ISSN 2278-3091
Volume 8, No.1.3, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse3881.32019.pdf
https://doi.org/10.30534/ijatcse/2019/3881.32019

# A Framework for English and Malay Cross-lingual Document Alignment Method

**Nurul Amelina Nasharuddin[1], Muhamad Taufik Abdullah[2], Azreen Azman[3], Rabiah Abdul Kadir[3]**
[1]Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia, nurulamelina@upm.edu.my
[2]Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia
[3]Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia
[4]Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Selangor, Malaysia

## ABSTRACT

Issues of information divide in multilingual information retrieval are usually being solved by translating users' queries to a language that the users understand. But dictionaries or other translation knowledge in some of the Asian languages are scarce. The objective of this study was to automatically align the English and Malay news documents to become a comparable corpus, which could contribute as a translation resource to improve the query translation in cross-lingual information retrieval. This study proposes a direct alignment framework by utilizing the textual features similarity of each document itself while attempting a novel approach of using the similarity of the documents sentiment in improving the effectiveness of the alignment method. The proposed sentiment-based approach outperformed existing alignment methods and improved the effectiveness in differentiating the related and unrelated documents. These aligned comparable documents can further be utilised in translation research for the English and Malay cross-lingual information retrieval tasks.

**Key words :** Cross-lingual information retrieval, Document alignment, Malay language, Sentiment-based approach.

## 1. INTRODUCTION

The classical cross-lingual information retrieval approach has one challenge for most researchers, which is the limitation of translation resources especially for some under-resourced languages such as the Malay language and Indonesian language [1]. Most of the cross-lingual research involve only top commonly-used languages in the world such as English, Mandarin, and Arabic. Also, research has been done on languages that have most influences in the economics and commerce of a country. Other than using bilingual dictionaries, corpus can also be used as a resource for the translation's tasks.

Parallel corpus which contains the exact documents in more than one languages in the Web usually is limited, especially for specific domains or those written with under-resourced languages [2]-[3]. To overcome these problems, the creation and use of comparable corpus are reasonable as it is easier to find bilingual texts with similar topics than texts that are translations of each other. Therefore, an efficient and straightforward method of automatic cross-lingual text document alignment is needed in order to extract high quality of translation knowledge from the aligned corpus.

Several other methods to create a bilingual comparable corpus have been proposed [4]. One of the methods is by automatically aligning text documents in different languages based on their lexical features [5]. Others have proposed the use of the concepts, synonyms and semantic text structures of the documents [3],[6],[7]. Path-based measures rely solely on the shortest path information between two concepts in a graph-like taxonomy although it requires a lot of effort in building rich and consistent taxonomies and maintaining them8. Feature-based document alignment methods compute the similarity score between two documents by extracting the individual features of each of the document itself. While information content-based methods calculate the probability similarity of the concepts or information occurring in the documents.

Some of the existing features and information extraction techniques focus on extracting the document title, length, date of publication and terms in the document content, to name a few [9]-[10]. This information is common among all types of documents thus make them suitable candidates to be used as the similarity features of the proposed alignment framework, compared to the path-based similarity. In user-generated contents such as in the news documents used in this study, there are non-syntactic features that can be used as an alignment feature. One of the textual features is the subjectivity or sentiments of a document. The most popular way to represent text documents is to isolate particular features in a document, such as writing style, parts of speech or author sentiment [11]. Document subjectivity in news articles is crucial as these kinds of documents naturally carry opinions or sentiments of the authors.

Previous studies on sentiment analysis have proven that sentiments are one of the important features in user-generated

documents and extremely helpful in decision-making tasks based on the written words expressed by the writer [12]-[13]. In the news collections analysed in this study, almost 30% of the collections contain subjective perspectives as according to the language experts who are assessing the documents. Thus, it can be suggested that this sentiment feature may have an effect in distinguishing between objective and subjective documents as the number of differences between the categories is prominent.

