Barrón Cedeño, LA.; Gupta, PA.; Rosso ., P. (2013). Methods for cross-language plagiarism detection. Knowledge-Based Systems. 50:211-217. doi:10.1016/j.knosys.2013.06.018.

Methods for Cross-Language Plagiarism Detection

Alberto Barrón-Cedeño[a,b,*], Parth Gupta[c], Paolo Rosso[c]

[a] *Talp Research Center, Universitat Politècnica de Catalunya, Spain*
[b] *Facultad de Informática, Universidad Politécnica de Madrid, Spain*
[c] *NLE Lab-ELiRF, Universitat Politècnica de València, Spain*

Abstract

Three reasons make plagiarism across languages to be on the rise: (*i*) speakers of under-resourced languages often consult documentation in a foreign language, (*ii*) people immersed in a foreign country can still consult material written in their native language, and (*iii*) people are often interested in writing in a language different to their native one. Most efforts for automatically detecting cross-language plagiarism depend on a preliminary translation, which is not always available.

In this paper we propose a freely available architecture for plagiarism detection across languages covering the entire process: heuristic retrieval, detailed analysis, and post-processing. On top of this architecture we explore the suitability of three cross-language similarity estimation models: Cross-Language Alignment-based Similarity Analysis (CL-ASA), Cross-Language Character n-Grams (CL-CNG), and Translation plus Monolingual Analysis (T+MA); three inherently different models in nature and required resources.

*Corresponding author
Email addresses: albarron@lsi.upc.edu (Alberto Barrón-Cedeño), pgupta@dsic.upv.es (Parth Gupta), prosso@dsic.upv.es (Paolo Rosso)

The authors appear in alphabetical order. A. Barrón and P. Gupta contributed equally to this work and should both be considered as first authors.

The three models are tested extensively under the same conditions on the different plagiarism detection sub-tasks—something never done before. The experiments show that T+MA produces the best results, closely followed by CL-ASA. Still CL-ASA obtains higher values of precision, an important factor in plagiarism detection when lesser user intervention is desired.
Keywords: automatic plagiarism detection, cross-language similarity

1. Introduction

Automatic plagiarism detection (PD) entails identifying plagiarised text fragments together with their source. The task is defined as follows. Let d_q be a suspicious document. Let D be a set of potential source documents. Determine whether a fragment $s_q \in d_q$ was borrowed from $s \in d$ ($d \in D$) (Potthast et al., 2009). Once $\{s_q, s\}$ are identified, an expert can determine whether a case of text re-use is indeed plagiarism (for instance, if no proper citation is provided). From a cross-language (CL) perspective, $d_q \in L$ and $d' \in L'$, where $L \neq L'$ are two languages. This is known as cross-language plagiarism detection (CLPD). Up to date, diverse approaches for PD in free text exist. However, few approaches are focused on CLPD.

For the first time, we empirically analyse how the different steps of CLPD can use Cross-Language Alignment-based Similarity Analysis (CL-ASA) (Barrón-Cedeño et al., 2008), Cross-Language Character n-grams (CL-CNG) (Mcnamee and Mayfield, 2004), or Translation plus Monolingual Analysis (T+MA) (Oberreuter et al., 2011). On the one hand, both CL-ASA and CL-CNG had been tested against collections of parallel documents that were "assumed" to contain cases of plagiarism (Barrón-Cedeño et al., 2010; Pot-

thast et al., 2011a). However, Barrón-Cedeño et al. (2010) just tried to identify sentences' translations from translation memories, whereas Potthast et al. (2011a) aimed at retrieving document translations from a parallel corpus. On the other hand, comparing two documents after translation is *in vogue* in the plagiarism detection community (Potthast et al., 2011b).

The focus of our research is two-fold. Firstly, we propose a freely available software architecture for cross-language plagiarism detection. Secondly, we explore the performance and suitability of three similarity models over different types of CL plagiarism in terms of length and kind of translation (automatic translation and automatic translation plus manual paraphrasing). For the first time, this comparison is carried out on top of a common detection architecture, allowing for a better appreciation of strengths and weaknesses —an analysis hardly carried out before. CL-ASA, CL-CNG, and T+MA are challenged with the different scenarios of CLPD, i.e., looking for: (*i*) entirely plagiarised documents and their source (document-level detection), (*ii*) plagiarised and source fragments within document pairs (fragment-level detection), and (*iii*) plagiarised and source fragments within an entire collection of suspicious and potential source documents (entire detection process). To the best of our knowledge, no research work has analysed these scenarios in depth before. CL-ASA's results are encouraging: it is competitive, even using limited dictionaries, and outperforms CL-CNG when facing different kinds of plagiarism. It is roughly comparable to T+MA, without relying on a translation module, but with much higher precision, causing the work load of the human reviewer to decrease.

The rest of the paper is structured as follows. Section 2 describes the

prototypical steps for CLPD and offers an overview of the state of the art of plagiarism detection systems, paying special attention to CL similarity measures. In Section 3 we define the architecture of our CLPD system and the three similarity measures we explored. Section 4 describes the experimental framework: corpus, evaluation measures, and proposed experiments. Results and discussion are included in Section 5. Finally, we draw some conclusions and discuss further work in Section 6.

