A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection

Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, Didier Schwab

A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection

Jérémy Ferrero[1,2], Frédéric Agnès[1], Laurent Besacier[2], Didier Schwab[2]
[1] Compilatio, 276 rue du Mont Blanc, 74540 Saint-Félix, France
[2] LIG-GETALP, Univ. Grenoble Alpes, France
{jeremy, frederic}@compilatio.net, {jeremy.ferrero, laurent.besacier, didier.schwab}@imag.fr

Abstract

In this paper we describe our effort to create a dataset for the evaluation of cross-language textual similarity detection. We present pre-existing corpora and their limits and we explain the various gathered resources to overcome these limits and build our enriched dataset. The proposed dataset is multilingual, includes cross-language alignment for different granularities (from chunk to document), is based on both parallel and comparable corpora and contains human and machine translated texts. Moreover, it includes texts written by multiple types of authors (from average to professionals). With the obtained dataset, we conduct a systematic and rigorous evaluation of several state-of-the-art cross-language textual similarity detection methods. The evaluation results are reviewed and discussed. Finally, dataset and scripts are made publicly available on GitHub: `http://github.com/FerreroJeremy/Cross-Language-Dataset`.

Keywords: Dataset, Cross-language dataset, Evaluation, Cross-language similarity detection, Cross-language plagiarism detection

1. Introduction

Guibert and Michaut (2011) state that 34.5% of European students have already copied all or part of a document to present it as their own work. This confirms the work of the Josephson Institute (2011) and McCabe (2010) who estimate that more than 30% of American and Canadian students have already re-used Web sentences without citing their source; this is considered as plagiarism. *"Plagiarism is an act of fraud to steal and pass off (the ideas or words of another) as one's own without crediting the source to present as new and original an idea or product derived from an existing source"* (Plagiarism.org, 2014).
In addition, Internet expansion facilitates access to documents in foreign languages and to increasingly efficient machine translation tools. Consequently, a new kind of plagiarism is becoming frequent: the *Cross-Language Plagiarism*. It involves plagiarism by translation, i.e. a text has been plagiarized while being translated (manually or automatically). The challenge in detecting this kind of plagiarism is that the suspicious document is in a language different from its source.
In this relatively new field of research, no complete evaluation framework has been carried out and no sufficiently diversified reference dataset has been made available to enable more systematic and rigorous evaluations.

Contributions. This paper presents our methodology to collect and build a reference dataset for the evaluation of cross-language textual similarity detection (made available to the research community). More precisely, the characteristics of our dataset are the following:

- it is multilingual: French, English and Spanish;
- it proposes cross-language alignment information at different granularities: document-level, sentence-level and chunk-level;
- it is based on both parallel and comparable corpora;
- it contains both human and machine translated text;
- part of it has been altered (to make the cross-language similarity detection more complicated) while the rest remains without noise;
- documents were written by multiple types of authors: from average to professionals.

The major contribution, in addition to merge and enrich existing corpora, has been to provide the various textual granularities and perform an evaluation of state-of-the-art methods on the proposed dataset.

Outline. After presenting the state-of-the-art methods, we first present the pre-existing corpora for the cross-language plagiarism detection and their limits, then we describe how we have gathered and enriched these corpora in a single dataset and we describe the characteristics of the dataset. Finally, we evaluate the main state-of-the-art methods on our dataset.

2. State-of-the-art

Textual similarity detection methods are not exactly methods to detect plagiarism. Plagiarism is a statement that someone copied text deliberately without attribution, while these methods only detect textual similarities. There is no way of knowing why texts are similar and thus to assimilate these similarities to plagiarism.

For the moment, there are five classes of approaches for cross-language similarity detection. Figure 1 presents the taxonomy (Potthast et al., 2011) of the different cross-language textual similarity detection methods grouped by class of methodology (in bold, the methods that we have evaluated on our dataset and which are detailed bellow).

