Language Resources and Evaluation
DOI: 10.1007/s10579-009-9114-z

Cross-Language Plagiarism Detection

**Martin Potthast · Alberto Barrón-Cedeño ·
Benno Stein · Paolo Rosso**

Published: January 30, 2010

Abstract Cross-language plagiarism detection deals with the automatic identification and extraction of plagiarism in a multilingual setting. In this setting, a suspicious document is given, and the task is to retrieve all sections from the document that originate from a large, multilingual document collection. Our contributions in this field are as follows: (*i*) a comprehensive retrieval process for cross-language plagiarism detection is introduced, highlighting the differences to monolingual plagiarism detection, (*ii*) state-of-the-art solutions for two important subtasks are reviewed, (*iii*) retrieval models for the assessment of cross-language similarity are surveyed, and, (*iv*) the three models CL-CNG, CL-ESA and CL-ASA are compared.

Our evaluation is of realistic scale: it relies on 120 000 test documents which are selected from the corpora JRC-Acquis and Wikipedia, so that for each test document highly similar documents are available in all of the 6 languages English, German, Spanish, French, Dutch, and Polish. The models are employed in a series of ranking tasks, and more than 100 million similarities are computed with each model. The results of our evaluation indicate that CL-CNG, despite its simple approach, is the best choice to rank and compare texts across languages if they are syntactically related. CL-ESA almost matches the performance of CL-CNG, but on arbitrary pairs of languages. CL-ASA works best on "exact" translations but does not generalize well.

Keywords
cross-language · plagiarism detection · similarity · retrieval model · evaluation

This work was partially supported by the TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 project and the CONACyT-Mexico 192021 grant.

M. Potthast and B. Stein
Web Technology and Information Systems (Webis)
Bauhaus-Universität Weimar, Germany
E-mail: {martin.potthast | benno.stein}@uni-weimar.de

A. Barrón-Cedeño and P. Rosso
Natural Language Engineering Lab - ELiRF
Universidad Politécnica de Valencia, Spain
E-mail: {lbarron | prosso}@dsic.upv.es

1 Introduction

Plagiarism, the unacknowledged use of another author's original work, is considered as one of the biggest problems in publishing, science, and education. Texts and other works of art have been plagiarized all throughout history, but with the advent of the World Wide Web text plagiarism is observed at an unprecedented scale. This observation is not surprising since the Web makes billions of texts, code sources, images, sounds, and videos easily accessible, that is to say, copyable.

Plagiarism detection, the automatic identification of plagiarism and the retrieval of the original sources, is developed and investigated as a possible countermeasure. Although humans can identify cases of plagiarism in their areas of expertise quite easily, it requires much effort to be aware of all potential sources on a given topic and to provide strong evidence against an offender. The manual analysis of text with respect to plagiarism becomes infeasible on a large scale, so that automatic plagiarism detection attracts considerable attention.

The paper in hand investigates a particular kind of text plagiarism, namely the detection of plagiarism across languages, sometimes called translation plagiarism. The different kinds of text plagiarism are organized in Figure 1. Cross-language plagiarism, shown encircled, refers to cases where an author translates text from another language and then integrates the translated text into his/her own writing. It is reasonable to assume that plagiarism does not stop at language barriers since, for instance, scholars from non-English speaking countries often write assignments, seminars, theses, and papers in their native languages, whereas current scientific discourse to refer to is often published in English. There are no studies which directly assess the amount of cross-language plagiarism, but in 2005 a broader study among 18 000 students revealed that almost 40% of them admittedly plagiarized at least once, which

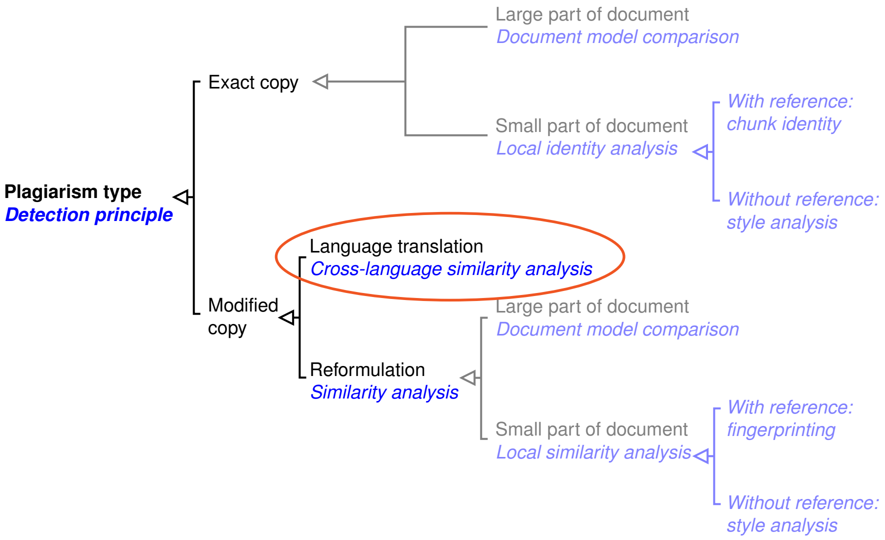


Figure 1 Taxonomy of text plagiarism types, along with approaches to detect them [19].

also includes cross-lingual cases [16]. Apart from being an important practical problem, the detection of cross-language plagiarism also poses a research challenge, since the syntactical similarity between source sections and plagiarized sections found in the monolingual setting is more or less lost across languages. Hence, research on this task may help to improve current methods of cross-language information retrieval as well.

1.1 Related Work

The authors of [8, 15] survey plagiarism detection approaches; here, we merely extend these surveys by recent developments. All of the different kinds of plagiarism shown in Figure 1 are addressed in the literature: the detection of exact copies [5, 12], the detection of modified copies [27, 28], and, for both of the former, their detection without reference collections [18, 19, 30]. Cross-language plagiarism detection has also attracted attention [2, 7, 22, 24, 26]. However, the mentioned research still focuses on a subtask of the retrieval task, namely text similarity computation across languages. I.e., the part is mistaken for the whole and it is overlooked that there are other subtasks that must also be tackled in order to build a practical solution. We also observe that the different approaches are not evaluated in a comparable manner.

