Linking English and Hindi news by IDF, Reference Monotony and Extended Contextual N-grams IR Engine

Worknotes for PAN @ CL!NSS 2013

Diego Antonio Rodríguez Torrejón[1] – José Manuel Martín Ramos[2]

Universidad de Huelva
[1]dartsystems@gmail.com [2]jmmartin@dti.uhu.es

Abstract: In this paper a new approach is shown to arrange the task of the cross-lingual linking between English and Hindi paper news. Due that it's a very similar problem to part of the Plagiarism Detection Process, to tackle it, part of the CoReMo Plagiarism Detector technology has been used: The HAIRS (High Accuracy Information Retrieval System) engine, which indexes the Hindi documents modeled to Extended Contextual N-grams and selects the best similar for every chunk of the Hindi translated versions (by Google Translate external service) of the English news, filtered by the Reference Monotony Prune Strategy to avoid chance matching. In this way it is achieved a very short and selective group of linked pairs instead of a long rank, enabling a very fast posterior comparison. The matching n-grams containment ratio is used as similarity to sort the pairs, and available to decide a most selective prune even.

Keywords: Information Retrieval, contextual n-grams, cross-lingual text reuse

1 Introduction

The cross-lingual news linking is a very similar problem to the candidate source retrieval in the plagiarism detection process. We have used our experience in 4 former PAN Plagiarism Detection Competitions[1] through the CoReMo System [1-4] development, however the difficulty for this task is much higher than for the cross-lingual plagiarism detection of former PAN editions, due to the special difficulties in so different language pairs, the very short documents length, its biggest amount in the local corpus, and the lack of filtered n-grams matching documents in the corpus to enable coexistence of different versions of same focal event or several different news related to same news event: these are almost the worst conditions to get reliable plagiarism detections.

[1] http://pan.webis.de

The CoReMo technology has now been improved to use the full UTF-8 char-set, a now new C++ version of the Lightweight Hindi Stemmer [5] (there was not any available from Snowball Project), and the ability to use non English base language.

As CoReMo has two ways to try the cross-lingual detection tasks, to try local translations, it was developed two special Hindi to English stem oriented dictionaries [2], [6], named direct2stemHi2En and stem2stemHi2En. However, the tests carried out for this languages pair had show that only about a 30% of non empty words could be translated by this technique (for German or Spanish to English it was got about 55%), getting very noisy translations, almost useless to look matching in so short text conditions. A best quality translation was needed, and the external Google translation service[2] was used to get Hindi versions of the English news, and used to compare to the collection. As this external service is not free for big requests, it was not experienced with the English to Hindi translation to use English model base comparisons.

To arrange pairs document search in local collections, CoReMo has its own High Accuracy Information Retrieval System (HAIRS), which combines idf[3] detection of the best matching Hindi new for each chunk of the translated from English one, with the Reference Monotony prune strategy to discard chance matching. HAIRS indexes and uses very especial n-grams named Extended model of Contextual N-grams (XCTNG).

2 The CoReMo Plagiarism Detection System

The CoReMo Plagiarism Detection System was developed to take part in the PAN Plagiarism Detection Competitions, and it has been proved from 2010 edition for many of the different tasks/problems every year, being the current reference for text alignment as winner in PAN 2013, having the best balanced text alignments into much lower runtime than the other competitors approaches. The main force for its ability to get matching for strong paraphrases conditions, included translations, are due to the new extended model of Contextual N-grams.

However, for local source collections has not been officially tested since 2011 edition, as PAN was focused to use external web search attack to get candidates. However, HAIRS has been unofficially tested together with the new extended model of Contextual N-grams and Google Translations over the PAN-PC-2011[7], getting also the best unofficial score. It demonstrates that it's a good idea to combine both HAIRS and xCT3G model for this task.

---

[2] http://translate.google.com

[3] Inverse Document Frequency: the inverse of the amount of documents having a concrete n-gram/term in the corpus. I's a way to measure how much rare is the term for the corpus.

2.1 Extended Contextual N-Grams

The Contextual N-gram [1] is a useful way to get n-grams to find matching when small changes in the words (derivatives)or words re-sorting due to paraphrase or translation happens. These n-grams are got by 6 steps process: case folding, stop-words and short length words removal, stemming and internal sort of remaining consecutive unigrams to get the final n-gram.