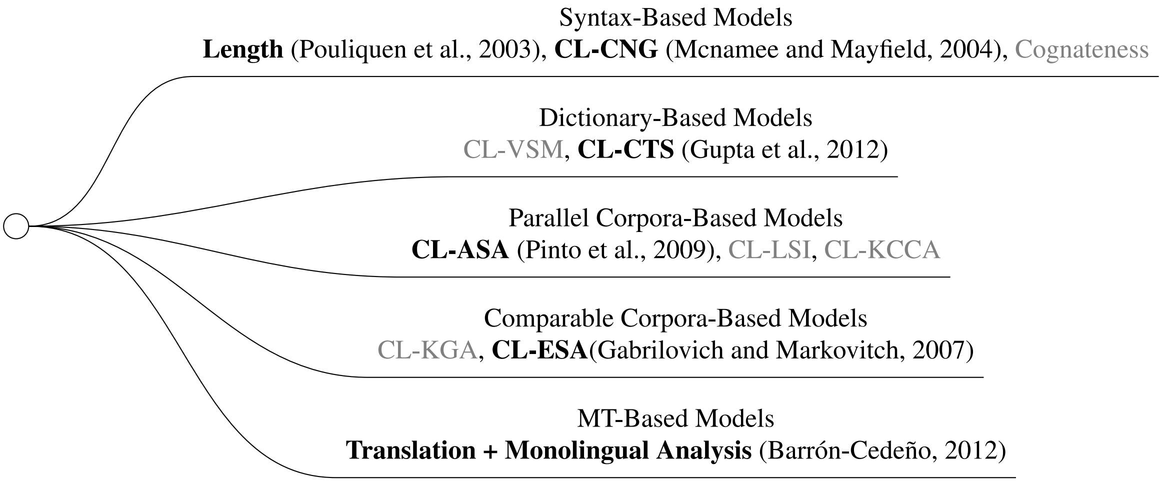

Figure 1: Taxonomy of different approaches for cross-language similarity detection (Potthast et al., 2011).

2.1. Length Model

Length Model aims to compare the size of two texts in an attempt to predict if they express the same thing or not. Though it is unlikely to find both documents d and d' written in two different languages L and L' with the same meaning, such as $|d| = |d'|$, i.e. having exactly the same length, it is assumed that their length are closely linked by a factor. Pouliquen et al. (2003) observe that there is a different factor for each language pair. They expressed the following formula, included in the work of Potthast et al. (2011):

$$\varrho(d, d') = exp\left(-0.5\left(\frac{|d'|/|d| - \mu}{\sigma}\right)^2\right) \qquad (1)$$

where μ is the average and σ is the standard deviation of the lengths (in characters) between the original documents and their translations, from L to L'. Table 1 represents the values of μ and σ which are used in the evaluation of Potthast et al. (2011).

Parameter	en-de	en-es	en-fr	en-nl	en-pl
μ	1.098	1.138	1.093	1.143	1.216
σ	0.268	0.631	0.157	1.885	6.399

Table 1: Coefficients of the average and the standard deviation between the languages pairs (Potthast et al., 2011).

2.2. Cross-Language Character N-Gram (CL-CNG)

CL-CNG is based on the Mcnamee and Mayfield (2004) work which is used in the information retrieval. It compares two texts under their n-grams vectors representation. The method achieve a good performance in information retrieval task for languages with the same origin because of common root words.

Let d and d', two documents in two different languages (respectively L and L'). First, the alphabet of these documents is normalized on a space $\sum = \{a - z, 0 - 9, \ \}$, so only spaces and alphanumeric characters are retained. Any other diacritic or symbol is deleted. The uppercase are passed into lowercase. The texts are then segmented into n-grams (sequences of n contiguous characters). The variable n is previously optimized (according to studies, $n = [3, 5]$, Mcnamee and Mayfield (2004) use a *CL-C4G* when Potthast et al. (2011) prefer to use a *CL-C3G* model). The texts are thus transformed into *tf.idf* vectors of character n-grams. The similarity between the vectors may be calculated by a cosine similarity.

2.3. Cross-Language Conceptual Thesaurus-based Similarity (CL-CTS)

CL-CTS aims to measure the semantic similarity between two vectors of concepts. The model consists in representing documents as vectors and compare them. The method also involves no explicit translation, the matching is performed using internal connections in the used ontology.

For example, Gupta et al. (2012) represent the documents using *Eurovoc* (1995) thesaurus concepts vectors. They use a stop words filter, a stemming step and a term frequency weighting to build the vectors. A cosine similarity between these vectors is associated with named entities matching, and the *Length Model* of Pouliquen et al. (2003), seen in section 2.1, is also used to compare the vectors. Česka et al. (2008) proceed similarly with *EuroWordNet*[1] while Pataki (2012) prefers use a synonym dictionary because, according to her, use an ontology raises two problems. The first is the data limitations and the second is the asymmetry of available data in the different languages.

Let S a sentence of length n, the n words of the sentence are represented by w_i as:

$$S = \{w_1, w_2, w_3, ..., w_n\} \qquad (2)$$

S_x and S_y are two sentences in two different languages. A bag of words vector V from each sentence S is built, by filtering stop words and by using a function that returns for a given word all possible translations. The vectors V_x and V_y are respectively the conceptual representations of S_x and S_y.

[1]http://www.illc.uva.nl/EuroWordNet/

To calculate the similarity between S_x and S_y, the most common method is to calculate the intersection between V_x and V_y:

$$sim(S_x, S_y) = \mid V_x \cap V_y \mid \quad (3)$$

If a sentence has sufficient common concepts with an other, then it is considered as the possible translation of the other. But Pataki (2012) uses a more discriminant formula taking into account the size of the compared bag of words:

$$sim(S_x, S_y) = min(\mid V_x \cap V_y \mid - \mid V_x \setminus V_y \mid, \mid V_y \cap V_x \mid - \mid V_y \setminus V_x \mid) \quad (4)$$

2.4. Cross-Language Alignment-based Similarity Analysis (CL-ASA)

CL-ASA is introduced for the first time by Barrón-Cedeño et al. (2008) and developed subsequently by Pinto et al. (2009). The aim of the method is to determinate how a textual unit d written in the language L is potentially the translation of an other textual unit d' written in a language L'. *CL-ASA* involves the creation of a bilingual unigram dictionary which contains the statistical probabilities of translations pairs determined from a parallel corpus. The IBM-1 model (Brown et al., 1993) can be adopted using only the lexical translations. Pinto et al. (2009) proposed a formula that factored the alignment function.

