# Feature-based Method for Document Alignment in Comparable News Corpora

**Thuy Vu, Ai Ti Aw, Min Zhang**
Department of Human Language Technology, Institute for Infocomm Research
1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632
{tvu, aaiti, mzhang}@i2r.a-star.edu.sg

## Abstract

In this paper, we present a feature-based method to align documents with similar content across two sets of bilingual comparable corpora from daily news texts. We evaluate the contribution of each individual feature and investigate the incorporation of these diverse statistical and heuristic features for the task of bilingual document alignment. Experimental results on the English-Chinese and English-Malay comparable news corpora show that our proposed Discrete Fourier Transform-based term frequency distribution feature is very effective. It contributes 4.1% and 8% to performance improvement over Pearson's correlation method on the two comparable corpora. In addition, when more heuristic and statistical features as well as a bilingual dictionary are utilized, our method shows an absolute performance improvement of 23.2% and 15.3% on the two sets of bilingual corpora when comparing with a prior information retrieval-based method.

## 1 Introduction

The problem of document alignment is described as the task of aligning documents, news articles for instance, across two corpora based on content similarity. The groups of corpora can be in the same or in different languages, depending on the purpose of one's task. In our study, we attempt to align similar documents across comparable corpora which are bilingual, each set written in a different language but having similar content and domain coverage for different communication needs.

Previous works on monolingual document alignment focus on automatic alignment between documents and their presentation slides or between documents and their abstracts. Kan (2007) uses two similarity measures, Cosine and Jaccard, to calculate the candidate alignment score in his SlideSeer system, a digital library software that retrieves documents and their narrated slide presentations. Daumé and Marcu (2004) use a phrase-based HMM model to mine the alignment between documents and their human-written abstracts. The main purpose of this work is to increase the size of the training corpus for a statistical-based summarization system.

The research on similarity calculation for multilingual comparable corpora has attracted more attention than monolingual comparable corpora. However, the purpose and scenario of these works are rather varied. Steinberger et al. (2002) represent document contents using descriptor terms of a multilingual thesaurus EUROVOC[1], and calculate the semantic similarity based on the distance between the two documents' representations. The assignment of descriptors is trained by log-likelihood test and computed by $TFIDF$, Cosine, and Okapi. Similarly, Pouliquen et al. (2004) use a linear combination of three types of knowledge: cognates, geographical place names reference, and map documents based on the EUROVOC. The major limitation of these works is the use of EUROVOC, which is a specific resource workable only for European languages.

Aligning documents across parallel corpora is another area of interest. Patry and Langlais (2005) use three similarity scores, Cosine, Normalized Edit Distance, and Sentence Alignment Score, to compute the similarity between two parallel documents. An Adaboost classifier is trained on a list of scored text pairs labeled as parallel or non-parallel. Then, the learned classifier is used to check the correctness of each alignment candidate. Their method is simple but effective. However, the features used in this method are only suitable for parallel corpora as the measurement is mainly based on structural similarity. One goal of document alignment is for parallel sentence extraction for applications like statistical machine translation. Cheung and Fung (2004) highlight that most

[1] EUROVOC is a multilingual thesaurus covering the fields in which the European Communities are active.

*Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 843–851,
Athens, Greece, 30 March – 3 April 2009. 

of the current sentence alignment models are applicable for parallel documents, rather than comparable documents. In addition, they argue that document alignment should be done before parallel sentence extraction.

Tao and Zhai (2005) propose a general method to extract comparable bilingual text without using any linguistic resources. The main feature of this method is the frequency correlation of words in different languages. They assume that those words in different languages should have similar frequency correlation if they are actually translations of each other. The association between two documents is then calculated based on this information using Pearson's correlation together with two monolingual features $BM25$, a term frequency normalization (Stephan et al., 1994), and $IDF$. The main advantages of this approach are that it is purely statistical-based and it is language-independent. However, its performance may be compromised due to the lack of linguistic knowledge, particularly across corpora which are linguistically very different. Recently, Munteanu (2006) introduces a rather simple way to get the group of similar content document in multilingual comparable corpus by using the Lemur IR Toolkit (Ogilvie and Callan, 2001). This method first pushes all the target documents into the database of the Lemur, and then uses a word-by-word translation of each source document as a query to retrieve similar content target documents.

