Crosslingual CoReMo System
Notebook for PAN at CLEF 2011

Diego Antonio Rodríguez Torrejón * º, José Manuel Martín Ramos º

* I.E.S. "José Caballero", º Universidad de Huelva (Spain)
diego@dartsystems.es jmmartin@dti.uhu.es

Abstract. This paper shows an extended version of external CoReMo System (Contextual Reference Monotony, ranked 6th in PAN2010), now with crosslingual capability (ranked 5th in PAN2011 / Plagdet 0,2340). It's not the best ranked system for translated plagiarism (ranked 3th / Plagdet 0,3587), but it has high reliability and speed (global results in 30 minutes), low computer requirements and its own internal translation system.

Keywords: crosslingual, plagiarism detection, n-gram, contextual n-gram, Referential Monotony, Information Retrieval.

1 Introduction

This paper shows an extended version of "CoReMo" System [2], adding crosslingual capability. It's also based on *Contextual n-grams* (CTNG) and *Referential Monotony* (RM) prune strategy to get high performance, fast response and low requirements.

If external translation services are used to get English version of the source documents, biased results could be obtained for crosslingual subcorpus: PAN-PC corpora are mainly getting automated plagiarism cases by using those algorithms. The task could then be compared to a poor obfuscation strategy for only English sources.

External translation services are slow methods exposed to availability for high amount of documents, large documents or changes in the API. Because of that, this proposal uses its own translated CTNG modeling, fast and simple, by two specially designed dictionaries: direct2stem and stem2stem.

2 External Plagiarism Detection (CoReMo System)

CoReMo System [2] is a high performance, low requirements and high speed External Plagiarism Detection System. It is mainly based on CTNG and RM concepts, joined to a quick search engine:

CTNGs are the basics for indexing and modeling documents for PD purposes. These n-grams are obtained by performing case folding, stopwords removal, Porter stemming [3] and internal sort. CTNGs are featured to be enough representative for their immediate context, and to act as discriminative fingerprint for their document/*split* into wide collections (unique CT3N in corpus > 90%).

The **Quick Search Engine** is based on a CTNGs inverted index, oriented to receive a *split*[1] as query, and returning the only best matching source document. Implemented in C++, it runs in 64 bits GNU-Linux distribution. It needs one only core and low memory requirements.

RM is a **prune strategy** to discard low likelihood suspicious sections: comparative is not performed until at least a minimum of consecutive *splits* are found pointing to same source. Then a dual border search is arranged by looking for CTNG matching between whole source document and the consecutive splits detected.

Separated analyses are performed **for monolingua**l (only English) **and cross-lingual** (non English) source collections. Both analyses are based on English only CTNG. After mixing results, an only final report is obtained.

2.1 Monolingual Analysis

For monolingual analysis, former CoReMo System was used without significant changes, excepting the newer tuned parameters obtained from PAN-PC-2010 [4].

2.2 Cross-lingual Analysis

For suspicious documents, direct CTNG modeling is used. Non English source documents processing (including stem translation) and settings are the differences.

2.2.1 Special Translation Dictionaries Focused to Contextual n-grams.

Two new resources were developed for every language to arrange a fast crosslingual detection: the ***direct2stem*** and the ***stem2stem*** dictionaries. Both dictionaries were created for direct return of stemmed translation. The first one has full word entries, and one single stemmed translated output. The second one is similar, but entries are stemmed words, being only used when the first one has not directly found the word, in the hope to get a translation at least for the root.

These dictionaries were extracted from *Wiktionary* [4] and *Wikipedia interlanguage links* dictionaries [5], by discarding composed entries, composed returns, and selecting the more frequent return stem when multiple output is available (compared to PAN-PC-2009 source-documents stem term frequency).

At present, both dictionary kinds are available at [6] under GPL terms, to get English stem words for German, Spanish and French entries.

---

1 Fix amount of consecutive CTNG in the document. Different sentences could be mixed.

2.2.2 Translated Contextual n-grams Modeling

Non English source documents are modeled to English CTNG (to get inverted index and locating plagiarized sections) in this way after (non English) stopwords removal:

- ***direct2stem*** specific dictionary (for source document language) is searched for full word: if it's found (19%), the most used stemmed translation is directly returned (faster to get CT1G), else the next step is processed.
- ***stem2stem*** specific dictionary for source document language is searched for former stem word (stemmed in original language): if it's found (34%), the most used stemmed translation is returned, else the third step is processed.
- The **original word is stemmed by English rules** (47%), in the hope to get matching for propper names, and used as possible stem translation.

