***0.69304***	***0.78245***	***0.79901***
CoReMo 2.1 (Soft. 4)	0.84667	***0.85120***	0.82827	0.82709	0.66816	0.77451	0.79255

The tuning approach choosing to be submitted to compete for the ranking, based on which we would use for a production system, was the Wide-Tuning version, although the Early-Bird evaluation forecasts and the features of the competition corpus get to feel that the best fit to win would be respectively the previous year tuning or the 2013 corpora optimized one.
During the last PAN edition, it was compared the combination of the different current and previous year's corpora and approaches, and it was noted that the 2012 corpus was significantly the hardest to detect for any approach, having lower Plagdet scores and needing much larger run-times. The WideTuning approach, allowed us to obtain for 2012 up to 5% more Plagdet score than the others, while on the 2013 corpora it falls less than 0.5% below the optimum one. It's not the first time we prefer showing our research than our ranking progress, as we did in previous years using local translation systems instead of the best external Google Service.

6 Performance Analysis

The CoReMo 2.3 performance is still amazing with a 0.85 Plagdet score, precision 0.90, recall 0.80, and an optimal 1.00 granularity. Everything is very closed to the winning approaches, but with the fastest runtime of only 31 seconds: at least 4 times faster than the next one, and 100 times faster than the winning approach. Nevertheless, currently the TIRA system does not offer the ability to operate in the multi-core mode, and CoReMo is optimized to use it, allowing to run the same job in just 4 seconds using an AMD FX8120 based machine.

The Table 4 shows that, as expected, the best tuning approach was the PAN2013 optimized (but it gets about 0.04 less Plagdet for PAN2012 test corpus than our choice), and the selected WideTuning version gets almost same best performance, having only 0,001 below score.

Table 4. PAN2014 Corpus 3 (new) performance

	Plagdet	**Precision**	**Recall**	**Granularity**
CoReMo 21	0.84667	0.89179	***0.80590***	***1.00000***
CoReMo 23 PAN13 optimized	***0.85006***	***0.90770***	0.79931	***1.00000***
CoReMo 23 WideTuning	0.84870	0.90032	0.80267	***1.00000***

Both CoReMo 23 tuning versions get to overpass the last year's tuning performance, however, a possible overfitting effect may be present on the best PAN13 optimized due to its lowest PAN2012 test corpus performance (see Table 3). We recommend to check the PAN2012 test corpus performance for all the PAN14 participants approaches which could arrange it in a reasonable time, to look for any possible overfittings.

7 Conclusions and Future Works

Although these auto-settings are in very good direction and achieving very good results at best runtime for any corpus, the authors think that the self-tuning rules for granularity reduction could be significantly improved by another function different to simple thresholds to change fixed parameters values.

As we mentioned in the previous editions, these methods could be combined with others such as the use of thesaurus to minimize the effect of replacing words by other synonyms.

It would be wise to check the performance of all participant approaches, looking for any possible overfittings, on the PAN2012 competition corpus (or at least which could arrange it in a reasonable time), which has proved to be the most difficult ever to analyze, and, at least in our case, has demonstrated to require very different settings that for 2013 or 2014 corpora.

8 References

1. 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 2010. Association for Computational Linguistics.
2. Diego A. Rodríguez Torrejón and José Manuel Martín Ramos. Text Alignment Module in CoReMo 2.1 Plagiarism Detector—Notebook for PAN at CLEF 2013. In Forner et al. [6]
3. Rodríguez-Torrejón, D.A., Martín-Ramos, J.M.: N-gramas de contexto cercano para mejorar la detección de plagio (Surrounding Context N-grams to Improve the Plagiarism Detection) In [7]
4. Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Overview of the 5th International Competition on Plagiarism Detection. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, Working Notes Papers of the CLEF 2013 Evaluation Labs, September 2013. ISBN 978-88-904810-3-1.
5. Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Recent Trends in Digital Text Forensics and its Evaluation. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative (CLEF 13), September 2013. Springer. ISBN 978-3-642-40801-4.
6. Pamela Forner, Roberto Navigli, and Dan Tufis, editors. CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, 2013. URL http://www.clef-initiative.eu/publication/working-notes.
7. II Congreso Español de Recuperación de Información (CERI 2012). 17-18 June, Valencia (2012). http://users.dsic.upv.es/grupos/nle/ceri/index.html