This paper will leverage on previous work, and propose and explore diverse range of features in our system. Our document alignment system consists of three stages: candidate generation, feature extraction and feature combination. We verify our method on two set of bilingual news comparable corpora English-Chinese and English-Malay. Experimental results show that 1) when only using Fourier Transform-based term frequency, our method outperforms our re-implementation of Tao (2005)'s method by 4.1% and 8% for the top 100 alignment candidates and, 2) when using all features, our method significantly outperforms our implementation of Munteanu's (2006) method by 23.2% and 15.3%.

The paper is organized as follows. In section 2, we describe the overall architecture of our system. Section 3 discusses our improved frequency correlation-based feature, while Section 4 describes in detail the document relationship heuristics used in our model. Section 5 reports the experimental results. Finally, we conclude our work in section 6.

## 2 System Architecture

Fig 1 shows the general architecture of our document alignment system. It consists of three components: candidate generation, feature extraction, and feature combination. Our system works on two sets of monolingual corpora to derive a set of document alignments that are comparable in their content.

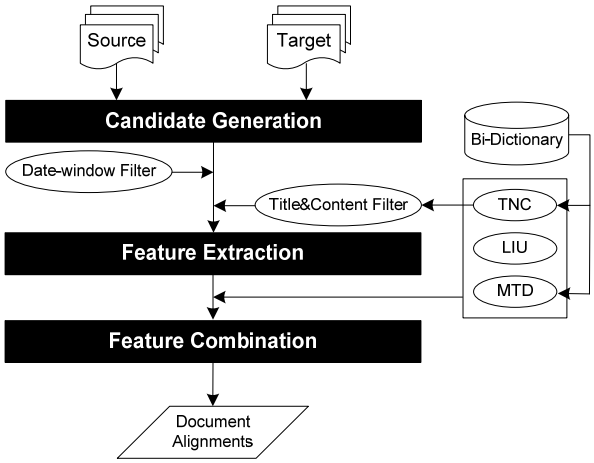


Fig 1. Architecture for Document Alignment Model.

### 2.1 Candidate Generation

Like many other text processing systems, the system first defines two filtering criteria to prune out "clearly bad" candidates. This will dramatically reduce the search space. We implement the following filers for this purpose:

**Date-Window Filter:** As mentioned earlier, the data used for the present work are news corpora—a text genre that has very strong links with the time element. The published date of document is available in data, and can easily be used as an indicator to evaluate the relation between two articles in terms of time. Similar to Munteanu's (2006), we aim to constrain the number of candidates by assuming that documents with similar content should have publication dates which are fairly close to each other, even though they reside in two different sets of corpora. By imposing this constraint, both the complexity and the cost in computation can be reduced tremendously as the number of candidates would be significantly reduced. For example, when a 1-day window size is set, this means that for a given source document, the search for its target candidates is set within 3 days of the source document: the same day of publication, the day after, and the day before. With this filter, using the data of one-month in our experiment, a reduction of 90% of all possible alignments can be achieved (section 5.1). Moreover, with our evaluation data,

after filtering out document pairs using a 1-day window size, up to 81.6% for English-Chinese and 80.3% for English-Malay of the golden alignments are covered. If the window size is increased to 5, the coverage is 96.6% and 95.6% for two language pairs respectively.

**Title-n-Content Filter:** previous date window filter constrains the number of candidates based purely on temporal information without exploiting any knowledge of the documents' contents. The number of candidates to be generated is thus dependent on the number of published articles per day, instead of the candidates' potential content similarity. For this reason, we introduce another filter which makes use of document titles to gauge content-wise cross document similarity. As document titles are available in news data, we capitalize on words found in these document titles, favoring alignment candidates where at least one of the title-words in the source document has its translation found in the content of the other target document. This filter can reduce a further 47.9% (English-Chinese) and 26.3% (English-Malay) of the remaining alignment candidates after applying the date-window filter.

### 2.2 Feature Extraction

The second step extracts all the features for each candidate and computes the score for each individual feature function. In our model, the feature set is composed of the Title-n-Content score ($TNC$), Linguistic-Independent-Unit score ($LIU$), and Monolingual Term Distribution similarity ($MTD$). We will discuss all three features in sections 3 and 4.

