ISSN 2277-3061

Feature-based Similarity Method for Aligning the Malay and English News Documents

Nurul Amelina Nasharuddin, Muhamad Taufik Abdullah, Azreen Azman,
Rabiah Abdul Kadir, Enrique Herrera-Viedma
Department nurulamelina@upm.edu.my
Department of Multimedia, Universiti Putra Malaysia, UPM Serdang, Malaysia
mta@upm.edu.my
Department of Multimedia, Universiti Putra Malaysia, UPM Serdang, Malaysia
azreenazman@upm.edu.my
Department of Computer Science, Universiti Putra Malaysia, UPM Serdang, Malaysia
rabiah@upm.edu.my
Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
viedma@decsai.ugr.es
of Multimedia, Universiti Putra Malaysia, UPM Serdang, Malaysia

ABSTRACT

Corpus-based translation approach can be used to obtain reliable translation knowledge in addition to the use of dictionaries or machine translation. But the availability of such corpus is very limited especially for the low-resources languages. Many works have been reported for the alignments of multilingual documents especially among the European languages, but less focusing on the languages with less linguistics resources. One of the challenges is to align the available multilingual documents for the creation of comparable corpus for these kinds of languages. This article describes an alignment method that utilized the statistical features of the documents such as the documents' titles, texts of the contents, and also the named entities present in each document. This method will be focusing on the English and Malay news documents, in which in which the Malay language is considered as a low-resource language. Source and target documents were then compared in a pair. Accuracy, precision, and recall measurements were used in evaluating the results with the inclusion of three relevance scales; Same story, Shared aspect and Unrelated, to assess the alignment pairs. The results indicate that the method performed well in aligning the news documents with the accuracy of 96% and average precision of 81%.

Indexing terms/Keywords

Document alignment; feature-based method; algorithm; Malay text processing; corpus-based information retrieval

Academic Discipline And Sub-Disciplines

Computer Science, Library Science

ISSN 2277-3061

1 INTRODUCTION

One of the simplest and effective methods for query translation in cross-language information retrieval is to perform dictionary look-up based on a bilingual dictionary. But at times it is not adequate to use the dictionary alone for translating some terms in user's query because of the coverage of the dictionary. Lack of the dictionary coverage will cause two problems: translation of specific terms and ambiguity [1, 2]. Additional source of knowledge such as the usage of corpus can help to overcome these problems and further help to increase the number of the relevant documents retrieved.

Parallel corpus, which contains the exact documents in more than one language is scarce, not readily available and usually have limited domain especially for languages of low resources such as Malay language [3]. On the other hand, the creation and use of the comparable corpus are reasonable as it is easier to find bilingual texts with similar topics rather than texts, which are exact translations of each other. The problem in the creation of the comparable corpus is described as the task of aligning documents based on their content similarity. See Figure 1 and 2.

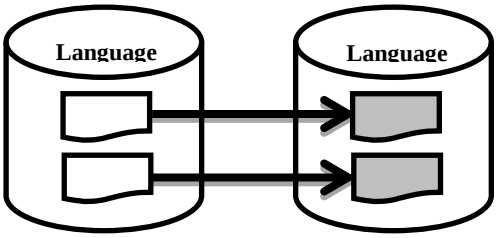

Figure 1: Unidirectional mapping between parallel documents

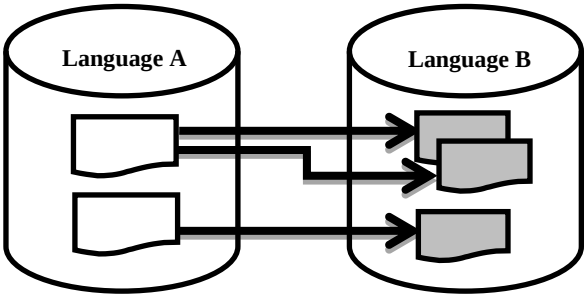

Figure 2: Unidirectional mapping between comparable documents

Document alignment is a technique to find at least two documents, which are exact translations of each other in different languages (or known as parallel documents) or documents which are topically similar without being parallel (or known as comparable documents). This leads to a unidirectional mapping between texts in the collections. It is very useful in automatic corpus construction of multilingual documents [4 - 7].

A high quality translation knowledge can be obtained from these kinds of corpora and further improved the query translation approach. Some of the examples of the comparable documents are the multilingual news feeds by agencies such as Cable News Network (CNN), Agence France-Presse (AFP), Reuters, Al-Jazeera as well as Bernama, which cover about the same events but in different countries and languages. The texts usually contain sentence pairs that are fairly good translations of each other, as are widely available online [3, 8].

With the availability of the documents, it is important to align the documents based on their content similarity. The more similar and larger the aligned corpus is, the more high quality translation knowledge can be obtained from the corpus. The aligned test collections that are readily available in the Web (for example, the JRC-ACQUIS[1]) usually have a limited domain and cover only a few languages. To overcome these problems, a method for finding the similarity of two bilingual documents will be suggested, in which the focus languages in this study are English and Malay. To the authors' knowledge, there are only a few structured and readily available corpora of documents in Malay language that can be used to aid the Malay Information Retrieval task, mostly in the domain of classical Malay literature [9, 10]. Also there are not many research studies focusing on the Malay language compared to other European language, thus labeling the language as one of the low-resource languages [11, 12].