2. Related Work

Recently, Potthast et al. (2011a) offered an overview of the prototypical CLPD process; referred as the entire plagiarism detection architecture:

(***i***) **Heuristic retrieval.** A set of candidate documents D^* is retrieved from D' ($|D^*| \ll |D'|$). D^* contains the most similar documents to d_q and, therefore, the most likely to contain the source of potential re-use.

(***ii***) **Detailed analysis.** d_q is compared against every $d' \in D^*$ section-wise. If a pair $\{s_q, s'\}$ is identified to be more similar than expected for independently generated texts, a potential case of plagiarism is located.

(***iii***) **Heuristic post-processing.** Plagiarism candidates that are not long or similar enough are discarded. Additionally, heuristics are applied to merge nearby candidates.[2]

[2]This stage was originally intended to filter false positives, such as cases of borrowing with proper citation (Stein et al., 2007).

Below we give an overview of a set of CL similarity assessment strategies that can be applied at steps (*i*) and (*ii*) . Our analysis extends that of Potthast et al. (2011a), who analysed a number of models for the detailed analysis stage and provided some hints on strategies for the heuristic retrieval stage, but did not go further in the analysis of the entire CLPD task, including an entire plagiarism detection architecture. We identify five model families:

a) Lexicon-based systems. They rely on lexical similarities between languages (e.g. English–French) and linguistic influence (e.g. English *computer* → Spanish *computadora*) between languages. Similarities across words in different languages can be reflected when composing short terms; e.g. character n-grams or prefixes. Probably two of the first similarity models of this kind are *cognateness* —based on prefixes and other tokens— (Simard et al., 1992) and *dot-plot* —based on character 4-grams (Church, 1993). While originally proposed to align bitexts, these models are useful to detect re-use across languages (Potthast et al., 2011a), with some limitations (Barrón-Cedeño et al., 2010). The Cross-Language Character N-Grams (CL-CNG) model, from this family, is considered in this research (cf. Section 3.1).

b) Thesaurus-based systems. These systems map words or concepts, such as named entities, into a common representation space by means of a multilingual thesaurus (e.g. Eurovoc (Steinberger et al., 2002) or EuroWordnet (Vossen, 1998)). However, multilingual thesauri are not always available; and Ceska et al. (2008) found that the incompleteness of the thesaurus (in that case EuroWordnet) may limit the detection capabilities.

c) Comparable corpus-based systems. These systems are trained over comparable corpora. One example is cross-language explicit semantic analysis (CL-ESA) (Potthast et al., 2008). d_q and d' are represented by a vector of similarities to the documents of a so-called CL index collection C_I, i.e., $\vec{d_q} = \{sim(d_q, c_1), \ldots sim(d_q, c_I)\}$, $\vec{d'} = \{sim(d', c'_1), \ldots sim(d', c'_I)\}$ ($c_i \in L, c'_i \in L'$), where sim is a monolingual similarity model, such as the cosine measure. $\vec{d_q}$ and $\vec{d'}$ are then compared to compute $sim(d_q, d')$.

d) Parallel corpus-based systems. These systems are trained on parallel corpora, either to find cross-language co-occurrences (Littman et al., 1998) or to obtain translation modules. The principles and resources of machine translation (MT) are applied, but no actual translation is performed. We consider one model of this family in the current research: CL-ASA (cf. Section 3.2).

e) Machine translation-based systems. These models are *in vogue* in CLPD (e.g. Corezola Pereira et al. (2010); Kent and Salim (2009); Nawab et al. (2010); Oberreuter et al. (2011)) and simplify the task by turning it into a monolingual problem. The prototypical process is as follows: (*i*) a language detector is applied to determine the most likely language of d_q; (*ii*) d_q is translated if not written in the comparison language; and (*iii*) a monolingual comparison is carried out between d'_q and d'. We also consider a model of this type, which we call T+MA (cf. Section 3.3).

We are particularly interested in the performance of this model. Systems that exploited Google Translator achieved good results in PAN 2011 (e.g., Grman and Ravas (2011)), but maybe because the same machine translator was used for both generation and detection. Such performance is optimistic and could deteriorate in a realistic setting, where the translations are generated

Algorithm 1: Given d_q and d'

$S_q \leftarrow \{split(d_q, w, t)\}$ $\quad$ $S' \leftarrow \{split(d', w, t)\}$ $\quad$ // **Detailed analysis**

for every $s_q \in S_q$:

$\quad P_{s_q,s'} \leftarrow \arg\max^5_{s' \in S'} sim(s_q, s')$

until no change: // **Post-processing**

$\quad$ for every combination of pairs $p_i, p_j \in P_{s_q,s'}$:

$\quad\quad$ if $\delta(p_i, p_j) < thres_1$:

$\quad\quad\quad merge\_fragments(p_i, p_j)$

return $\{p \in P_{s_q,s'} \mid |p| > thres_2\}$

Figure 1: Algorithm. Cross-language detailed analysis and post-processing.

by different (translation) systems. Here we apply a different machine translator to see whether different translation systems produce text different enough to make the monolingual comparison process more difficult: roughly equivalent to the detection of cases with a high density of paraphrasing, whose proper detection remains an open issue (Potthast et al., 2011b).