1.2 Outline and Contributions

Section 2 introduces a comprehensive retrieval process for cross-language plagiarism detection. The process is derived from monolingual plagiarism detection approaches, while two important subtasks that are different in a multilingual setting are discussed in detail: Section 3 is about the heuristic retrieval of candidate documents, and Section 4 surveys retrieval models for the detailed comparison of documents. With respect to the latter, Section 5 presents a large-scale evaluation of three retrieval models to measure the cross-language similarity of texts: the CL-CNG model [17], the CL-ESA model [24], and the CL-ASA model [2]. All experiments were repeated on test collections sampled from the parallel JRC-Acquis corpus and the comparable Wikipedia corpus. Each test collection contains aligned documents written in English, Spanish, German, French, Dutch, and Polish.

2 Retrieval Process for Cross-Language Plagiarism Detection

Let d_q denote a suspicious document written in language L, and let D' denote a document collection written in another language L'. The detection of a text section in d_q that is plagiarized from D' can be organized within three steps (see Figure 2):

1. *Heuristic Retrieval.* From D' a set of candidate documents D'_q is retrieved where each document is likely to contain sections that are very similar to certain sections in d_q. This step requires methods to map the topic or genre of d_q from L to L'.

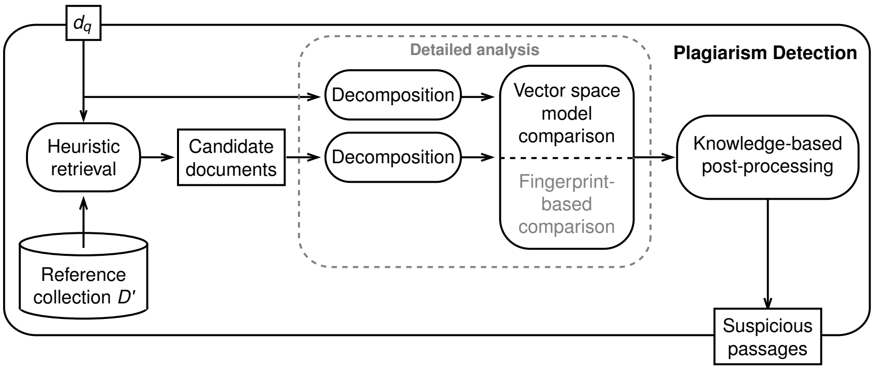


Figure 2 Retrieval process of cross-language plagiarism detection, inspired by [32].

2. *Detailed Analysis.* Each document in D'_q is compared section-wise with d_q, using a retrieval model to measure the cross-language similarity between documents from L and L'. If for a pair of sections a high similarity is measured, a possible case of cross-language plagiarism is assumed.
3. *Knowledge-based Post-Processing.* The candidates for cross-language plagiarism are analyzed in detail in order to filter false positives, e.g., if the copied sections have been properly cited.

At first sight this process may appear rather generic, but the underlying considerations become obvious when taking the view of the practitioner: since plagiarists make use of the World Wide Web, a plagiarism detection solution has to use the entire indexed part of the Web as reference collection D'. This requires the retrieval of candidate documents D'_q with $|D'_q| \ll |D'|$, since a comparison of d_q against each Web document is infeasible. The following sections discuss particularities of step 1 and 2 with respect to a multilingual setting. Note that the third step requires no language-specific treatment.

3 Heuristic Retrieval of Candidate Documents

We identify three alternatives for the heuristic retrieval of candidate documents across languages. They all demonstrate solutions for this task, utilizing well-known methods from cross-language information retrieval (CLIR), monolingual information retrieval (IR), and hash-based search. Figure 3 shows the alternatives. The approaches divide into methods based on a focused search and methods based on hash-based search. The former reuse existing keyword indexes and well-known keyword retrieval methods to retrieve D'_q, the latter rely on a fingerprint index of D' where text sections are mapped onto sets of hash codes.

Approach 1. Research in cross-language information retrieval addresses keyword query tasks in first place, where for a user-specified query q in language L documents are to be retrieved from a collection D' in language L'. By contrast, our task is a so-called “query by example task”, where the query is the document d_q, and documents similar to d_q are to be retrieved from D'. Given a keyword extraction algorithm

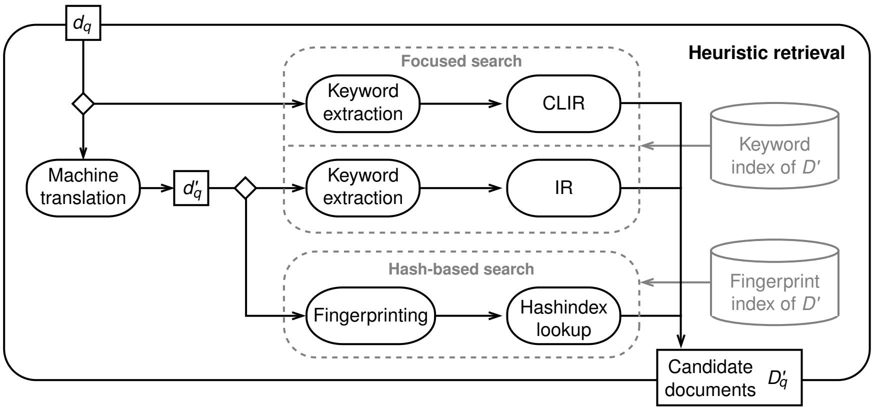


Figure 3 Retrieval process of the heuristic retrieval step for cross-language plagiarism detection.

both tasks are solved in the same way using standard CLIR methods: translation of the keywords from L to L' and querying of a keyword index which stores D'.

Approach 2. In this approach d_q is translated from L to L' with machine translation technology, this way obtaining d'_q. Afterwards keyword extraction is applied to d'_q, which is similar to Approach 1, and the keyword index of D' is queried with the extracted words in order to retrieve D'_q. This approach compares to the first one in terms of retrieval quality, however, Approach 3 provides a faster solution if d_q is translated to d'_q.

Approach 3. A fingerprinted document d_q is represented as small set of integers, called fingerprint. The integers are computed with a similarity hash function h_φ which operationalizes a similarity measure φ and which maps similar documents with a high probability onto the same hash code. Given d_q's translation d'_q, the set of candidate documents is retrieved in virtually constant time by querying the fingerprint index of D' with $h_\varphi(d'_q)$. An alternative option, which has not been investigated yet, is the construction of a cross-language similarity hash function. With such a function at hand the task of translating d_q to d'_q can be omitted.