Let x and y, two sentences, such as x_j is the j^{th} word of the sentence x and y_i, the i^{th} word of the sentence y. We want to know the probability $p(x, y)$ that x is the translation of y.

$$p(x, y) = \prod_{j=1}^{|x|} p(x_j|y) \quad (5)$$

where

$$p(x_j|y) = \sum_{i=1}^{|y|} \frac{1}{|y|+1} p(x_j|y_i) \quad (6)$$

Improvements of the method were later proposed. For example, consider for each word x, only the best translations y, above a minimum probability (threshold of 0.4 according to the work of Barrón-Cedeño et al. (2010)) or also filter the stop words to minimize the number of operations. Barrón-Cedeño et al. (2008) (2012) also propose to replace the language model, usually used, by the *Length Model* of Pouliquen et al. (2003) seen in section 2.1. In this case, the final formula for *CL-ASA* becomes:

$$sim(d, d') = \varrho(d, d') \cdot p(d, d') \quad (7)$$

2.5. Cross-Language Explicit Semantic Analysis (CL-ESA)

CL-ESA is based on the explicit semantic analysis model introduced for the first time by Gabrilovich and Markovitch (2007), which represents the meaning of a document by a vector based on concepts derived from *Wikipedia*, to find a document within a corpus. It was reused by Potthast et al. (2008) in the context of cross-language document retrieval. In *ESA*, a document d is represented by its similarities with the documents of a collection D, represented by a similarity vector **d** of n dimensions, such as:

$$\mathbf{d} = (\varphi(v, v_1^*), ..., \varphi(v, v_n^*))^T \quad (8)$$

where v is the terms vector of d, v_i^* is the terms vector of the i^{th} document in D and n is the number of documents in D. Any terms vector can be used but in the state-of-the-art, v is usually a *tf.idf* character n-grams vector. If $\varphi(v, v_i^*)$ is smaller than a fixed threshold, it can be reduced to zero in order to minimize noise and facilitate calculations. In the state-of-the-art, the function φ is a cosine similarity. Let **d'** a vectorial representation of another document d' relating to D. Thus the similarity between d and d' can be defined as $\varphi(\mathbf{d}, \mathbf{d'})$.

For the cross-language task, we now consider d and d', two documents in two different languages (respectively L and L'), and D and D' two different collections containing a large number of documents in the respective languages of d and d'. If the documents inside D are one to one parallel or comparable with the documents inside D', then the representations of ESA in both languages become comparable. We build **d**, the vectorial representation of d, where each dimension i in **d** represents the similarity between d and each document D_i of the corpus D. For the second document d', we proceed the same way, building a vector **d'** using the collection D'. The two vectors **d** and **d'** are a representation of d and d' related to the collections D and D'. The similarity between d and d' can be expressed as:

$$sim(d, d') = \varphi(\varphi(d, D), \varphi(d', D')) = \varphi(\mathbf{d}, \mathbf{d'}) \quad (9)$$

2.6. Translation + Monolingual Analysis (T+MA)

T+MA is a rather intuitive method that has been updated by Barrón-Cedeño (2012). It consists in translating the texts in the same language in order to operate a monolingual comparison between them.

Let d and d', two documents written in two different languages (respectively L and L'). The first step of the method consists in translating the document d in language L', the document d' in language L or the two documents in a third language L'', which is called hub or pivot language. To do that, Kent and Salim (2010) directly use the Google Translate API, while Muhr et al. (2010) replace each word of one text by its most likely translations in the language of the other text.

After the translation step, a monolingual comparison of both documents is now possible. According to Barrón-Cedeño et al. (2010) and Muhr et al. (2010), it is better to use methods such as bags of words that show better results on monolingual textual comparisons (Barrón-Cedeño et al., 2009). Because machine translation tools can give too multiple translations (all correct but being substantially different) and therefore it is not advisable to make a monolingual alignment with lexical or syntactic methods (Barrón-Cedeño et al., 2010).

3. Dataset for the cross-language plagiarism detection task

3.1. Existing corpora

There are many multi-language and cross-language dataset listed by OPUS[2] website. One example of these most

[2] http://opus.lingfil.uu.se/

used corpora is undoubtedly *Europarl*[3] (Koehn, 2005). It is a widely used corpus in cross-language text analysis and machine translation. It is a parallel corpus consisting of the European Parliament exchanges transcriptions, about nearly 10,000 parallel documents in more than 21 languages spoken across the European Union. Similarly, *JRC-Acquis*[4] is also often used in cross-language NLP or translation research. It is a parallel corpus, representing extracts of *Acquis Communautaire* (applicable laws in the European Union states), available in over 20 languages. As well, *Wikipedia* is often used as a comparable corpus in multiple languages. These last two, i.e. *JRC-Acquis* and *Wikipedia*, were used by Potthast et al. (2011) for cross-language plagiarism detection. Finally, another interesting collection of documents is the one gathered by Prettenhofer and Stein (2010) who collected *Amazon Product Reviews* (books, DVD and music albums) for a cross-language sentiment analysis task (Google Translate was used to build the parallel corpus).