However, the existing document alignment framework does not take into consideration of the subjectivity of a document. The emotions and sentiments of the news' authors toward an event or incident correlate to the choices of words and tone used. The analysis of these direction-based texts determines whether the texts are objective or subjective. These choices of words and documents' subjectivity can be the alignment features in order to find the similarity between two documents. Thus, there should be a new cross-lingual document alignment method to align the English and Malay news documents using the combination of lexical features and sentiment analyses techniques. This new document alignment approach could reduce incorrect documents pairs to be aligned and produce better accuracy in the alignment method. Since the approach has never been done, it is believed that the new implementation is required to automatically develop an English and Malay cross-lingual comparable corpus.

## 2. RESEARCH METHODOLOGY

Fig. 1 shows the framework for the cross-lingual document alignment method for the English and Malay news documents. The framework focuses on assessing the relation between two bilingual text documents based on the similarity of their sentiments and lexical features. The framework starts with the preparation step for the news documents in the English and Malay languages. Then, two main phases of the framework are included which are the lexicon-based sentiment analysis phase (refer as Phase 1 in the diagram) and the document alignment phase (Phase 2) for news documents in both languages.

In sentiment analysis phase, it is an essential requirement to identify the sentiment orientation, which is a list of positive, negative and neutral words with the sentiment score, of a document using a lexicon-based approach. The process had been explained in detail in previous research by Nasharuddin and colleagues [14]. The document alignment phase comprises of three main processing analyses namely, (i) lexical document features extraction, (ii) sentiment and lexical features combination, and (iii) document alignment categorisation. A comparable corpus was automatically developed as a result of the previous phases of categorisation.

### 2.1. English and Malay News Documents Preparation

The news documents in test collection are processed uniformly before being analysed. They are manually annotated with the XML tags describing the news' author, publication date, headline as well as the contents. No change or correction performed on the contents of the documents. The three main steps for both English and Malay documents preparation are lexical analysis, stopwords removal and stemming and part-of-speech tagging. In the lexical analysis step, all the digits in the documents are removed and the hyphens are replaced by spaces. The case of letters is usually not important when selecting the index terms in retrieval systems. But in this study, all the capital letters are left unchanged, as they are important indicators during the part-of-speech tagging.

### 2.2. Lexicon-Based Sentiment Analysis Phase

The main objective of Phase 1 is to detect the sentiment orientations of the news documents by using an English-Malay cross-lingual sentiment lexicon that is automatically constructed from the readily-available lexical databases. Towards this objective, an automatic news document sentiment classification approach is proposed for documents in Malay and English languages [14]. The lexicon is constructed based on the similarity between the SentiWordNet, a sentiment lexical database for English language and Wordnet Bahasa, a Malay lexical knowledge. It is worth to mention that no translation is needed when constructing the cross-lingual sentiment lexicon but for untranslated word, the original word is being used without change. Once the document's sentiment orientation has been identified, the document was classified into one of these three categories, which are 'Positive', 'Negative' or 'Neutral'.

### 2.3 Document Alignment Phase

In the document features extraction of Phase 2, the main analysis of the document features which includes a combination of types of lexical and sentiment features will be performed. The results from the previous phase of the pre-processing are used as the input documents for the features extraction processes. The proposed framework consists of three main processing stages.

The similarity score of the documents' titles and contents for both English and Malay news documents are extracted in the first stage [9]. The basic idea is to find the similarity of two document pairs based on the presence of the same words in both documents. Next, the similarity of named entities in both documents is identified and the similarity score is calculated using the Damerau-Levenshtein measure, which is a popular method in string alignment task [15]. For documents in the Malay language, an extension of contextual rules for recognising Malay named entities is presented for identifying a person, an organisation and a location [16]-[17].