### 2.3 Feature Combination

The final score for each alignment candidate is computed by combining all the feature function scores into a unique score. In literature, there are many methods concerning the estimation of the overall score for a given feature set, which vary from supervised to unsupervised method. Supervised methods such as Support Vector Machine (SVM) and Maximum Entropy (ME) estimate the weight of each feature based on training data which are then used to calculate the final score. However, these supervised learning-based methods may not be applicable to our proposed issue as we are motivated to build a language independent unsupervised system. We simply take a product of all normalized features to obtain one unique score. This is because our features are probabilistically independent. In our implementation, we normalize the scores to make them less sensitive to the absolute value by taking the logarithm $ln(.)$ as follows:

$$norm(x) = \begin{cases} ln(x+T), & x > (e-T) \\ 1, & else \end{cases} \quad (1)$$

$(e-T)$ is a threshold for $x$ to contribute positively to the unique score. In our experiment, we empirically choose $T$ be $2.2$, and the threshold for $x$ is $0.51828$ (as $e \approx 2.71828$).

## 3 Monolingual Term Distribution

### 3.1 Baseline Model

The main feature used in Tao and Zhai (2005) is the frequency distribution similarity or frequency correlation of words in two given corpora. It is assumed that frequency distributions of topically-related words in multilingual comparable corpora are often correlated due to the correlated coverage of the same events.

Let $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ and $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ be the frequency distribution vectors of two words $x$ and $y$ in two documents respectively. The frequency correlation of the two words is computed by Pearson's Correlation Coefficient in (2).

$$r(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right)}} \quad (2)$$

The similarity of two documents is calculated with the addition of two features namely Inverse Document Frequency ($IDF$) and $BM25$ term frequency normalization shown in the equation (3).

$$s(d_1,d_2) = \sum_{x \in d_1, y \in d_2} IDF(x) \cdot IDF(y) \cdot r(x,y) \cdot BM25(x,d_1) \cdot BM25(y,d_2) \quad (3)$$

Where $BM25(w,d)$ is the word frequency normalization for word $w$ in document $d$, and $AveDocLen$ is the average length of a document.

$$BM25(w,d) = \frac{k_1 c(w,d)}{c(w,d) + k_1\left(1 - b + b\frac{|d|}{AveDocLen}\right)} \quad (4)$$

It is noted that the key feature used by Tao and Zhai (2005) is the $r(x,y)$ score which depends purely on statistical information. Therefore, our motivation is to propose more features to link the source and target documents more effectively for a better performance.

### 3.2 Study on Frequency Correlation

We further investigate the frequency correlation of words from comparable sets of corpora comprising three different languages using the above-defined model.

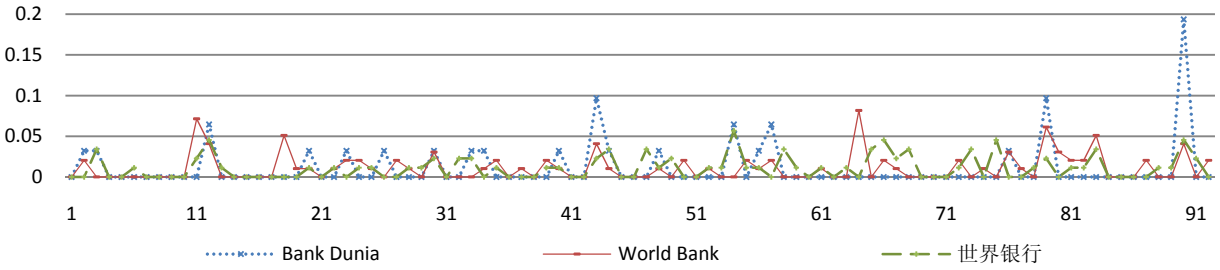


Fig 2. Sample of frequency correlation for "Bank Dunia", "World Bank", and "世界银行".

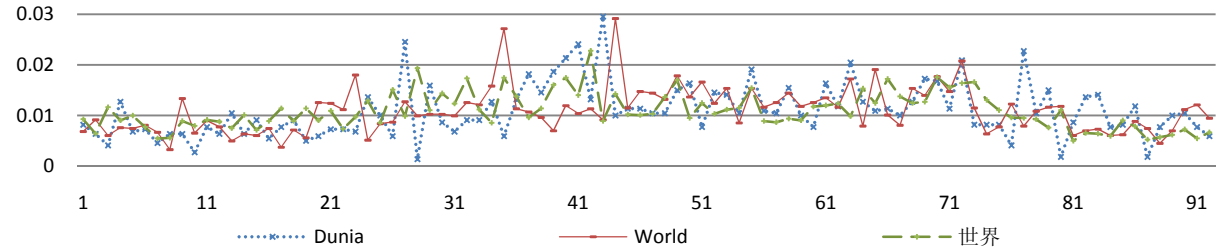


Fig 3. Sample of frequency correlation for "Dunia", "World", and "世界".