Remarks. Given the choice among the outlined alternatives the question is "Which way to go?". Today we argue as follows: there is no reason to disregard existing Web indexes, such as the keyword indexes maintained by the major search engine providers. This favors Approach 1 and 2, and it is up to the developer if he trusts the CLIR approach more than the combination of machine translation and IR, or vice versa. Both approaches require careful development and adjustment in order to work in practice. However, if one intends to index portions of the Web in order to build a dedicated index for plagiarism detection purposes, hash-based search (Approach 3) is the choice. It provides near-optimum retrieval speed at reasonable retrieval quality and a significantly smaller index compared to a keyword index [23, 28, 31].

4 Detailed Analysis: Retrieval Models to Measure Cross-Language Similarity

This section surveys retrieval models which can be applied in the detailed analysis step of cross-language plagiarism detection; they measure the cross-language similarity between sections of the suspicious document d_q and sections of the candidate documents in D'_q. Three retrieval models are described in detail, the cross-language character 3-gram model, the cross-language explicit semantic analysis model, and the cross-language alignment-based similarity analysis model.

4.1 Terminology and Existing Retrieval Models

In information retrieval two real-world documents, d_q and d', are compared using a retrieval model $\mathcal{R}$, which provides the means to compute document representations $\mathbf{d}_q$ and $\mathbf{d}'$ as well as a similarity function φ. $\varphi(\mathbf{d}_q, \mathbf{d}')$ maps onto a real value which indicates the topical similarity between d_q and d'. A common retrieval model is the vector space model, VSM, where documents are represented as term vectors whose similarity is assessed with the cosine similarity.

We distinguish four kinds of cross-language retrieval models (see Figure 4): (*i*) models based on language syntax, (*ii*) models based on dictionaries, gazetteers, rules, and thesauri, (*iii*) models based on comparable corpora, and (*iv*) models based on parallel corpora. Models of the first kind rely on syntactical similarities between languages and on the appearance of foreign words. Models of the second kind can be called cross-language vector space models. They bridge the language barrier by translating single words or concepts such as locations, dates, and number expressions from L to L'. Models of the third and fourth kind have to be trained on an aligned corpus that contains documents from the languages to be compared. The two approaches differ with respect to the required degree of alignment: comparable alignment refers to documents in different languages, which describe roughly the same topic, while parallel alignment refers to documents that are translations of each other and whose words or sentences have been mapped manually or heuristically to their respective translations. Obviously the latter poses a much higher requirement than the former. The following models have been proposed:

- CL-CNG represents documents by character n-grams (CNG) [17].
- CL-VSM and Eurovoc-based models build a vector space model [13, 25, 33].
- CL-ESA exploits the vocabulary correlations of comparable documents [24, 36].
- CL-ASA is based on statistical machine translation technology [2].
- CL-LSI performs latent semantic indexing [10, 14].
- CL-KCCA performs a kernel canonical correlation analysis [35].

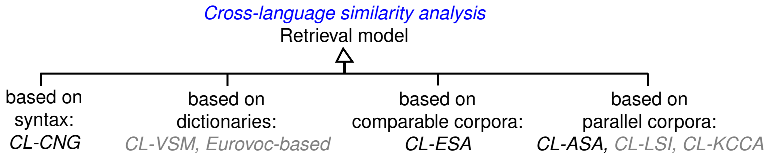

Cross-language similarity analysis
Retrieval model
based on syntax: CL-CNG
based on dictionaries: CL-VSM, Eurovoc-based
based on comparable corpora: CL-ESA
based on parallel corpora: CL-ASA, CL-LSI, CL-KCCA

Figure 4 Taxonomy of retrieval models for cross-language similarity analysis.

The alternatives imply a trade-off between retrieval quality and retrieval speed. Also, the availability of necessary resources for all considered languages is a concern. CL-CNG can be straightforwardly operationalized and requires only little language-specific adjustments, e.g., alphabet normalization by removal of diacritics. The CL-VSM variants offer a retrieval speed comparable to that of the VSM in monolingual information retrieval, but the availability of handmade translation dictionaries depends on the frequency of translations between the respective languages. Moreover, this model requires significant efforts with respect to disambiguation and domain-specific term translations [1, 33]. CL-LSI and CL-KCCA are reported to achieve a high retrieval quality, but their runtime behavior disqualifies them for many practical applications: at the heart of both models is a singular value decomposition of a term-document matrix which has cubic runtime. This is why we chose to compare CL-CNG, CL-ESA, and CL-ASA. All of them are reported to provide a reasonable retrieval quality, they require no manual fine-tuning, pretty few cross-language resources, and they can be scaled to work in a real-world setting. A comparison of these models is also interesting since they operationalize different paradigms for cross-language similarity assessment.

4.2 Cross-Language Character n-Gram Model (CL-CNG)

Character n-grams for cross-language information retrieval achieve a remarkable performance in keyword retrieval for languages with syntactical similarities [17]. We expect that this approach extends to measuring the cross-language document similarity between such languages as well. Given a pre-defined alphabet Σ and an $n \in [1, 5]$, a document d is represented as a vector $\mathbf{d}$ whose dimension is in $O(|\Sigma|^n)$. Obviously $\mathbf{d}$ is sparse, since only a fraction of the possible n-grams occur in any d. In analogy to the VSM, the elements in $\mathbf{d}$ can be weighted according to a standard weighting scheme, and two documents d and d' can be compared with a standard measure $\varphi(\mathbf{d}, \mathbf{d}')$. Here we choose $\Sigma = \{a, \ldots, z, 0, \ldots, 9\}$, $n = 3$, *tf·idf*-weighting, and the cosine similarity as φ. In the following we refer to this model variant as CL-C3G.