Many alignment techniques have been proposed extensively, each with different objectives and characteristics [13]. Some of the algorithms need additional linguistics resources such as bilingual dictionary or thesaurus and other algorithms only focusing on the individual statistical features that can represent the documents. Different algorithms align the corpus at different levels and each of them implements different evaluation measures. One of the preferred methods focused on the documents' statistical features similarity irrespective of the documents' languages [14, 15]. An alternative approach is based on the multilingual keyword extraction and translation method, by taking the advantage of the availability of terms or

[1] JRC-ACQUIS's website: http://optima.jrc.it/Acquis/index_2.2.html

keywords extraction algorithms [8, 16]. The available parallel aligned documents pairs can also become a resource for the alignment but these kinds of approaches rely heavily on training the data and further prevent them from generalizing well with documents from different domains.

In this study, we will discuss the possibility of aligning two bilingual documents using only the individual statistical features of each of the document itself. Encouraged by the simplicity and direct approach of the so-called feature-based techniques, this method will make use of the availability of the text features inside the documents individually such as the documents' titles, the text contents as well as the named entities, in order to find the relationship between the documents. The rest of the article is organized as follows. In Section 2, a brief overview of related works will be presented. The preprocessing techniques and the details of the algorithm will be described in Section 3. Section 4 contains the results of the experiments and discussion on some possibilities for the future works. Finally the conclusion will be in Section 5.

2 RELATED WORK

Document alignment is a technique to map one document to at least one or many other documents either in the same or different languages. It can associates the documents that cover similar topics or stories. Besides alignment at the document level, the alignment can be made on the paragraph, sentence, word or in character level [17, 18]. In the last few years, there have been many studies reported in textual alignments using different kinds of techniques. These alignment approaches mainly focusing on multilingual parallel or comparable corpus composed of different types of documents. But most of the research studies focusing on aligning the documents in European languages and none on the Malay text. A simple but effective method for aligning documents is by using similarity scores calculations such as cosine similarity, normalized edit distance and sentence alignment but these scores are only suitable for parallel documents as the measurement focusing on the similarity of the document structure [19].

One of the approaches of aligning the multilingual comparable documents is the usage of self-organizing map (SOM). In Yang, Lee, and Tsai [20], the SOM is being used to represent the associations between entities in different languages. The multilingual documents were first clustered using the SOMs and then the feature maps will be created for each language. Next, a hierarchy generation process will be applied to these maps to create bilingual hierarchies. This algorithm does not need dictionaries as the alignment of the hierarchies will be developed independently for each language but the giveaway of this method is the SOMs need to be trained using sample data to produce the representation of the data.

A close second is by extracting keywords from the documents and uses these keywords as the query in a retrieval system to retrieve the potential alignment candidates [8, 16]. It is a popular and straightforward approach to build a comparable corpus. Usually, the framework will consists of the source and target document sets, either from the available documents collections or crawled from the web. Potential keywords will then be extracted from the source documents, translated using dictionaries and combined to be used as the queries. The queries are then run against the target documents. The alignment pairs between source and target documents are developed according to the value of similarity calculated by the retrieval tool. Thus the selection and extraction of keywords to represent the important contents of the documents are very important in order to achieve high quality correlations mapping, especially when selecting multiword phrases as the keywords.

The third approach is by using the individual statistical features of the document itself, either locally or globally in a corpus. Some of the features that can be incorporated are document titles, publish date of documents, document lengths, document themes, linguistics independent unit (e.g., numbers, dates, and monetary values), punctuations, quotations as well as the terms distribution throughout the documents. Each of the individual features can contribute to the final score for the alignment candidates. These feature-based similarity methods were being proposed extensively throughout the years [14, 15, 21 - 23].

In Rasooli, Kashefi, and Minaei-Bidgoli [14], an approach to align translated documents at paragraph and sentence levels for parallel documents was proposed. The method included the length-based, punctuation-based and semantic-based similarity score calculation. They experimented using the Persian-English parallel documents. Each score was weighted by a coefficient that was derived by genetic algorithm. The study showed that the similarity based on cognate really depends on the languages involved in the corpora, in which they need to be of the same alphabet and linguistic family. A study by Vu, Aw, and Zhang [22] showed the possibility of aligning multilingual documents using only the individual features. They included statistical features such as document titles, monolingual terms distributions, and linguistic independent units. This method is adaptable to any language-pairs without relying heavily on linguistics resources, except for bilingual dictionaries for the languages being experimented.

The terms distribution feature calculates the frequency distribution of words in documents. Their use is commonly based on the assumption that the words that are topically-related and translations of each other tend to appear and correlate in documents with similar contexts. The correlation of words or else known as word co-occurrence has been used as a feature in numbers of studies [22, 24]. Co-occurrence matrices, such as co-citation, co-word and co-link matrices, provide us with useful data for mapping and understanding the strength of association between keywords in textual data, despite of the language differences of the document [25]. One of the methods to measure the association of words is by using the word co-occurrence frequency matrices [26, 27]. The idea is to calculate the frequency of words using the Hyperspace Analogue to Language theory with premise that words with similar meaning will occur closely, repeatedly (also known as co-occurrence). Previous study showed that the usage of these kinds of matrices is effective in finding the similarity between different documents without any language restrictions [27].