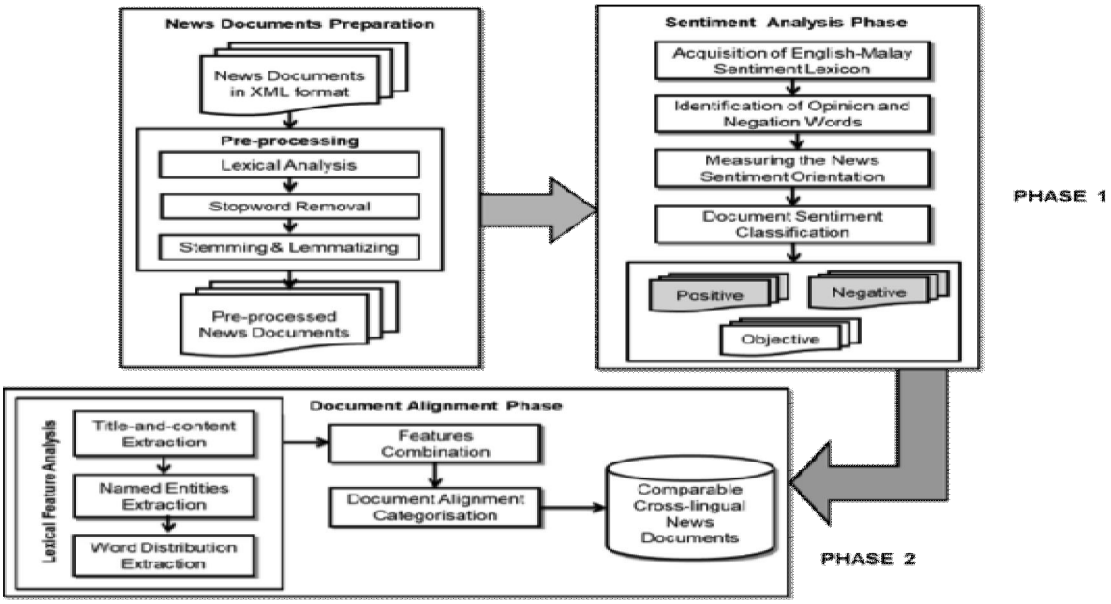


**Figure 1** : The cross Lingual Document Alignment Framewoer

The similarity score of word frequency distributions in both documents is calculated in the final stage of the alignment method [18]. The distribution matrix for each of the document will be constructed using the Hyperspace Analogue to Language framework [19]. The similarity score is calculated using the Cosine Similarity measure for the distribution matrix of the two documents.

These three stages are created for handling the features analyses separately. Each of the stages produced a similarity score as the representative of each feature, which then is combined into a single score that represents the relatedness of the document pair. The major advantage of using the proposed framework is its simplicity as it only utilises the local characteristics of the documents. Local characteristics are defined as the characteristics that are available in the document itself or intra-document, without the need of information from other documents.

The second phase will combine the individual scores for all the features including the sentiment similarity score to produce one final similarity score. In this phase, a weighted linear evidence (or feature) combination has been proposed in order to produce the final similarity score for each pair of documents in the English ($d_s$) and Malay ($d_t$) languages. The formula as in (1).

$$wcomb(d_s, d_t) = \sum_{i=1}^{f} w_i s_i(d_s, d_t) \quad (1)$$

where $s_i$ is the number of different scores for each of the $f$ individual features and $w_i$ is a non-negative normalised weight such that $\sum_{i=1}^{f} w_i = 1$. In this study, all the features are given the same weight to show that all of them have equal importance. This document alignment phase is one of the important contributions because it implements the sentiment analysis of each of the documents (as an additional feature) to find the documents similarity, which has not been implemented as a similarity feature in existing alignment method.

## 3. EXPERIMENT

### 3.1. Test Document Collection

The test collection that was used in the experiments was a sample of 883 Malay and English news articles from a national news agency of Malaysia, where they consist of 439 Malay and 444 English articles. The collection contains various contents, such as politics, nature, sports, economics, business as well as finance to avoid the limitations of the domain-specific corpus.