3.2. Limits of existing corpora

The above mentioned cross-language corpora present the following shortcomings:

- They propose only one alignment granularity (document or sentence) whereas plagiarism can occur at different levels (sub-sentence level for instance);
- Taken separately, these corpora are very specific: parallel or comparable documents, manual or automatic translations, average or professional translators;
- Taken separately, these corpora only cover a specific domain (e.g. law or politics) which questions the validity of an evaluation done on a single dataset.

Ideally, a dataset allowing a rigorous evaluation of the cross-language similarity detection methods should not contain these limitations and be as diversified as possible.

4. Our dataset

4.1. Merged data

So far, our dataset only focuses on French, English and Spanish languages. The collections in these three languages, presented in the section 3.1, were first gathered in our dataset. The result is that the *JRC-Acquis* corpus (10,000 documents per language), *Europarl* corpus (close to 9,500 documents for each language), *Wikipedia* collections (10,000 documents per language) and *Webis-CLS-10* corpus also know as *Amazon Product Reviews (APR)* corpus (6,000 documents per language) have been reused. To enrich these corpora, we also used:

- **The corpus used for the *PAN 2011* evaluation (Potthast et al., 2010) of the CLEF campaign**. The corpus was designed for mono-language plagiarism detection task but it contains excerpts of texts of same books in different languages. These texts come from books freely available on the Gutenberg Project website[5]. The extraction process involves analyzing XML files containing the metadata of each document in the corpus. Then, using this information, parallel English-Spanish pairs are extracted. The process led to nearly 3,000 document pairs.
- **Conference papers**. So far, no corpus includes scientific texts, this is why we collected conference papers that were initially published in one language and then translated by their authors to be published in another language. For practical reasons, we focused exclusively on articles published first in French and then in English. The BibTeX file of French speaking conferences in NLP (the 1997-2014 *TALN archives*, made available in the works of Boudin (2013)[6] and the 2006-2011 RNTI collection made available by the challenge of the EGC 2016 conference[7]) were parsed to extract the names of the authors of each article. Then, names were used as queries in Google Scholar and Google Search Engine. Papers in PDF format corresponding to the most relevant search results were downloaded. We detected the language of each downloaded file according to the Cavnar and Trenkle (1994) classification algorithm and each English candidate file was manually checked to see if a significant part of it was related to one of the French original documents cited in the BibTeX. A total of 35 pairs of French-English conference papers were collected this way.

4.2. Multiple alignment granularities

To allow a rigorous evaluation of the state-of-the-art methods, we wanted a corpus with multiple granularities of aligned textual units. Thus alignment of our dataset at both sentence- and chunk- level was also needed in order to evaluate the performance of different methods on different types of texts but also on different sizes of texts.
To begin, each document in the dataset was split into sentences. To align sentences by pairs or triplets (depending of the languages present in the collections), we use HunAlign (Varga et al., 2005), whose dictionary for alignment has been enriched with DBNary[8] entries (Sérasset, 2015). The use of HunAlign is coupled with the *Length Model* described in Pouliquen et al. (2003). An ad-hoc threshold was used to filter the HunAlign's ouput to ensure the best possible ratio between the number of alignments achieved and their quality.
For the lower granularity, i.e. chunk level, we decided to focus on noun chunks because they are considered as the most meaningful elements in a sentence. To obtain these aligned noun chunks, we use the part-of-speech tagger TreeTagger (Schmid, 1994) followed by a post-processing step concatenating tokens according to their part-of-speech tag to build

[3] http://www.statmt.org/europarl
[4] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
[5] http://www.gutenberg.org/
[6] http://github.com/boudinfl/taln-archives
[7] http://www.egc.asso.fr/Manifestations_dEGC/71-FR-Defi_EGC_2016_Communaute_EGC_quelle_histoire_et_quel_avenir
[8] http://kaiko.getalp.org/about-dbnary/

Sub-corpus	Alignment	Authors	Translations	Alteration	NE (%)
JRC-Acquis	Parallel	Politicians	Professional translators	No	3.74
Europarl	Parallel	Politicians	Professional translators	No	7.74
Wikipedia	Comparable	Anyone	-	Noise	8.37
PAN (Gutenberg Project)	Parallel	Professional authors	Professional authors	Yes	3.24
Amazon Product Reviews	Parallel	Anyone	Google Translate	No	6.04
Conference papers	Comparable	Computer scientists	Computer scientists	Noise	9.36

Table 2: Characteristics by sub-corpus of our dataset. The percentages of named entities present in the last column are calculated with Stanford Named Entity Recognizer: `http://nlp.stanford.edu/software/CRF-NER.shtml`.