The CTNG modeling is finished as usual, by n-grouping CT1Gs (n-1 overlapped) alphabetically ordered. This saves the changed word distribution happened after translations.

3 Training and Evaluation

The system was trained using PAN-PC-2010. The best parameters found for monolingual analyses (CT3N, split length 17, RM threshold 3, feedback disabled, split overlapping 0 and border compensation 0), getting Plagdet increased from 0.5851 to 0.6026 by PAN-PC-2010 rules (without crosslingual analysis).

Using a single inverted index with former parameters for all English and non English source documents, a lower global performance was obtained.

Analyzing non English source subcorpora separately, the best parameters can be noticed very different for non English sources: split length 120 and RM threshold 3. The performance(PD 0.36, R 0.23, P 0.80, G 1.00) was good enough to be mixed.

Similar results (PD 0.36, R 0.24, P 0.69, G 1.00) were got with PAN-PC-2011.

The big split length has the annoyance that small and medium plagiarized sections (< 480 words | 2600 chars approx.) could pass unnoticed. That's the reason to have separated analysis to take the opportunity of lower split length for monolingual use.

The fast stem translation system disambiguates by most used stem term. Using CT2G, chance matching is frequent (low precision). By CT3G, chance matching is less usual, being selective enough however if large split length is used.

Final external training results after mixing: **PD 0.71, R 0.59, P 0.89, G 1.00**.

3.1 Comparative Results

The new CoReMo version has improvements for crosslingual plagiarism only, however, compared to last year competition, performance is only better for translated plagiarism, similar (but a bit lower) for non paraphrased plagiarism, and in general much lower performance for any other subcorpus. Compared to values obtained in training, similar values were got for automated translation plagiarism however.

3.2 Used Equipment and Timing Features.

The biggest inverted index needed to analyze PAN-PC-2011 cannot be fully loaded, for a single analysis, into the same 4 GB RAM laptop used last year. Former partitioning and mixing experiences got a bit lower *Plagdet* mark and about 1/3 larger time. So, a standard PC was used instead (P4 3.1 GHz, 8GB RAM, 64GB SSD HD), running Ubuntu 10.04/64 bits once again. The results are 2.8 times faster compared to former laptop: PAN-PC-2010 crosslingual analysis in 45 minutes. **PAN-PC-2011 is analyzed faster (only 30 minutes)** due to the smaller suspicious collection to analyze (half size), in spite of a larger (40%) source collection.

4 Conclusions

External CoReMo rank is better than last year's, but it got lower Plagdet (0.234) than training (0.710). Harder obfuscation in new PAN-PC-2011 corpus may be the reason.

CTNG has demonstrated feasibility to attack cross-language plagiarism.

English documents CTNG model would change (focused on increasing matching by synonymy) to improve hard obfuscation and crosslingual results, and a possible integration of monolingual and crosslingual analysis into a single one. To avoid precision loss, other changes would be necessary. Larger dictionaries will also help.

5 Bibliography

1. Rodríguez-Torrejón D.A., Martín-Ramos J.M: CoReMo System (Contextual Reference Monotony) A Fast, Low Cost and High Performance Plagiarism Analyzer System: *Lab Report for PAN at CLEF 2010. In* Braschler M., Harman D., Pianta E., editors. Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 *September, Padua, Italy, 2010. ISBN 978-88-904810-0-0.*
2. Porter M. F.: An algorithm for suffix stripping.(Porter stemmer) *Program*, 14(3):130-137. (1980) http://tartarus.org/~martin/PorterStemmer/index.html
3. Potthast M., Stein B., Eiselt A., Barrón-Cedeño A., Rosso P: An Evaluation Framework for Plagiarism Detection. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, August 2010. Association for Computational Linguistics.
4. Wiktionary Collaborative Project. http://www.wiktionary.org
5. Dicts.info: Free Dictionaries Project. http://www.dicts.info
6. Free resources from D'ART Systems homepage (author). http://www.dartsystems.es/downloads/resources