Journal of Computer Science 2012, 8 (12), 1970-1978
ISSN 1549-3636

doi:10.3844/jcssp.2012.1970.1978 Published Online 8 (12) 2012 (http://www.thescipub.com/jcs.toc)

OPTIMIZING CLUSTERS ALIGNMENT FOR BILINGUAL MALAY-ENGLISH CORPORA

[1]Rayner Alfred, [1]Chan Chen Jie, [1]Ng Zhen Wei, [1]Asni Tahir and [2]Joe Henry Obit

[1]School of Engineering and Information Technology,
Universiti Malaysia Sabah, Malaysia
[2]School of Informatics Science Labuan, Universiti Malaysia Sabah, Malaysia

Received 2012-07-13, Revised 2012-09-20; Accepted 2012-11-26

ABSTRACT

Bilingual corpora, containing the same documents in two different languages, are becoming an essential resource for natural language processing. Clustering bilingual corpora provides us with an insight into the differences between languages when term frequency-based Information Retrieval (IR) tools are used. It also allows one to use the Natural Language Processing (NLP) and IR tools in one language to implement IR for another language. This study reports on our work on applying Hierarchical Agglomerative Clustering (HAC) to a large corpus of documents where each appears both in Malay and English languages. These documents are clustered for each language and both results are compared with respect to the content of clusters produced. Further, the effects of using different methods of computing the inter-clusters distance on the cluster results is also studied. These methods include Single, Complete and Average links. Finally, this study describes an experiment employing a genetic algorithm to fine-tune individual term's weight in order to reproduce more closely a predefined set of clusters. In this way, clustering becomes a supervised learning technique that is trained to better reproduce known clusters in Malay language when applied to the corresponding documents in English language. On the data available, the results of clustering one language resemble the other, provided the number of clusters required is relatively small. The method used to compute the inter-clusters distance also influences the cluster results. The result actually showed an increase in the percentage of aligned clusters, when we applied the genetic algorithm to fine-tune weights of terms considered in clustering the bilingual Malay-English corpora. This study concludes that with a smaller number of clusters, k = 5, all of the clusters from English texts can be mapped into the clusters of Malay texts, by using the Complete link distance measure in clustering the bilingual parallel corpus. In contrast, with a large size of clusters, fewer clusters from English texts can be mapped into the clusters of Malay texts.

Keywords: Bilingual Corpora, Hierarchical Agglomerative Clustering, Parallel Clustering, Genetic Algorithm, Malay-English Corpora, Knowledge Management

1. INTRODUCTION

In labeling articles in both languages, an appropriate clustering technique must be applied in order to have an efficient and effective representation of articles in both languages. In particular, clustering algorithms that build illustrative and meaningful hierarchies (e.g., hierarchical agglomerative clustering technique) out of large document collections are ideal tools for their interactive visualization and exploration, as they provide data-views that are consistent, predictable and contain multiple levels of granularity. Thus, effective and efficient document clustering algorithms are required in order to provide efficient and effective intuitive navigation and browsing mechanisms by categorizing large amount of information into a small number of meaningful clusters.

There has been a lot of researches in clustering text documents. However, there are few experiments that

Corresponding Author: Rayner Alfred, School of Engineering and Information Technology, Universiti Malaysia Sabah, Malaysia

examine the impacts of clustering corpora when the weights of terms are tuned by using a genetic algorithm in order to optimize the clustering results. In our previous works, we found that by reducing the number of terms when clustering bilingual Bulgarian-English articles in parallel, the percentage of aligned clusters can be improved (Alfred, 2009). In contrast, applying clustering algorithm to a set of documents based on the fine-tuned weights of terms that exist in the documents can be attractive compared to a clustering algorithm for the same documents based on all equally weighted terms. For instance, clustering the corpora, based on the fine-tuned weights of terms that exist in the documents, may increase the quality of clustering results, since the weights of terms are fine-tuned according to a predefined fitness function implemented in the optimization algorithm (e.g., evolutionary algorithm).

The aim of the experiments presented in this study is to investigate the effects of applying a clustering technique to parallel bilingual texts on the cluster results, based on the fine-tuned weights of terms that exist in the documents. Specifically, the aim is to introduce the tools necessary for this task and display a set of experimental results and issues which have become apparent. In this experiment, it is interesting to look at the similarities and differences of two main areas: Malay-English cluster mapping alignments and the most representative terms extracted for Malay-English clusters. In this study, we provide the results of clustering parallel corpora of Malay-English texts based on the fine-tuned weights of terms that exist in the documents. In addition to that, we also present some findings obtained on the mapping the Malay-English clusters and also on the most representative terms extracted for Malay-English clusters.

This study is organized as followed. First, we explain some of the background knowledge related to the vector space model representation of documents, the hierarchical agglomerative clustering method, genetic algorithm and the semi-supervised clustering technique. Then, we describe the experimental design set-up. Finally, the experimental evaluation is discussed and then the conclusion section summarizes the study and presents some ideas for future research.

2. MATERIALS AND METHODS

2.1. Vector Space Model Representation

In this experiment, a vector space model (Salton *et al.*, 1975) is used to represent a document as a vector in n-dimensional space (where n is the number of different terms in the Bag of Words (BOW)). Here, documents are categorized by the words they contain and their weights. Before computing the weights for all terms extracted from documents, pre-process tasks that include stemming and stop word removal are performed. Stopword removal eliminates irrelevant terms (e.g., those from the closed vocabulary) and thus reduces the number of dimensions in the term-space. Then, the weight of term can be computed by counting the frequency of each term across the corpus and weighting them using Term Frequency-Inverse Document Frequency (TF-IDF) (Salton *et al.*, 1975), as shown in (1).

Weights are assigned to give an indication of the importance of a word in characterizing a document as distinct from the rest of the corpus. In summary, each document is viewed as a vector whose dimensions correspond to words or terms extracted from the document. The component magnitudes of the vector are the tf-idf weights of the terms. In this model, tf-idf, as described in equation (1), is the product of term frequency tf(t,d), which is the number of times term t occurs in document d and the inverse document frequency, equation (2), where |D| is the number of documents in the complete collection and df(t) is the number of documents in which term t occurs at least once. To account for documents of different lengths, the length of each document vector is normalized so that it is of unit length (Rijsbergen, 1979).

2.2. Hierarchical Agglomerative Clustering (HAC)

In this study, we concentrate on the hierarchical agglomerative clustering technique. A Hierarchical agglomerative algorithm builds the solution by initially assigning each document to its own cluster and then repeatedly selecting and merging pairs of clusters, to obtain a single all-inclusive cluster, generating the cluster tree from leaves to root (Zhao *et al.*, 2005). The main parameters in agglomerative algorithms are the metric used to compute the similarity of documents and the method used to determine the pair of clusters to be merged at each step.

In these experiments, the cosine distance, equation (3), is used to compute the similarity between two documents d_i and d_j. This widely utilized document similarity measure becomes 1 if the documents are identical and 0 if they share no words. The two clusters, to merge at each step, are found by using either, the Single link, Complete link or Average link method (Khalilian and Mustapha, 2010; Torres *et al.*, 2009). In this scheme, the two clusters to merge are those with the greatest minimum (Single link), maximum (Complete link) or average (Average link) similarity distances between the documents in one cluster and those in the other (Khalilian and Mustapha, 2010; Torres *et al*., 2009). Given a set of documents D, one can measure how consistent the results of clustering for each of the languages to which these documents are translated in the following way. The clusters produced for one language are used as the 'gold standard', a source of annotation assigning each document in the set D a cluster label L from the list L_{ALL} of all clusters for that particular language. Clustering in the other language is then carried out and purity (Pantel and Lin, 2002), equation (5), is used to compare each of the resulting clusters $C \in C_{ALL}$ to its closest match among all clusters L_{ALL}.

2.3. Genetic Algorithm

A Genetic Algorithm (GA) is a computational abstraction of biological evolution that can be used to some optimization problems (Holland, 1992). In its simplest form, a GA is an iterative process applying a series of genetic operators such as selection, crossover and mutation to a population of elements. These elements, called chromosomes, represent possible solutions to the problem. Initially, a random population is created, which represents different points in the search space. An objective and fitness function is associated with each chromosome that represents the degree of goodness of the chromosome. Based on the principle of the survival of the fittest, a few of the chromosomes are selected and each is assigned a number of copies that go into the mating pool. Biologically inspired operators like crossover and mutation are applied on these strings to yield a new generation of strings. The process of selection, crossover and mutation continues for a fixed number of generations or till a termination condition is satisfied. More details survey of Genetic Algorithms can be found in (Filho *et al*., 1994).

In this study, we examine the clustering algorithm that minimizes some objective functions applied to k-cluster centers. In our case, we consider the cluster dispersion. Before the clustering task, each term is assigned with a specific weight that is normalized across all terms. The main objective is to choose the best weight for all terms considered that minimize some measure of cluster dispersion. Typically cluster dis-persion metric is used, such as the Davies-Bouldin Index (DBI) (Davies and Bouldin, 1979). DBI uses both the intra-cluster and inter-clusters distances to measure the cluster quality. Let $d_{centroid}(Q_k)$, defined in (8), denotes the average link distances within-cluster Q_k, where $x_i \in Q_k$, N_k is the number of samples in cluster Q_k, c_k is the center of the cluster and $k \le K$ clusters. Let $d_{between}(Q_k, Q_l)$, defined in (10), denotes the distances inter-clusters Q_k and Q_l, where c_k is the centroid of cluster Q_k and c_l is the centroid of cluster Q_l. In this study, we also compute the inter-clusters distance based on the minimum (Single link), maximum (Complete link) and average (Average link) distance methods between clusters Eq. 1-10:

$$tf - idf = tf(t,d) \cdot idf(t) \quad (1)$$

$$idf(t) = \log\left(\frac{|D|}{df(t)}\right) \quad (2)$$

$$sim(d_i, d_j) = \frac{(d_i d_j)}{(\|d_i\| \cdot \|dj\|)} \quad (3)$$

$$Precision\ (C,L) = \frac{|C \cap L|}{|C|}, C \in C_{ALL}, L \in L_{ALL} \quad (4)$$

$$Purity = \sum_{CEC_{ALL}} \frac{|C|}{|D|}.P(C,L) \quad (5)$$

$$Precision\ (EMM) = \frac{C(E) \cap C(M)}{C(E)} \vartriangle ABC \quad (6)$$

$$Precision\ (MEM) = \frac{C(M) \cap C(E)}{C(M)} \quad (7)$$

$$d_{centroid}(Q_k) = \frac{\sum_i \|x_i - c_k\|}{N_k} \quad (8)$$

$$(c_k = 1/N_k \left(\sum_{X_{i \in Q_k}} x_i\right) \quad (9)$$

$$d_{between}(Q_k, Q_l) = \|c_k - c_l\| \quad (10)$$

Therefore, given a partition of the N points into K-clusters, DBI is defined in (11). This cluster dispersion measure can be incorporated into any clustering algorithm in order to evaluate a particular segmentation of data Eq. 11 and 12:

$$DBI = \frac{1}{K} \sum_{k=1}^{K} \max_{l \neq k} \left\{ \frac{d_{centroid}(Q_k) + d_{centroid}(Q_l)}{d_{between}(Q_k, Q_l)} \right\} \quad (11)$$

$$f(N,K) = \text{Cluster Dispersion} = \text{DBI} \quad (12)$$

In general, the objective function is defined in (12). By minimizing the objective function that minimizes the cluster dispersion measure (DBI), a better quality of clusters can be produced. More specifically, given N points and K-clusters, select the weight of each terms in the documents so that the objective function defined in (12) can be minimized.

2.4. Clustering Parallel Corpora

In the first stage of the experiment, there are two set of parallel corpora in two different languages; Malay and English languages. In both corpora, each English document E corresponds to a Malay document M with the same content.

The process of stemming English corpora is relatively simple due to the low inflectional variability of English. However, for morphologically richer languages, such as Malay language, where the impact of stemming is potentially greater, the process of building an accurate algorithm becomes a more challenging task (Sembok and Bakar, 2011; Abdullah *et al.*, 2009). In this experiment, the Malay texts are stemmed by using the Rules Frequency Order (RFO) stemmer (Abdullah *et al.*, 2009). Documents in each language are clustered separately using hierarchical agglomerative clustering. The output of each run consists of two elements: a list of terms characterizing each cluster and the cluster members. A detailed comparison of the results for the two languages looking at each of these elements will be discussed in this study.

2.5. Fine-Tuning Weights of Terms using a Genetic Algorithm

The second stage of the experiment is clustering the documents based on a set of optimized weights of the terms that exist in the document in order to best cluster the documents according to the fitness function of the GA, defined in (12). Here, we describe the representation of the problem in the Genetic Algorithm setting.

A population of X strings of length m is randomly generated, where m is the number of unique terms (e.g., cardinality of terms) that exist in the corpus. X strings are generated with continuous numbers (0.5, 1.0 and 1.5) representing the term's weights that will be used to adjust the tf-idf weight.

The computation of the objective or fitness function is based on the Cluster Dispersion. In order to get clusters of better quality, DBI value must be minimized which is defined in (11). In other words, the Objective Fitness Function (OFF) that we want to maximize will be, OFF = 1/DBI.

For the selection process, a rouleete wheel with slots sized according to the fitness is used. First, the fitness value for each chromosome, f_i and $i \leq X$, is calculated and the total overall fitness for X strings of chromosome, $T_{Fitness}$, is obtained. Then, the probability of a selection p_i for each chromosome, $i \leq X$, $p_i = f_i/T_{Fitness}$, is calculated. Finally, the cumulative probability q_i for each chromosome, $q_i = \sum_{j=1}^{i} p_j$, is calculated. The selection process is based on spinning the roulette wheel, X times. Each time we select a single chromosome for a new population, a random number r from the range of [0..1] is generated and the i-th chromosome such that $q_{i-1} < r \leq q_i$, is selected.

For the crossover process, a pair of chromosome, c_i and c_j, are chosen for applying the crossover operator. One of the parameters of a genetic system is probability of crossover p_c. In this experiment, we set $p_c = 0.25$. This probability gives us the expected number $p_c \bullet X$ of chromosomes, which undergo the crossover operation. We proceed by generating a random number of r from the range [0..1]. Then, we perform the crossover if $r < p_c$. For each pair of coupled chromosomes we generate a random integer number pos from the range [1..(m-1)] (where m is the length of the chromosome), which indicates the position of the crossing point.

Finally, the mutation operator performs a weight-by-weight basis with values 0.5, 1.0 and 1.5. Another parameter of the genetic system, probability of permutation p_m gives the expected number of mutated weights. In this experiment, we set $p_m = 0.01$. In the mutation process, for each chromosome and for each weight within the chromosome by generating a random number of r from the range [0..1] and performing the mutation of each bit if $r < p_m$.

3. RESULTS

3.1. Mapping of Malay-English Cluster Alignment

In the first experiment, every cluster in Malay is paired with the English cluster with which it shares the most documents. The same is repeated in the direction of English to Malay mapping. There are 200 pairs of Malay-English documents obtained from the Malaysia News

(The Star Online) that cover 6 categories; Business, Feature, General, Politics and Sport news from the year of 2009 until 2010. Two precision values of these pairs are then calculated, the precision of the Malay-English Mapping (MEM) and that of the English-Malay Mapping (EMM).

Figure 1-3 show the precisions for the EMM and MEM for the cluster pairings obtained with k = 5 (numbers of clusters) and also with three different inter-clusters distance method used (Single link, Complete link and Average link), for each of the two set of documents in two different languages, Malay and English. The X axis label indicates the identification number of the cluster whose nearest match in the other language is sought, while the Y axis indicates the precision of the best match found. For example, English cluster 2 (E2) is best matched with Malay cluster 3 (M3) with the EMM mapping precision equal to 66.67% and MEM precision equal to 100.00%, as shown in **Fig. 2. Table 1** shows the percentage of aligned clusters, using the Single, Complete and Average link methods.

Table 1. Percentage of Malay-English clusters alignment

Inter-cluster distance	Cluster alignment (%) k = 5	k = 10	k = 15	Average
Single link	80.0	60.0	80.0	73.3
Complete link	100.0	90.0	86.7	92.3
Average link	80.0	80.0	86.7	82.3
Average	86.7	76.7	84.5	82.6