3. Detection and Similarity Analysis

Here we describe our cross-language plagiarism detection architecture[3] and the three CL similarity models we explore. Our strategy follows the schema depicted in Section 2. For CL heuristic retrieval, we select the top 50 $d' \in D'$ for each d_q according to $sim(d_q, d')$. CL detailed analysis and post-processing are performed as explained in Fig. 1.

[3]It is freely available for research purposes at `http://www.dsic.upv.es/grupos/nle/resources/clpd-code.tar.gz`

At detailed analysis, d_q (d') is split into chunks of length w with step t. We use $w = 5$ sentences and $t = 2$ aiming at considering chunks close to paragraphs. $sim(s_q, s')$ computes the similarity between the text fragments either on the basis of CL-ASA (Section 3.2), CL-CNG (Section 3.1), or T+MA(Section 3.3). $\arg\max^5_{s \in S}$ retrieves the 5 most similar fragments $s \in S$ with respect to s_q. The resulting candidate pairs $\{s_q, s\}$ are stored into $P_{s_q,s'}$, and they are the input for the post-processing step. If the distance in characters between two (highly similar) candidate pairs $\delta(p_i, p_j)$ is lower than threshold $thres_1 = 1,500$, p_i and p_j are merged. Only those candidates that are composed of at least three of the identified fragments ($thres_2$) are considered potentially plagiarised (thresholds defined empirically).

This is the core algorithm for our approach to plagiarism detection. The similarities between the texts can be based on any (cross-language) similarity estimation model. We explore three: CL-ASA, CL-CNG, and T+MA.

3.1. Cross-Language Character n-Grams

CL-CNG was originally proposed by Mcnamee and Mayfield (2004) for CLIR (with clear roots in bitext alignment (Church, 1993)). The text is case-folded, punctuation marks and diacritics are removed. Multiple white-space and new-line characters are replaced by a single white-space. Moreover, a single white-space is inserted at the beginning and end of the text. Finally, the resulting text strings are encoded into character n-grams as depicted below, where "-" should be considered as white-space and $n = 4$:

"El espíritu" $\rightarrow$ "-el-","el-e","␣l-es","-esp","espi","spir","piri","irit","ritu","itu-".

Similarity $sim(d_q, d')$ is estimated by the unigram language model:

$$sim(d_q, d') = P(d'|d_q) = \prod_{q \in d_q} [\alpha P(q|d') + (1 - \alpha) P(q|C)] \quad (1)$$

where $P(q|d')$ is the document level probability of term q in document d' and C denotes the entire collection. We use $n = 4$ and $\alpha = 0.7$ as these values yielded the best results for English–Spanish in the original work.

3.2. Cross-Language Alignment-based Similarity Analysis

Similarity $sim(d_q, d')$ is computed by estimating the likelihood of d' of being a translation of d_q. It is an adaptation of Bayes' rule for MT (Brown et al., 1993) that Barrón-Cedeño et al. (2008) defines as:

$$\varphi(d_q, d') = \varrho(d')\ p(d_q \mid d'). \quad (2)$$

where, $\varrho(d')$ is known as *length model* (λM). The length of the d's translation into d' is closely related to a *translation length factor*, defined as:

$$\varrho(d') = e^{-0.5\left(\frac{\frac{|d'|}{|d_q|} - \mu}{\sigma}\right)^2}, \quad (3)$$

where μ and σ are the mean and standard deviation of the character lengths between actual translations from L into L' (Pouliquen et al., 2003). If the length of d' is unexpected given d_q, it receives a low likelihood. We use the values estimated by Potthast et al. (2011a) for English–Spanish: $\mu = 1.138$, $\sigma = 0.631$. (At heuristic retrieval, λM is neglected, as the lengths of d_q and d' are independent from those of the specific borrowed fragments within them.)

In statistical MT, the conditional probability $p(d_q \mid d')$ is known as *translation model probability* (TM), computed on the basis of a statistical bilingual dictionary. The adaptation of this model is defined as:

$$\rho(d \mid d') = \sum_{x \in d} \sum_{y \in d'} p(x, y) \ , \tag{4}$$

which no longer represents a probability measure. The dictionary $p(x, y)$ defines the likelihood of word x of being a valid translation of y. We consider three strategies to estimate $p(x, y)$: (*i*) JRC —a dictionary computed from the JRC–Acquis parallel corpus (Steinberger et al., 2006), on the basis of the IBM M1 (Brown et al., 1993); (*ii*) INF —a dictionary of inflectional forms produced from a "traditional" bilingual dictionary, where all the possible inflectional forms of a word are considered (Sidorov et al., 2010); and (*iii*) STEM —a stemmed version of INF, where the weights are accumulated and distributed over the stems. We also explore the impact of considering only the k best translations for each word (those with the highest probabilities) up to a minimum probability mass of 0.20. These dictionaries are called [JRC|INF|STEM].*pm* where *pm* stands for the considered probability mass.