The Extended model of Contextual N-grams [4] is the combination of the former ones with other 3 types obtained by skipping in the process (l_SCnG, r_SCnG and OEnG), obtaining as many 4 n-grams than words or single Contextual n-grams. It gets much more possibilities of matching for other obfuscation types like words changed, deleted or inserted, but in a discriminative way to avoid the chance matching obtained when a lower grade n-gram is used for the same goal. It has been the best way to get reliable seeds in the text alignment task for PAN Plagiarism Detection competition.

Let's see an example to best understanding from the well-known sentence:

"~~The~~ **quick brown fox jump**s ~~over~~ ~~the~~ **laz**y **dog**"

1. 1 _2_3 QUICK BROWN FOX → BROWN _FOX_QUICK (CTnG)
2. 1 _2_4 QUICK BROWN JUMPS → BROWN _JUMP_QUICK (SC3G)
3. 1 _3_4 QUICK FOX JUMPS → FOX_JUMP_QUICK (SC3G)
4. 1 _3_5 QUICK FOX LAZY → LAZ_FOX_QUICK (OEnG)
5. 2_3_4 BROWN FOX JUMPS → BROWN _FOX_JUMP (CTnG)
6. 2_3_5 BROWN FOX LAZY → BROWN _FOX_LAZ (SC3G)
7. 2_4_5 BROWN JUMPS LAZY → BROWN _JUMP_LAZ (SC3G)
8. 2_4_6 BROWN JUMPS DOG → BROWN _DOG _JUMP (OEnG)
9. 3_4_5 FOX JUMPS LAZY → FOX_JUMP_LAZ (CTnG)
10. 3_4_6 FOX JUMPS DOG → DOG _FOX_JUMP (SC3G)
11 . 3_5_6 FOX LAZY DOG → DOG _LAZ_FOX (SC3G)...

CoReMo can use different xCTnG grade operation modes, but the best results for plagiarism detection as for this current CL!NSS approach uses 3th grade (xCT3G).

2.2 High Accuracy Information Retrieval System (HAIRS)

The xCT3G also acts as a signature of a document or a text chunk for most of cases. An df study (see Table 1)from PAN-PC-2010 or CL!NSS 2012/2013 Hindi corpora is shown to understand that in fact, about 80% of xCT3G in the corpus are exclusives, 90% belong at most to two news and 96% at most to 5. Only 1 of each 5 new xCT3G will be repeated at any of the 50691 Hindi docs, and the probability to match to an arbitrary document in this corpus, was calculated of 0.003762%.

HAIRS indexes all the xCT3G obtained from Hindi news, but it only registers a fixed maximum of references for each one in order to optimize speed and memory.

Table 1. n-gram document frequency study on CL!NSS and PAN-PC-2011 source subcorpora

df	n-grams quantity	ratio	n-grams quantity	ratio
	CL!NSS 2012/13 Hindi xCT3G		PAN-PC-2011 English xCT3G	
--	26829851	1.0000	537613396	1.0000
01	21086302	0.7856	481407991	0.8955
02	2887507	0.1076	34537949	0.0642
03	989264	0.0369	9974359	0.0186
04	494964	0.0184	4327470	0.0080
...				
97	414	0.0000	265	0.0000
98	393	0.0000	260	0.0000
99	420	0.0000	261	0.0000
> 99	26686	0.0010	8626	0.0000

Each translated English Document is modeled to Hindi xCT3G and splitted in chunks of constant xCT3G length (for this task we tuned the system to chunk length of 10 xCT3G). Each chunk is used as query to be sent to HAIRS, which returns a only best matching Hindi document for that chunk based on the amount of matching xCT3G pondered by its *idf (1 / df)*.

It's difficult to think in persistent chance matching (remember the probability of 0,003762% for each xCT3G), even when not immediately else in the closed n-grams, if there is not a real relation in the content of both documents. In fact, to get the most improvable chance matching, CoReMo uses the **Reference Monotony** prune strategy. This prune strategy, basically consists in discarding apparent detections when the same Hindi document is not referenced at least in a threshold (monotony) of consecutive xCT3G chunks.