Another important feature that usually exists in documents especially among the news corpus is the named entities. Named entity (NE) is basically a word or phrase that belongs to a predefined category. Examples of NEs are names, geographic locations, numbers as well as name of organizations. Named entity recognition (NER) is an important step in many natural language processing tasks. The frequency information of NEs may be helpful in distinguishing documents for the creation of the alignment pairs. The study by Montalvo, Martínez, Casillas, and Fresno [28] showed that NEs were a good source of knowledge for news clustering. The main advantage of using the NEs is that no translation is required, with the condition that each language has their own NER system.

The objective of this study is to develop a method to align two English and Malay news documents using the feature-based similarity methods similar to Vu, Aw, and Zhang [22] but exploiting different features. The alignment technique will be developed based on three most important features of a news article. The first is the similarity of the news titles and the contents. The algorithm will incorporate the co-words ratio of the titles and text contents to represent the relations between a pair of documents. This feature was introduced in the studies of Tao and Zai [24] and Vu, Aw, and Zhang [22]. Then, the inter-word distance or co-occurrence frequency of the words is utilized to reflect the semantic information in each document. Another feature in the algorithm is the similarity among NEs between documents, getting to know that the NEs are useful in revealing the relations among document pairs. The contributions of each individual feature will then be combined into one final score of similarity.

Figure 3 illustrates the overall framework of the proposed method. The advantages of applying this method are it is purely statistical-based and also adaptable to any language pairs. Although it requires little additional knowledge resources, it still gives good results in finding better alignments. The feature which requires dictionaries or word lists may take the advantages of the availability of simple, inexpensive bilingual dictionaries or word lists that can be easily obtained nowadays. In the next section, the detailed approach of the proposed document alignment technique will be described.

3 ALIGNING THE NEWS DOCUMENTS

This section will describe the steps implemented by the alignment method. The preprocessing step for each document will be defined first. Then the features extraction step including the calculation of the similarity score of the title-and-content, word co-occurrence as well as NEs will be detailed out. Finally the combination of these features will be explained.

3.1 Document Preprocessing

Document preprocessing which is also known as text normalization is a procedure of morphological analysis to prepare the text documents before being analyzed [29, 30]. The goal behind preprocessing is to separate the text into individual words as a representation of each document. This step is crucial in determining the quality of the alignment stage. It is important to select the significant keywords that carry the meaning and discard the words that do not contribute in distinguishing the documents.

In this study, the first step during the preprocessing for both English and Malay documents, was to do the lexical analysis of the texts. Some of the operations being done were treating digits, hyphens, punctuation marks and the case of letters. All the digits in the documents were removed and the hyphens were replaced by space. The case of letters is usually not important when selecting the index terms in retrieval systems. But in this study, all the capital letters were being changed to small case.

Next step was to remove stopwords. Stopwords are the most frequent words often do not carry much meaning. Examples of such words in English include ‘the’, ‘of’, ‘and’, and ‘to.’ This present study used the SMART stop word list for the English documents [31] and a stoplist of 321 words for the Malay documents.

Stemming techniques were used to discover the root or stem of a word. Although usually it is included in the preprocessing step, we did not include in our studies, getting to know that it is quite hard to find a word stemmer for some languages such as Malay. But for the source documents, we included the lemmatization technique to remove the inflectional endings and return the base of the words in the documents. The English collection was lemmatized with Stanford CoreNLP Java library, a natural language analysis tool[2].