CL-ASA is considered a parallel corpus-based system. Firstly, its parameters are learnt from a parallel corpus. Secondly, every potential translation of a word participates in the similarity assessment, making it flexible.

3.3. Translation plus Monolingual Analysis

The first step of this approach is the translation of all the documents into a common language. We translate the documents from Spanish into English with Apertium, an open-source machine translation framework (Armentano-Oller et al., 2005). In the second step we discard stopwords and stem the documents (both translated and originally in English) with the Snowball

stemmer[4]. Afterwards, we weight the documents' terms with tf-idf and compare the texts using the cosine measure over a bag-of-words representation. When identifying specific plagiarised fragments, we use the original offsets of the documents in Spanish.

4. Evaluation Framework

In this section we describe the corpus, evaluation metrics, and proposed experiments.

4.1. Corpus

We use the PAN-PC-11 corpus (Potthast et al., 2011b), a standard corpus for PD. It contains a wide range of (synthetic) cases, from verbatim copy to different levels of paraphrasing up to translated cases. For the first time we explore how CL-ASA, CL-CNG, and T+MA perform on the PAN-PC-11 corpus and we do it with focus on the Spanish–English cases. German–English cases are not explored because we are not aware of any inflectional dictionary for this pair (and one of our objectives is to investigate the behaviour of CL-ASA with this kind of resource). The corpus partition comprises 304 suspicious and 202 potential source documents, including two types of borrowing: automatic translation (*auto*) and manually paraphrased automatic translation (*manual*). It is worth noting that, $s_q \in d_q$, the borrowed fragment, is in general on a different topic to that of d_q.

For experimental purposes, we use three corpus partitions (x represents the experiment C_x the partition is used in): **(*i*)** C_A is composed of the

[4] `http://snowball.tartarus.org/`

specific $\{s_q, s'\}$ pairs, considered as entire documents. This partition includes $2,920$ source and $2,920$ plagiarised documents (fragments). ***(ii)*** C_B includes the entire set of 304 suspicious and 202 potential source documents, with plagiarised fragments within them. The document d which d_q's borrowings come from is identified. ***(iii)*** C_C is composed as C_B, but no preliminary information about the source documents exist. These three partitions allow for analysing the performance of our model on all the scenarios of CLPD.[5]

4.2. Evaluation Metrics

We define three experiments to evaluate the different configurations of our detection system (cf. Section 4.3). In order to evaluate the retrieval-by-example task of Experiment A, we use recall at rank k ($rec@k$).[6] In experiments B and C we use versions of recall and precision fitted for evaluating whether a specific text fragment s_q (s) has been correctly labelled as plagiarised (source) (Potthast et al., 2010). Plagiarised fragments are treated as basic retrieval units, with $s_i \in S$ defining a query for which a detection system returns a result set $R_i \subseteq R$. Recall and precision are defined as:

$$prec(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|r \sqcap \cup_{s \in S} S|}{|r|} \ , \text{ and} \tag{5}$$

$$rec(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|s \sqcap \bigcup_{r \in R} r|}{|s|} \ , \tag{6}$$

[5]These three partitions are available for download at `http://www.dsic.upv.es/grupos/nle/resources/clpd-data.tar.gz`

[6]The only precision value that makes sense in the framework of experiment A is precision at rank 1 ($prec@1$), as only one relevant document exists for each query document (note that $rec@1 = prec@1$).

where $\sqcap$ computes the positional overlapping characters. The overall performance is evaluated with the harmonic mean of *prec* and *rec*, i.e., F_1-measure.

4.3. Proposed Experiments

We designed three experiments to investigate the performance of CL-ASA, CL-CNG, and T+MA on the different CLPD steps and scenarios:

Experiment A. We are given d_q and D'; d_q is entirely plagiarised from $d' \in D'$, and the task is to find d'. This depicts the scenario where almost the whole document is plagiarised from one source. Moreover, it is an approximation to the scenario where fragment $s_q \in d_q$ and d_q are on the same topic (this does not occur in PAN-PC-11). This experiment is used to tune the parameters of CL-ASA: exploring different dictionaries, probability masses, and the inclusion of the λM length model. CL-CNG and T+MA are used as proposed in the related literature.

Experiment B. We are given d_q and d' and the task is finding $s_q \in d_q$ and $s' \in d'$ such that s_q is a plagiarised fragment from s'. This depicts the scenario where d_q and d' are already identified and we aim at locating the borrowed text fragments, i.e., the detailed analysis stage.

Experiment C. We are given d_q and D' and the task is finding $s_q \in d_q$ and $s' \in d'$ ($d' \in D'$) where s_q is plagiarised from s'. This depicts the scenario of the overall PD process. We use the same heuristic retrieval process for all the three models so that we can better analyse them when there is noise in the candidate list.