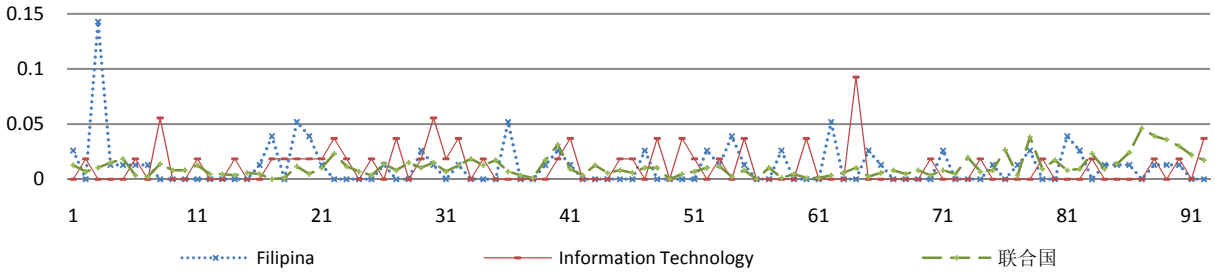


Fig 4. Sample of frequency correlation for "Filipina", "Information Technology", and "联合国".

Using three months - May to July, 2006 – of daily newspaper in Strait Times[2] (in English), Zao Bao[3] (in Chinese), and Berita Harian[4] (in Malay), we conduct the experiments described in the following Fig 2, Fig 3, and Fig 4 showing three different cases of term or word correlation. In these figures, the $x$-axis denotes time and the $y$-axis shows the frequency distribution of the term or word.

**Multi-word versus Single-word:** Fig 2 illustrates that the distributions for multi-word term such as "World Bank", "世界银行(*World Bank* in Chinese)", and "Bank Dunia (*World Bank* in Malay)" in the three language corpora are almost similar because of the discriminative power of that phrase. The phrase has no variance and contains no ambiguity. On the other hand, the distributions for single words may have much less similarity.

**Related Common Word:** we also investigate the similarity in frequency distribution for related common single words in the case of "World", "世界 (*world* in Chinese)", and "Dunia (*world* in Malay)" as shown in Fig 3. It can be observed that the correlation of these common words is not as strong as that in the multi-word sample illustrated in Fig 2. The reason is that there are many variances of these common words, which usually do not have high discriminative power due to the ambiguities presented within them. Nonetheless, among these variances, there is still a small similar distribution trends that can be detected, which may enable us to discover the associations between them.

**Unrelated Common Word:** Fig 4 shows the frequency distribution of three unrelated common words over the same three-month period. No correlation in distribution is found among them.

[2] http://www.straitstimes.com/ an English news agency in Singapore. Source © Singapore Press Holdings Ltd.
[3] http://www.zaobao.com/ a Chinese news agency in Singapore. Source © Singapore Press Holdings Ltd.
[4] http://cyberita.asia1.com.sg/ a Malay news agency in Singapore. Source © Singapore Press Holdings Ltd.

### 3.3 Enhancement from Baseline Model

#### 3.3.1 Monolingual Term Correlation

Due to the inadequacy of the baseline's purely statistical approach, and our studies on the correlations of single, multiple and commonly appearing words, we propose using "term" or "multi-word" instead of "single-word" or "word" to calculate the similarity of term frequency distribution between two documents. This presents us with two main advantages. Firstly, the smaller number of terms compared to the number of words present in any document would imply fewer possible document alignment pairs for the system. This increases the computation speed remarkably. To extract automatically the list of terms in each document, we use the term extraction model from Vu et al. (2008). In corpora used in our experiments, the average ratios of word/term per document are 556/37, 410/28 and 384/28 for English, Chinese, and Malay respectively. The other advantage of using terms is that terms are more distinctive than words as they contain less ambiguity, thus enabling high correlation to be observed when compared with single words.

#### 3.3.2 Bilingual Dictionary Incorporation

In addition to using terms for the computation, we observed from equation (3) that the only mutual feature relating the two documents is the frequency distribution coefficient $r(x,y)$. It is likely that the alignment performance could be enhanced if more features relating the two documents are incorporated.

Following that, we introduce a linguistic feature, $DicScore(x,y)$, to the baseline model to enhance the association between two documents. This feature involves the comparison of the translations of words within a particular term in one language, and the presence of these translations in the corresponding target language term. If more translations obtained from a bilingual dictionary of words within a term are found in the term extracted from the other language's document, it is more likely that the 2 bilingual terms are translations of each other. This feature counts the number of word translation found between the two terms, as described in the following. Let $t_1$ and $t_2$ be the term list of $d_1$ and $d_2$ respectively, the similarity score in our model is:

$$s_{dic}(d_1,d_2) = \sum_{x\in t_1, y\in t_2} IDF(x)\cdot IDF(y)\cdot r(x,y)\cdot DicScore(x,y)\cdot BM25(x,d_1)\cdot BM25(y,d_2) \quad (5)$$

#### 3.3.3 Distribution Similarity Measurement using Monolingual Term

Finally, we apply the results of time-series research to replace Pearson's correlation which is used in the baseline model, in our calculation of the similarity score of two frequency distributions. A popular technique for time sequence matching is to use Discrete Fourier Transform ($DFT$) (Agrawal et al, 1993). More recently, Klementiev and Roth (2006) also use F-index (Hetland, 2004), a score using $DFT$, to calculate the time distribution similarity. In our model, we assume that the frequency chain of a word is a sequence, and calculate $DFT$ score for each chain by the following formula:

$$H_n = \sum_{k=0} h_k . e^{2\pi i k n / N} \quad (6)$$

In time series research, it is proven that only the first few $k$ coefficients of a $DFT$ chain are strong and important for comparison (Agrawal et al, 1993). Our experiments in section 5 show that the best value for $k$ is 7 for both language pairs.

$$R(x,y) = \left( \sqrt{\sum_{i=0}^{k} \left(H_{xi} - H_{yi}\right)^2} \right)^{-1} \quad (7)$$

The $r(x,y)$ in equation (5) is replaced by $R(x,y)$ in equation (8) to calculate the Monolingual Term Distribution ($MTD$) score.

$$s_{MTD}(d_1,d_2) = \sum_{x\in t_1, y\in t_2} IDF(x)\cdot IDF(y)\cdot R(x,y)\cdot DicScore(x,y)\cdot BM25(x,d_1)\cdot BM25(y,d_2) \quad (8)$$

## 4 Document Relationship Heuristics

Besides the $MTD$, we also propose two heuristic-based features that focus directly on the relationship between two multilingual documents, namely the *Title-n-Content* score– $TNC$, which measures the relationship between the title and content of a document pair, and *Linguistic Independent Unit* score – $LIU$, which make use of orthographic similarity between unit of words for the different languages.

### 4.1 Title-n-Content Score ($TNC$)

Besides being a filter for removing bad alignment candidates, $TNC$ is also incorporated as a feature in the computation of document alignment score. In the corpora used, in most documents, "title" does reveal the main topic of a document. The use of words in a news title is

typically concise and conveys the essence of the information in the document. Thus, a high $TNC$ score would indicate a high likelihood of similarity between two bilingual documents. Therefore, we use $TNC$ as a quantitative feature in our feature set. Function $TR(w, c)$ checks whether the translation of a word in a document's title is found in the content of its aligned document:

$$TR(w, c) = \begin{cases} 1, & \text{translation of } w \text{ is in } c \\ 0, & \text{else} \end{cases} \quad (9)$$

The $TNC$ score of document $d_s$ and $d_t$ is calculated by the following formula:

$$TNC(d_s, d_t) = \sum_{w_i \in T_s} TR(w_i, c_t) + \sum_{w_j \in T_t} TR(w_j, d_s) \quad (10)$$

Where $c_t$ and $c_s$ are the content of document $d_t$ and $d_s$; and $T_s$ and $T_t$ are the set of title words of two documents.

In addition, this method speeds up the alignment process without compromising performance when compared with the calculation based only on contents on both sides.

### 4.2 Linguistic Independent Unit (*LIU*)

Linguistic Independent Unit score (LIU) is defined as the piece of information, which is written in the same way for different languages. The following highlight the number 25, 11, and 50 as linguistic-independent-units for the two sentences.