### 3.2. Experiment Settings

The document alignment framework produces its output in the form of alignment pairs. For example, an English document, which covers exactly the same story with a Malay document, will be suggested as an alignment pair. For evaluation of performance, the experiment results produced by the proposed algorithms are compared to the manual assessments produced by three language experts as there are no available test collections specifically for alignment purposes that can be used to determine the relevance of Malay collection documents in relation to each English document (shared the same topic or at least some vocabulary). The alignment was judged by the three experts with respect to the five-level relatedness categories [20]. The categories are 'Highly

related', 'Fairly related', 'Marginally related', 'Weakly related' and 'Unrelated'.

In addition, the proposed document alignment framework was also being compared with the existing method (referring to as baseline) to observe the impact of the inclusion of sentiment analysis in the proposed alignment framework. Like in most of the text alignment algorithms especially in text classification research, the framework performances are measured in terms of accuracy, precision, recall and F-measure [5],[21].

### 3.3. Experiment Results

For the overall accuracy, the proposed framework managed to correctly align 135 pairs or 31.4% with almost 11% increase when compared to the results of the baseline. The most striking result to emerge from the data is that up to 20% of 'Weakly related' and 'Unrelated' document pairs are correctly produced by the proposed alignment framework. Previously there were no unrelated document pairs that can be classified using the baseline method. These results may be explained by the fact that by using additional similarity scores of the sentiment analysis, the document sentiment orientation scores can help to differentiate between related and unrelated documents. Although the scores could not significantly distinguish the top-three related categories. Fig. 2 compares the improvements or decrements of the total numbers of correct alignments pairs between the baseline and proposed framework.

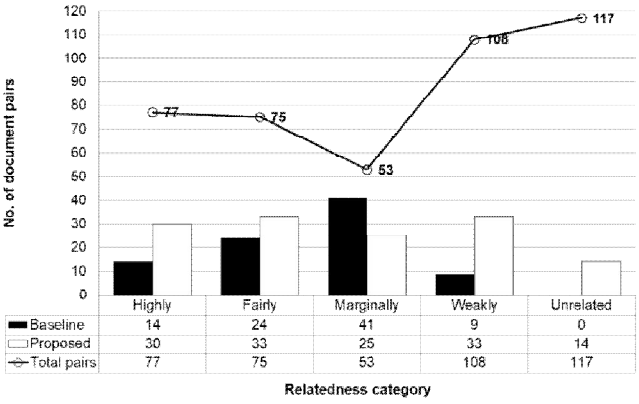


**Figure 2:** Comparison of Total Numbers Of Correct Alignments Pairs Between The Baseline And Proposed Method

Statistical *t*-test analysis is being employed to decide whether the improvement by proposed framework over the baseline is significant. The *t*-test calculates a *p*-value based on the performance of the baseline and proposed framework and will conclude that the improvement is statistically significant if the *p*-value is less than 0.05. As the result, it is found that there was a statistically significant difference between the baseline method (M = 0.45, SD = 0.21) and the proposed method (M = 0.51, SD = 0.21), $t(429) = 8.62$, $p \leq .05$ (CI.95 0.04, 0.09) and this confirmed that the proposed framework has better effects than the baseline method.

The method by Vu, Aw and Zhang is the most relevant to this work [9]. The performance of the proposed document alignment framework was compared with the results of Vu's method using the macro average precision of the similarity score. The performance measure is checked in the top-50 (precision at 50 or *P@50*), top-100 (precision at 100 or *P@100*), and top-200 (precision at 200 or *P@200*) of the list of document pairs for all the relatedness categories. The results are compared with two language pairs, which are the English-Chinese, and English-Malay experimented by Vu and colleagues. The comparative results are shown in Table 1.