5. Results and Discussion

This study mainly aims to compare a wide variety of cross-language similarity models across different scenarios of plagiarism detection. The resources required by the models are inherently different: T+MA needs a complete MT system, CL-ASA requires parallel corpora to estimate a bilingual dictionary and a length model, and CL-CNG is a crude model which does not depend on any resource.[7]

Experiment A. The tuning results for CL-ASA are shown in Figure 2. The best option is using the JRC dictionary with only 20% of the probability mass. That is, JRC.20, with limited (potentially biased) vocabulary, performs better than the vocabulary-rich dictionaries INF and STEM.[8] The best results with INF and STEM come with $mass = 1.0$; i.e., the entire dictionary. On the one hand, JRC is empirically generated from a parallel corpus. As a result, noisy entries (with low probabilities) are included. Reducing the probability mass is roughly equivalent to discarding such noisy entries. On the other hand, INF and STEM come from traditional dictionaries, and every entry is presumably a correct translation. As expected, considering λM empowers the similarity assessment capabilities of CL-ASA. Therefore, the best CL-ASA parameters are JRC.20 with λM.

Figure 3 displays the curves obtained with the three similarity models:

[7] Transliteration may be required if the languages do not share the same script, though.

[8] CL-ASA does not contemplate any senses discrimination among the potential translations as it aims to make a flexible comparison: every potential translation is considered. This decision could introduce noise to the similarity estimation, but the context of the fragment helps to minimise the impact of potential ambiguities.

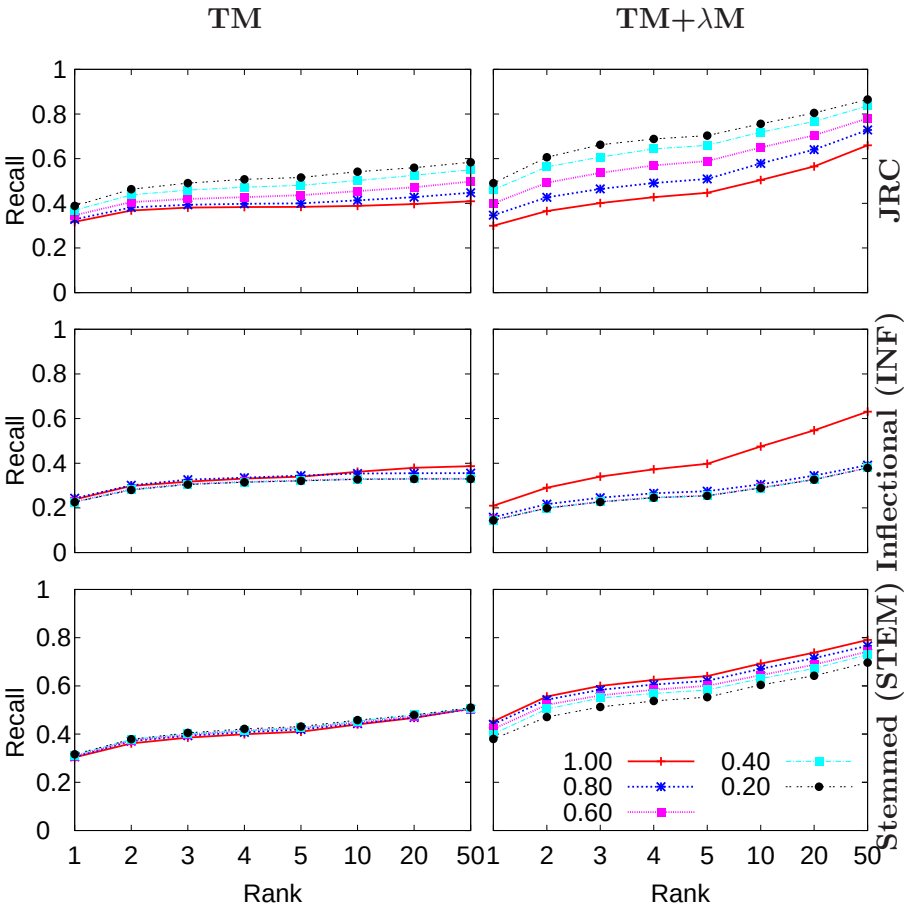


Figure 2: Results for CL-ASA tuning with the three dictionaries: JRC, Inflectional and Stemmed. In the top (bottom) plots the TM (TM and λM) is (are) applied. The best CL-ASA parameters are JRC dictionary, probability mass of 0.20, and length model.