[2] Stanford CoreNLP’s website: http://nlp.stanford.edu/software/corenlp.shtml

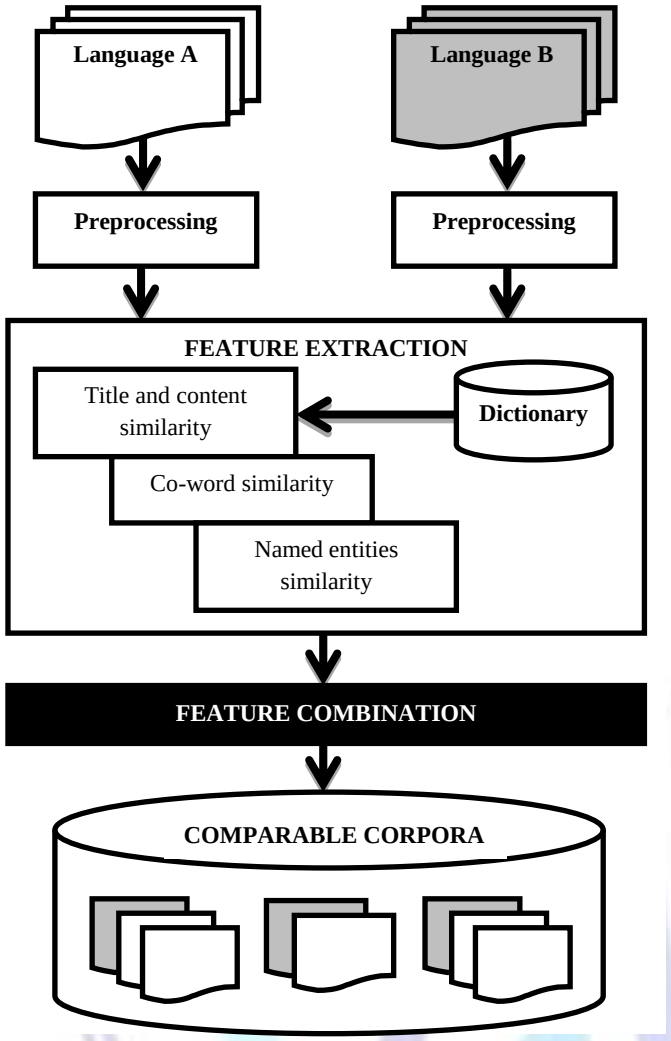

Figure 3: The proposed news document alignment framework

3.2 Basic Approach

In creating the alignment, the usage of three different features of each news document: the relationship between the title and content, the likelihood of words' combination, and the NE similarity between pairs of documents, were being proposed. This combination of features will be calculated by comparing a pair of source and target documents with the inclusion of a small bilingual word list for translations.

1. Title-and-content similarity score: Document title is one of the important features in most documents, especially in news articles. Getting to know that documents titles usually give the reader clues about the main topic, we decided to implement this feature. It was also being used in the study of Vu, Aw, and Zhang [22]. The titles of news documents are usually concise and convey the information essence in the document. The title-and-content similarity score between source document D_s and target document D_t is calculated as in Equation (1) where c_s and c_t are the contents of document D_s and D_t respectively. T_s and T_t are the sets of title words of the two documents and *trans(w, c)* is defined as in Equation (2).

$$tnc(D_s, D_t) = \sum_{w_i \in T_s} trans(w_i, c_t) + \sum_{w_j \in T_t} trans(w_j, c_s) \qquad (1)$$

$$trans(w, c) = \begin{cases} 1, & \text{if translation of word } w \text{ is in content } c \\ 0, & \text{else} \end{cases} \qquad (2)$$

   Therefore, a high title-and-content score will indicate a high likelihood of similarity between two bilingual documents.

2. Word co-occurrence similarity score: In this study, the co-occurrence feature is defined as the frequency of two words with the basis that words which appear (or co-occur) together frequently will have larger strength values than the words which are not [27]. The co-occurrence frequency between these words is calculated via a *w*-sized

sliding window and all the words occurring within the window are considered as co-occurring with each other. An accumulated co-occurrence matrix for all the words is being produced by moving the window across the document. The strength of association between two words is inversely proportional to their distance.

Based on the definition, a word co-occurrence matrix is developed for each document. Let *w* be the total number of the significant keywords in the documents and each document can described by a *w*w* matrix, as in Equation (3) where $s_{i,j}$ is the frequency between word x_i and x_j.

$$S = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,w} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,w} \\ \vdots & \vdots & \ddots & \vdots \\ s_{w,1} & s_{w,2} & \cdots & s_{w,w} \end{pmatrix} \quad (3)$$

And further to calculate the similarity between two documents, the method implements the cosine similarity measure that suitable for comparing vectors of different lengths, as in Equation (4). The result of this calculation will always be a value between 0 and 1. Hence, are normalized.

$$cosim(D_s, D_t) = \frac{\sum_{i=1}^{w}\sum_{j=1}^{w} s_{i,j} * t_{i,j}}{\sqrt{\sum_{i=1}^{w}\sum_{j=1}^{w} s_{i,j}^{\,2}} * \sqrt{\sum_{i=1}^{w}\sum_{j=1}^{w} t_{i,j}^{\,2}}} \quad (4)$$

In the case of our method where both of the documents were in different languages, the keywords of the target documents (in Malay language) will be translated first to the source documents' language (English language). All the translation alternatives will be included in the sliding window and further used to construct the co-occurrence matrix. Then, the cosine similarity measure is computed between the source document and the translated word co-occurrence matrix.

3. Named entity similarity score: In this feature, the same NER system was being used to ta both NEs in English and Malay news documents although they are not cognate languages. It is assumed that the NEs are mostly the same in both languages because they implement the same writing system. Furthermore, most of the NEs used are the same in Malay and English languages especially for persons' names, locations as well as numbers.

   For measuring the similarity of the NEs in the document pairs, a readily available software called Stanford NER system is being used to tag the NEs [32]. The software labels sequences of words such as the names of things (e.g. person or organization names) or gene and protein names in a text. It implements a linear chain Conditional Random Field (CRF) models for extracting the features for the recognition. All the tagged NEs are collected for source and target documents and then the similarities between the two sets of NEs are calculated.