4.3 Cross-Language Explicit Semantic Analysis (CL-ESA)

The CL-ESA model is an extension of the explicit semantic analysis model [11, 24, 36]. ESA is a collection-relative retrieval model, which means that a document d is represented by its similarities to the documents of a so-called index collection D_I. These similarities in turn are computed with a monolingual retrieval model such as the VSM [29]:

$$\mathbf{d}_{|D_I} = A_{D_I}^T \cdot \mathbf{d}_{\text{VSM}},$$

where $A_{D_I}^T$ denotes the matrix transpose of the term-document matrix of the documents in D_I, and $\mathbf{d}_{\text{VSM}}$ denotes the term vector representation of d. Again, various term weighting schemes are applicable in this connection.

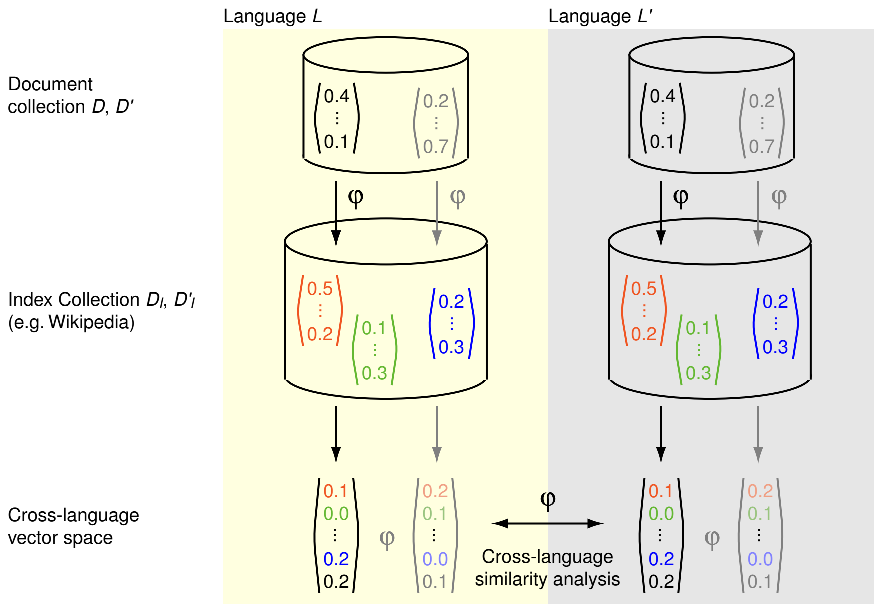

Language L
Language L'
Document collection D, D'
Index Collection D_I, D'_I (e.g. Wikipedia)
Cross-language vector space
φ
Cross-language similarity analysis
0.4
0.1
0.2
0.7
0.5
0.2
0.1
0.3
0.2
0.3
0.1
0.0
0.2
0.2
0.2
0.1
0.0
0.1

Figure 5 Illustration of the cross-language explicit semantic analysis model.

If a second index collection D'_I in another language is given such that the documents in D'_I have a topical one-to-one correspondence to the documents in D_I, the ESA representations in both languages become comparable. I.e., the cross-language similarity between d and d' can be expressed as $\varphi(\mathbf{d}_{|D_I}, \mathbf{d}'_{|D'_I})$. Figure 5 illustrates this principle for two languages. CL-ESA naturally extends to multiple languages; moreover, the approach gets by without translation technology, be it dictionary-based or other. The model requires merely a comparable corpus of documents written in different languages about similar topics. These documents may still be written independently of each other. An example for such a corpus is the Wikipedia encyclopedia where numerous concepts are covered in many languages.

4.4 Cross-Language Alignment-based Similarity Analysis (CL-ASA)

The CL-ASA model is based on statistical machine translation technology; it combines a two-step probabilistic translation and similarity analysis [2]. Given d_q, written in L, and a document d' from a collection D' written in L', the model estimates the probability that d' is a translation of d_q according to Bayes' rule:

$$p(d' \mid d_q) = \frac{p(d')\ p(d_q \mid d')}{p(d_q)} \tag{1}$$

$p(d_q)$ does not depend on d' and hence is neglected. From a machine translation viewpoint $p(d_q \mid d')$ is known as *translation model probability*; it is computed using a statistical bilingual dictionary. $p(d')$ is known as *language model probability*; it describes the target language L' in order to obtain grammatically acceptable text in the translation [6].

Our concern is the retrieval of possible translations of d_q written in L' (and not translating d_q into L'), and against this background we propose adaptations for the two sub-models: (*i*) the adapted translation model is a non-probabilistic measure $w(d_q \mid d')$, and (*ii*) the language model is replaced by a *length model* $\varrho(d')$, which depends on document lengths instead of language structures. Based on these adaptations we define the following similarity measure:

$$\varphi(d_q, d') = s(d' \mid d_q) = \varrho(d')\, w(d_q \mid d') \tag{2}$$

Unlike other similarity measures this one is not normalized; note that the partial order induced among documents resembles the order of other similarity measures. The following subsections describe the adapted translation model $w(d_q \mid d')$ and the length model $\varrho(d')$.

4.4.1 Translation Model

The translation model requires a statistical bilingual dictionary. Given the vocabularies of the corresponding languages $\mathcal{X} \in L$ and $\mathcal{Y} \in L'$, the bilingual dictionary provides estimates of the translation probabilities $p(x, y)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. This distribution expresses the probability for a word x to be a valid translation of a word y. The bilingual dictionary is estimated by means of the well-known IBM M1 alignment model [6, 20], which has been successfully applied in monolingual and cross-lingual information retrieval tasks [4, 21]. In order to generate a bilingual dictionary, M1 requires a sentence-aligned parallel corpus.[1] The translation probability of two texts d and d' is originally defined as:

$$p(d \mid d') = \prod_{x \in d} \sum_{y \in d'} p(x, y), \tag{3}$$

where $p(x, y)$ is the probability that the word x is a translation of the word y. The model was demonstrated to generate good sentence translations, but since we are considering entire documents of variable lengths, the formula is adapted as follows:

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

The weight $w(d \mid d')$ increases if valid translations (x, y) appear in the implied vocabularies. For a word x with $p(x, y) = 0$ for all $y \in d'$, $w(d \mid d')$ is decreased by $\epsilon = 0.1$.

[1] The estimation is carried out on the basis of the EM algorithm [3, 9]. See [6, 22] for an explanation of the bilingual dictionary estimation process.