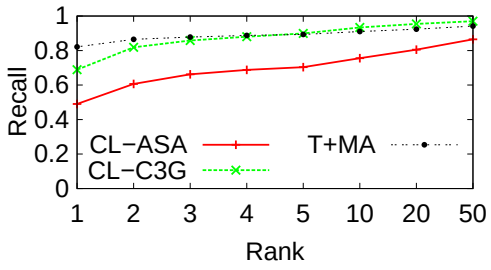
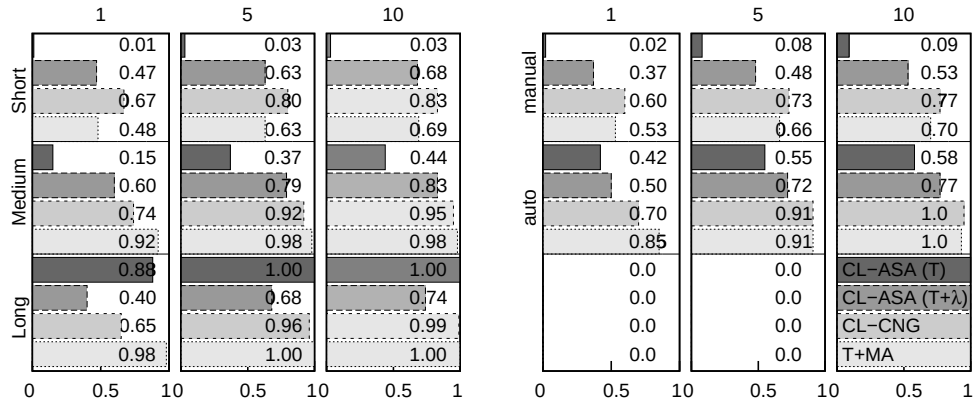


Figure 3: Experiment A. Comparison of CL-ASA, CL-CNG, and T+MA.

CL-ASA, CL-CNG, and T+MA. It is worth noting that the ceiling of R@1 is around 0.8, the value obtained by T+MA. When looking for entirely plagiarised documents, T+MA shows to be the best option.

We further analyse the results of the three models for different lengths and types of plagiarism. The results for short, medium, and long documents are displayed in Figure 4(a). A few aspects are worth noting: (*i*) as expected, the small amount of information available for CLPD systems when considering short texts harm their performance; (*ii*) for both CL-CNG and T+MA, the longer the document the better the performance; and (*iii*) the length model causes the accuracy of CL-ASA to decrease for long documents. The last point is in agreement with the results obtained by Barrón-Cedeño et al. (2010): CL-ASA is sensitive to the amount of information it can exploit. The more text, the better the TM performs, but not so for the λM length model. This can be caused by the length model original purpose: grabbing potential translations at the sentence level. When facing sentences, the length variations are somehow limited by relatively short texts. When analysing long documents (in our case from a bunch of paragraphs up to entire documents), the length of the resulting translations can cause the "expected" length to



(a) Length-wise (b) Type-wise

Figure 4: Performance analysis in terms of $rec@k$, considering $k = \{1, 5, 10\}$ (experiment A). CL-ASA applied with JRC.20 and either translation model only (T, darkest bar) or translation and length models (T+λ, dark bar). CL-CNG and T+MA are displayed with clearer bars (see legends at bottom right).

become very different. Still this is an issue for further research.

Our last comparison of experiment A regards to determine how the models perform when dealing with *auto* and *manual* —i.e., further paraphrased— cases (cf. Section 4.1). As Figure 4(b) shows, *manual* cases are harder to detect. CL-CNG emerges to be the best when looking for further paraphrased cases. As the best CL-ASA results are in general obtained with JRC.20 and including λM, we use this version in experiments B and C.

Experiment B. Figure 5 shows the results of the detailed analysis experiment. T+MA obtains the best recall both overall and type-wise, with CL-ASA close behind in most cases. The precision of CL-ASA outperforms the other models regardless of the length or nature of the cases. The best results of CL-ASA are obtained with long plagiarism cases, which does not contradict

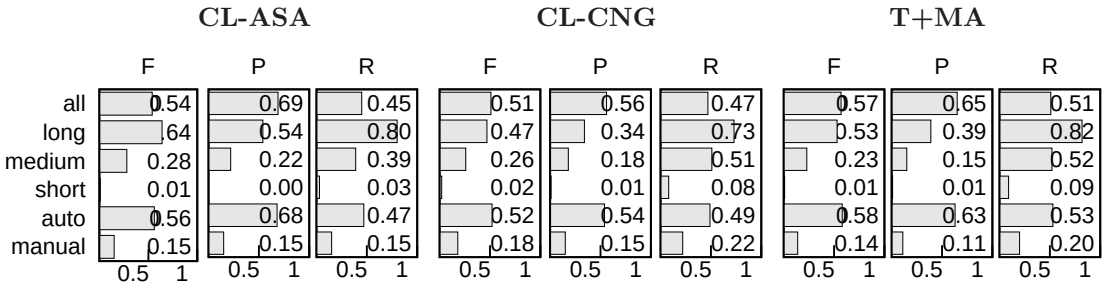


Figure 5: Performance of CL-ASA, CL-CNG, and T+MA on the detailed analysis problem in terms of F_1 measure, Precision, and Recall (experiment B).

