Cross-lingual Text Fragment Alignment using Divergence from Randomness

Sirvan Yahyaei, Marco Bonzanini, and Thomas Roelleke

Queen Mary, University of London
Mile End Road, E1 4NS London, UK
{sirvan,marcob,thor}@eecs.qmul.ac.uk

Abstract. This paper describes an approach to automatically align fragments of texts of two documents in different languages. A text fragment is a list of continuous sentences and an aligned pair of fragments consists of two fragments in two documents, which are content-wise related. Cross-lingual similarity between fragments of texts is estimated based on models of divergence from randomness. A set of aligned fragments based on the similarity scores are selected to provide an alignment between sections of the two documents. Similarity measures based on divergence show strong performance in the context of cross-lingual fragment alignment in the performed experiments.

Keywords: fragment alignment, divergence from randomness, summarisation

1 Introduction

A notable portion of the information available on the Internet is given by documents which are obtainable from more than one source. For example, the same web page might be published on different mirror web sites, or the same piece of news could be reported, in slightly different versions, possibly in different languages. This phenomenon has several implications.

In the context of web search, data redundancy in the search results has already been shown to be an issue [4]. For example, even if a document is considered to be relevant to an information need, when shown after a number of redundant documents, it does not provide the user any additional information. In other words, showing redundant documents does not benefit the user for the purpose of satisfying an information need.

Given the dynamic nature of the Web, it is common to find different versions of the same document. The task of identifying versioned or plagiarised documents, with a distinction between real plagiarism and mere topic similarity, is not trivial. Both versioning and plagiarism might affect a document as a whole, or just portions (e.g. sections, paragraphs, or more in general fragments) of it. An intelligent tool which helps in recognising duplicate text fragments could benefit editors and authors.

To tackle one aspect of these implications, this paper investigates the possibility of aligning text fragments between documents written in two different languages. The main focus is identifying pairs of fragments with a strong content-based similarity. Figure 1 shows an example of aligning fragments of texts, which do not necessarily have the same length. Our approach, starts with measuring similarity at sentence level between the documents and then extract aligned fragments of texts based on the sentence similarities. The outcome will be a set of disjoint aligned fragments with the highest score based on the previously estimated sentence similarities.

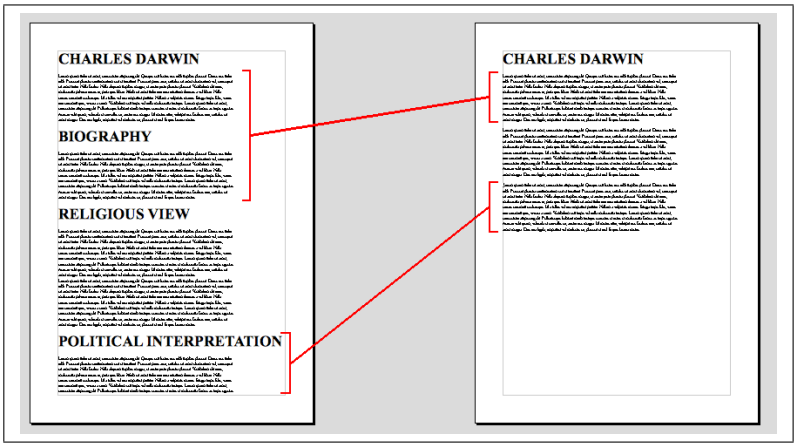

Fig. 1. An example of aligned text fragments.

The main component of our method is measuring the similarity between two text fragments. We have chosen models of information retrieval based on divergence from randomness to estimate the similarities and examine the best performing model in the context of cross-lingual text alignment. An advantage of models based on divergence consists in having multiple choices of randomness models, and hence the opportunity to evaluate many IR models for this task. In addition, these models are non-parametric and do not require parameter tuning and training data to perform well.