Table 1 Estimated length factors for the language pairs L-L', measured in characters. A value of $\mu > 1$ implies $|d| < |d'|$ for d and its translation d'.

Parameter	en-de	en-es	en-fr	en-nl	en-pl
μ	1.089	1.138	1.093	1.143	1.216
σ	0.268	0.631	0.157	1.885	6.399

4.4.2 Length Model

Though it is unlikely to find a pair of translated documents d and d' such that $|d| = |d'|$, we expect that their lengths will be closely related by a certain length factor for each language pair. In accordance with [26] we define the length model probability as follows:

$$\varrho(d') = \exp\left(-0.5\left(\frac{(|d'|/|d|) - \mu}{\sigma}\right)^2\right), \tag{5}$$

where μ and σ are the average and the standard deviation of the character lengths between translations of documents from L to L'. Observe that in cases where a translation d' of a document d_q has not the expected length, the similarity $\varphi(d_q, d')$ is reduced.

Table 1 lists the values for μ and σ that are used in the evaluation for the considered language pairs; these values have been estimated using the JRC-Acquis training collection. The variation of the length between a document d_q and its translation d' approximates a normal distribution (cf. Figure 6 for an illustration).

5 Evaluation of Retrieval Models for the Detailed Analysis

In our evaluation we compare CL-C3G, CL-ESA, and CL-ASA in a ranking task. Three experiments are conducted on two test collections with each model and over all language pairs whose first language is English and whose second language is one of Spanish, German, French, Dutch, and Polish. In total, more than 100 million similarities are computed with each model.

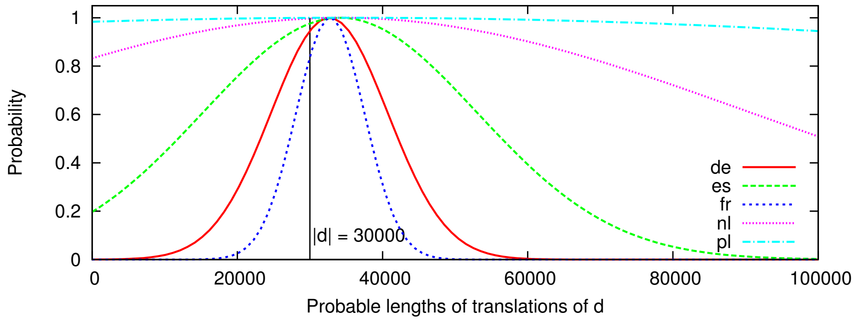


Figure 6 Length model distributions that quantify the likelihood whether the length of the translation of d into the considered languages is larger than $|d|$. In this example, d is an English document of 30 000 characters (vertical line), corresponding to 6 600 words.

5.1 Corpora for Model Training and Evaluation

To train the retrieval models and to test their performance we extracted large collections from the parallel corpus JRC-Acquis and the comparable corpus Wikipedia. The JRC-Acquis Multilingual Parallel Corpus comprises legal documents from the European Union which have been translated and aligned with respect to 22 languages [34]. The Wikipedia encyclopedia is considered to be a comparable corpus since it comprises documents from more than 200 languages which are linked across languages in case they describe the same topic [24]. From these corpora only those documents are considered for which aligned versions exist in all of the aforementioned languages: JRC-Acquis contains 23 564 such documents, and Wikipedia contains 45 984 documents, excluding those articles that are lists of things or which describe a date.[2]

The extracted documents from both corpora are divided into a training collection that is used to train the respective retrieval model, and a test collection that is used in the experiments (4 collections in total). The JRC-Acquis test collection and the Wikipedia test collection contain 10 000 aligned documents each, and the corresponding training collections contain the remainder. In total, the test collections comprise 120 000 documents: 10 000 documents per corpus × 2 corpora × 6 languages. As described above, CL-ESA requires the comparable Wikipedia training collection as index documents, whereas CL-ASA requires the parallel JRC-Acquis training collection to train bilingual dictionaries for all of the considered language pairs. Note that CL-C3G requires no training.

5.2 Experiments and Methodology

The experiments are based on those of [24]: let d_q be a query document from a test collection D, let D' be the documents aligned with those in D, and let d'_q denote the document that is aligned with d_q. The following experiments have been repeated for 1 000 randomly selected query documents with all three retrieval models on both test collections, averaging the results.

Experiment 1: Cross-Language Ranking. Given d_q, all documents in D' are ranked according to their cross-language similarity to d_q; the retrieval rank of d'_q is recorded. Ideally, d'_q should be on the first or, at least, on one of the top ranks.

Experiment 2: Bilingual Rank Correlation. Given a pair of aligned documents $d_q \in D$ and $d'_q \in D'$, the documents from D' are ranked twice: (*i*) with respect to their cross-language similarity to d_q using one of the cross-language retrieval models, and, (*ii*) with respect to their monolingual similarity to d'_q using the vector space model. The top 100 ranks of the two rankings are compared using Spearman's ρ, a rank correlation coefficient which measures the disagreement and agreement of rankings as a value between -1 and 1. This experiment relates to "diagonalization:" a monolingual reference ranking is compared to a cross-lingual test ranking.

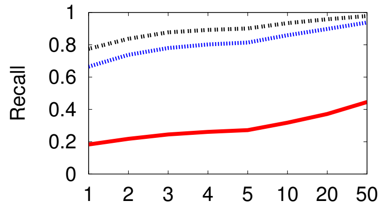
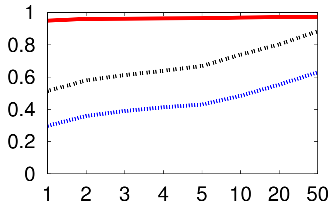
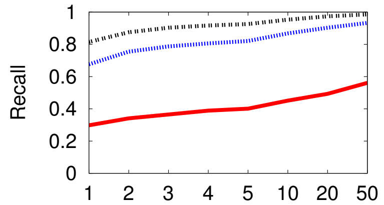
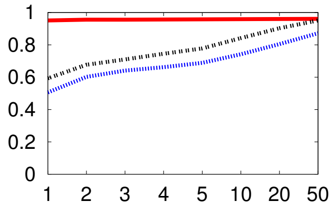
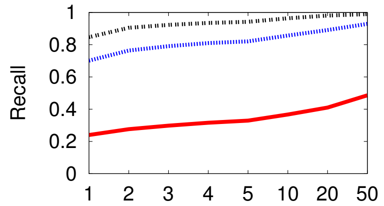
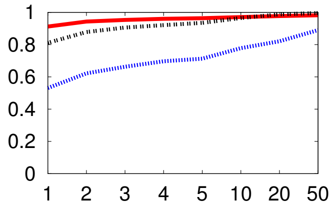
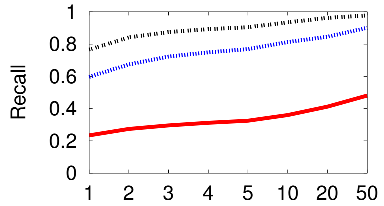
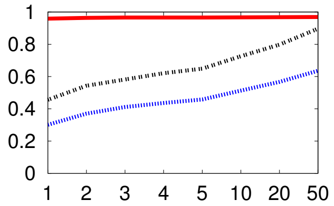
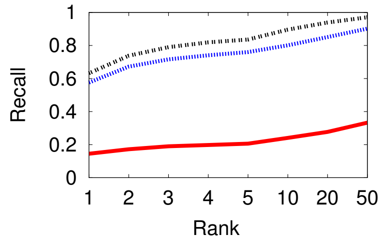
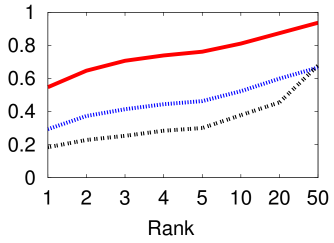
Experiment 3: Cross-Language Similarity Distribution. This experiment contrasts the similarity distributions of comparable documents and parallel documents.

[2] If only pairs of languages are considered, many more aligned documents can be extracted from Wikipedia, e.g., currently more than 200 000 between English and German.

5.3 Results and Discussion

Experiment 1: Cross-Language Ranking. This experiment resembles the situation of cross-language plagiarism in which a document (a section) is given and its translation has to be retrieved from a collection of documents (of sections). The results of the experiment are shown in Table 2 as recall-over-rank plots.

Table 2 Results of Experiment 1 for the cross-language retrieval models.

Experiment 1: Cross-Language Ranking		
Wikipedia	JRC-Acquis	Language Pair
CL-ASA ▬	CL-ESA ┈	CL-C3G ┈
		en-de
		en-es
		en-fr
		en-nl
		en-pl

Observe that CL-ASA achieves near-perfect performance on the JRC-Acquis test collection, while its performance on the Wikipedia test collection is poor for all language pairs. CL-ESA achieves between a medium and a good performance on both collections, dependent on the language pair, and so does CL-C3G, which outperforms CL-ESA in most cases. With respect to the different language pairings all models vary in their performance, but, with the exception of both CL-ASA and CL-C3G on the English-Polish portion of JRC-Acquis (bottom right plot), the performance characteristics are the same on all language pairs.

It follows that CL-ASA has in general a large variance in its performance, while CL-ESA and CL-C3G show a stable performance across the corpora. Remember that JRC-Acquis is a parallel corpus while Wikipedia is a comparable corpus, so that CL-ASA seems to be working much better on "exact" translations than on comparable documents. Interestingly, CL-ESA and CL-C3G work better on comparable documents than on translations. An explanation for these findings is that the JRC-Acquis corpus is biased to some extent; it contains only legislative texts from the European Union and hence is pretty homogeneous. In this respect both CL-ESA and CL-C3G appear much less susceptible than CL-ASA, while the latter may perform better when trained on a more diverse parallel corpus. The Polish portion of JRC-Acquis seems to be a problem for both CL-ASA and CL-C3G, but less so for CL-ESA, which shows that the latter can cope with less related languages.

Experiment 2: Bilingual Rank Correlation. This experiment can be considered as a standard ranking task where documents have to be ranked according to their similarity to a document written in another language. The results of the experiment are reported as averaged rank correlations in Table 3.

As in Experiment 1, CL-ASA performs good on JRC-Acquis and unsatisfactory on Wikipedia. In contrast to Experiment 1, CL-ESA performs similar to both CL-CNG and CL-ESA on JRC-Acquis with respect to different language pairs, and it outperforms CL-ASA on Wikipedia. Again, unlike in the first experiment, CL-C3G is outperformed by CL-ESA. With respect to the different language pairings all models show weaknesses, e.g., CL-ASA on English-Polish and, CL-ESA as well as CL-C3G on English-Spanish and English-Dutch. It follows that CL-ESA is more applicable as a general purpose retrieval model than are CL-ASA or CL-C3G, while special care needs to be taken with respect to the involved languages. We argue that the reason for the varying performance is rooted in the varying quality of the employed language-specific indexing pipelines and not in the retrieval models themselves.

Table 3 Results of Experiment 2 for the cross-language retrieval models.