Sub-corpus	Languages	# Aligned documents	# Aligned sentences	# Aligned noun chunks
JRC-Acquis	EN, FR, ES	≃10,000	≃150,000	≃10,000
Europarl	EN, FR, ES	≃10,000	≃475,000	≃25,600
Wikipedia	EN, FR, ES	≃10,000	≃5,000	≃150
PAN (Gutenberg Project)	EN, ES	≃3,000	≃90,000	≃1,400
Amazon Product Reviews	EN, FR	≃6,000	≃23,000	≃2,600
Conference papers	EN, FR	≃35	≃1,300	≃300

Table 3: Number of aligned documents, sentences and noun chunks by sub-corpus.

phrases that can be considered as chunks. We also consider a minimal size (empirically set to 3 words) to form each chunk. To align these units, we proceeded the same way as for sentences.
Table 3 summarizes the statistics of our dataset (number of aligned documents, sentences and noun chunks). Obviously, the alignment step yields better results (more parallel sentences obtained) on parallel sub-corpora than on comparable sub-corpora. Also, the bigger corpora obviously lead to more aligned sentences at the sentence-level granularity. Concerning the chunks, the HunAlign threshold is set to maximize the quality of aligned chunks, which can explain their reduced number compared with number of parallel sentences.

4.3. Final corpus characteristics

The different characteristics of our dataset are synthesized in Table 2 while Table 3 presents the number of aligned units, by sub-corpus and by granularity, of our final dataset. To summarize, our dataset is composed of texts:

- in French, English and Spanish;
- aligned at the document-, sentence- and chunk- level;
- aligned from parallel or comparable collections;
- covering various fields;
- translated by humans (professionals or not) or automatically;
- altered or without added noise.

A manual check of more than 1,300 randomly chosen aligned chunks has been performed (which represents more than 3% of the chunk-level sub-corpus), providing an alignment confidence greater than 92%. We could get more accuracy, but with a decrease amount of exploitable alignments.

5. Evaluation protocol and Experiments

For the evaluation, we build a distance matrix of size $N \times M$, with M = 1,000 and $N = |S|$ where S is the evaluated sub-corpus. Each textual unit of S is compared to itself and to M - 1 other units randomly selected from S. A matching score for each comparison performed is thus obtained, leading to the distance matrix. Thresholding on the matrix is applied to find the threshold giving the best F_1 score. The F_1 score is the harmonic mean of precision and recall. Precision is defined as the proportion of relevant matches retrieved among all the matches retrieved. Recall is the proportion of relevant matches retrieved among all the relevant matches to retrieve. Each method is applied on each EN-FR sub-corpus for the three granularities, except the *PAN* corpus, that do not have EN-FR collection. For each configuration (i.e. one method on one sub-corpus at one granularity), 10 folds are carried out by changing the M selected units. The same unit may be selected several times at each fold. The averages and the confidence intervals of the F_1 scores of the 10 related folds are reported in Table 4 for the chunk-level, Table 5 for the sentence-level and Table 6 for the document-level.

During the evaluation, the *Length Model* used is that of Pouliquen et al. (2003) and *CL-CNG* considered is the one described by Potthast et al. (2011). *CL-CTS* used is that of Pataki (2012) and *T+MA* is the one applied by Muhr et al. (2010), both using lexical data from DBNary. *CL-ASA* used is that of Pinto et al. (2009) with a lexical dictionary calculated from the concatenation of TED[9] (Cettolo et al., 2012) and News[10] parallel corpora. *CL-ESA* implemented is that of Potthast et al. (2008) with the comparable data of *Wikipedia* that are not used in the test data.

[9]`https://wit3.fbk.eu/`

[10]`http://www.statmt.org/wmt13/translation-task.html#download`

Methods	Wikipedia (%)	TALN (%)	JRC (%)	APR (%)	Europarl (%)	Overall (%)
Random Baseline	00.28 $\pm$0.046	00.23 $\pm$0.028	00.21 $\pm$0.019	00.22 $\pm$0.025	00.23 $\pm$0.040	00.23
Length Model	00.30 $\pm$0.000	00.20 $\pm$0.000	00.30 $\pm$0.000	00.29 $\pm$0.019	00.27 $\pm$0.028	00.27
CL-C3G	62.91 $\pm$0.815	40.90 $\pm$0.500	36.63 $\pm$0.826	80.30 $\pm$0.703	53.29 $\pm$0.583	54.81
CL-CTS	58.00 $\pm$0.519	33.71 $\pm$0.382	29.87 $\pm$0.815	67.51 $\pm$1.050	44.95 $\pm$1.157	46.81
CL-ASA	23.33 $\pm$0.724	23.39 $\pm$0.432	33.14 $\pm$0.936	26.49 $\pm$1.205	55.50 $\pm$0.681	32.37
CL-ESA	64.89 $\pm$0.664	23.78 $\pm$0.613	14.03 $\pm$0.997	23.14 $\pm$0.777	14.19 $\pm$0.590	28.01
T+MA	58.22 $\pm$0.756	39.13 $\pm$0.551	28.61 $\pm$0.597	73.14 $\pm$0.666	36.95 $\pm$1.502	47.21
Average	53.47	32.18	28.46	54.12	40.98	

Table 4: Average F_1 scores and confidence intervals of state-of-the-art methods applied on the chunk-level EN-FR sub-corpora. The last row is the average F_1 scores from *CL-C3G* to *T+MA*.

Methods	Wikipedia (%)	TALN (%)	JRC (%)	APR (%)	Europarl (%)	Overall (%)
Random Baseline	00.21 $\pm$0.019	00.22 $\pm$0.025	00.23 $\pm$0.029	00.22 $\pm$0.025	00.24 $\pm$0.030	00.22
Length Model	00.30 $\pm$0.000	00.30 $\pm$0.000	00.30 $\pm$0.000	00.30 $\pm$0.000	00.30 $\pm$0.000	00.30
CL-C3G	48.25 $\pm$0.349	48.08 $\pm$0.538	36.68 $\pm$0.693	61.10 $\pm$0.581	52.72 $\pm$0.866	49.37
CL-CTS	46.68 $\pm$0.437	38.67 $\pm$0.552	28.21 $\pm$0.612	50.82 $\pm$1.034	53.21 $\pm$0.601	43.52
CL-ASA	27.63 $\pm$0.330	27.25 $\pm$0.341	35.17 $\pm$0.644	25.53 $\pm$0.795	36.55 $\pm$1.139	30.43
CL-ESA	51.14 $\pm$0.875	14.25 $\pm$0.334	14.44 $\pm$0.341	13.93 $\pm$0.714	13.91 $\pm$0.618	21.53
T+MA	50.57 $\pm$0.888	37.79 $\pm$0.364	32.36 $\pm$0.369	61.94 $\pm$0.756	37.92 $\pm$0.552	44.12
Average	44.85	33.21	29.37	42.66	38.86	

Table 5: Average F_1 scores and confidence intervals of state-of-the-art methods applied on the sentence-level EN-FR sub-corpora. The last row is the average F_1 scores from *CL-C3G* to *T+MA*.

Methods	Wikipedia (%)	TALN (%)	JRC (%)	APR (%)	Europarl (%)	Overall (%)
Random Baseline	00.21 $\pm$0.019	00.21 $\pm$0.019	00.22 $\pm$0.025	00.23 $\pm$0.028	00.21 $\pm$0.019	00.22
Length Model	00.23 $\pm$0.028	00.32 $\pm$0.025	00.32 $\pm$0.046	00.37 $\pm$0.028	00.32 $\pm$0.037	00.31
CL-C3G	51.58 $\pm$1.942	48.67 $\pm$1.662	37.91 $\pm$1.096	57.55 $\pm$1.103	53.86 $\pm$1.330	49.91
CL-CTS	48.45 $\pm$1.867	38.33 $\pm$1.494	27.16 $\pm$0.699	50.60 $\pm$1.771	55.19 $\pm$1.376	43.95
CL-ASA	33.87 $\pm$1.181	26.42 $\pm$1.400	34.08 $\pm$0.944	34.43 $\pm$1.813	36.59 $\pm$1.236	33.08
CL-ESA	53.44 $\pm$1.516	18.03 $\pm$1.261	12.93 $\pm$1.074	13.67 $\pm$0.995	11.73 $\pm$0.963	21.96
T+MA	55.82 $\pm$2.344	34.84 $\pm$1.049	27.27 $\pm$0.771	47.49 $\pm$2.130	32.80 $\pm$1.340	39.64
Average	48.63	33.26	27.87	40.75	38.03	

Table 6: Average F_1 scores and confidence intervals of state-of-the-art methods applied on the document-level EN-FR sub-corpora. The last row is the average F_1 scores from *CL-C3G* to *T+MA*.

Method	Time
Random Baseline	$\simeq$ 3”
Length Model	$\simeq$ 12”
CL-C3G	$\simeq$ 9”
CL-CTS	$\simeq$ 6’14”
CL-ASA	$\simeq$ 3’18”
CL-ESA	$\simeq$ 41’58”
T+MA	$\simeq$ 20’02”

Table 7: Comparison of execution times for each method applied on 1,000 $\times$ 1,000 textual units sizing from 35 to 55 words.

Table 7 lists the execution times of methods for the comparison of 1,000 $\times$ 1,000 textual units sizing from 35 to 55 words. The methods which require access to external resources and those making numerous vector calculations are the most expensive in time in addition to being the most expensive in memory resources consumption.

The evaluation was parallelized with a queuing mechanism (which explains the relatively long time of the baseline methods) and carried out on a Linux Debian server[11].

6. Results and Discussion

The *Length Model* show very poor performance (close to the *Random Baseline* with $\leq 0.31\%$) due to the choice of a large M. The latter greatly increases the number of potential false positives and thus negatively affects accuracy of baseline methods. The rest of the results confirm the state-of-the-art (Franco-Salvador et al., 2016; Potthast et

[11] 16-core AMD Opteron clocked at 2,0 GHz with 3,0 Go of RAM

al., 2011). *CL-ESA* seems to show better results on comparable corpora, like *Wikipedia*. In contrast, *CL-ASA* obtains better results on parallel corpora such as *JRC*, *Europarl* or *APR* collections. *CL-C3G* is in general the most effective method, as long as the corpus includes named entities. *CL-CTS* and *T+MA* are pretty efficient and versatile too. *CL-ESA* is not very effective; it is the more time-consuming method (see Table 7) and it is highly dependent of the corpus used. Some irregularities are present in the results, due to the fact that to build the chunk- and sentence- level, we carried out, as explained in section 4.2, a realignment and a length limitation of textual units that transformed some comparable units in parallel units.

Note that the performances of the methods using external resources such as ontologies, dictionaries or corpora, are extremely dependent of these resources. It is also important to note that the confidence intervals are larger on the document-level (with an average of 1.37% against 0.61% for the sentence-level and 0.76% for the chunk-level) because during the evaluation of this granularity, the number N of evaluated units is such that $N \neq |S|$ but N = 2,000. There is a strong correlation between the results of methods on the three granularities (average of 0.938), except between the chunk- and sentence- level for *CL-CTS* (0.757) and between the sentence- and the document- level for *CL-ASA* (0.493). Some methods on some sub-corpora are more efficient on fairly small textual units (*CL-C3G* on *Wikipedia* sub-corpus) while other methods are more efficient on longer units (*CL-C3G* on TALN sub-corpus), although the average best results are obtained at the chunk-level. Generally, all the methods see their performances gradually deteriorate as the granularity of compared documents increases, however we also see that many methods see their performances stagnated between the sentence and document level (*CL-CTS* or *CL-ESA* for example). Also, the results tend to be better on *Wikipedia*, *APR* and *Europarl* corpora because the ratio of named entities present in these corpora is more important (see Table 2).

The trend of the results on parallel corpora commonly used in evaluation tasks (e.g. *JRC*, *APR* and *Europarl*), at the sentence- and document- level, correlate very well (0.875) with scientific papers sub-corpus (*TALN*). This suggests that a method efficient on *JRC* and *Europarl* corpora should be useful for cross-language similarity detection on scientific papers.

7. Conclusion and Perspectives

In conclusion, our results confirm that the different methods of the state-of-the-art behave differently depending on the characteristics of the compared texts but also that the granularity impacts their performances. Our dataset may be interesting for future evaluation tasks and is made available on GitHub (`http://github.com/FerreroJeremy/Cross-Language-Dataset`).

In future works, we would like to include in our dataset, sub-corpora with more extreme percentage of named entities (one sub-corpus close to 0% and another one with more than 10% for example) in order to verify the impact of this feature on the effectiveness of the detection methods. We would also like to add an intermediate granularity, between the chunk-level and the sentence-level, that will not only consists of noun chunks but also includes verbal and adverbial phrases. Also, we have plans to develop a corpus builder tool, which generates, from cross-language dataset, a corpus to evaluate plagiarism detection and not just textual similarity detection, i.e. a corpus which will takes into account the granularity of the plagiarized excerpts as *PAN* corpus does (Potthast et al., 2010). Finally, our short term goal is to work on the improvement of the similarity detection methods by fusion, boosting or introduction of new approaches (using word embeddings for instance).

8. Bibliographical References

Barrón-Cedeño, A., Rosso, P., Pinto, D., and Juan, A. (2008). On Cross-lingual Plagiarism Analysis using a Statistical Model. In Benno Stein and Efstathios Stamatatos and Moshe Koppel, editor, *Proceedings of the ECAI'08 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse*, pages 9–13.

Barrón-Cedeño, A., Eiselt, A., and Rosso, P. (2009). Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions. In Sharma, et al., editors, *Proceedings of the 7th International Conference on Natural Language Processing (ICON'09)* , pages 29–38. Macmillan Publishers.

Barrón-Cedeño, A., Rosso, P., Agirre, E., and Labaka, G. (2010). Plagiarism Detection across Distant Language Pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 37–45. Association for Computational Linguistics.

Barrón-Cedeño, A. (2012). On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism. In *PhD thesis*, València, Spain.

Boudin, F. (2013). TALN Archives: a digital archive of French research articles in Natural Language Processing (TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue) [in French]). In *Proceedings of TALN 2013 (Volume 2: Short Papers)*, pages 507–514.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. In *Computational Linguistics*, volume 19, pages 263–311, Cambridge, MA, USA. MIT Press.

Cavnar, W. B. and Trenkle, J. M. (1994). N-Gram-Based Text Categorization. In *Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*, pages 161–175.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268.

Eurovoc. (1995). Thesaurus Eurovoc. In *Volume 2: Subject-Oriented Version*.