   This study focused on finding the similarity of the documents based on the PERSON, ORGANIZATION, LOCATION as well as NUMBER categories. The NEs similarity score between Ds and Dt is calculated using the Jaccard index as in Equation (5) where NE_s and NE_t are the sets representing unique NEs in documents D_s and D_t. Different similarity scores are calculated for each category and a combination score is achieved by calculating their average.

$$ne(D_s, D_t) = \frac{|NE_s \cap NE_t|}{|NE_s \cup NE_t|} \quad (5)$$

3.3 Features Combination

The score for each feature are normalized to a same scale in order to be combined. In finding the estimation of the overall score for the features, there are many ways that can be implemented, varying from unsupervised to supervised method [22]. Supervised methods will give weight for each feature calculated based on the training data, to be used in calculating the final score. In this study, the final score for the similarity between a document pair D_s and D_t was obtained by simply calculating the average of all normalized features to obtain one unique score as shown in Table 1. As the aim of this study was to build an unsupervised system and the features included are probabilistically independent. Thus, for each source document, a list of alignments with target documents will be obtained, ordered by the average of the score for each feature.

Table 1: Comparison of pairs of documents in two languages

pair	*tnc*	*cosim*	*ne*	average
D_s, D_t	…	…	…	…

4 METHOD'S EVALUATION AND EXPERIMENTATION

In this section, the way the source and target document test sets were acquired will be explained. Then the method was tested on the two test collections. The experimental results are reported and analyzed finally.

4.1 Experiments Setup

The experiments were conducted on an English-Malay comparable corpus. For the experiments' purposes, a total number of 205 Malay and English newspapers' articles had been randomly chosen as the source and target documents of the comparable collection. These articles were from a national news agency of Malaysia, the Bernama[3] which published in January 2005. The collections consisted of various contents, such as politics, disasters and accidents, sports, economics, business as well as finance so that they can avoid the limitations of domain-specific corpus. Table 2 shows the statistics of the comparable documents for the evaluation of the alignment method.

Table 2: Statistics on evaluation data

Language pair	ENG - MLY
Distinct source	94
Distinct target	111
Total alignment pair	5217

The comparable collection was divided into two test sets for the experiments: *S1* and *S2*. Test set *S1* consists of 110 news; 47 in English and 63 in Malay and set *S2* consists of 95 news; 47 in English and 48 in Malay. All the test sets were processed uniformly. They were annotated with the XML tags describing the authors, publication date, headline as well as the contents. The original contents were kept with no change or correction. The test sets are shared online so that it can be reuse in other research projects (for Internet URL, use [33]). Figure 4 displays a sample of the test data. For calculating the title-and-content similarity score, a unidirectional English-Malay bilingual dictionary which is a word list of 22,228 entries together with their translations from Rais, Abdullah, and Kadir [34] was being used.

<ARTICLE LANGUAGE="ENG" AUTHOR="NJ RM" DATE="01/01/2005" DOC_ID="B050102-00001">
<HEADLINE>18 MALAYSIANS REPORTED MISSING IN TSUNAMI-HIT COUNTRIES</HEADLINE>
<BODY>Eighteen Malaysians holidaying in several countries which were ravaged by the earthquake-triggered tsunami on Sunday have been reported missing, Foreign Ministry Secretary-General Tan Sri Ahmad Fuzi Abdul Razak said today. He said that 10 of them had been reported missing in Chennai, five in Phuket and three in Acheh. Wisma Putra received missing persons reports from their respective family members yesterday, he told reporters at Wisma Putra here. He earlier attended the presentation of assistance to 18 Malaysian students studying at Institut Agama Islam Negeri Ar-Raniry in Acheh who returned home safely yesterday. Domestic Trade and Consumer Affairs Minister Datuk Mohamed Shafie Apdal, who is chairman of the Umno Club overseas, presented them RM1,000 each. Ahmad Fuzi also handed over a total of RM400 presented by the Foreign Ministry's Sports and Welfare Club and the National Security Division. The students were accommodated at the Wisma Putra hostel where they were taken to on arrival at the Royal Malaysian Air Force (RMAF) station in Subang. Three of them, suffering from malaria, have been warded at the Universiti Malaya Medical Centre (UMMC) for treatment. At the same news conference, Mohamed Shafie said that he would discuss the position of the affected students with Higher Education Minister Datuk Dr Shafie Mohamed Salleh to ensure that they could continue with their studies. "What's important is that we will do something so that their pursuit of academic success will not be disrupted," he said. The minister also said that counselling would be given to any student suffering from trauma or was emotionally affected by the catastrophic event.
</BODY>
</ARTICLE>

Figure 4: An example of an XML tagged source document

For experimenting with alignment schemes, the evaluations could not rely on the traditional information retrieval test collections with test queries and relevance assessments to get the recall and precision values [11]. In addition, there were no available test collections specifically for alignment purposes that can determine the relevance of target collection documents in relation to each source document (e.g. shared the same topic or at least some vocabulary). As we were getting to know from previous research that making such relevance assessments for even a fraction of the source documents, would have been a huge task.

Therefore for these experiments, the alignments of the chosen source documents were manually assessed with the relevance scales adapted from Braschler and Schäuble [35]. From the five levels of relevance, we only adopted three of the levels in order to simplify the manual alignment tasks. The levels and the characteristics of their documents are

Same story: The two documents cover exactly the same story.

Shared aspect: The documents address various topics. They may share locations or persons.

[3] Bernama's website: http://www.bernama.com

Unrelated: The similarities between the documents are slight or nonexistent.