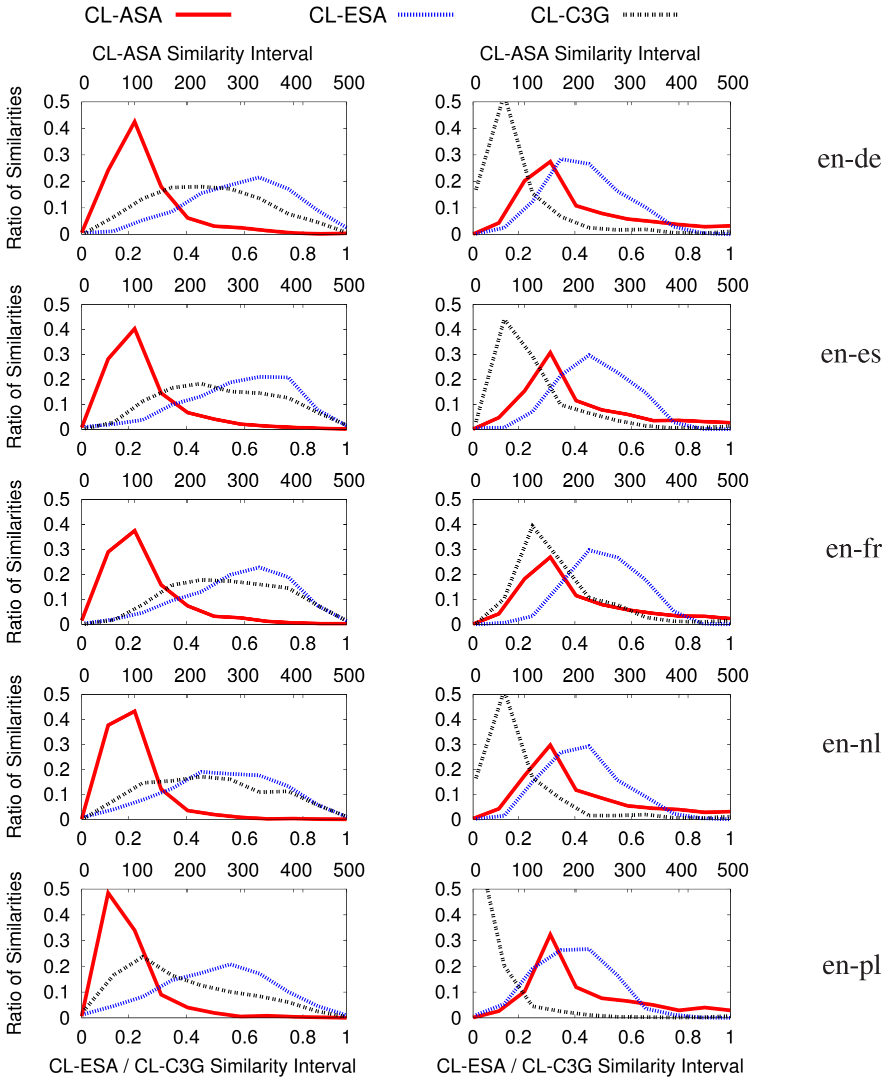
Language	**Experiment 2: Bilingual Rank Correlation**					
Pair	Wikipedia			JRC-Acquis		
	CL-ASA	CL-ESA	CL-C3G	CL-ASA	CL-ESA	CL-C3G
en-de	0.14	0.58	0.37	0.47	0.31	0.28
en-es	0.18	0.17	0.10	0.66	0.51	0.42
en-fr	0.16	0.29	0.20	0.38	0.54	0.55
en-nl	0.14	0.17	0.11	0.58	0.33	0.31
en-pl	0.11	0.40	0.22	0.15	0.35	0.15

Experiment 3: Cross-Language Similarity Distribution. This experiment shall give us an idea about what can be expected from each retrieval model; the experiment cannot directly be used to compare the models or to tell something about their quality. Rather, it tells us something about the range of cross-language similarity values one will measure when using the model, in particular, which values indicate a high similarity and which values indicate a low similarity. The results of the experiment are shown in Table 4 as plots of ratio of similarities-over-similarity intervals.

Table 4 Results of Experiment 3 for the cross-language retrieval models.

Experiment 3: Cross-language Similarity Distribution

Wikipedia JRC-Acquis Language Pair



Observe that the similarity distributions of CL-ASA has been plotted on a different scale than those of CL-ESA and CL-C3G: the top x-axis of the plots shows the range of similarities measured with CL-ASA, the bottom x-axis shows the range of similarities measured with the other models. This is necessary since the similarities computed with CL-ASA are not normalized. It follows that the absolute values measured with the three retrieval models are not important, but the order they induce among the compared documents is. In fact, this holds for each of retrieval models, be it cross-lingual or not. This is also why the similarity values computed with two models cannot be compared to one another: e.g., the similarity distribution of CL-ESA looks "better" than that of CL-C3G because it is more to the right, but in fact, CL-C3G outperforms CL-ESA in Experiment 1.

6 Summary

Cross-language plagiarism is an important direction of plagiarism detection research but is still in its infancy. In this paper we pointed out a basic retrieval strategy for this task, including two important subtasks which require special attention: the heuristic multilingual retrieval of potential source candidates for plagiarism from the Web, and the detailed comparison of two documents across languages. With respect to the former, well-known and less well-known state-of-the-art research is reviewed. With respect to the latter, we survey existing retrieval models and describe three of them in detail, namely the cross-language character n-gram model (CL-CNG), the cross-language explicit semantic analysis (CL-ESA) and the cross-language alignment-based similarity analysis (CL-ASA). For these models we report on a large-scale comparative evaluation.

The evaluation covers three experiments with two aligned corpora, the comparable Wikipedia corpus and the parallel JRC-Acquis corpus. In the experiments the models are employed in different tasks related to cross-language ranking in order to determine whether or not they can be used to retrieve documents known to be highly similar across languages. Our findings include that the CL-C3G model and the CL-ESA model are in general better suited for this task, while CL-ASA achieves good results on professional and automatic translations. CL-CNG outperforms CL-ESA and CL-ASA. However, unlike the former, CL-ESA and CL-ASA can also be used on language pairs whose alphabet or syntax are unrelated.

References

1. Lisa Ann Ballesteros. *Resolving Ambiguity for Cross-Language Information Retrieval: A Dictionary Approach.* PhD thesis, University of Massachusetts Amherst, USA, 2001. Bruce Croft.
2. Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, and Alfons Juan. On Cross-Lingual Plagiarism Analysis Using a Statistical Model. In Benno Stein, Efstathios Stamatatos, and Moshe Koppel, editors, *ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 08)*, pages 9–13, Patras (Greece), July 2008.
3. Leonard E. Baum. An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process. *Inequalities*, 3:1–8, 1972.