**Table 1:** Macro-averaged precision results comparison between the proposed framework and previous method by Vu, Aw and Zhang [9]

| Method | Lang. | P@50 | P@100 | P@200 |
|---|---|---|---|---|
| Vu, Aw & Zhang | En-Mly | 0.82 | 0.76 | **0.69** |
| Vu, Aw & Zhang | En-Chi | 0.43 | 0.43 | 0.40 |
| Proposed | En-Mly | **1.00** | **0.94** | 0.68 |

The experiment results show that the proposed document alignment framework can perform better compared to the previous method, in terms of the precision and number of correct alignments. Among the top-50 and top-100, the proposed framework is good in suggesting correct alignment pairs. The previous method does not take into consideration acquiring the similarity of named entities and sentiments of the documents. The method introduced the combination of title-and-content, linguistics independent unit (LIU) and term distribution features to compute the similarity score. Although the feature LIU contributed to the performance of the method, they do not help in distinguishing valuable information that can help in future development of the comparable corpus. In contrast to Vu's method, the proposed framework can acquire the relevant named entities and used these as valuable information in the alignment.

## 4. CONCLUSION

In this study, a new approach for cross-lingual document alignment framework was presented through three major processing phases on English and Malay languages news collection which are the document's lexical and sentiment features extraction, features combination and document alignment classification. The features extraction phase was implemented based on the individual features of the document itself. One of the contributions of the proposed framework is on the features that were selected in finding the similarity of the English and Malay document. The experiments showed that the extraction of titles and contents, analysis of word distribution and recognition of named entities could still achieve promising results even though it depended entirely on morphological information. Apart from the existing framework, the sentiment analysis was included for improving the alignment without additional data or knowledge. The inclusion of the sentiment feature helps in suggesting the correct relatedness among related and

unrelated categories. The combination of these stages is the main contribution of the proposed work.

Therefore, the proposed method is useful and able to provide a positive impact in finding the relatedness between cross-lingual document pairs and in a recent search, it is the first work to include the sentiment analysis in facilitating the alignment method. An aligned comparable corpus can be automatically developed as an output from this framework, thus it can serves as a basis to create a translation knowledge for the English and Malay languages in any cross-lingual information retrieval tasks later on.

## ACKNOWLEDGEMENT

This work was supported by the Fundamental Research Grant Scheme (FRGS), under the Ministry of Education Malaysia (Project No.: 08-01-14-1479FR).