As the documents were randomly chosen, each source document may or may not have their comparable documents. If they have, they may have one-to-one or one-to-many alignment with the target documents in our corpus. Thresholding was applied for all pairs within a given set of documents, where document pairs that had higher scores than the predetermined threshold were judged as comparable text. In the context of our evaluations, the *'Same story'* level of relevance was defined as the similarity score equals to 1. Any two bilingual articles with the similarity score more than 0.4 were considered containing *'Shared aspect'* and therefore were judged as comparable. In overall, the evaluations were conducted as follow: 1) bilingual speakers of both source and target languages judged whether each pair of documents in the test sets are comparable or not; 2) every document pairs were then aligned using the proposed method; 3) the similarity scores generated by the system were compared against the human judgments; and 4) the threshold was used for the document alignment process.

4.2 Experiments Results

In most of the text alignment algorithms especially in parallel research, the performances are usually measured in terms of accuracy, fall-out, precision, and recall [14, 15, 21, 22]. The performance of this method was done using the same measurements by running through the comparable test sets (with at least 5217 document pairs) and checking on how many of these alignments were correctly aligned. Then the accuracy and fall-out of the system were computed.

The accuracy of the alignment is the number of correct alignments for the source alignment candidates as in Equation (6). The correctness is based on the levels of relevance as defined before. In a formal way, let *aligned_correct* be the total number of document pairs in the test set that the system correctly aligned and *true_alignment* be the total number of alignment pairs in the test set.

$$accuracy = \frac{aligned\_correct}{true\_alignment} \qquad (6)$$

Table 3 shows the average accuracy and fall-out results for both test sets using the proposed method. Overall, the average accuracy of the method is quite high at 96% with the average number of fall-out only at 3%.

Table 3: Accuracy and fall-out of the alignment method

	Accuracy	Fall-out
S1	0.94	0.05
S2	0.98	0.01
AVG	0.96	0.03

We also assessed the performance of a ranked list that returned more correct alignment for comparable documents in the top 10 alignment results by computing the precision and recall. Precision is the ratio of document pairs correctly judged as comparable (Level 1 & 2) to the total number of pairs judged as comparable by the system. Recall is the ratio of document pairs correctly identified as comparable by the system to the total number of truly comparable pairs (or the number of pairs in the comparable corpus used to generate the test dataset).

Let *aligned_comparable* be the total number of document pairs from our test sets that the system judged as comparable, *aligned_well* be the number of document pairs that the system correctly judged as comparable and *true_comparable* be the total number of comparable pairs in the test set. Thus, the precision and recall as in Equation (7) and Equation (8).

$$precision = \frac{aligned\_well}{aligned\_comparable} \qquad (7) \qquad recall = \frac{aligned\_well}{true\_comparable} \qquad (8)$$

As the documents in the test sets were randomly chosen, we need to remove the documents that do not have relevant pairs in the test sets. After the removal, there were 197 documents that had at least one relevant document in the test sets. Table 4 shows the non-interpolated average precision over all relevant documents at cutoff depth of 10 and the resulting mean average value. The 11-point precision and recall curves corresponding to the evaluations on the two sets are shown in Figure 5.

Table 4: Average and mean average precisions of the alignment method

	Average
S1	0.78
S2	0.84
MAP	0.81

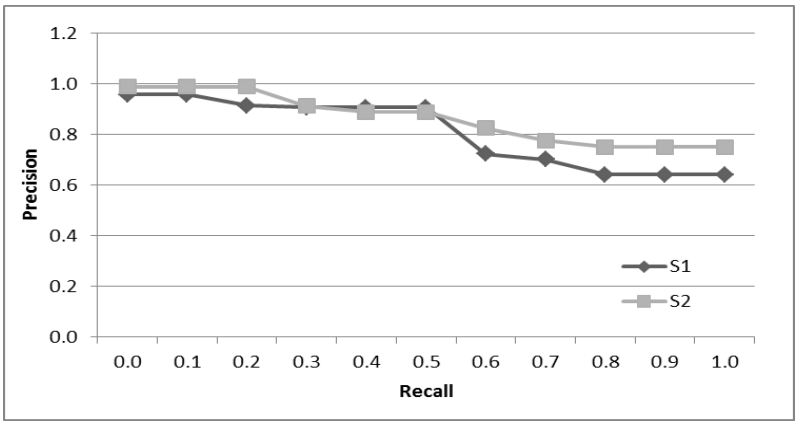

Fig 5: Average interpolated P-R curves for the test sets *S1* and *S2*

4.3 Discussion and Future Works

Based on the results, they indicated that the proposed method performed satisfactorily. The alignments were in most cases correct, which can be explained by the similarity of the two languages. The accuracy was slightly better than the mean average precision, as can be seen in Table 3 and 4. In this task, this is an acceptable situation as the method was comparing in pairwise and there was possibility of returning incorrect alignments for each source document, getting to know that the importance of this evaluation is the correctness of the alignment of each document pairs.

If categorized according to the features, the score of the co-occurrence of words was the lowest compared to the similarity score for the title-and-content and NEs. The relatively poor performance of the similarity score could be partly attributed to the small dictionary used to translate the words in the experiments. In future studies, our aim is to use a more complete dictionary and possibly with the inclusion of proper nouns.