In the figure 1, the detection process and search space reduction by RM is shown: dark gray emphasized chunks give direct detection (5 consecutive splits pointing to reference doc #91)due to pass RM threshold (3 in the example). Light gray and all the other references are discarded as does not pass the monotony threshold.

73	-1	6	49	11	-1	31	**91**	**91**	**91**	**91**	**91**	6	92	5	7	98	57	57	-1	-1	-1	61

Fig. 1: Document's Chunk Map, with single source candidate for each (basic for RM)

3 Ranking the Pair Lists

After getting by HAIRS the detected candidates, it was arranged a text alignment as int the case of PAN 2013, because by that way it can be got the involved almost duplicated zones, and the xCT3G matching ratio (also named containment)from both, the Hindi and English documents. I was used the average ob both ratios as similarity to rank the list.

4 Training Phase

In the training phase, it was experienced by changing the reference monotony threshold (RM), chunk length and the index maximum references per document registration parameters.

The best values for the training were ever got by the chunk length of 10 xCT3G, RM threshold of 3 chunks and, depends of the goal, indexing at most 1, 2, 5 or 25 possible source Hindi news per indexed xCT3G.

Really we had many problems to get conclusions as the way our system runs, does not return a big ranked list of documents, else a single best Hindi new for every chunk.

The evaluation program was expecting to receive 100 pairs for each English document, and it evaluated in a wrong and rare way with our short full lists which had from 11 to abut 190 pairs at most, joining all the 50 training inputs. Our late arrival to the task, joined to this problem, were stopping our advance, sending us to obfuscated decisions until it was discovered, even when the organization sent us almost immediately a fixed evaluation version.

We had some more problems with the new adaptation of CoReMo for this task: i.e. we could not get try lower chunk length than 10 xCT3G due to any yet unresolved bug, so may be we could get better results after fixing it.

Another bug in the adaptation, did not enabled to really try the RM of only 2 chunks. It has been detected an fixed after the competition, and for same tuning as for the competition run (RM threshold of 3 chunks), it was got a significant improvement for all the NDCG@x scores.

As expected, the biggest RM threshold you put, less documents you will get, and the lower references registered in the HAIRS index (only 1), the most precise pairs is got for NDCG@1, but less pairs to compare are lost however due that after passing the maximum of registered references, they are ignored and so they are useless to detect. It happens mostly when same news event but different focus event.

The most pairs are detected by HAIRS, more text alignments are necessary to rank the results. The results obtained by monotony 3 and max index list of references of 1, got a good score for a really low list of pairs. However, it's not a problem for

CoReMo Technology, as it's a optimized and parallelized algorithm feasible to align more than 5000 documents pairs per second in an AMD FX8120 8 cores machine.

5 Evaluation Results

We have sent 3 runs, with same tuning excepting that it was used different index type. The best result for NDCG@1 was got by the single reference index for each n-gram, which has several advantages: it's much less memory hungry than the other options, and there are to align less documents to get it. The other two runs were got by 2 and 5 maximum references registered for each indexed xCT3G. As more references are indexed, more documents are recovered from same news events and more recall is got, but less precision however. It implies best scores as NDCG@1 values are far from 1.

The table 2 shows the results obtained for our official and non-official extra runs.

The dark positions were obtained by the officially sent runs. Together, it can be seen the results of later tests carried out by 25 references lists index limit and by RM of only two chunks. The amount of pairs returned by HAIRS for the 25 documents are closed in brackets for each run.

6 Conclusions and Future works

Cross-lingual English Hindi News Text Reuse is a really difficult task, due to the reasons explained at Introduction.

Would be interesting to study the another cross-lingual direction, to see the influence and the scores differences, but translate all the Hindi corpora by Google Translate is expensive for us. However, when many different Indian Languages are involved, would be more interesting having a single pivot language to generate the IR index. Next time, would be interesting that the organization would offer directly the Indian news translations to English.

We are now considering different options as similarity to rank the pairs: as a composition of the publishing date distance, the English xCT3G matching ratio and the amount of detection involved in the text alignment.

Fixing the bug to test lower chunk length could get some more surprise.

Another possibility is to test directly the comparison to all the possible pairs in the corpora. It will not use IR methods, and so it is not so interesting for FIRE. Our systems expect less than 10 seconds per English document, so all the CL!NSS'12 problem could be analyzed in about 8 minutes and CL!NSS'13 in about 4. The difference of IR search methods and brute strength would be interesting however to analyze again, as in [8].

Table 2. Scores obtained form official runs (gray shaded) and other later runs after improvements.

Monotony 3 - Chunk length 10	NDCG@1	NDCG@2	NDCG@3	NDCG@4	NDCG@5	NDCG@10	NDCG@20	NDCG@50
Max. Refs. = 25 (177 pairs)	0.5200	0.3867	0.3822	**0.3747**	**0.3607**	**0.3499**	**0.3448**	**0.3443**
Max. Refs. = 5 (164 pairs) **run 2**	0.5200	0.3867	**0.3918**	0.3697	**0.3555**	**0.3408**	0.3355	0.3351
Max. Refs. = 2 (129 pairs) **run 1**	0.5200	**0.4267**	0.3865	0.3586	0.3450	0.3293	0.3224	0.3217
Max. Refs. = 1 (95 pairs) **run 3**	**0.5208**	0.3958	0.3538	0.3317	0.3201	0.3075	0.3030	0.3030

Monotony 3 - Chunk length 10	NDCG@1	NDCG@2	NDCG@3	NDCG@4	NDCG@5	NDCG@10	NDCG@20	NDCG@50
Max. Refs. = 25 (2028 pairs)	0.6000	**0.5333**	**0.4837**	**0.4741**	0.4531	**0.4746**	**0.4742**	**0.4769**
Max. Refs. = 5 (1921 pairs)	0.6000	0.5200	0.4823	0.4661	0.4507	0.4659	0.4638	0.4709
Max. Refs. = 2 (1698 pairs)	0.6000	**0.5333**	**0.4837**	**0.4741**	**0.4580**	0.4527	0.4536	0.4557
Max. Refs. = 1 (1047 pairs)	**0.6400**	0.5000	0.4767	0.4550	0.4361	0.4426	0.4391	0.4414

7 Acknowledgements

I've learnt a lot in the way, and the CoReMo technology has got interesting improvements from it. I'd like to give thanks to the CL!NSS organization for enforcing us to take part in the task and its full collaboration to solve all the inconveniences found. Thanks to all the participants as our contribution gives a elevated dimension for all of our jobs, and get the task more interesting every year. Finally thanks to my family to support again my non-remunerated research at home, discounted from family time, in specially critic times.

8 References

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.
2. 3. Rodríguez-Torrejón, D.A., Martín-Ramos, J.M.: Crosslingual CoReMo System: Notebook for PAN at CLEF 2011. In 10. Vivien Petras and Paul Clough (Eds.): Notebook Papers of CLEF 2011 Labs and Workshops, 19-22 September, Amsterdam, The Netherlands (2011).
3. Rodríguez-Torrejón. D.A. and Martín-Ramos, J.M.: Detailed Comparison Module In CoReMo 1.9 Plagiarism Detector —Notebook for PAN at CLEF 2012. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors. CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy, 2012. URL http://www.clef-initiative.eu/publication/working-notes.
4. 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 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.
5. Ramanathan, A. and Rao, D : A Lightweight Stemmer for Hindi. Workshop on Computational Linguistics for South-Asian Languages, EACL (2003)
6. 7. Rodríguez-Torrejón, D.A., Barrón-Cedeño, A., Sidorov, G., Martín-Ramos, J.M., Rosso, P.: "Influencia del diccionario en la traducción para la detección de plagio translingüe". (Dictionary Influence in Crosslingual Plagiarism Detection). in II Congreso Español de Recuperación de Información (CERI 2012). 17-18 June, Valencia (2012). http://users.dsic.upv.es/grupos/nle/ceri/index.html
7. 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.
8. Barrón-Cedeño, A., Rosso P.: On the Relevance of Search Space Reduction in Automatic Plagiarism Detection. Procesamiento del Lenguaje Natural, 43:141-149. (2009)