## REFERENCES

1. V.K. Sharma, N. Mittal, **Cross lingual information retrieval (CLIR): review of tools, challenges and translation approache**s, Advances in Intelligent Systems and Computing, Springer, New Delhi, 2016 pp. 699-708. https://doi.org/10.1007/978-81-322-2755-7_72
2. T. Talvensaari, M. Juhola, J. Laurikkala, K. Jarvelin, **Corpus-based cross-language information retrieval in retrieval of highly relevant documents**: research articles, J Assoc Inf Sci Technol, 58, 3, 2007. https://doi.org/10.1002/asi.20515
3. K. Wołk, K. Marasek, **Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs**, Proc Tech, 18, 2014. https://doi.org/10.1016/j.protcy.2014.11.024
4. M.L. Paramita, D. Guthrie, E. Kanoulas, R. Gaizauskas, P. Clough, M. Sanderson, **Methods for collection and evaluation of comparable documents, Building and Using Comparable Corpora, Springer, Heidelberg**, 2013 pp. 93-112. https://doi.org/10.1007/978-3-642-20128-8_5
5. M.S. Rasooli, O. Kashefi, B. Minaei-Bidgoli**, Extracting parallel paragraphs and sentences from English-Persian translated documents, Information Retrieval Technology** AIRS 2011, Springer, Heidelberg, 2011 pp. 574-583. https://doi.org/10.1007/978-3-642-25631-8_52
6. M. Boháč, K. Blavka, **Text-to-speech alignment for imperfect transcriptions, Text, Speech, and Dialogue, Springer, Heidelberg**, 2013 pp. 536–543. https://doi.org/10.1007/978-3-642-40585-3_67
7. G. Ke, P.F. Marteau, **Co-clustering of bilingual datasets as a mean for assisting the construction of thematic bilingual comparable corpora**, LREC 2014: The 9th Edition of the Language Resources and Evaluation Conference, ELRA, Reykjavik, 2014 pp. 1992–1999.
8. B.T. McInnes, T. Pedersen, Y. Liu, G.B. Melton, S.V. Pakhomov**, U-path: An undirected path-based measure of semantic similarity**, AMIA Annual Symposium Proceedings, American Medical Informatics Association, New York, 2014 pp. 882-891.
9. T. Vu, A.T. Aw, M. Zhang, **Feature-based method for document alignment in comparable news corpora, EACL '09: The 12th Conference of the European Chapter of the Association for Computational Linguistics**, EACL '09, ACL, Stroudsburg, 2009 pp. 843–851.
10. W. Xu, M. Esteva, **Finding stories in the archive through paragraph alignment, Lit Linguist Comput**, 26, 3, 2011. https://doi.org/10.1093/llc/fqr017
11. R.P. Schumaker, Y. Zhang, C.N. Huang, H. Chen, **Evaluating sentiment in financial news articles**, Decis Support Syst, 53, 3, 2012. https://doi.org/10.1016/j.dss.2012.03.001
12. N. Hollenstein, M. Amsler, M. Bachmann, M. Klenner, Sa-uzh: **Verb-based sentiment analysis, SemEval** 2014: The 8th International Workshop on Semantic Evaluation, ACL and Dublin City University, Dublin, 2014 pp. 503–507. https://doi.org/10.3115/v1/S14-2087
13. S. Trinh, L. Nguyen, M. Vo, P. **Do, Lexicon-based sentiment analysis of Facebook comments in Vietnamese language,** Recent Developments in Intelligent Information and Database Systems, Springer, Cham, 2016 pp. 263-276. https://doi.org/10.1007/978-3-319-31277-4_23
14. N.A. Nasharuddin, M.T. Abdullah, A. Azman, R. Abdul Kadir, **English and Malay cross-lingual sentiment lexicon acquisition and analysis, Information Science and Applications 2017**, Springer, Singapore, 2017**,** 424, pp. 467-475. https://doi.org/10.1007/978-981-10-4154-9_54
15. F.J. Damerau, **A technique for computer detection and correction of spelling errors, Commun. ACM**, 7, 3, 1960. https://doi.org/10.1145/363958.363994
16. S.F. Yong, B. Ranaivo-Malançon, A.Y. Wee, NERSIL: **The named-entity recognition system for Iban language, PACLIC 2011**: The 25th Pacific Asia Conference on Language, Information and Computation, Institute of Digital Enhancement of Cognitive Processing, Waseda University, 2011**,** pp. 549-558.
17. R. Alfred, L.C. Leong, C.K. On, P. Anthony, T.S. Fun, M.N. Razali, M.H. Ahmad Hijazi, **A rule-based named-entity recognition for Malay articles**, ADMA 2013: Advanced Data Mining and Applications, Springer, Heidelberg, 2013**,** pp. 299-299. https://doi.org/10.1007/978-3-642-53914-5_25
18. X. Cheng, X. Yan, Y. Lan, J. Guo, BTM: **Topic modeling over short texts**, IEEE Trans. Knowl. Data Eng., 26, 12, 2014. https://doi.org/10.1109/TKDE.2014.2313872
19. N. Gali, R. Mariescu-Istodor, P. Fränti, Similarity measures for title matching, 23rd International

Conference on Pattern Recognition, ICPR 2016, IEEE, 2016**,** pp. 1548-1553. https://doi.org/10.1109/ICPR.2016.7899857

20. T. Talvensaari, J. Laurikkala, K. Järvelin, M. Juhola, **A study on automatic creation of a comparable document collection in cross-language information retrieval**, J Doc, 62, 3, 2006. https://doi.org/10.1108/00220410610666510
21. H. Mohamed Hanum, Z. Abu Bakar, **Detection of Malay phrase breaks using energy and duration, International Journal of Simulation**: Systems, Science and Technology, 17, 32, 2016.