One of the giveaways in this study is that the NER was not being trained for documents in Malay language. The study implemented the NER mapping of only the regular expressions to recognize Malays' NEs. It is likely that the performance of the NER would possibly be better with a Malay-trained NER system instead of relying only on the regular expression. Furthermore, the NER can also be enhanced by also measuring the context similarities among pairs of NEs in future work. Future work will also be focusing on the alignment using multiword expression co-occurrence as it is essential for alignment of specialized domains [21]. Generally, multiword phrases can capture the main idea of a document more efficiently, compared to a single word. One of the ways it could be done is with the inclusion of the weighted n-grams approach.

As aforementioned, there are many ways to improve the algorithm by implementing different techniques focusing on the same objective; aligning documents in many different languages. The combined use of different algorithms as well as methods in the future could allow us to develop a more complete document alignment algorithm. We think that a complete document alignment algorithm could be carried out using the powerful features of each algorithm. For example, the number of candidates can be reduced by filtering them using the publish date of the documents. This document alignment could be very time-consuming if the given pools of documents are very large because they were being compared in pairs. The date-windows technique can be implemented to improve the alignment speed in future. This technique will determined on how many days of documents should be compared, instead of comparing for the whole collections.

The compilation of more news corpus can also be included in order to confirm these results and conclusions. In the future, we hope to collect few other corpus of several millions words in multiple languages and develop better algorithms for such improvement. This could be achieved by mining the Web to collect the comparable documents. Further alignments at sentence level can also be done once a better alignment method is achieved and further used for building a similarity thesaurus. Also, we will compare with some state-of-the-art methods as a benchmark to further justify the results.

5 CONCLUSION

In this article, a feature-based alignment method at document level has been presented. It incorporated the contributions of the individual features of the documents which are title-and-content similarity, word co-occurrence, and named entities. The presented algorithm for creating the alignment does not rely heavily on external resources compared to certain other algorithms although a simple and inexpensive bilingual word list is needed.

The results showed that the method can aligned similar and comparable documents and returned high accurate results. Additionally, the empirical evidence obtained, suggested that the method can be applied to corpora in different domains with good precision. The results indicated that this method can further help to create comparable corpora to be used as the translation source in query translation. From the resulting alignment pairs, the terms can be effectively used as entries in a similarity thesaurus.

As in Malaysia, currently there are on-going researches of query translation among Malaysian languages especially on the Malay-English language pair [34, 36]. But to the author knowledge, there are none focusing on aligning the English documents to the Malay and vice versa, in which the results can be used as a translation resource. It is hope that with this

study as a basis, more research that focusing on creating comparable corpora among the Malaysian languages will be conducted in the future.

ACKNOWLEDGMENTS

This work was funded by the Feder Financing of Projects FuzzyLIng-II TIN2010-17876, Andalucian Excellence Projects TIC5299 and TIC-5991 and Fundamental Research Grant Scheme (FRGS), MOHE Malaysia FRGS/2/2010/SG/UPM/02/26.

REFERENCES

[1] Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., and Laurikkala, J. 2008. Focused web crawling in the acquisition of comparable corpora. Inf. Retr. 11, 427-445.

[2] Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. 2001. Dictionary-based cross-language information retrieval: problems, methods and research findings. Inf. Retr. 4, 209-230.

[3] Munteanu, D. S. and Marcu, D. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In Proceedings of the 21st International Conference on Computational Linguistics, pp. 81-88.

[4] Nie, J-Y. Cross-language information retrieval. 2010. Synthesis Lectures on Human Language Technologies, vol. 3, pp. 1-125.

[5] Maeda, K., Ma, X., and Strassel, S. 2008. Creating sentence-aligned parallel text corpora from a large archive of potential parallel text using bits and Champollion. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, pp. 3066-3069.

[6] Abusalah, M., Tait, J., and Oakes, M. 2005. Literature review of cross language information retrieval. World Academy of Science, Engineering and Technology. 4, 175-177.

[7] Ma, X. and Liberman, M.Y. 1999. BITS: A method for bilingual text search over the web. In Proceedings of the Machine Translation Summit VII, pp. 538-542.

[8] Talvensaari, T., Juhola, M., Laurikkala, J., and Järvelin, K. 2007. Corpus-based cross-language information retrieval in retrieval of highly relevant documents. J. Am. Soc. Inf. Sci. Technol. 58, 322-334.

[9] LDC. 1994. ECI multilingual text. Linguistic Data Consortium, Philadelphia.

[10] Proudfoot, I. n.d. Malay Concordance Project. Australian National University. http://mcp.anu.edu.au/Q/mcp.html. Accessed 10 February 2013

[11] Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., and Keskustalo, H. 2007. Creating and exploiting a comparable corpus in cross-language information retrieval. ACM T. Inform. Syst. 25, 1-21.

[12] Yang, C.C. and Li, K.W. 2004. Building parallel corpora by automatic title alignment using length-based and text-based approaches. Inform. Process. Manag. 40, 939-955.

[13] Nasharuddin, N.A., Abdullah, M.T., Abdul Kadir, R., Azman, A., and Herrera-Viedma, E. 2012. A review on document alignment algorithms in corpus-based information retrieval. Manuscript submitted for publication.