The information about the fragments of the documents produced by the alignment algorithm, can be used later for specific applications. Such applications include the possibility of automatically creating training data sets for machine translation or document summarisation, as well as automatically synchronising complex multi-lingual web sites (e.g. Wiki-based encyclopedias, or other user-driven sites). Previous work in this area has explored both novelty detection for improving search effectiveness, and the use of fingerprinting techniques for identifying redundant documents [4], but mainly in a monolingual environment.

The remainder of this paper is organised as follows: Section 2 provides a review of current research and methods in fields related to cross-lingual text

alignment. Section 3 describes the alignment of text fragments algorithm and similarity measures to perform the sentence alignment. Construction of the test collection and experiments are reported in Section 4 and Section 5 concludes the paper.

2 Related Work

This work lays on the overlap between the two areas of document summarisation and machine translation. Despite their differences in concepts and techniques, both summarisation and translation systems are mostly built on top of statistical methods, which require training data to acquire statistical patterns. [6] propose an approach to automatically align documents to their respective summaries and extract transformation rules to shorten phrases to produce shorter and more informative summaries. Their algorithm is an extension to the standard HMM model and learns word-to-word and phrase-to-phrase alignment in an unsupervised manner.

In case of machine translation, availability of training data set is more crucial. Statistical machine translation, uses manually translated data in the forms of parallel sentences to learn translation patterns by statistical means. There has been extensive work focusing in finding parallel documents [14] and aligning sentences in fairly parallel corpora [8] and even non-parallel corpora [9]. [10] presents an approach to find sub-sentential segments from comparable corpora. Despite previous work, [14] propose a method that solely relies on textual content of the documents instead of meta-data or document structure to find near-duplicate documents. All documents are automatically translated and n-gram features are extracted to construct a small set of candidate documents in a very large collection of documents. One-by-one comparison is performed using idf-weighted cosine similarity among the documents in the candidate set. They report that incorporating term frequency or other retrieval ranking functions degrade the performance compared to the mentioned similarity measure. Our approach is also based on textual content only, but the alignment is performed on fragments (see Section 3) rather than sentences or entire documents.

In cross-lingual plagiarism, the aim is finding fragments of text that have been plagiarised from the source document written in a different language. [2] describe an statistical approach based on IBM model 1 [5] to retrieve the plagiarised fragment among a list of candidate fragments. The statistical approach is proposed to perform cross-lingual retrieval, bilingual classification and cross-lingual plagiarism and it focuses on the retrieval aspect of plagiarism. [12] investigates the performance and effectiveness of different models of cross-lingual retrieval for the purpose of plagiarism detection. They compare retrieval models based on parallel and comparable corpora to models based on dictionaries and syntax of the languages involved. Similarly to [2], IBM model 1 probabilities are used as translation probabilities in the statistical models and a length component is introduced to take into account the ration of length differences between the two languages.

Similar work, in a mono-lingual environment, involves the identification of redundant [4] and co-derivative [3] documents, using fingerprinting techniques. Fingerprints are compact representations of text chunks. In these approaches, hash functions are used to calculate fingerprints of documents. Different documents are then identified as redundant, or as co-derivative, according to the fingerprint similarities. In our approach, the similarity is calculated on a fragment level, based on the content of the fragments.

3 Text Fragment Alignment

We define a text fragment as a list of continuous sentences in a document. Ideally, the content of a fragment is semantically coherent (i.e. it can be considered to be about a single topic). The aim of the proposed fragment alignment is to find fragment pairs in two documents, which are written in two different languages. Assume $\boldsymbol{d_e} =< s_{e_1}, s_{e_2}, \ldots, s_{e_n} >$ and $\boldsymbol{d_f} =< s_{f_1}, s_{f_2}, \ldots, s_{f_m} >$ are two documents in languages e and f, which contain n and m number of sentences respectively. We want to find a set of paired fragments that contains aligned text fragments that are related:

$$\{(\boldsymbol{\epsilon_i^{i'}}, \boldsymbol{\phi_j^{j'}}) | 1 \leq i \leq i' \leq n \wedge 1 \leq j \leq j' \leq m\} \tag{1}$$

where, $\boldsymbol{\epsilon_i^{i'}}$ represents a fragment that contains sentences i to i' from $\boldsymbol{d_e}$ and $\boldsymbol{\phi_j^{j'}}$ is a fragment that contains sentences j to j' from $\boldsymbol{d_f}$. Based on these definitions, fragments of a document can consist of different number of sentences and even relatively different number of sentences for each fragment in an aligned one. Since considering all the possible fragments in a document and aligning them with all the possible fragments in the other document is computationally very expensive, we restrict extracting the fragments by initial information about the alignment of sentences. The initial information is acquired by aligning sentences in the two documents and finding a few strong links between some of the sentences. A paired fragment can not contain a link to sentences outside the pair. This restriction significantly reduces the number of fragments that can be extracted.

Figure 2 sketches the text fragment alignment algorithm. The first step is to score all the sentence pairs and find a few links between the sentences. Next, all the fragments which are compatible with the links are extracted and sorted according to their scores. Finally, a set of non-overlap fragment pairs are selected as the output. It is important to note that the algorithm takes two documents as input and the computational cost only depends on the length of the documents. In other words, the algorithm of Figure 2 is run on a set of paired documents and does not depend on the document collection size.

3.1 Similarity Measures and Divergence from Randomness

A major step in finding aligned fragments of two documents is estimating similarity between sentences. As pointed out in the introduction, we have chosen

Input: $\boldsymbol{d_e}$ and $\boldsymbol{d_f}$ {$\boldsymbol{d_e}$ is English document, $\boldsymbol{d_f}$ is foreign document}
Input: similarity threshold *min_score*
for all s_{e_i} in $\boldsymbol{d_e}$ **do**
  for all s_{f_j} in $\boldsymbol{d_f}$ **do**
    $score[i][j] \leftarrow$ estimate similarity between s_{e_i} and s_{f_j}
    $link[i][j] \leftarrow (score[i][j] > min\_score)$
  end for
end for
aligned $\leftarrow$ extract fragment pairs compatible with *link*
chosen $\leftarrow$ {}
for all *fragment* in (sort *aligned*) **do**
  if *fragment* overlaps with no member of *chosen* **then**
    $chosen \leftarrow chosen \cup fragment$
  end if
end for

Fig. 2. Text fragment alignment algorithm. *aligned* is the set of all aligned fragments and *chosen* is the final set of selected fragments.

a set of probabilistic models of information retrieval based on divergence from randomness [1]. A basic assumption of DFR (Divergence from Randomness) models is that non-informative words are randomly distributed in the collection. In DFR, a randomness model M is chosen to compute the probabilities and there are many ways to choose M, such as Bose-Einstein distribution or Inverse Document Frequency model. $Prob_1(\text{tf})$ is defined as the probability of observing tf occurrences of a term in a randomly selected document according to M. Thus, if $Prob_1$ is relatively small for a term, then the term is an informative one. Another probability, $Prob_2$, is defined as the probability of occurrence of a term within a document with regard to a set of documents that contain the term.

The term weight, under the above definitions is the product of two factors: Firstly, information content of the term with respect to the whole collection, which is formulated as $Inf_1 = -\log_2 Prob_1$. Secondly, $Inf_2 = 1 - Prob_2$, information gain of the term with respect to its elite set, which is the set of documents that contain the term.

$$w = Inf_1 \times Inf_2 = (-\log_2 Prob_1) \times (1 - Prob_2) \tag{2}$$

Here, we are computing the similarity between two sentences in two different languages, s_e and s_f. Terms in s_f are translated based on a lexical translation model and converted to a bag-of-word with, s'_f, translation probabilities for each term. The lexical translation model is based on the IBM model 1 [5], that does not take into account the order of words in calculating the translation probabilities. The similarity between two sentences s_e and s_f is calculated as follows:

$$\text{sim}(s_e, s_f) = \text{sim}(s_e, s'_f) = \sum_{t \in \{s_e \cap s'_f\} \wedge \tau \in s_f} w_M(t, s_e) \times p(t|\tau) \tag{3}$$

where, $w(t, s_e)$ is the weight if term t in sentence s_e according to similarity model M and $p(t|\tau)$ is the translation probability of translating τ to t. The collection for equation 3 is $\boldsymbol{d_e}$, which is the document that contains s_e and all the collection statistics in the similarity measures are computed based on $\boldsymbol{d_e}$. Table 1 shows a list of all the models used in this work to estimate the sentence similarity between two documents.

Table 1. Similarity measures used to estimate the similarity between sentences. For detailed information on each model, please refer to [1].

	Name	Description
1	TF-IDF	The tf . idf weighting function, where tf is the total term frequency and *idf* is Sparck-Jones' formulation
2	TF_k-IDF	Same as above but with the BM25 tf quantification $\frac{tf}{tf+k}$
3	$I(n)L2$	Model with Inverse document frequency, with Laplace after-effect and 2nd normalisation
4	$I(F)B2$	Model with Inverse of the term frequency, with Bernoulli after-effect and 2nd normalisation
5	$I(n_e)B2$	Model with Inverse of the expected document frequency, with Bernoulli after-effect and 2nd normalisation in base 2
6	$I(n_e)C2$	Model with Inverse of the expected document frequency, with Bernoulli after-effect and 2nd normalisation in base e
7	$BB2$	Limiting form of Bose-Einstein, with Bernoulli after-effect and 2nd normalisation
8	$PL2$	Poisson approximation of the binomial model, with Laplace after-effect and 2nd normalisation
9	BM25b	BM25 probabilistic model
10	OkapiBM25	Okapi formulation of BM25; the same as BM25b with within-query term frequency (k_3) set to 0

3.2 Extraction of Fragments

After scoring all the sentence pairs, only those with similarity score higher than a certain threshold are aligned. Aligned fragments are extracted by an algorithm adopted from phrase-based statistical machine translation [11]. Simply, two fragments are aligned if no sentence inside them is aligned to sentences outside the fragments and there is at least one link between the two fragments. Fragments in an extracted fragment pair are only aligned to each other and not to any fragment outside the fragment pair.

Many of the extracted aligned fragments overlap and there are sentences which belong to more than one fragment. Therefore, we sort all the aligned fragments according to their similarity score and drop those with lower scores and overlap. The score of an aligned fragment is estimated by averaging the similarity scores of its sentence pairs computed before. The remaining aligned fragments are the result of the algorithm.

4 Experimental Study

Since we did not have a manually annotated documents with aligned fragments, a pseudo-collection is constructed to perform the experiments. A collection of documents and their summaries in English and Italian is built by crawling the web-site of the Press releases of the European Union[1] and pseudo-documents are created by randomly concatenating documents and summaries to each other. For the English side, x documents are randomly chosen and concatenated to create a document with multiple topics. On the Italian side, y documents are randomly chosen, added to the set of x aligned summaries of the chosen documents and randomly concatenated. As a result, we have an English document consisting of x documents and an Italian document consisting of $x + y$ summaries, including the summaries of the English documents. The task is now defined as aligning all the sentences of the summaries to their correct English documents or to not-align those with no corresponding document. In other words, in the English side there are x documents and in the Italian side there summaries with y more summaries mixed with them. Our algorithm tries to align the summaries to their corresponding documents. Table 2 shows statistics of the corpus. All the documents and summaries in the collection are processed by tokenisation, lower-casing and sentence splitting.

Table 2. English-Italian corpus statistics

	English	**Italian**	**Average**
Mean Document Length (sentences)	34.66	35.29	34.96
Mean Summary Length (sentences)	5.09	4.87	4.98
Mean Compression Ratio (sentences)	14.68%	13.81%	14.26%
Mean Document Length (words)	794.85	874.73	834.79
Mean Summary Length (words)	106.08	118.74	112.43
Mean Compression Ratio (words)	13.35%	13.58%	13.47%
Number of document/summary pairs		192	

4.1 Document-Summary Association

As a basic task compared to finding aligned fragments of text, we examine the problem of associating documents to their summaries. Association is the process of finding two related structures in a collection of structures. In a collection of documents and summaries, the aim is to find the most related summary to each document. We assume that there is a one-to-one association between the summaries and the documents.

[1] Available at http://europa.eu/rapid

The association process can be performed in two ways. Firstly, a two-stage method which translates and summarises the document and computes the similarity between the summaries. Secondly, a one-stage cross-lingual association approach that directly calculates the similarity between the document and the summary in different languages. An illustration of English-to-Italian association is drawn in Figure 3, which shows the two ways that the association can be performed in. The one-stage approach estimates the similarity between the document and the summary according to equation 3, but instead of similarity between sentences, its the similarity between documents and summaries.

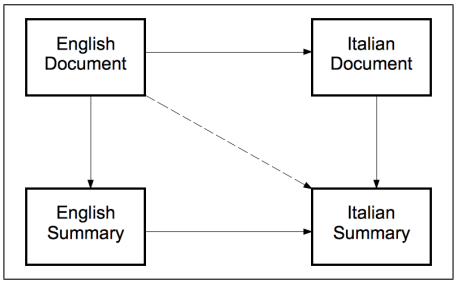

Fig. 3. Cross-lingual Summarisation Pipelines: Two-Stage vs. One-Stage

In the two-stage approach, the summarisation component relies on MEAD [13], which is an extractive summariser. The machine translation system used for translation form Italian to English is a phrase-based statistical MT system with translation model and language model as its main components. The full detail of the system is described in [15]. The training data for the SMT system is taken from the Europarl corpus [7]. 1.6 million parallel sentences were used for building the translation model and 50 million sentences to train the English language model. For both approaches, lexical probabilities are estimated based on IBM model 1 and the parallel training data mentioned before.

The scores for the one-stage system, which associates English documents to Italian summaries, are shown in Table 3, where one can observe that the OkapiBM25 function is performing the best. The best scores for the two-stage method are P@1= 78.1% and MRR= 82.0 and the results of the two-stage approach are in all the cases substantially lower than the one-stage one.

In the two-stage approach approach, the summarisation and translation tasks lead to a loss of information which cannot be adequately captured by the association functions we have examined. After performing the association of English summaries and MEAD generated summaries from the documents, a basic similarity measure such as TF-IDF achieved a P@1 score of 98.0 and MRR of 99.2. This means that the translation component is the major source of precision loss in the two-stage method. The translation component translates each Italian sentence to exactly one English sentence. For translating each sentence, it selects the translation with highest score according to its model to produce a fluent

Table 3. Results of document-to-summary association of the one-stage approach with different similarity measures.

Similarity	P@1	MRR	Similarity	P@1	MRR
TF-IDF	88.6	92.1	$I(n)L2$	90.1	93.1
TF_k-IDF	89.1	92.4	$I(F)B2$	81.8	86.9
IDF	86.0	89.8	$I(n_e)B2$	86.5	90.6
BM25b	89.6	93.0	$I(n_e)C2$	86.0	89.9
OkapiBM25	**91.7**	**94.3**	$PL2$	90.1	93.2

English. The produced sentence only contains one possible translation for each word or phrase. On the other hand, the one-stage approach considers all the possible translations in the lexical model for each word, hence having a higher chance of finding a match between document words and summary words. The 91% success rate of the one-stage approach, shows it is possible to associate the majority of the summaries to their documents in this collection. The results of the text fragment alignment show the difficulty of finding the same summaries, while they are mixed with other summaries.

4.2 Text Fragment Alignment Evaluation

To find out the cross-lingual effect of the task, we performed the text fragment alignment algorithm on mono-lingual data as well as the cross-lingual data. For each word only the top 5 translations based on the their translation weights are picked. The threshold is set to the average score of the alignment links, therefore alignment links with score less than the average are discarded. For each similarity measure, the alignment algorithm is run $2,000$ times to select different variations of the documents and summaries.

The goal of text fragment alignment is to find the longest relevant fragments of text on each side, without including irrelevant sentences. Therefore, both recall and precision are important in evaluating the algorithm. F-measure combines the two, to give one single score to demonstrate the performance of the algorithm. To calculate the F-measure, each sentence on the e side is labelled true positive if it belongs to a fragment, which is fully or partially correctly aligned. The sentence is labelled false positive if it belongs to a fragment which is incorrectly aligned. It is a false positive instance, if it is not aligned and it should not have been. A false negative instance is an unaligned sentence, which should have been aligned. F-measure is calculated based on these labels for both sides, English to foreign and foreign to English.

Table 4 shows the results of both mono-lingual and cross-lingual text fragment alignment experiments. As expected, the results of the mono-lingual text fragment alignment are higher than the cross-lingual runs. In all settings and in both directions (source to target and target to source), models based on DFR substantially outperform TF-IDF weighting methods. In both mono-lingual and

Table 4. The results of text fragment alignment, for mono-lingual and cross-lingual. For mono-lingual, source and target (src2trg and trg2src) are both English documents and summaries. In cross-lingual settings, source is English documents and target is Italian summaries.

	Mono-lingual				Cross-lingual			
	μF_1	μF_1	MF_1	MF_1	μF_1	μF_1	MF_1	MF_1
Similarity	**src2trg**	**trg2src**	**src2trg**	**trg2src**	**src2trg**	**trg2src**	**src2trg**	**trg2src**
TF-IDF	33.5	77.0	34.9	77.0	23.0	28.5	23.6	27.0
TF_k-IDF	35.2	76.4	35.4	76.3	22.4	28.7	21.5	26.6
$I(n)L2$	35.8	80.2	35.7	80.3	30.0	**32.9**	29.4	**31.8**
$BB2$	34.5	**88.1**	34.9	**88.0**	27.6	32.3	28.2	31.4
$I(F)B2$	35.0	81.9	35.2	81.9	27.4	31.6	27.9	30.4
$I(n_e)B2$	34.9	74.3	35.4	74.2	27.9	31.9	28.4	30.7
$I(n_e)C2$	**38.3**	71.7	**38.2**	71.5	29.0	31.4	28.7	30.3
$PL2$	35.8	79.1	35.5	79.0	29.6	32.5	28.8	31.2
BM25b	36.7	72.4	36.7	72.1	30.8	32.5	30.1	31.3
OkapiBM25	37.3	71.3	37.5	71.1	**31.5**	31.9	**31.0**	31.0

cross-lingual runs OkapiBM25 performs consistently very well compared to others. It has been pointed out by [1] that BM25 formula can be derived from the model $I(n)L2$, which has the highest score in the target to source cross-lingual runs and it is very close to other BM25 scores. Substantial drop of F-measure score of the target to source direction of the cross-lingual runs compared to mono-lingual ones, shows that the summary to document alignment is more prone to translation than the other direction.

Two important components of all similarity methods used in these experiments, are document length and average document length in the collection. These factors are considered to reduce the effect of variance in document length in text collections. However, since in our experiments, a document is the collection and its sentences are the documents, the variance of document length does not exist. To see the effect of this fact, we investigated two other ways to estimate sentence length and used them instead of the default method, which was number of tokens. One is sum of the term frequency in the document for each term in the sentence[2] and the other one, the sum of their selectivity (inverse sentence frequency)[3]. Both methods produced different results for all the runs, however, they were most of the times slightly worse than the number of tokens, and in general the differences were negligible. Only for TF-IDF similarity, the sum of the selectivity of the terms performs slightly better than the number of tokens, but in all other cases it was behind the latter. We concluded that even though there is a difference between sentence length variation and document

[2] $\text{len\_tf}(s, \boldsymbol{d}) := \sum_{t \in s} \text{tf}(t, \boldsymbol{d})$, where s is a sentence in document $\boldsymbol{d}$.

[3] $\text{len\_isf}(s, \boldsymbol{d}) := \sum_{t \in s} \text{sf}(t, \boldsymbol{d})^{-1}$, where s is a sentence in document $\boldsymbol{d}$ and $\text{sf}(t, \boldsymbol{d})$ is the number of sentences in $\boldsymbol{d}$ that contain t.

length variation in large collections, the DFR models perform well, regardless of length estimation method, in the context of sentence similarity.

5 Conclusion and Future Work

We developed an algorithm to perform cross-lingual text fragment alignment and ran a series of experiments with different similarity measures based on models of divergence from randomness. The results show that term statistics based on divergence models are consistently superior to TF-IDF schemes. Despite the fact that sentences tend to be similar in length, we discovered that other ways of estimating sentence length does not improve the quality of the alignment compared to the basic method of counting the number of the tokens. In addition, for the source to target alignment the cross-lingual scores were not substantially lower than the mono-lingual ones, which shows that the translation component performs well enough not to degrade the overall performance considerably.

Preliminary investigation of cross-lingual association of documents and their summaries showed that a one-stage direct computation of similarity using a probabilistic dictionary (lexical probabilities) significantly outperforms a method that translates and summaries the documents and estimates a mono-lingual similarity between the documents. Experiments on mono-lingual associating of generated summaries and manual summaries showed that the low performance of the two-stage method is mainly due to the selective nature of the translation component. One translation is chosen among a list of possible translations based on the context of the sentence and the rest of the candidates are discarded, therefore, the chance of a match between the words of the two documents are heavily degraded.

Although the scores of the basic similarity measures were lower than most of the models of DFR in the association task, the difference was not substantial. In other words, even the basic models of similarity performed well in finding the corresponding summary for a document in our experiments.

These research results can be used to align multi-lingual content in resources such as Wikipedia, or other Wiki-based web sites, where the documents are often not parallel in the different languages.

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20, 357–389 (October 2002)
2. Barrón-Cedeño, A., Rosso, P., Pinto, D., Juan, A.: On cross-lingual plagiarism analysis using a statistical model. In: Proceedings of the ECAI'08 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse. pp. 9–13. Patras, Greece (July 2008)
3. Bernstein, Y., Zobel, J.: A scalable system for identifying co-derivative documents. In: Proceedings of 11th International Conference on String Processing and Information Retrieval (SPIRE). pp. 55–67. Padova, Italy (October 2004)

4. Bernstein, Y., Zobel, J.: Redundant documents and search effectiveness. In: Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management. pp. 736–743. Bremen, Germany (November 2005)
5. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Comput. Linguist. 19(2), 263–311 (June 1993)
6. Daumé III, H., Marcu, D.: A phrase-based HMM approach to document/abstract alignment. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 119–126. Barcelona, Spain (July 2004)
7. Koehn, P.: Europarl: A parallel corpus for statistical machine translations. In: MT Summit X. pp. 79–86. Phuket, Thailand (September 2005)
8. Ma, X.: Champollion: A robust parallel text sentence aligner. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC). Genova, Italy (May 2006)
9. Munteanu, D.S., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Comput. Linguist. 31, 477–504 (December 2005)
10. Munteanu, D.S., Marcu, D.: Extracting parallel sub-sentential fragments from non-parallel corpora. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL). pp. 81–88. Sydney, Australia (July 2006)
11. Och, F.J., Tillmann, C., Ney, H.: Improved alignment models for statistical machine translation. In: Proceedings of the Joint SIGDAT Conference of Empirical Methods in Natural Language Processing and Very Large Corpora. pp. 20–28. College Park, MD (1999)
12. Pouliquen, B., Steinberger, R., Ignat, C.: Automatic identification of document translations in large multilingual document collections. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP). pp. 401–408 (September 2003)
13. Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., Zhang, Z.: MEAD - a platform for multidocument multilingual text summarization. In: LREC 2004. Lisbon, Portugal (2004)
14. Uszkoreit, J., Ponte, J.M., Popat, A.C., Dubiner, M.: Large scale parallel document mining for machine translation. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING). pp. 1101–1109. Beijing, China (August 2010)
15. Yahyaei, S., Monz, C.: The QMUL system description for IWSLT 2010. In: Proceedings of the Seventh International Workshop on Spoken Language Translation (IWSLT). pp. 157–162. Paris, France (December 2010)