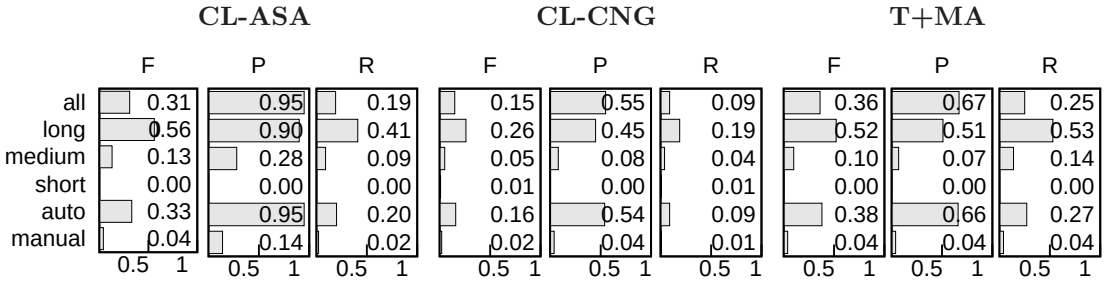


Figure 6: Performance of CL-ASA and CL-CNG on the overall plagiarism detection task in terms of F_1 measure, Precision, and Recall (experiment C).

the previous results: the compared chunks are of fixed length: five sentences.

Our heuristic to determine if an actual case of plagiarism is at hand causes short cases to go unnoticed regardless of the similarity model; since the algorithm needs evidence in terms of matching consecutive chunks (cf. Section 3). Moreover, as already observed in experiment A, shorter cases are the most difficult to uncover and most paraphrased translation cases (*manual*; by far the hardest to detect) in PAN-PC-11 are short.

Experiment C. In this experiment we want to analyse how the models behave when facing a noisy set of potential source documents. Hence, the

heuristic retrieval stage —fetching 50 candidate documents from the source collection for each suspicious document—, for the three models is performed with CL-ASA. The performance of the CL-ASA-based heuristic retrieval, i.e., properly including the source document of a case within the 50 retrieved documents, is 31%. The source and plagiarised documents in the PAN-PC-11 are not on common topics; the corpus contains cases of plagiarism inserted in randomly selected documents (something unexpected in real scenarios). In the candidate retrieval step, the system considers the contents of the entire document. As a result, the accuracy is affected in experiment C, in contrast to the other experiments, where no candidate retrieval is performed. The expected results of this task would be better in real scenarios, where d_q and d are documents on similar topics (as supported by the results of experiment A).

The results obtained after the overall process (including detailed analysis and post-processing) are presented in Fig. 6. The results are similar to those of experiments A and B: T+MA obtains a better F_1 in all but long cases, where CL-ASA outperforms it. Interestingly in this case CL-CNG falls far behind, regardless of the length or nature of the case. As expected after experiment A, neither CL-ASA, CL-CNG, nor T+MA detect short cases. Our strategy to identify the correct boundaries of the plagiarism cases (cf. Section 3) works generally well if the detailed analysis strategy is able to properly retrieve the plagiarised–source fragments, leading to CL-ASA to achieve high values of precision. The relatively lower precision levels of T+MA may be caused by mistaken translations during the language normalisation stage.

A revelation of this work is the suitability of these models for application-oriented necessities. T+MA and CL-CNG show the signs of a recall-oriented system, whereas CL-ASA is more suitable when precision is more important. Indeed, CL-ASA shows very robust precision in experiment C—a realistic scenario, with noisy source candidates. In general CL-ASA obtains a low amount of false-positives which is possible when the model shows high confidence in estimating similarities. After this study a user could set preferences for the similarity model based on the application at hand.

Another major contribution of this work is the fragment–level plagiarism detection algorithm, which works on the principles of maximum votes (i.e., neighbour text fragments have to be "voted" together as plagiarism suspicion to consider them a potential case). The algorithm is very robust which is supported by the high precision achieved in *experiment-C* for CL-ASA when the similarities for consecutive plagiarised paragraphs are estimated with high confidence.

6. Conclusions and Future Work

Automatic plagiarism detection models aim to provide experts (e.g. forensic linguists and professors) with evidence for taking decisions about potential cases of unauthorised text re-use. In this paper, we studied the performance of a cross-language plagiarism detection architecture when relying on different similarity estimation models.

Different similarity estimation models can be plugged into our freely-available architecture. In particular, we experimented with three: cross-language alignment-based similarity analysis, cross-language character n-

grams, and translation plus monolingual analysis. Our strategy was tested extensively on a set of experiments reflecting different steps and scenarios of cross-language plagiarism detection: from the detection of entirely plagiarised documents to the identification of specific borrowed text fragments. The similarity models showed a remarkable performance when detecting plagiarism of entire documents, including further paraphrased translations. When aiming at detecting specific borrowed fragments and their source, both short and further paraphrased cases caused difficulties. Still the precision of cross-language alignment-based similarity analysis was always high (for some types higher than 0.9). As a result, if it identifies a potential case of plagiarism, it is certainly worth analysing it.

As future work, we aim to improve our heuristic retrieval module, i.e., retrieving good potential source documents for a possible case of plagiarism. This is a complicated task as, to the best of our knowledge, no large scale cross-language corpus with the necessary characteristics exists.

References

Armentano-Oller, C., Corbí-Bellot, A., Forcada, M., Ginestí-Rosell, M., Bonev, B., Ortiz-Rojas, S., Pérez-Ortiz, J., Ramírez-Sánchez, G., Sánchez-Martínez, F., September 2005. An Open-Source Shallow-Transfer Machine Translation Toolbox: Consequences of its Release and Availability. In: OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X. pp. 23–30.