[14] Rasooli, M.S., Kashefi, O., and Minaei-Bidgoli, B. 2011. Extracting parallel paragraphs and sentences from English-Persian translated documents. In Proceedings of the 7th Asia Conference on Information Retrieval Technology, pp. 574-583.

[15] Xu, W. and Esteva, M. 2011. Finding stories in the archive through paragraph alignment. Lit. Ling. Comput. 26, 359-363.

[16] Huang, D., Zhao, L., Li, L., and Yu, H. 2010. Mining large-scale comparable corpora from Chinese-English news collections. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 472-480.

[17] Talvensaari, T., Laurikkala, J., Järvelin, K., and Juhola, M. 2006. A study on automatic creation of a comparable document collection in cross-language information retrieval. J. Doc. 62, 372-387.

[18] Church, K.W. 1993. Char align: a program for aligning parallel texts at the character level. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93), pp. 1-8.

[19] Patry, A. and Langlais, P. 2005. Automatic Identification of parallel documents with light or without linguistics resources. In Proceedings of the 18th Annual Conference on Artificial Intelligence, pp. 354-365.

[20] Yang, H-C., Lee, C-H., and Tsai, H-T. 2009. Multilingual hierarchy generation and alignment using self-organizing maps. In Proceedings of the 2009 International Conference on Education Technology and Computer, pp. 326-330.

[21] Nazar, R. 2012. Bifid: un alineador de corpus paralelo a nivel de documento, oración y vocabulario. Linguamática. 4, 45-56.

[22] Vu, T., Aw, A.T., and Zhang, M. 2009. Feature-based method for document alignment in comparable news corpora. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 843-851.

[23] Woon, W.L. and Wong, K-S.D. 2009. String alignment for automated document versioning. Knowl. Inf. Syst. 18, 293-309.

[24] Tao, T. and Zai, C.X. 2005. Mining comparable bilingual text corpora for cross-language information integration. In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 691-696.

[25] Coulter, N., Monarch, I., and Konda, S. 1998. Software engineering as seen through its research literature: a study in co-word analysis. J. Am. Soc. Inf. Sci. Technol. 49, 1206–1223.

[26] Chen, Z. and Lu, Y. 2011. A word co-occurrence matrix based method for relevance feedback. J. Comput. Inf. Sys. 7, 17-24.

[27] Lund, K. and Burgess, C. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Meth. Ins. C. 28, 203-208.

[28] Montalvo, S., Martínez, R., Casillas, A., and Fresno, V. 2007. Bilingual news clustering using named entities and fuzzy similarity. In Proceedings of the 10th International Conference on Text, Speech and Dialogue, pp. 107-114.

[29] Baeza-Yates, R. and Rebeiro-Neto, B. 2011. Modern Information Retrieval. ACM Press/Addison-Wesley.

[30] Srividhya, V. and Anitha, R. 2011. Evaluating preprocessing techniques in text categorization. Int. J. C. Sci. Appl. Issue, 11, 49-51.

[31] Salton, G. 1989. Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, Reading, Pennsylvania.

[32] Finkel, J.R., Grenager, T., and Manning, C. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 363-370. (2005)

[33] English-Malay Alignment Corpus. 2012. Research project. Available at: https://dl.dropbox.com/u/1196480/bernama05-dataset.zip. Accessed 11th March 2013)

[34] Rais, N.H., Abdullah, M.T., and Abdul Kadir, R. 2011. Multiword phrases indexing for Malay-English cross-language information retrieval. Inf. Tech. J. 10, 1554-1562.

[35] Braschler, M. and Schäuble, P. 1998. Multilingual information retrieval based on document alignment techniques. In Nikolaou, C.; Stephanidis, C. (eds.) Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, pp. 183–197. Springer, Heidelberg.

[36] Rais, N.H., Abdullah, M.T., and Abdul Kadir, R. 2011. Malay-English cross-language information retrieval: compound words and proper names handling. In Yonazi, J.J.; Sedoyeka, E.; Ariwa, E.; El-Qawasmeh, E. (eds.) e-Technologies and Networks for Development, pp. 309-317. Springer, Heidelberg.

Author' biography

Nurul Amelina Nasharuddin is currently Ph.D candidate in Information Retrieval at Universiti Putra Malaysia (UPM), Malaysia. Her research interests are in the fields of multimedia information retrieval, artificial intelligence, and computational linguistics.

Muhamad Taufik Abdullah is a senior lecturer at FCSIT, Universiti Putra Malaysia (UPM), Malaysia. His research interests are in the fields of information retrieval, cross-language information retrieval and multimedia information system.

Azreen Azman is a senior lecturer at FCSIT, Universiti Putra Malaysia (UPM), Malaysia. His research interests are in the fields of browsing model, web information retrieval, data mining and knowledge discovery, relevance feedback learning and decision modelling.

Rabiah Abdul Kadir is a senior lecturer at FCSIT, Universiti Putra Malaysia (UPM), Malaysia. Her research interests are in the fields of computational linguistics and natural language processing.

Enrique Herrera-Viedma is a Professor at DECSAI, University of Granada, Spain. His research interests are in the fields of linguistic modeling, fuzzy decision making, aggregation, consensus, information retrieval, recommender systems, digital library, web retrieval, web quality and bibliometric.