Franco-Salvador, M., Rossoa, P., and y Gómez, M. M. (2016). A systematic study of knowledge graph analysis for cross-language plagiarism detection. In *Information Processing and Management*.

Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, pages 1606–1611. Morgan Kaufmann Publishers Inc.

Guibert, P. and Michaut, C. (2011). Le plagiat étudiant. In *Education et sociétés*, volume 28, page 214. De Boeck Supérieur.

Gupta, P., Barrón-Cedeño, A., and Rosso, P. (2012). Cross-language High Similarity Search using a Conceptual Thesaurus. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, pages 67–75. Springer Berlin Heidelberg.

Josephson Institute. (2011). WHAT WOULD HONEST ABE LINCOLN SAY? In *Installment 2: Honesty and Integrity - The Ethics of American Youth: 2010, study by Josephson Institute of Ethics' Report Card on American Youth's Values and Actions*.

Kent, C. K. and Salim, N. (2010). Web Based Cross Language Plagiarism Detection. In *Second International Conference on Computational Intelligence, Modelling and Simulation (CIMSiM)*, pages 199–204. IEEE.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand.

McCabe, D. (2010). Students' cheating takes a high-tech turn. In *Rutgers Business School*.

Mcnamee, P. and Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. In *Information Retrieval Proceedings*, volume 7, pages 73–97. Kluwer Academic Publishers.

Muhr, M., Kern, R., Zechner, M., and Granitzer, M. (2010). External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System - Lab Report for PAN at CLEF 2010. In Martin Braschler, et al., editors, *CLEF Notebook*.

Pataki, M. (2012). A New Approach for Searching Translated Plagiarism. In *Proceedings of the 5th International Plagiarism Conference*, Newcastle, UK.

Pinto, D., Civera, J., Juan, A., Rosso, P., and Barrón-Cedeño, A. (2009). A Statistical Approach to Crosslingual Natural Language Tasks. In *CEUR Workshop Proceedings*, volume 64 of *Journal of Algorithms*, pages 51–60.

Plagiarism.org. (2014). What is plagiarism? Website. `http://www.plagiarism.org/plagiarism-101/what-is-plagiarism`.

Potthast, M., Stein, B., and Anderka, M. (2008). A Wikipedia-Based Multilingual Retrieval Model. In *30th European Conference on IR Research*, volume 4956 of *LNCS of Lecture Notes in Computer Science*, pages 522–530. Springer.

Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. (2010). An Evaluation Framework for Plagiarism Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, Beijing, China.

Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011). Cross-Language Plagiarism Detection. In *Language Ressources and Evaluation*, volume 45, pages 45–62.

Pouliquen, B., Steinberger, R., and Ignat, C. (2003). Automatic Identification of Document Translations in Large Multilingual Document Collections. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'03)*, pages 401–408.

Prettenhofer, P. and Stein, B. (2010). Cross-language Text Classification Using Structural Correspondence Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL'10, pages 1118–1127, Uppsala, Sweden.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Sérasset, G. (2015). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. In *Semantic Web Journal (special issue on Multilingual Linked Open Data)*, volume 6, pages 355–361.

Varga, D., Hálacsy, P., Nagy, V., Németh, L., Kornai, A., and Trón, V. (2005). Parallel corpora for Medium Density Languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596.

Česka, Z., Toman, M., and Jezek, K. (2008). Multilingual Plagiarism Detection. In *Artificial Intelligence: Methodology, Systems, and Applications*, volume 5253 of *Lecture Notes in Computer Science*, pages 83–92. Springer Berlin Heidelberg.

9. Language Resource References

Dániel Varga and Péter Hálacsy and Viktor Nagy and László Németh and András Kornai and Viktor Trón. (2005). *Hunalign sentence aligner*. Centre for Media Research and Education, version 1.1.

Florian Boudin. (2013). *TALN Archives: a digital archive of French research articles in Natural Language Processing*. version 18.11.2014.

Gilles Sérasset. (2012). *DBnary*.

Helmut Schmid. (1994). *TreeTagger - a language independent part-of-speech tagger*. TC project at the Institute for Computational Linguistics of the University of Stuttgart.

Martin Potthast and Alberto Barrón-Cedeño and Benno Stein and Paolo Rosso. (2009). *CL-PL-09 corpus*. Natural Language Engineering Lab, version 1.0.

Martin Potthast and Alberto Barrón-Cedeño and Benno Stein and Paolo Rosso. (2011). *PAN-PC-11 corpus*. Bauhaus-Universität Weimar & Universidad Politécnica de Valencia, version 1.0.

Peter Prettenhofer and Benno Stein. (2010). *Webis-CLS-10 corpus*. Bauhaus-Universität Weimar, version 11.5.2010.

Philipp Koehn. (2005). *Europarl: European Parliament Proceedings Parallel Corpus*. European Language Resources Association (ELRA), version 6.0.

Ralf Steinberger. (2011). *JRC-ACQUIS Multilingual Parallel Corpus*. European Commission's Joint Research Centre, version 3.0, ISLRN 821-325-977-001-1.