4. Adam Berger and John Lafferty. Information Retrieval as Statistical Translation. In *SIGIR'99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 4629, pages 222–229, Berkeley, California, United States, 1999. ACM.
5. Sergey Brin, James Davis, and Hector Garcia-Molina. Copy Detection Mechanisms for Digital Documents. In *SIGMOD '95*, pages 398–409, New York, NY, USA, 1995. ACM Press.
6. Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.
7. Zdenek Ceska, Michal Toman, and Karel Jezek. Multilingual Plagiarism Detection. In *AIMSA'08: Proceedings of the 13th international conference on Artificial Intelligence*, pages 83–92, Berlin, Heidelberg, 2008. Springer-Verlag.
8. Paul Clough. Old and New Challenges in Automatic Plagiarism Detection. National UK Plagiarism Advisory Service, `http://ir.shef.ac.uk/cloughie/papers/pas_plagiarism.pdf`, 2003.
9. Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1):1–38, 1977.
10. Susan T. Dumais, Letsche Todd, A., Michael, L. Littman, and Thomas K. Landauer. Automatic Cross-language Retrieval Using Latent Semantic Indexing. In D. Hull and D. Oard, editors, *AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval*, pages 18–24, Stanford University, March 1997. American Association for Artificial Intelligence.
11. Evgeniy Gabrilovich and Shaul Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of The 20th International Joint Conference for Artificial Intelligence*, Hyderabad, India, 2007.
12. Timothy C. Hoad and Justin Zobel. Methods for Identifying Versioned and Plagiarised Documents. *American Society for Information Science and Technology*, 54(3):203–215, 2003.
13. Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. Dictionary-based Techniques for Cross-language Information Retrieval. *Inf. Process. Manage.*, 41(3):523–547, 2005.
14. Michael Littman, Susan T. Dumais, and Thomas K. Landauer. Automatic Cross-language Information Retrieval Using Latent Semantic Indexing. In *Cross-Language Information Retrieval, chapter 5*, pages 51–62. Kluwer Academic Publishers, 1998.
15. Hermann Maurer, Frank Kappe, and Bilal Zaka. Plagiarism - A Survey. *Journal of Universal Computer Science*, 12(8):1050–1084, 2006.
16. Donald McCabe. Research Report of the Center for Academic Integrity. `http://www.academicintegrity.org`, 2005.
17. Paul Mcnamee and James Mayfield. Character N-Gram Tokenization for European Language Text Retrieval. *Inf. Retr.*, 7(1-2):73–97, 2004.
18. Sven Meyer zu Eissen and Benno Stein. Intrinsic Plagiarism detection. In Mounia Lalmas, Andy MacFarlane, Stefan M. Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, editors, *Proceedings of the European Conference on Information Retrieval (ECIR 2006)*, volume 3936 of *Lecture Notes in Computer Science*, pages 565–569. Springer, 2006.
19. Sven Meyer zu Eissen, Benno Stein, and Marion Kulig. Plagiarism Detection without Reference Collections. In Reinhold Decker and Hans J. Lenz, editors, *Advances in Data Analysis*, pages 359–366. Springer, 2007.
20. Franz J. Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.
21. David Pinto, Alfons Juan, and Paolo Rosso. Using Query-Relevant Documents Pairs for Cross-Lingual Information Retrieval. In V. Matousek and P. Mautner, editors, *Proceedings of the TSD-2006: Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Artificial Intelligence*, pages 630–637, Pilsen, Czech Republic, 2007.
22. David Pinto, Jorge Civera, Alberto Barrón-Cedeño, Alfons Juan, and Paolo Rosso. A Statistical Approach to Cross-lingual Natural Language Tasks. *J. Algorithms*, 64(1):51–60, 2009.
23. Martin Potthast. Wikipedia in the Pocket - Indexing Technology for Near-Duplicate Detection and High Similarity Search. In Charles Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij, and Arjen de Vries, editors, *30th Annual International ACM SIGIR Conference*, pages 909–909. ACM, July 2007.
24. Martin Potthast, Benno Stein, and Maik Anderka. A Wikipedia-Based Multilingual Retrieval Model. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *30th European Conference on IR Research, ECIR 2008, Glasgow*, volume 4956 LNCS of *Lecture Notes*

in Computer Science, pages 522–530, Berlin Heidelberg New York, 2008. Springer.

25. Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In *Proceedings of the Workshop 'Ontologies and Information Extraction' at the Summer School 'The Semantic Web and Language Technology - Its Potential and Practicalities' (EUROLAN'2003)*, pages 9–28, Bucharest, Romania, August 2003.
26. Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. Automatic Identification of Document Translations in Large Multilingual Document Collections. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'2003)*, pages 401–408, Borovets, Bulgaria, September 2003.
27. Benno Stein. Fuzzy-Fingerprints for Text-Based Information Retrieval. In Klaus Tochtermann and Hermann Maurer, editors, *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05), Graz*, Journal of Universal Computer Science, pages 572–579. Know-Center, July 2005.
28. Benno Stein. Principles of Hash-based Text Retrieval. In Charles Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij, and Arjen de Vries, editors, *30th Annual International ACM SIGIR Conference*, pages 527–534. ACM, July 2007.
29. Benno Stein and Maik Anderka. Collection-Relative Representations: A Unifying View to Retrieval Models. In A. M. Tjoa and R. R. Wagner, editors, *20th International Conference on Database and Expert Systems Applications (DEXA 09)*, pages 383–387. IEEE, September 2009.
30. Benno Stein and Sven Meyer zu Eissen. Intrinsic Plagiarism Analysis with Meta Learning. In Benno Stein, Moshe Koppel, and Efstathios Stamatatos, editors, *SIGIR Workshop Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07)*, pages 45–50. CEUR-WS.org, July 2007.
31. Benno Stein and Martin Potthast. Construction of Compact Retrieval Models. In Sándor Dominich and Ferenc Kiss, editors, *Studies in Theory of Information Retrieval*, pages 85–93. Foundation for Information Society, October 2007.
32. Benno Stein, Sven Meyer zu Eissen, and Martin Potthast. Strategies for Retrieving Plagiarized Documents. In Charles Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij, and Arjen de Vries, editors, *30th Annual International ACM SIGIR Conference*, pages 825–826. ACM, July 2007.
33. Ralf Steinberger, Bruno Pouliquen, and Camelia Ignat. Exploiting Multilingual Nomenclatures and Language-Independent Text Features as an Interlingua for Cross-lingual Text Analysis Applications. In *Proceedings of the 4th Slovenian Language Technology Conference. Information Society 2004 (IS'2004)*, 2004.
34. Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. The JRC-Acquis: A multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, May 2006.
35. Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS-02: Advances in Neural Information Processing Systems*, pages 1473–1480. MIT Press, 2003.
36. Yiming Yang,, Jaime G. Carbonell,, Ralf D. Brown,, and Robert E. Frederking,. Translingual Information Retrieval: Learning from Bilingual Corpora. *Artif. Intell.*, 103(1-2):323–345, 1998.