*English*: *Between Feb **25** and March **11** this year, she used counterfeit $50 notes **10** times to pay taxi fares ranging from $2.50 to $4.20.*

*Chinese*:被告使用伪钞的控状，指她从 2 月 **25** 日至 3 月 **11** 日，以 **50** 元面额的伪钞，缴付介于 2 元 5 角至 4 元 2 角的德士费。

## 5 Experiment and Evaluation

### 5.1 Experimental Setup

The experiments were conducted on two sets of comparable corpora namely English-Chinese and English-Malay. The data are from three news publications in Singapore: the Strait Times (ST, English), Lian He Zao Bao (ZB, Chinese), and Berita Harian (BH, Malay). Since these languages are from different language families[5], our model can be considered as language independent.

[5] English is in Indo-European; Chinese is in Sino-Tibetan; Malay is in Austronesian family [Wikipedia].

The evaluation is conducted based on a set of manually aligned documents prepared by a group of bilingual students. It is done by carefully reading through each article in the month of June (2006) for both sets of corpora and trying to find articles of similar content in the other language within the given time window. Alignment is based on similarity of content where the same story or event is mentioned. Any two bilingual articles with at least 50% content overlapping are considered as comparable. This set of reference data is cross-validated between annotators. Table 1 shows the statistics of our reference data for document alignment.

| Language pair | ST – ZB | ST – BH |
|---|---|---|
| Distinct source | 396 | 176 |
| Distinct target | 437 | 175 |
| Total alignments | 438 | 183 |

Table 1. Statistics on evaluation data.

Note that although there are 438 alignments for ST-ZB, the number of unique ST articles are 396, implying that the mapping is not one-to-one.

### 5.2 Evaluation Metrics

Evaluation is performed on two levels to reflect performance from two different perspectives. "Macro evaluation" is conducted to assess the correctness of the alignment candidates given their rank among all the alignment candidates. "Micro evaluation" concerns about the correctness of the aligned documents returned for a given source document.

**Macro evaluation:** we present the performance for macro evaluation using average precision. It is used to evaluate the performance of a ranked list and gives higher score for the list that returns more correct alignment in the top.

**Micro evaluation:** for micro evaluation, we evaluate the F-Score, calculated from recall and precision, based on the number of correct alignments for the top of alignment candidates for each source document.

### 5.3 Experiment and Result

First we implement the method of Tao and Zhai (2005) as the baseline. Basically, this method does not depend on any linguistic resources and calculates the similarity between two documents purely by comparing all possible pairs of words. In addition to this, we also implement Munteanu's (2006) method which uses Okapi scoring function from the Lemur Toolkit (Ogilvie and

Callan, 2001) to obtain the similarity score. This approach relies heavily on bilingual dictionaries. To assess performances more fairly, the result from baseline method of Tao and Zhai are compared against the results of the following list of incremental approaches: the baseline **(A)**; the baseline using term instead of word **(B)**; replacing $r(x, y)$ by $R(x, y)$ for $DFT$ feature, with and without bilingual dictionaries in **(C)** and **(D)** respectively; and including $LIU$ and $TNC$ for our final model in **(E)**. Our model is also compared our model with results from the implementation of Munteanu (2006) using Okapi **(F),** and the results from a combination of our model with Okapi **(G)**. Table 2 and Table 3 show the experimental results for two language pairs English – Chinese (ST-ZB) and English – Malay (ST-BH), respectively. Each row displays the result of each experiment at a certain cut-off among the top returned alignments. The "Top" columns reflect the cut-off threshold.

The first three cases **(A)**, **(B)** and **(C),** which do not rely on linguistic resources, suggest that our new features lead to better performance improvement over the baseline. It can be seen that the use of term and $DFT$ significantly improves the performance. The improvement indicated by a sharp increase in all cases from **(C)** to **(D)** shows that dictionaries can indeed help $DFT$ features.