Barrón-Cedeño, A., Rosso, P., Agirre, E., Labaka, G., August 2010. Pla-

giarism detection across distant language pairs. In: Huang and Jurafsky (2010).

Barrón-Cedeño, A., Rosso, P., Pinto, D., Juan, A., 2008. On Cross-Lingual Plagiarism Analysis Using a Statistical Model. In: Stein, B., Stamatatos, E., Koppel, M. (Eds.), ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008). Vol. 377. CEUR-WS.org, Patras, Greece, pp. 9–13, http://ceur-ws.org/Vol-377.

Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R., 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics 19 (2), 263–311.

Ceska, Z., Toman, M., Jezek, K., 2008. Multilingual Plagiarism Detection. In: Proceedings of the 13th International Conference on Artificial Intelligence (ICAI 2008). Springer-Verlag, Varna, Bulgaria, pp. 83–92.

Church, K., 1993. Char_align: A Program for Aligning Parallel Texts at the Character Level. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL 1993). ACL, Columbus, OH, USA, pp. 1–8.

Corezola Pereira, R., Moreira, V., Galante, R., 2010. A New Approach for Cross-Language Plagiarism Analysis. Multilingual and Multimodal Information Access Evaluation, International Conference of the Cross-Language Evaluation Forum LNCS (6360), 15–26, Springer-Verlag.

Grman, J., Ravas, R., Sep. 2011. Improved Implementation for Finding Text

Similarities in Large Collections of Data - Notebook for PAN at CLEF 2011. In: Petras et al. (2011).

Huang, C.-R., Jurafsky, D. (Eds.), August 2010. Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). COLING 2010 Organizing Committee, Beijing, China.

Kent, C., Salim, N., Dec. 2009. Web Based Cross Language Plagiarism Detection. Journal of Computing 1 (1), 39–43.

Littman, M., Dumais, S., Landauer, T., 1998. Cross-Language Information Retrieval, chapter 5. Kluwer Academic Publishers, Ch. Automatic Cross-language Information Retrieval Using Latent Semantic Indexing, pp. 51–62.

Mcnamee, P., Mayfield, J., 2004. Character N-Gram Tokenization for European Language Text Retrieval. Information Retrieval 7 (1-2), 73–97.

Nawab, R., Stevenson, M., Clough, P., Sep. 2010. University of Sheffield lab report for PAN at CLEF 2010. In: Braschler, M., Harman, D. (Eds.), Notebook Papers of CLEF 2010 LABs and Workshops. Padua, Italy.

Oberreuter, G., L'Huillier, G., Ríos, S. A., Velásquez, J. D., 2011. Approaches for intrinsic and external plagiarism detection - notebook for pan at clef 2011. In: Petras, V., Forner, P., Clough, P. D. (Eds.), CLEF (Notebook Papers/Labs/Workshop).

Petras, V., Forner, P., Clough, P. (Eds.), Sep. 2011. Notebook Papers of CLEF 2011 LABs and Workshops. Amsterdam, The Netherlands.

Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P., 2011a. Cross-language plagiarism detection. Language Resources and Evaluation (LRE), Special Issue on Plagiarism and Authorship Analysis 45 (1), 1–18.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P., Sep. 2011b. Overview of the 3rd International Competition on Plagiarism Detection. In: Petras et al. (2011).

Potthast, M., Stein, B., Anderka, M., 2008. A Wikipedia-Based Multilingual Retrieval Model. Advances in Information Retrieval, 30th European Conference on IR Research LNCS (4956), 522–530, springer-Verlag.

Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P., August 2010. An evaluation framework for plagiarism detection. In: Huang and Jurafsky (2010), pp. 997–1005.

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P., 2009. Overview of the 1st international competition on plagiarism detection. Vol. 502. CEUR-WS.org, San Sebastian, Spain, pp. 1–9, http://ceur-ws.org/Vol-502.

Pouliquen, B., Steinberger, R., Ignat, C., 2003. Automatic Identification of Document Translations in Large Multilingual Document Collections. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003). Borovets, Bulgaria, pp. 401–408.

Sidorov, G., Barrón-Cedeño, A., Rosso, P., 2010. English-Spanish Large Statistical Dictionary of Inflectional Forms. In: Calzolari, N., Choukri, K.,

Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., Tapias, D. (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010). Valletta, Malta.

Simard, M., Foster, G. F., Isabelle, P., 1992. Using Cognates to Align Sentences in Bilingual Corpora. In: Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation.

Stein, B., Meyer zu Eissen, S., Potthast, M., 2007. Strategies for Retrieving Plagiarized Documents. In: Clarke, C., Fuhr, N., Kando, N., Kraaij, W., de Vries, A. (Eds.), Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Amsterdam, The Netherlands, pp. 825–826.

Steinberger, R., Pouliquen, B., Hagman, J., 2002. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. Computational Linguistics and Intelligent Text Processing. Proceedings of the CICLing 2002 LNCS (2276), 415—424, springer-Verlag.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D., 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In: Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. (Eds.), Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy.

Vossen, P. (Ed.), 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers.