Information Processing and Management 52 (2016) 1004–1017

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information

Nava Ehsan[a], Azadeh Shakery[a,b,*]

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
[b] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Iran

ARTICLE INFO

Article history:
Received 21 February 2015
Revised 11 April 2016
Accepted 18 April 2016
Available online 29 April 2016

Keywords:
Candidate document retrieval
Cross-language plagiarism detection
Text segmentation
Proximity-based retrieval

ABSTRACT

The rapid growth of documents in different languages, the increased accessibility of electronic documents, and the availability of translation tools have caused cross-lingual plagiarism detection research area to receive increasing attention in recent years. The task of cross-language plagiarism detection entails two main steps: candidate retrieval and assessing pairwise document similarity. In this paper we examine candidate retrieval, where the goal is to find potential source documents of a suspicious text. Our proposed method for cross-language plagiarism detection is a keyword-focused approach. Since plagiarism usually happens in parts of the text, there is a requirement to segment the texts into fragments to detect local similarity. Therefore we propose a topic-based segmentation algorithm to convert the suspicious document to a set of related passages. After that, we use a proximity-based model to retrieve documents with the best matching passages. Experiments show promising results for this important phase of cross-language plagiarism detection.

1. Introduction

Plagiarism refers to unauthorised use of text, code and ideas (Potthast, Barrón-Cedeño, Stein, & Rosso, 2011). In automatic cross-language plagiarism detection, the task is to retrieve plagiarized text written in language L that has originated from another document in a language other than L. With the rapid growth of documents in different languages, the increased accessibility of electronic documents, and the availability of translation tools, cross-language plagiarism has become a serious problem and its detection requires more attention.

Given a suspicious document s' and a set of potential source documents D, we should determine whether a fragment of the suspicious document, $s'_{f'} \in s'$, was borrowed from a source document. This task comprises two main steps: candidate retrieval and detailed analysis. Candidate retrieval entails the identification of source documents that contain suspicious fragments. Detailed analysis requires closer comparison of the subject document with each suspected source and retrieval of plagiarized fragments. In this paper we focus on the first step, candidate document retrieval. Since a second phase will follow this step to eliminate false positive matches, we are more interested in high recall than in high precision in this research.

[*] Corresponding author. Tel.: +00982182089722.
E-mail addresses: n.ehsan@ece.ut.ac.ir (N. Ehsan), shakery@ut.ac.ir (A. Shakery).

http://dx.doi.org/10.1016/j.ipm.2016.04.006
0306-4573/© 2016 Elsevier Ltd. All rights reserved.

In cross-language plagiarism detection the languages of source and suspicious documents differ. To date only a few approaches have been focused on cross-language plagiarism detection (Barrón-Cedeño, Gupta, & Rosso, 2013a). Most previous methods are based on translating the whole suspicious or source documents coupled with monolingual techniques (Barrón-Cedeño et al., 2013a). Document translation depends on the existence and quality of machine translators. Translating documents in languages with low quality translation tools may cause poor quality documents. In this paper we propose an approach for the candidate retrieval phase of cross-language plagiarism detection which only considers a set of representative words and phrases extracted from each document as its content representation, instead of using the whole text. Since documents are represented by some extracted words and phrases, this approach is insensitive to punctuation, extra white space, and permutation of the document context and requires less translation time rather than translating the entire document. Our approach is therefore less dependent on the quality of machine translation between two languages, and if there is not a high quality translation tool available, any other translation resources such as dictionaries, parallel or comparable corpora could be used for translating representative words. Thus, our approach is applicable in languages with even limited translation resources.

Since plagiarism usually happens in parts of the text, there is a requirement to segment the texts into fragments to detect local similarity. There are some previous works that break the document into constituent parts such as sections, paragraphs (Nawab, 2012), or a fixed number of sentences (Pereira, Moreira, & Galante, 2010). In this paper a topic-based text segmentation approach is proposed in order to break the document based on its topical structure. Thus, a set of topically related passages from the suspicious document are used to retrieve potential sources.

In our proposed candidate retrieval process, after segmentation, we use a second level for considering proximity in retrieval of candidate documents. For each segment the word proximity is measured with positional language modelling (PLM) (Lv & Zhai, 2009). We believe this to be the first use of PLM in cross-language plagiarism detection.

We present the results of no-segmentation with a non-proximity-based language model as a baseline. According to the candidate document retrieval experiments, the segmentation technique increased F_2 measure about 0.11 (21% improvement) over the baseline. Accompanying the segmentation technique with the positional language model increased F_2 measure about 0.13 (25% improvement) over the baseline. These results are further compared with CL-CNG (Mcnamee & Mayfield, 2004) and a combination of translation and monolingual analysis. The proposed approach with text segmentation and using proximity-based retrieval outperforms these approaches with respect to F_2.

The rest of the paper is organized as follows: Section 2 outlines related work in cross-language plagiarism detection. Section 3 describes the candidate document retrieval process in which the text segmentation approach, representative word extraction, and retrieval model are explained. Finally the experimental framework and results are discussed in Section 4, and our conclusion and future work are reported in Section 5.

2. Related work

Plagiarism detection methods can be classified into two approaches, intrinsic and external (Potthast et al., 2012). Intrinsic detection methods are those that use style analysis to detect parts of the text that are inconsistent in terms of writing style (Meyer zu Eißen & Stein, 2006; Oberreuter & Velásquez, 2013). The aim of external plagiarism detection methods is not only finding the suspicious text, but also finding the source for the plagiarized text. In monolingual plagiarism detection, parts of the suspicious text could be an exact copy or a modified copy, and those parts should be large enough to be more than just a coincidence. In cross-language plagiarism, the suspicious document in language L originates from another document in a language other than L. The plagiarized fragment could be the exact translation or a paraphrased translation.

In order to detect cross-lingual plagiarism, either cross-language similarity is used or the document is translated and monolingual similarity is used. There are five kinds of cross-language similarity assessment models that have been proposed in the literature (Franco-Salvador, Gupta, & Rosso, 2013; Potthast et al., 2011), (1) syntax-based, (2) dictionary-based, (3) parallel corpus-based, (4) comparable corpus-based, and (5) multilingual semantic network based approaches.

Syntax-based methods rely on lexical similarities between languages. Cross-Language Character N-Gram (CL-CNG) is a syntax-based method (Mcnamee & Mayfield, 2004). In this model, documents are represented by overlapping character n-grams. Defining the alphabet $\sum$ and n, the texts will be coded into character n-grams. The resulting texts are compared by means of similarity measures. The model is useful for comparing multilingual documents without translation, and is applicable for languages with similar syntax (Potthast et al., 2011), but it is ineffective when the languages differ syntactically.

Pouliquen et al. propose a dictionary based method to find similar documents in a multilingual document collection (Pouliquen, Steinberger, & Ignat, 2006). They map the document content to a vector of descriptors from the Eurovoc thesaurus and measure the semantic similarity between the resulting vectors. In this work the authors assume that the documents are completely similar, whereas in plagiarism detection the similarity could happen in parts of the text only.

CL-ASA (Barrón-Cedeño, Rosso, Pinto, & Juan, 2008), LSI (Dumais, Letsche, Littman, & Landauer, 1997) and KCCA (Vinokourov, Cristianini, & Shawe-taylor, 2002) use parallel corpora in order to find cross-language similarity, while Cross-Language Explicit Semantic Analysis (CL-ESA) (Potthast, Stein, & Anderka, 2008), which is the cross-lingual generalization of ESA (Gabrilovich & Markovitch, 2007), uses comparable corpora for this purpose. CL-ASA (Barrón-Cedeño et al., 2008) and CL-CNG (Mcnamee & Mayfield, 2004) are compared in (Barrón-Cedeño et al., 2013a) for document-level retrieval when suspicious documents are entirely plagiarised from sources.

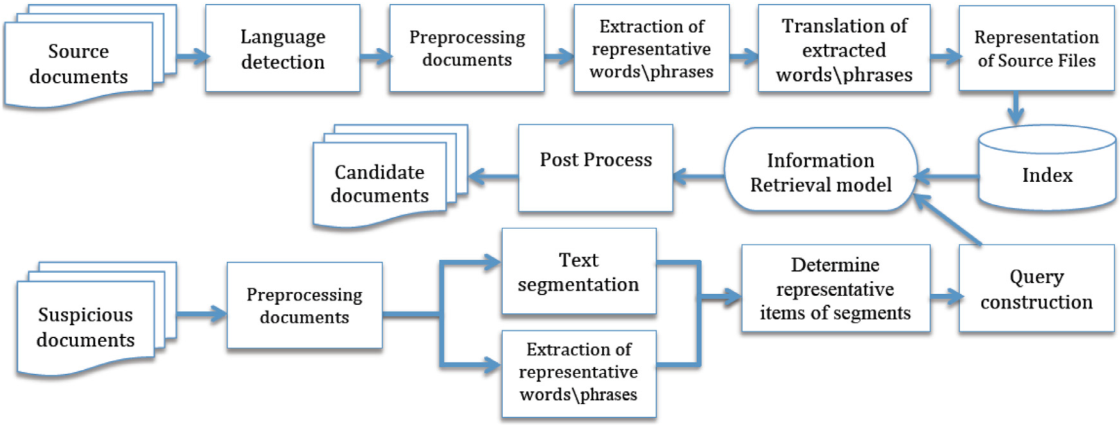
Source documents
Language detection
Preprocessing documents
Extraction of representative words\phrases
Translation of extracted words\phrases
Representation of Source Files
Index
Information Retrieval model
Post Process
Candidate documents
Suspicious documents
Preprocessing documents
Text segmentation
Extraction of representative words\phrases
Determine representative items of segments
Query construction

Fig. 1. Process of cross-language plagiarism candidate retrieval.

The BabelNet[1] multilingual semantic network, can be used for mapping between words in different languages. The performance of this approach is reported for the detailed analysis phase (comparing source and suspicious pairs) (Franco-Salvador et al., 2013).

There have been some attempts to translate suspicious documents and use monolingual approaches. The models of this type are called T+MA (Translation followed by Monolingual Analysis) (Barrón-Cedeño et al., 2013a). One method for detecting monolingual plagiarism is comparing fragments of suspicious and source documents using fingerprint indexing. For example, the Winnowing approach (Schleimer, Wilkerson, & Aiken, 2003), which is used in the widely used plagiarism detection tool MOSS, is based on fingerprint indexing. Pereira et al. propose a classification method in which the text is first translated, normalized and divided into sub-documents (Pereira et al., 2010). Then using a training collection some features are selected to build a decision tree classification model. Anguita et al. detected plagiarism in Spanish from English sources by translating suspicious fragments, using a search engine to extract similar documents, and then using cosine similarity to compare sentences of highly ranked documents (Anguita, Beghelli, & Creixell, 2011). There have also been some attempts to detect paraphrased sentences in monolingual texts (Androutsopoulos & Malakasiotis, 2010). According to monolingual experiments (Barrón-Cedeño, Vila, Martí, & Rosso, 2013b), paraphrasing could make plagiarism detection more difficult.

In recent years the PAN competition has provided an evaluation environment for plagiarism detection algorithms. This competition also offers evaluation corpora for plagiarism detection[2].

3. Proximity-based candidate document retrieval

The problem of candidate retrieval is defined as follows: given a suspicious document s' and a set of potential source documents D, retrieve those source documents $s \in D$ that likely contain source texts of some fragments of the suspicious document $s'_{f'} \in s'$.

Our proposed process of cross-language candidate retrieval is depicted in Fig. 1. Since source documents may be in different languages, the process on source documents starts with language detection. Language detection process uses the frequency of stopwords of each language to select the appropriate language (Johnson, 1993). This step can be omitted if all the source documents are in a single known language. In the next step, the text is pre-processed by converting it to lower case, removing stopwords, punctuation marks and extra white spaces and normalizing diacritic characters. After pre-processing representative words and phrases of the source and suspicious documents are extracted. The representative words and phrases of source documents are translated to the target language (suspicious document language). Representative word and phrase extraction is described in Section 3.2.

Suspicious documents are divided into segments. The segmentation process is the first level of proximity consideration. The goal is to split the text into meaningful units and create queries from these units in order to capture local similarities. Each segment is represented by the containing words and phrases. Text segmentation is explained in Section 3.1.

The next step is calculating the similarity between texts with an information retrieval model. This is our second level of proximity consideration where we use a method considering positional language models as the retrieval model to reward documents containing query words close to each other. This step is described in Section 3.3. A post process is applied on the retrieved results to avoid coincidental similarity between texts. The post-processing is described in Section 3.4.

[1] http://babelnet.org/

[2] http://www.uni-weimar.de/medien/webis/events/pan-12/pan12-web/plagiarism-detection.html

3.1. Text segmentation

A plagiarism detection method should be able to detect local similarities where only a short passage is common to both documents. Since plagiarism usually happens in parts of the text, there is a requirement to segment the texts into meaningful fragments. Text segmentation helps to retrieve those documents which contain not only similar words, but also the words that are placed in close proximity.

A language-model based text segmentation is proposed in (Stolcke & Shriberg, 1996). In this model the Viterbi algorithm is used to identify the most likely locations of segment boundaries according to the language model. This text segmentation method is implemented in the SRILM toolkit[3]. The language model used by SRILM considers the words in the text without the semantic relatedness. Thus, this approach lacks the advantage of using semantic coherence between parts of the text. There are some methods that use the similarity between the words of sentences to segment the text into sub-topic passages (Hearst, 1997; Utiyama & Isahara, 2001). The TopicTiling method (Riedl & Biemann, 2012), uses the topical similarity between the sentences in order to segment the text into sub-topic passages. In this model each block is represented as a T-dimensional vector and the coherence score between adjacent vectors is calculated by cosine similarity. The local minima in coherence scores are considered as possible segmentation boundaries.

In the following subsection we will propose a topic-based text segmentation algorithm. The algorithm is a hierarchical clustering method. The advantages of the proposed hierarchical topic-based algorithm are using topical probability distributions and performing a probability based similarity measure for comparing the probability distributions, instead of using a vector-based model. This means that instead of comparing the similarity of two vectors by cosine similarity, we evaluate whether the two chunks have similar topical distributions. The other advantage is its hierarchical characteristic. As the chunks are merged, the probability distributions and similarities are updated. Thus, the final probability distribution of each segment is computed by an update formula, which considers the topical probability distribution of the members of that segment.

3.1.1. Topic-based text segmentation

Topic-based segmentation is proposed with the objective of creating semantically related sequences of text. The proposed model is a hierarchical clustering method which uses topical probability distributions obtained from Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003).

In the first step, it is necessary to eliminate stopwords, digits, and punctuation in each document. This preprocessing is needed because the eliminated words do not contain any special topic and assigning topics to these words by LDA process will have an unfavourable effect on the topical probability distribution of text. The text is then split into the intended chunks. These chunks (e.g. the sentences of a document) are the smallest units for calculating topical probability distribution. The topical distribution probability of the chunks of each document are then obtained by the LDA model. Each chunk of document is represented by a specific topic distribution. The segmentation model uses these topical distribution probabilities to merge semantically related parts.

The process of the proposed topic-based segmentation algorithm is as follows: after obtaining the topic-based probability distribution for each chunk of text with the LDA model, the similarity of consecutive chunks are evaluated with the Jensen–Shannon (JS) divergence method (Lin, 1991). Jensen–Shannon divergence is a symmetric method of measuring the similarity between two probability distributions, and it is defined as follows for chunks A and B:

$$JS(A||B) = -(\frac{1}{2}KL(A||M) + \frac{1}{2}KL(B||M)) \tag{1}$$

where $KL(A||M)$ represents the Kullback–Leibler divergence of A and M (Kullback & Leibler, 1951), $M = \frac{A+B}{2}$, and t represents the topics.

The maximum similarity is considered for selecting topic-based similar sequences. The sequences with highest similarity are merged and the probability distribution of the new chunk is updated using the Dirichlet distribution. The generated topic for the words in documents are used for updating distributions. Considering a chunk d_j with n_{d_j} words containing $n^i_{d_j}$ words with topic i and a chunk d_k with n_{d_k} words containing $n^i_{d_k}$ words with topic i, the probability distribution of the resulting chunk from merging d_j and d_k is computed as follows:

$$p(t_i|d_j \odot d_k) = \frac{n^i_{d_j} + n^i_{d_k} + \alpha}{n_{d_j} + n_{d_k} + T * \alpha} \tag{2}$$

where $d_j \odot d_k$ reflects merging the chunks d_j and d_k and $p(t_i|d_j \odot d_k)$ is the probability of topic i in the resulting chunk. T is the number of topics and α is the parameter of Dirichlet distribution and is recommended to be $^{50}/_T$ (Griffiths & Steyvers, 2004). According to the new probability distribution, similarities between the next and previous chunks are also updated using Eq. 1. Thus, the algorithm starts with the preliminary probability distribution of each chunk obtained from the LDA model and progressively updates the statistical topical distribution of chunks and similarities between the chunks. In this

[3] http://www.speech.sri.com/projects/srilm/

case if one chunk does not have sufficient context to represent its topic, its topical distribution will be updated by merging with other chunks. The iterations continue as long as the similarity is above $\mu - {}^{\sigma}/_{2}$, where μ is the mean and σ is the standard variation of the scores. According to the non-deterministic nature of LDA model, the algorithm is repeated k times. In order to identify the best segmentation, we used Davies–Bouldin Index (DBI) (Davies & Bouldin, 1979) as an internal metric for evaluating clusters. The segmentation with minimum DBI is returned as the final segmentation.

3.2. Extraction of representative words and phrases

Given a document, a preprocessing phase is performed and then representative items are extracted. The reason for extracting representative items is to select only those words (phrases) which give us the chance to retrieve source documents (fragments) matching the suspicious document (fragments). These items should also be a good representative for the document. Retrieving representative items from documents for translation, rather than using the whole document, causes this approach not to be very dependent on the quality of machine translators.

A subset of representative words are extracted based on two common heuristics in information retrieval, term frequency (TF) and inverse document frequency (IDF). 'Tf' factor provides a measure of how well the term describes the document context and 'IDF' factor is used for quantification of inter-document dissimilarity (Manning, Raghavan, & Schütze, 2008).

Another subset of words are extracted by analogy with mixture model feedback method (Zhai & Lafferty, 2001a) where each document is considered to be generated from a combination of its topical language model and a collection language model. The words with highest probabilities in the topical language model are considered as representative words.

We also extract some keyphrases, because they are less ambiguous than words and translation of phrases is more accurate than words translations. For example we would like to deal with 'bank account' as a phrase. Since, our task is in cross-language domain accurate translation would help to retrieve more relevant documents. Furthermore, many complex or technical concepts or product names are expressed with multi-word compounds. Words that often appear close to each other may be more relevant. For this reason we also select bigrams and trigrams with high 'TF' factor within the document as representative phrases. These phrases represent the sequences of words that an author uses frequently. Also, the trigrams with high language model probability in the entire collection for each language could represent phrases in that language. Existence of phrase translations in suspicious sentences could be a signal of plagiarism. Trigram language model probabilities are evaluated with SRILM toolkit in which Witten-Bell discounting is used for smoothing n-gram counts (Stolcke, 2002). Extracted words and phrases from source document are translated using the Google translation tool. For suspicious documents, in addition to words and phrases described above, the words which occur only once in the suspicious document are also considered as representative words for candidate retrieval.

3.3. Retrieval model

In the retrieval step, the representative words and phrases from segments of suspected documents are used as a query against an index of source documents to retrieve candidate documents. In the proposed retrieval model, we retrieve the candidate documents in two steps. First, we compare the representative words of queries and the source documents with the Jelinek–Mercer (JM) smoothing method (Zhai & Lafferty, 2001b) and the KL-divergence (Lafferty & Zhai, 2001) measure. Second a proximity-based model is used to rerank the results of the previous step. The KL-divergence score for query Q and document D, $S(Q, D)$ is calculated as follows (Lafferty & Zhai, 2001):

$$S(Q, D) = -\sum_{w\in V} p(w|Q) log \frac{p(w|Q)}{p(w|D)} \tag{3}$$

where V is the vocabulary set and $p(w|.)$ is an estimated language model.

We did not use the proximity-based approach in the first step, because we wanted to retrieve the documents with a great diversity of common words with the query. The proximity-based approach rewards those documents which contain query terms close to each other. If only few query terms appear close to each other, or there is close repetition of a query word, that document would be rewarded which is not desired.

The proximity-based model is used as a reranker in the second step. In considering the proximity of words, translations of representative items of source documents are replaced with original ones in the source documents. Other words are replaced with a $ sign to preserve the distance between the words.

The reranking step using the proximity-based model is as follows. Documents are compared with KL-divergence using the positional language model (Lv & Zhai, 2009), which rewards documents where the query terms, representative items of each segment, appear close to each other and thus retrieves documents with the best matching passage. The positional language model at position i is calculated as follows (Lv & Zhai, 2009):

$$S(Q, D, i) = -\sum_{w\in V} p(w|Q) log \frac{p(w|Q)}{p(w|D, i)} \tag{4}$$

where $p(w|.)$ is an estimated language model. In this model, each word in each position in the document propagates its occurrence. Thus, $P(w|D, i)$ is a positional language model in position i which is smoothed using the JM smoothing method.

The propagation function $k(i, j)$ is the propagated count to position i from a term in position j with parameter σ that restricts the propagation scope of each term. The Gaussian, Passage and Circle kernels are defined as follows (Lv & Zhai, 2009):

Gaussian kernel:

$$k(i,j) = exp\left[\frac{-(i-j)^2}{2\sigma^2}\right] \tag{5}$$

Passage kernel:

$$k(i,j) = \begin{cases} 1 & \text{if } |i-j| \le \sigma \\ 0 & \text{Otherwise} \end{cases} \tag{6}$$

Circle kernel:

$$k(i,j) = \begin{cases} \sqrt{1-(\frac{|i-j|}{\sigma})^2} & \text{if } |i-j| \le \sigma \\ 0 & \text{Otherwise} \end{cases} \tag{7}$$

3.4. Post process

The objective of the post processing step is to avoid retrieving those documents with coincidental similarity between the texts. The plagiarized part of the suspicious document could be a subset of the entire document. Thus, we have some segments that do not contain any plagiarized part, but may have coincidental similarity to some source files. For this reason we need to consider the difference between the retrieval scores to prune the results. This means that, if the retrieval score between a segment and some source files slightly differs from the retrieval scores between the segment and other source files, the similarity could be coincidental, but if the retrieval score between a segment and a source file greatly exceeds other similarities, it would be a sign of plagiarism. Thus for each segment of a suspicious document, we have a list of potential similar source files. The best result for each segment is reported as a potential source document if the retrieval score between that segment and the source file considerably differs from other retrieval scores. To make this determination, the average of differences of retrieval scores between other documents is considered as a threshold.

4. Experimental framework

The aim of this section is to analyse and discuss the proposed approaches in text segmentation and plagiarism candidate retrieval. For this reason, we performed two sets of experiments. The first set of experiments evaluate the performance of the text segmentation algorithm (reported in Section 4.1) and the second set of experiments evaluate the performance of the cross-lingual plagiarism candidate retrieval (reported in Section 4.2).

4.1. Evaluation of text segmentation

To study the performance of our topic-based text segmentation algorithm, we used the Choi dataset (Choi, 2000), which is commonly used in the field of text segmentation (Riedl & Biemann, 2012). The corpus is artificially generated from the Brown corpus (Francis & Kucera, 1979), which consists of segments with sentence counts of 3–5, 6–8, 9–11 and 3–11. In each range of segment lengths, we used 50 files, where each file includes 10 segments.

The evaluation metric we use is boundary similarity, which overcomes the flaws of other existing text segmentation metrics (Fournier, 2013). In computing the boundary similarity, edit operations are considered to model segmentation differences. The edit operations are insertion/deletion A_e and n-wise transposition T_e. Since in our task the type of segments are the same, there is no substitution operation. The boundary similarity between two segmentations s_1 and s_2 is evaluated as follows (Fournier, 2013):

$$Boundary - Similarity(s_1, s_2, n_t) = 1 - \frac{|A_e| + w_{t\_span}(T_e, n_t)}{|A_e| + |T_e| + |B_M|} \tag{8}$$

$$w_{t\_span}(T_e, n_t) = \sum_{j=1}^{|T_e|}(w_t + \frac{abs(T_e[j][1] - T_e[j][2]}{n_t - 1}) \tag{9}$$

where $|B_M|$ is the number of all matches, and n_t is the largest distance between boundary positions that could be considered as a near miss. The transposition severity is calculated by Eq. 9 in terms of the distance between paired boundaries over n_t plus w_t, where w_t is zero by default (Fournier, 2013). The results are evaluated using the SegEval[4] toolkit.

[4] http://segeval.readthedocs.org/en/latest/

Table 1
Boundary Similarity values of text segmentation evaluation on the Choi dataset.

Method	Range of segment lengths			
	3-5	6-8	9-11	3-11
TopicTiling	0.26	0.26	0.17	0.2
Topic-based segmentation	0.53	0.34	0.26	0.35

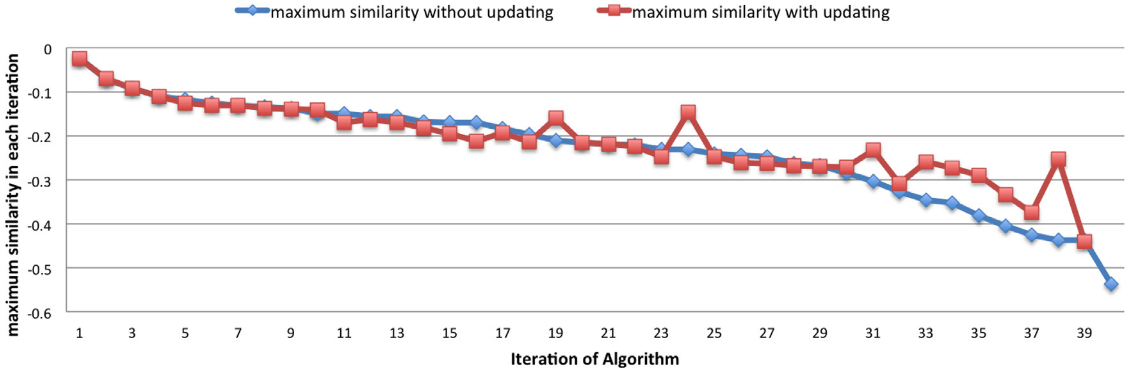

Fig. 2. Effect of updating the similarities and probability distributions in each iteration.

In the experiments, we used JGibbsLDA (Phan, Nguyen, & Horiguchi, 2008) for executing LDA and we followed the standard parameters for the number of topics, and Dirichlet priors (α and β), which were set to 100, 0.5 and 0.01, respectively (Griffiths & Steyvers, 2004; Riedl & Biemann, 2012). We compare the proposed text segmentation algorithm with another topic-based text segmentation called TopicTiling (Riedl & Biemann, 2012)[5]. In (Riedl & Biemann, 2012) it is shown that TopicTiling outperforms previous text segmentation algorithms reported in (Choi, 2000), (Utiyama & Isahara, 2001), (Galley, McKeown, Fosler-Lussier, & Jing, 2003), (Fragkou, Petridis, & Kehagias, 2004) and (Misra, Yvon, Jose, & Cappe, 2009). Assuming an LDA model with *T* topics, TopicTiling represents each block as a T-dimensional vector, and the coherence score is calculated by cosine similarity for adjacent vectors. The local minima in coherence scores are considered as possible segmentation boundaries. The advantage of our proposed topic-based algorithm over TopicTiling involves using the JS similarity measure for comparing the topical probability distributions, which is motivated by probability theory, rather than using cosine similarity. The other important characteristic of our algorithm is its hierarchical characteristic and updating the probability distributions and similarities in each iteration of the algorithm. In this case the probability distributions are updated after merging with other sentences in order to refine the probability distributions of segments and similarities between consecutive segments. Table 1 shows the results of these two algorithms using boundary similarity evaluation. The tables show that in all ranges of segment lengths we had improvements over TopicTiling. The improvements are about 103, 30, 53 and 75% for ranges of 3–5, 6–8, 9–11 and 3–11 sentence length of segments, respectively. The reason that we obtained better result in segments with 3 to 5 sentences, is that, with fewer sentences in a segment, the topic is more specific and the algorithm is more successful in determining the sentences of the segment.

Fig. 2 shows the effect of updating the similarities and probability distributions in each iteration in a document. We can see that, by updating the probability distribution of chunks, we can obtain higher similarities in some iterations. For example, this increase could be seen in iterations 19 and 24. This means that, by gradually increasing the number of sentences, we will overcome the problem of low context of a sentence; thus, quantifying the similarity between consecutive segments will be more reliable.

We also calculate the entropy of the probability distribution of sentences before and after applying the proposed text segmentation algorithm. This is shown in Fig. 3. The consecutive sentences with equal entropy are those located in a segment. The reduction of entropy after segmentation shows that the topics of the segments are more specific compared to the topics of the chunks before segmentation. We compared the entropy of the resulting segmentation with a case of random segmentation, TopicTiling, and correct segmentation. In random segmentation we merged the consecutive sentences randomly, until the number of clusters was the same as our proposed segmentation method. We can see that the entropy of segments using the proposed method is close to entropies with correct segmentation. Thus, by taking advantage of topical model and using similarities between the topical probability distributions and updating the similarities gradually, we could find sentences that are more likely to be on the same topic, which would help us to segment the text according to its topics.

[5] http://sourceforge.net/projects/topictiling/

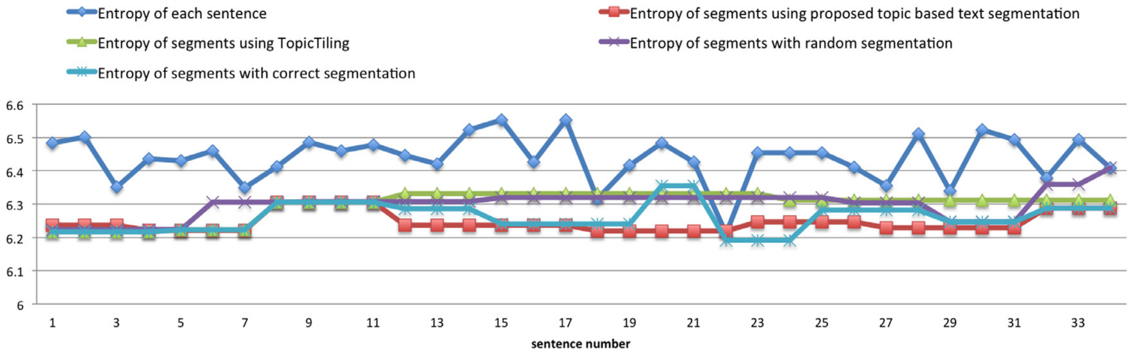

Fig. 3. Entropy of probability distribution.

Table 2
The effect of text segmentation using non-proximity based approach.

	Precision	Recall	F_1	F_2
No segmentation	0.6075	0.5195	0.5600	0.5350
Topic-based text segmentation	0.3130	0.6722	0.4271	0.5468
Language model text segmentation	0.3546	0.7259	0.4765	0.6002
Topic-based and language model text segmentation	0.4432	0.7365	0.5534	0.6504

4.2. Candidate document retrieval experiments

We chose the PAN-PC-12 corpus for evaluating the task of candidate retrieval. The construction principles of the corpus are defined in (Potthast, Stein, Barrón-Cedeño, & Rosso, 2010). The source retrieval corpus is monolingual, but the corpus for detailed analysis in 2012 contain 1000 document pairs where the plagiarized passages are obfuscated by translation into a different language[6] (Potthast et al., 2012). The corpus contains 496 source documents in German and Spanish and 386 suspicious documents in English. Without any candidate retrieval approach, we should compare 191,456 pair documents for pairwise document similarity which is a time consuming task. The average length of the source and suspicious documents are 28,629 and 26,889 words, respectively. The average number of unique words is 4500 for source documents and 3500 for suspicious documents. There are 1000 plagiarized document pairs that should be detected, in which some sentences in a suspicious document are plagiarized by translation from a source document. The average number of sources used for plagiarism for each document is about 2.6. We evaluate the performance with the macro average precision and recall metrics. Macro average recall metric is defined as follows:

$$Recall_{avg} = \frac{1}{|N|} \sum_{i=1}^{N} R_i \tag{10}$$

where R_i is the recall score of the i^{th} suspicious document. Precision and recall are combined into F_1 score, the equally weighted harmonic mean of precision and recall. Since, the source documents missed in this step will not be examined later, it is important in the candidate retrieval to obtain high recall. Thus, the F_2 score is also calculated that weights recall 2 times more than precision.

In these experiments source documents are indexed with the Lemur toolkit[7] and segments of the suspicious documents are queried in order to retrieve the potential source documents. The following subsections describe the effect of different parts of the proposed model for candidate retrieval task.

4.2.1. Text segmentation in candidate document retrieval

First, we will evaluate the effect of text segmentation without using proximity-based retrieval. The details of the non-proximity based language model with respect to precision, recall, F_1, and F_2 measures are reported in Table 2. After retrieving representative words and phrases from the source and suspicious documents, in the first set of our experiments we compare source and suspicious documents without any segmentation and without the proximity based retrieval. In other words, representative items of the source and suspicious documents are compared using the KL-divergence retrieval model. In this case we have 386 queries with average length of 1294 words. The resulting F_2 measure is 0.5350, which is shown in

[6] http://www.uni-weimar.de/medien/webis/corpora/corpus-pan-labs-09-today/pan-12/pan12-data/pan12-text-alignment-training-corpus-2012-03-16.zip
[7] www.lemurproject.org/

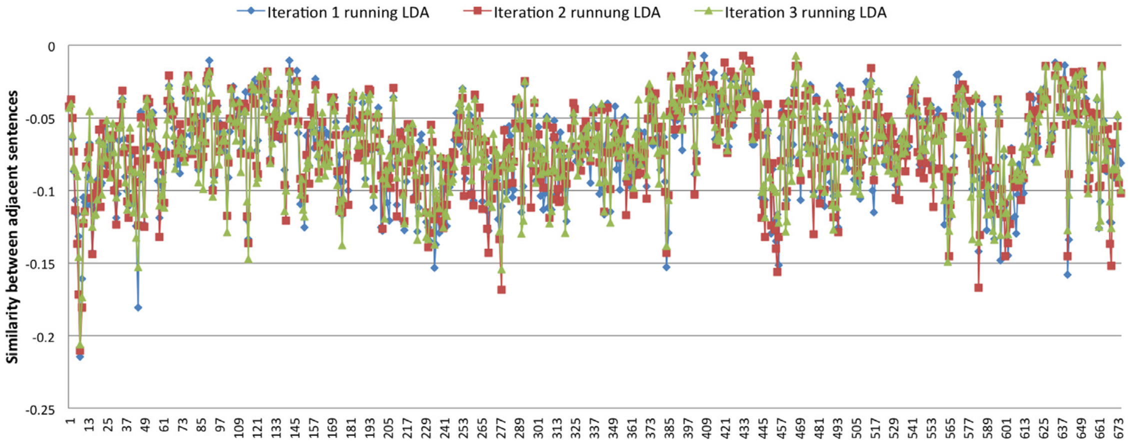

Fig. 4. Similarities between probability distributions of adjacent sentences for one document.

the first row of Table 2. In this case using the ranked retrieved list of documents using the KL-divergence model, the potential source files for each suspicious document are reported as follows: the one-best result for each suspicious document is reported as a potential source document and the filtering post process is performed on documents retrieved in ranks 2–4. The average of differences of retrieval scores between documents of rank i and rank $i+1$ is calculated as a threshold. The documents in ranks 2–4 are also reported as potential source documents until the difference of retrieval score between two following documents is above the threshold.

In the next set of experiments, the text is split into files with the topical text segmentation algorithm described in Section 3.1.1. The result of this part is shown in the second row of Table 2. We used the sentences of each document as the chunks for obtaining topical probability distribution for text segmentation. In order to avoid clusters with only one sentence, we add the criteria that the number of chunks in each cluster should be more than one. The result of text segmentation algorithm is about 90,000 segments for all documents with average length of 8.7 words. The potential source files for each suspicious document are retrieved as follows: for each segment of each suspicious document we have a list of potential similar source files. The one-best result for each segment is reported as a potential source document if it meets the condition of Eq. 11, where d_i represents the difference of retrieval scores between documents of rank i and rank $i+1$

$$\begin{aligned} & d_1 > \alpha * \frac{\sum_{i=2}^{4} d_i}{3} \\ & Score(q,d) = \sum_{t \in q} \frac{idf * tf * (k_1 + 1)}{tf + k_1 * ((1-b) + b * (\frac{l_d}{l_{avg}}))} \end{aligned} \tag{11}$$

Without using proximity based retrieval model, the resulting F_2 measure is 0.5468 which increased about 0.01 and has 2.2% improvement over the baseline without text segmentation. We can see that both recall and F_2 measures are improved by performing text segmentation approach. Fig. 4 illustrates similarities between probability distributions of adjacent sentences in three runs of LDA for one document, and it shows the non-deterministic characteristic of the LDA approach. The figure shows that similarities between probability distributions of adjacent sentences slightly differ in three runs of LDA. Thus, the LDA approach is repeated 5 times and the best segmentation is considered according to the Davies–Bouldin Index metric.

The problem of using sentences as blocks of obtaining topical probability distribution is that the sentences may not have enough context to represent its topic. We also used a language model base text segmentation. A language model-based text segmentation algorithm is proposed in (Stolcke & Shriberg, 1996). A Viterbi algorithm is used to identify the most likely locations of segment boundaries according to the language model of the text. Text segmentation is performed using the SRILM toolkit (Stolcke, 2002). In this case we have about 51,000 segments with average length of 13.6 query terms. In this case the resulting F_2 measure is 0.6002 which increased about 0.06 and has 12.2 percent improvement over the baseline.

Since the language model text segmentation does not use any semantic or topical information, we augmented the approach with our topic-based text segmentation. There are two advantages for augmenting language model and topical-based text segmentation. First, one sentence may not have enough context to obtain its probability distribution. By using the language model we have more context to obtain better topical distributions. Second, since language model segmentation does not use any semantic and topical information of the text, the results would be improved by using the topical information in topic-based text segmentation algorithm. For this purpose, the preliminary chunks for evaluating the LDA model are the logical blocks obtained by the language model based approach rather than sentences. The number of segments in this case was reduced to 18,000 for all documents with 177 unique words on average. The average number of query terms was 34 words per query. The results are pruned using Eq. 11. The resulting F_2 measure from augmenting language model and topic-based

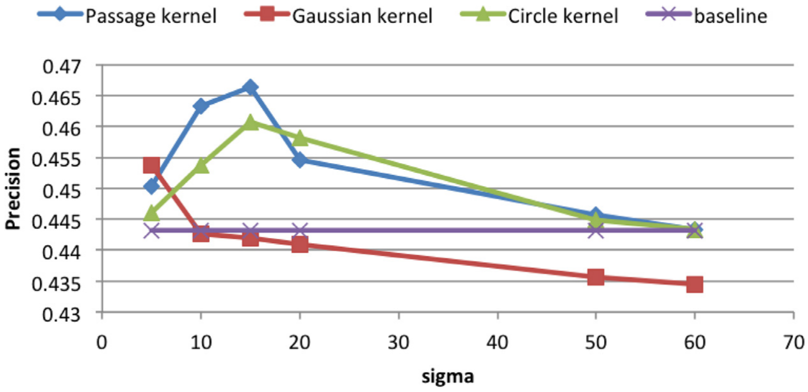

Fig. 5. The effect of changing parameter sigma on resulting precision measure.

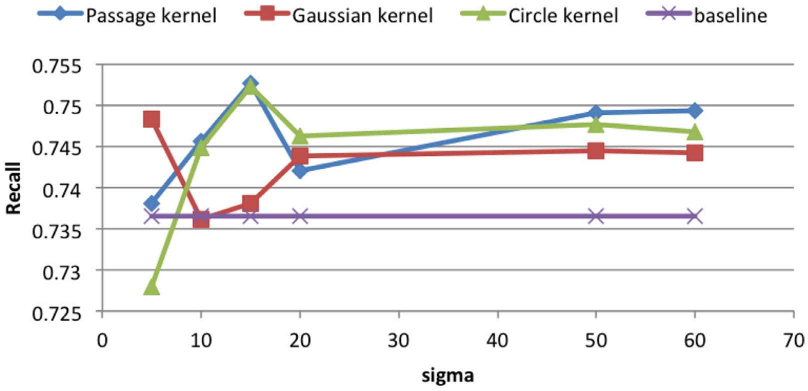

Fig. 6. The effect of changing parameter sigma on resulting recall measure.

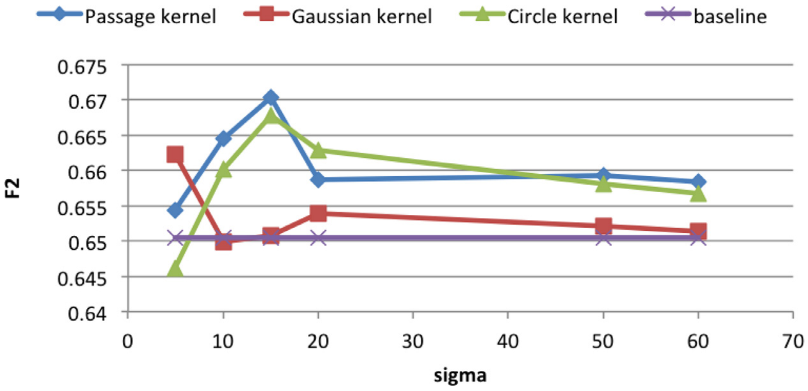

Fig. 7. The effect of changing parameter sigma on resulting F_2 measure.

model is 0.6504 which shows about 0.11 increase (21.5% improvement) over the baseline. We can see that by augmenting the language model information with our topic based text segmentation we obtained higher precision and recall metrics and significant improvement in F_2 measure as shown in Table 2.

4.2.2. Using proximity in candidate retrieval

In the next step we evaluate the results of using proximity-based approach for retrieval. The results obtained from the KL-divergence model in Subsection 4.2.1 are reranked using the positional language model (Lv & Zhai, 2009). The sensitivity to parameter σ on the resulting precision, recall and F_2 measure using the Gaussian, Passage and Circle kernels are shown in Figs. 5, 6 and 7. As we can see, by using Passage and Circle kernels, at the beginning precision and F_2 increase as σ increases for small values, which shows that we need to increase the amount of proximity, and then these metrics start decreasing by increasing σ or reducing the proximity effect. This confirms the importance of proximity in the retrieval

Table 3
Topic based text segmentation using sentence level information.

	Precision	Recall	F_1	F_2
Non-proximity-based language model	0.3130	0.6722	0.4271	0.5468
Proximity-based model	0.3179	0.6830	0.4339	0.5555

Table 4
Topic based text segmentation augmented with language model based text segmentation.

	Precision	Recall	F_1	F_2
Non-proximity-based language model	0.4432	0.7365	0.5534	0.6504
Positional Language modelling	0.4663	0.7526	0.5758	0.6703

Table 5
Summarized results of candidate retrieval experiments.

	Precision	Recall	F_1	F_2
No segmentation				
Non-proximity-based language model	0.6075	0.5195	0.5600	0.5350
Topic-based text segmentation				
Non-proximity-based language model	0.3130	0.6722	0.4271	0.5468
Positional Language modelling	0.3179	0.6830	0.4339	0.5555
Language model text segmentation				
Non-proximity-based language model	0.3546	0.7259	0.4765	0.6002
Language model and topic-based text segmentation				
Non-proximity-based language model	0.4432	0.7365	0.5534	0.6504
Positional Language modelling	0.4663	0.7526	0.5758	0.6703

of candidate documents. In these cases, when σ is set to 15, recall achieves a maximum value and increasing σ further corresponds to decreased recall. Also in the Gaussian kernel as we can see the precision decreases and the F_2 measure gets close to the baseline as σ increases. We can see that the Passage kernel works better in our experiments with σ equal to 15 for candidate document retrieval. Although (Lv & Zhai, 2009) reported that Gaussian kernel works better in information retrieval task, we can see that in plagiarism detection, Passage kernel performs better. The problem of using the Gaussian kernel in our task is that it will not cut the scope of propagation; it just reduces the effect of each term with increase in distance. This reinforces the importance of proximity in plagiarism detection task, and there is a requirement to restrict the propagation scope in order to find candidate documents of plagiarism.

The results of this part of experiments are reported in Table 3 and 4. With topic-based text segmentation using sentence level information, the best result with respect to F_2 is obtained on σ equal to 5 using the Passage kernel and it is 0.5555. The F_2 measure increased about 0.0087 which is 1.6% improvement over non proximity based approach. Also the result increased about 0.02 which is 3.8% improvement over non-proximity and no segmentation approach. This improvement in performance is statistically significant (Wilcoxon signed-rank test $p < 0.05$).

With topic-based text segmentation augmented with using language model information, the best results are obtained on σ equal to 15 using Passage kernel. In this case we realized that the best result is achieved when reporting the first ranked document of each segment, but not including those that their difference between the scores of the first and second retrieved documents is zero. The F_2 measure is 0.6703, which has 3% improvement (increase of 0.02) over non proximity based approach and 25.2% improvement (increase of 0.13) over non-proximity and no segmentation. Improvement in performance is again statistically significant (Wilcoxon signed-rank test $p < 0.05$). In this case the number of retrieved documents is 5698, just 3% of the possible pairs that would otherwise need to be tested. According to results reported in Tabels 2, 3 and 4 we can see that both text segmentation and considering proximity in retrieval improve the results of candidate retrieval in plagiarism detection task. The results of Table 2, 3 and 4 are summarized in Table 5.

The detection performance of other cross-language plagiarism approaches, CL-CNG and CL-ASA and a combination of translation and monolingual analysis (T+MA), are compared in (Barrón-Cedeño et al., 2013a). In their experiments, in which the suspicious document is an exact copy of a source document, CL-CNG and T+MA show better performance. For monolingual analysis, TF-IDF weighting for documents' terms was used in (Barrón-Cedeño et al., 2013a). Although, T+MA showed to be the best option, machine translation is a computationally expensive method and good automatic translators still do not exist for some language pairs (Danilova, 2013). We compare the detection performance of our cross-language candidate retrieval model with the detection performance of CL-CNG and T+MA (Translation + Monolingual Analysis) in candidate retrieval. For monolingual analysis we used both the probabilistic model (BM25) and the KL-divergence retrieval model, which are implemented in the Lemur toolkit, with considering bag of words. We used translation of the entire source document and we used all words (excluding stopwords) of source and suspicious documents for retrieval. The 1-best, 2-best and 3-best

Table 6
Comparison with other methods.

The proposed two-level proximity-based model	Precision	Recall	F_1	F_2
	0.4663	0.7526	0.5758	0.6703
CL-C3G	0.0065	0.0026	0.0037	0.0029
CL-C4G	0.0052	0.0017	0.0026	0.002
CL-C3G + segmentation	0.0174	0.0244	0.0204	0.0226
T+MA (BM25 (1-best))	0.6477	0.3630	0.4653	0.3980
T+MA (BM25 (2-best))	0.3938	0.4218	0.4073	0.4159
T+MA (BM25 (3-best))	0.2807	0.4397	0.3426	0.395
T+MA (BM25 + segmentation)	0.4209	0.6992	0.5252	0.6174
T+MA (KL-divergence (1-best))	0.8290	0.4785	0.6068	0.5227
T+MA (KL-divergence (2-best))	0.5777	0.5915	0.5845	0.5887
T+MA (KL-divergence (3-best))	0.4318	0.6273	0.5115	0.5752
T+MA (KL-divergence + segmentation)	0.5894	0.7974	0.6778	0.7448

retrieval results are reported in Table 6. The KL-divergence retrieval score is shown in Eq. 4. The BM25 retrieval score for query q and document d is shown in Eq. 11, where k_1 controls scaling of term frequency and set to 1, parameter b controls scaling of document length and it is set to 0.3, and l_d and l_{avg} are document length and average document length in the collection, respectively.

For CL-CNG detection method, source and suspicious documents are represented by character n-grams. A space is placed at the beginning and end of each pair of words; adjacent n-grams share all but one letter. For retrieval phase, we applied KL-divergence retrieval method with Jelinek-Mercer smoothing method on the extracted n-grams. We considered n of lengths 3 (CL-C3G) and 4 (CL-C4G). Table 6 shows precision, recall, F_1, and F_2 for the k best results of the cross-language candidate retrieval. As the table shows, CL-CNG delivers the weakest performance. The reason could be that in the experiments reported in (Barrón-Cedeño et al., 2013a) the suspicious document was the exact copy of its reference, but in these reported experiments the plagiarized part is just a fragment in a suspicious file.

In order to make the comparisons fair we also compared these techniques after applying text segmentation. The results are shown in rows 4, 8 and 12 of Table 6. The segments are the same segments used for our proposed approach, which are extracted by the combination of topic-based and language model based techniques described in Section 4.2.1. Before applying the text segmentation to these approaches, the proposed approach outperforms CL-CNG and T+MA with respect to recall and F_2 measure in the candidate retrieval phase of plagiarism detection. After applying the text segmentation algorithm the proposed approach outperforms CL-CNG and T+MA (using BM25 retrieval model). The table shows that using translation of the whole document in KL-divergence retrieval along with text segmentation and post-process, obtains highest F_2 measure in this phase of plagiarism detection. This indicates that if there is an available machine translation system and with the cost of translation time and effort for translating the entire document, using KL-divergence retrieval model along with the segmentation procedure has better performance in cross-language candidate document retrieval. The high precision achieved by the T+MA model indicates that this model will generally detect sources of plagiarism correctly. However, this model achieves lower recall before text segmentation compared to the proposed method, which indicates that the T+MA model fails to detect sources of plagiarism. These results would lead us to believe that considering the position of words in the documents could be beneficial in finding more potential sources of plagiarism in candidate retrieval phase of plagiarism detection.

5. Conclusion and future work

This paper proposes a candidate document retrieval technique for retrieving potential source documents as the first step of plagiarism detection across languages. The proposed approach is based on word proximity to retrieve the potential sources for each suspicious document. This is an important factor in plagiarism detection. We deal with proximity in two steps. First, since plagiarism usually happens in parts of the text, there is a requirement to segment the texts into fragments to detect local similarity. For this reason, a topic-based text segmentation approach is proposed. Second, we need to find documents that contain a passage close to query terms. For this reason, a proximity-based language model rather than the bag of words model is used.

A topic-based segmentation is proposed with the objective of creating semantically related sequences of text. The proposed model uses topical probability distributions obtained from the Latent Dirichlet Allocation model. We evaluated our topic-based text segmentation on a dataset used for this purpose, and we obtained better results in all ranges of segment lengths over another recent proposed topic-based segmentation model, TopicTiling. Then, the effect of text segmentation is evaluated on candidate document retrieval and shows 0.01 increase in F_2 measure (2.2% improvement) over the baseline when using sentences as chunks for obtaining topical probability distributions. In order to overcome the short context of sentences in obtaining the probability distributions, the topic-based model is augmented with a language model text segmentation. The result show about 0.11 increase in F_2 measure (21.5% improvement) over the baseline of no segmentation and no proximity-based model.

The results obtained from the KL-divergence model are reranked using the positional language model for considering proximity. The sensitivity of the approach to different kernels and scope of propagation for proximity is studied. We see that in the candidate document retrieval task, the best result is obtained by using the Passage kernel with propagation scope, $\sigma = 15$. The results increased about 0.0087 (1.6%) and 0.02 (3%) for using sentence level information and language model information respectively. Thus, at the end the result has about 0.13 increase in F_2 measure (25.2% improvement) over the baseline of no segmentation and no proximity-based model.

For further work, the effect of using other sources of translation such as dictionaries, parallel or comparable corpora could be evaluated. The post-processing step has rooms for improvement. Other proximity-based models could also be evaluated for this phase of plagiarism detection. The effect of extracting representative words and phrases could be compared with employing a multilingual semantic network to map the text into its representative concepts. The proposed candidate document retrieval approach for cross-language plagiarism detection could be compared with a graph-based approach (Franco-Salvador, Rosso, & Montes-y Gómez, 2016) after creating a knowledge graph by using a multilingual semantic network. Construction of multi-lingual plagiarism corpora in other languages would help to compare the results in different languages. In the next step we will use the results of candidate document retrieval as input for the pairwise document similarity task comprising the detailed analysis step. We would like to devise a method for this task that considers the proximities of words and also paraphrased translations in sentences.

Acknowledgement

This research was in part supported by a grant from the Institute for Research in Fundamental Sciences (No. CS1393-4-43). We would like to acknowledge the assistance and information provided by Mr. Mohammad-Hossein Motallebi.

We also gratefully acknowledge the editor and anonymous reviewers of our paper for the invaluable and constructive comments that help improving our manuscript.

References

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Artificial Intelligence Research*, 135–187.

Anguita, A., Beghelli, A., & Creixell, W. (2011). Automatic cross-language plagiarism detection. In *7th international conference on natural language processing and knowledge engineering (nlp-ke), 2011* (pp. 173–176). IEEE.

Barrón-Cedeño, A., Gupta, P., & Rosso, P. (2013a). Methods for cross-language plagiarism detection. *Knowledge-Based Systems, 45*(1), 45–62.

Barrón-Cedeño, A., Rosso, P., Pinto, D., & Juan, A. (2008). On cross-lingual plagiarism analysis using a statistical model. In *Workshop on uncovering plagiarism, authorship, and social software misuse PAN08* (pp. 9–13).

Barrón-Cedeño, A., Vila, M., Martí, M. A., & Rosso, P. (2013b). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics, MIT Press, 39*(4), 917–947.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Choi, F. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the association for computational linguistics conference* (pp. 26–33). Association for Computational Linguistics.

Danilova, V. (2013). Cross-language plagiarism detection methods. In *The student research workshop associated with recent advances in natural language processing, RANLP* (pp. 51–57).

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2), 224–227.

Dumais, S. T., Letsche, T. A., Littman, M. L., & Landauer, T. K. (1997). Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval: 15* (pp. 18–24).

Fournier, C. (2013). Evaluating text segmentation using boundary edit distance.. In *Association for computational linguistics, (ACL (1))* (pp. 1702–1712).

Fragkou, P., Petridis, V., & Kehagias, A. (2004). A dynamic programming algorithm for linear text segmentation. *Journal of Intelligent Information Systems, 23*(2), 179–197.

Francis, W. N., & Kucera, H. (1979). Brown Corpus manual. *Technical Report*. Department of Linguistics, Brown University, Providence, Rhode Island, US.

Franco-Salvador, M., Gupta, P., & Rosso, P. (2013). Cross-language plagiarism detection using a multilingual semantic network. In *Advances in information retrieval, proceedings of the 35th European conference on information retrieval (ECIR13): 7814* (pp. 710–713).

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

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *International joint conference on artificial intelligence, IJCAI: vol. 7* (pp. 1606–1611).

Galley, M., McKeown, K., Fosler-Lussier, E., & Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st annual meeting on association for computational linguistics: vol. 1* (pp. 562–569). Association for Computational Linguistics.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Suppl 1), 5228–5235.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics, 23*(1), 33–64.

Johnson, S. (1993). Solving the problem of language recognition. *Technical Report*. School of Computer Studies, University of Leeds.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 111–119). ACM.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory, 37*(1), 145–151.

Lv, Y., & Zhai, C. (2009). Positional language models for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval* (pp. 299–306). ACM.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*: vol. 1. Cambridge: Cambridge University Press.

Mcnamee, P., & Mayfield, J. (2004). Character n-gram tokenization for european language text retrieval. *Information Retrieval, 7*(1-2), 73–97.

Meyer zu Eißen, S., & Stein, B. (2006). Intrinsic plagiarism detection. In *Advances in information retrieval. 28th European conference on IR research (ECIR 06)*. In *Lecture notes in computer science: vol. 3936 LNCS* (pp. 565–569). Berlin Heidelberg New York: Springer.

Misra, H., Yvon, F., Jose, J. M., & Cappe, O. (2009). Text segmentation via topic modeling: an analytical study. In *Proceedings of the 18th ACM conference on information and knowledge management* (pp. 1553–1556). ACM.

Nawab, R. M. A. (2012). *Mono-lingual paraphrased text reuse and plagiarism detection*. University of Sheffield Ph.D. thesis..

Oberreuter, G., & Velásquez, J. D. (2013). Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications, 40*(9), 3756–3763.

Pereira, R. C., Moreira, V. P., & Galante, R. (2010). A new approach for cross-language plagiarism analysis. In *Multilingual and multimodal information access evaluation, international conference of the cross-language evaluation forum lncs (6360)* (pp. 15–26). Springer.

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on world wide web* (pp. 91–100). ACM.

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation, 45*(1), 45–62.

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., et al. (2012). Overview of the 4th international competition on plagiarism detection. *Clef (online working notes/labs/workshop)*.

Potthast, M., Stein, B., & Anderka, M. (2008). A wikipedia-based multilingual retrieval model. *Advances in Information Retrieval, 4956*, 522–530.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 997–1005). Association for Computational Linguistics.

Pouliquen, B., Steinberger, R., & Ignat, C. (2006). Automatic identification of document translations in large multilingual document collections. In *Recent advances in natural language processing, RANLP* (pp. 401–408).

Riedl, M., & Biemann, C. (2012). Text segmentation with topic models. *Journal for Language Technology and Computational Linguistics, 27*(1), 47–69.

Schleimer, S., Wilkerson, D. S., & Aiken, A. (2003). Winnowing: Local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOd international conference on management of data* (pp. 76–85). ACM.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit.. In *Interspeech* (pp. 901–904).

Stolcke, A., & Shriberg, E. (1996). Automatic linguistic segmentation of conversational speech. In *International conference on spoken language processing, ICSLP 96.: 2* (pp. 1005–1008). IEEE.

Utiyama, M., & Isahara, H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of the 39th annual meeting on association for computational linguistics* (pp. 499–506). Association for Computational Linguistics.

Vinokourov, A., Cristianini, N., & Shawe-taylor, J. S. (2002). Inferring a semantic representation of text via cross-language correlation analysis. In *Advances in neural information processing systems* (pp. 1473–1480).

Zhai, C., & Lafferty, J. (2001a). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on information and knowledge management*. In *CIKM '01*. ACM.

Zhai, C., & Lafferty, J. (2001b). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 334–342). ACM.