Based on the result of **(E)**, our final model significantly outperforms the model of Munteanu **(F)** in both macro and micro evaluation. It is noted that our features rely less heavily on dictionaries as it only makes use of this resource to translate term words and title words of a document while Munteanu (2006) needs to translate entire documents, exclude stopword, and relying on an IR system. It is also observed that the performance of **(G)** shows that although the incorporation of Okapi score in our final model **(E)** improves the average precision performance of ST-ZB slightly, it does not appear to be helpful for our ST-BH data. However, Okapi does help in the F-Measure on both corpora.

| Pair | | Strait Times – Zao Bao | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Level | Top | A | B | C | D | E | F | G |
| Ave/Precision Macro | 50 | 0.042 | 0.083 | 0.08 | 0.559 | 0.430 | 0.209 | 0.508 |
| | 100 | 0.042 | 0.069 | 0.083 | 0.438 | 0.426 | 0.194 | 0.479 |
| | 200 | 0.025 | 0.069 | 0.110 | 0.342 | 0.396 | 0.153 | 0.439 |
| | 500 | 0.025 | 0.054 | 0.110 | 0.270 | 0.351 | 0.111 | 0.376 |
| F-Measure Micro | 1 | 0.005 | 0.007 | 0.009 | 0.297 | 0.315 | 0.157 | 0.333 |
| | 2 | 0.006 | 0.005 | 0.013 | 0.277 | 0.286 | 0.133 | 0.308 |
| | 5 | 0.005 | 0.006 | 0.009 | 0.200 | 0.190 | 0.096 | 0.206 |
| | 10 | 0.005 | 0.005 | 0.007 | 0.123 | 0.119 | 0.063 | 0.126 |
| | 20 | 0.006 | 0.008 | 0.007 | 0.073 | 0.074 | 0.038 | 0.076 |

Table 2. Performance of Strait Times – Zao Bao.

| Pair | | Strait Times – Berita Harian | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Level | Top | A | B | C | D | E | F | G |
| Ave/Precision Macro | 50 | 0.000 | 0.000 | 0.000 | 0.514 | 0.818 | 0.000 | 0.782 |
| | 100 | 0.000 | 0.000 | 0.080 | 0.484 | 0.759 | 0.052 | 0.729 |
| | 200 | 0.000 | 0.008 | 0.090 | 0.443 | 0.687 | 0.073 | 0.673 |
| | 500 | 0.005 | 0.008 | 0.010 | 0.383 | 0.604 | 0.078 | 0.591 |
| F-Measure Micro | 1 | 0.000 | 0.000 | 0.005 | 0.399 | 0.634 | 0.119 | 0.650 |
| | 2 | 0.000 | 0.004 | 0.010 | 0.340 | 0.515 | 0.128 | 0.515 |
| | 5 | 0.002 | 0.005 | 0.010 | 0.205 | 0.270 | 0.105 | 0.273 |
| | 10 | 0.004 | 0.014 | 0.013 | 0.130 | 0.150 | 0.076 | 0.150 |
| | 20 | 0.006 | 0.017 | 0.017 | 0.074 | 0.078 | 0.043 | 0.078 |

Table 3. Performance of Strait Times – Berita Harian.

### 5.4 Discussion

It can be seen from Table 2 and Table 3 that by exploiting the frequency distribution of terms using Discrete Fourier Transform instead of words on Pearson's Correlation, performance is noticeably improved. Fig 5 shows the incremental improvement of our model for top-200 and top-2 alignments using macro and micro evaluation respectively. The sharp increase can be seen in Fig 5 from point **(C)** onwards.

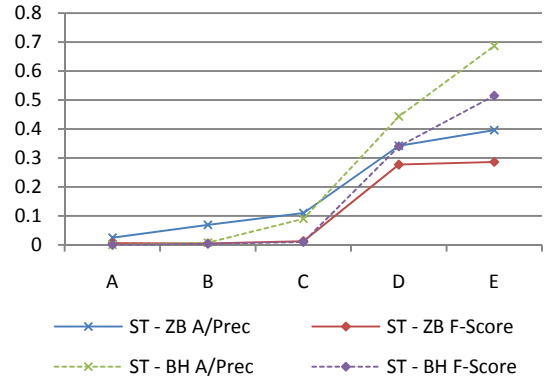


Fig 5. Step-wise improvement at top-200 for macro and top-2 for micro evaluation.

Fig 6 compares the performance of our system with Tao and Zhai (2005) and Munteanu (2006). It is shown that our systems outperform these two systems under the same experimental parameters. Moreover, even without the use of dictionaries, our system's performance on ST-BH data is much better than Munteanu's (2006) on the same data.

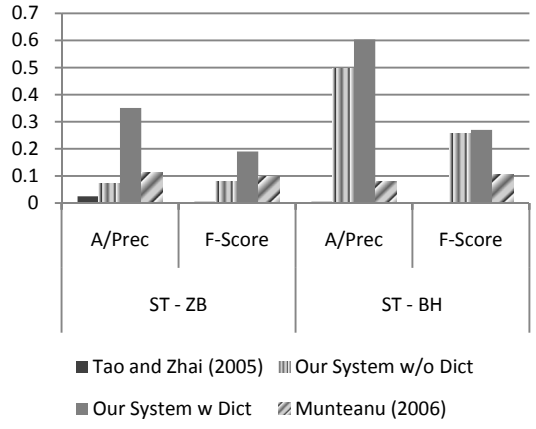


Fig 6. System comparison for ST-ZB and ST-BH at top-500 for macro and top-5 for micro evaluation.

We find that dictionary usage contributes much more to performance improvement in ST-BH compared to that in ST-ZB. We attribute this to the fact that the feature LIU already contributes markedly to the increase in the performance of ST-BH. As a result, it is harder to make further improvements even with the application of bilingual dictionaries.

## 6 Conclusion and Future Work

In this paper, we propose a feature based model for aligning documents from multilingual comparable corpora. Our feature set is selected based on the need for a method to be adaptable to new language-pairs without relying heavily on linguistic resources, unsupervised learning strategy. Thus, in the proposed method we make use of simple bilingual dictionaries, which are rather inexpensive and easily obtained nowadays. We also explore diverse features, including Monolingual Term Distribution ( $MTD$ ), Title-and-Content ($TNC$), and Linguistic Independent Unit ($LIU$) and measure their contributions in an incremental way. The experiment results show that our system can retrieve similar documents from two comparable corpora much better than using an information retrieval, such as that used by Munteanu (2006). It also performs better than a word correlation-based method such as Tao's (2005).

Besides document alignment as an end, there are many tasks that can directly benefit from comparable corpora with documents that are well-aligned. These include sentence alignment, term alignment, and machine translation, especially statistical machine translation. In the future, we aim to extract other valuable information from comparable corpora which benefits from comparable documents.

## Acknowledgements

We would like to thank the anonymous reviewers for their many constructive suggestions for improving this paper. Our thanks also go to Mahani Aljunied for her contributions to the linguistic assessment in our work.

## References

Percy Cheung and Pascale Fung. 2004. Sentence Alignment in Parallel, Comparable, and Quasi-comparable Corpora. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal.

Hal Daume III and Daniel Marcu. 2004. A Phrase-Based HMM Approach to Document/Abstract Alignment. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*. Spain.

Min-Yen Kan. 2007. SlideSeer: A Digital Library of Aligned Document and Presentation Pairs. In *Proceedings of the Joint Conference on Digital Libraries (JCDL)*. Vancouver, Canada.

Soto Montalvo, Raquel Martinez, Arantza Casillas, and Victor Fresno. 2006. Multilingual Document Clustering: a Heuristic Approach Based on Cognate Named Entities. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)*. Gaithersburg, USA.

Dragos Stefan Munteanu. 2006. Exploiting Comparable Corpora. PhD Thesis. Information Sciences Institute, University of Southern California. USA.

Ogilvie, P., and Callan, J. 2001. Experiments using the Lemur toolkit. In *Proceedings of the 10th Text REtrieval Conference (TREC)*.

Alexandre Patry and Philippe Langlais. 2005. Automatic Identification of Parallel Documents with light or without Linguistics Resources. In *Proceedings of 18th Annual Conference on Artificial Intelligent.*

Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Emilia Kasper, and Irina Temnikova. 2004. Multilingual and Cross-lingual news topic tracking. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. 2002. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. *Computational Linguistics and Intelligent Text Processing*.

Tao Tao and ChengXiang Zhai. 2005. Mining Comparable Bilingual Text Corpora for Cross-Language Information Integration. In *Proceedings of the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Thuy Vu, Ai Ti Aw and Min Zhang. 2008. Term extraction through unithood and termhood unification. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*. Hyderabad, India.

ChengXiang Zhai and John Lafferty. 2001. 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*. Louisiana, United States.

R. Agrawal, C. Faloutsos, and A. Swami. 1993. Efficient similarity search in sequence databases. In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*. Chicago, United States.

Magnus Lie Hetland. 2004. A survey of recent methods for efficient retrieval of similar time sequences. In *Data Mining in Time Series Databases*. World Scientific.

Alexandre Klementiev and Dan Roth. 2006. Weakly Supervised Named Entity Transliteration and Discovery from Multilingual Comparable Corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*.