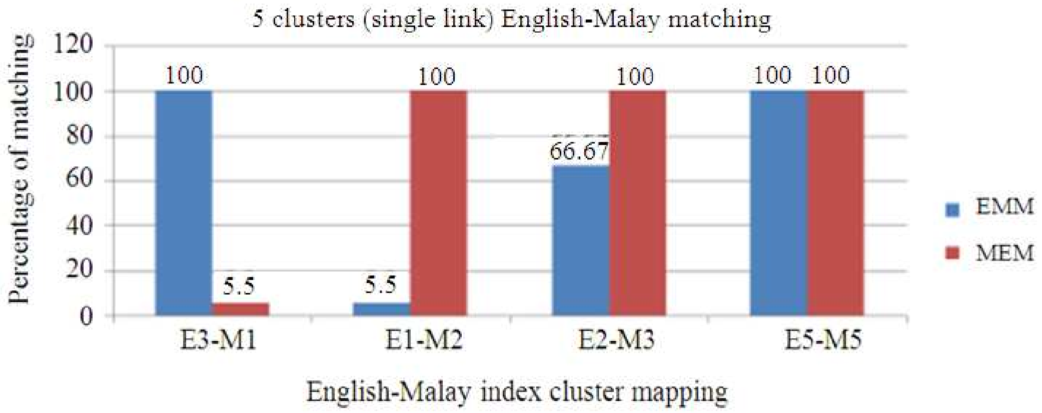

Fig. 1. Cluster mapping results-Single link with 5 clusters

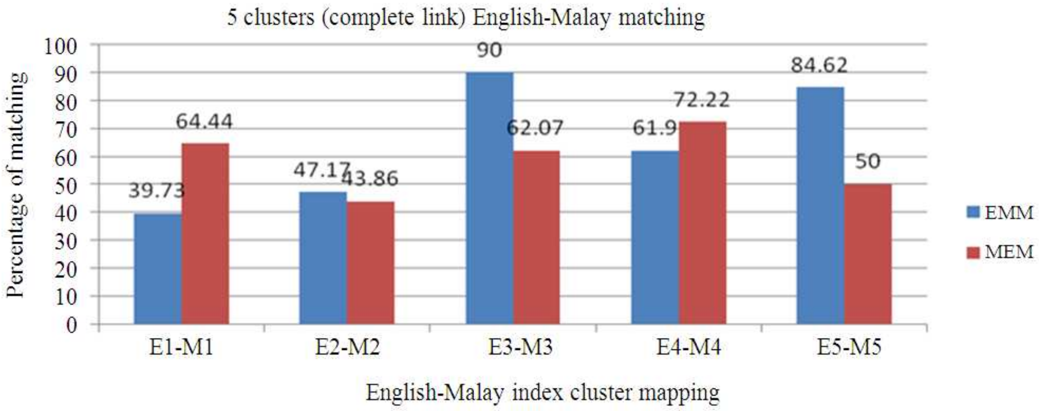

Fig. 2. Cluster mapping results-Complete link with 5 clusters

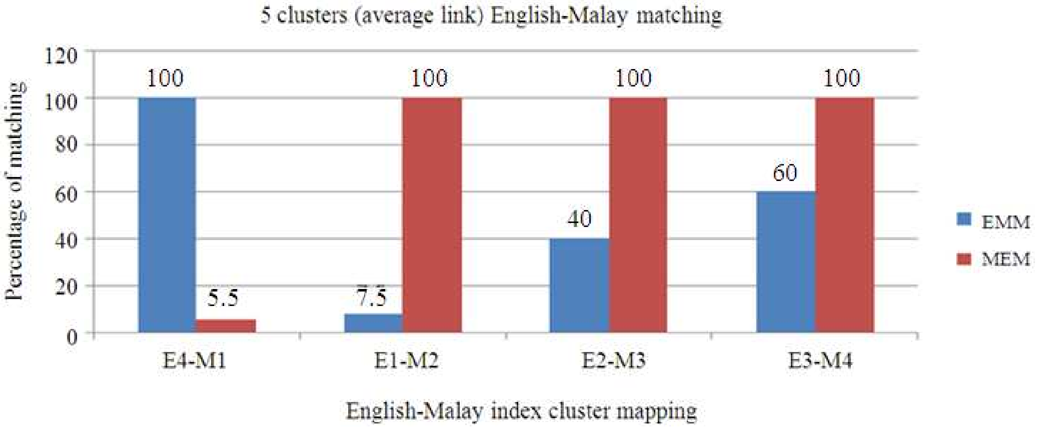

Fig. 3. Cluster mapping results-average link with 5 clusters

Table 2. Terms extracted for single link based clustering

Mapping	English cluster	Malay cluster
k = 5		
E3-M1	Grave, skelet, rebury, graveyard, kin, reloc, remain, pusara, burial, tier	Parti, bank, umno, bn, anggota, pkr, atus, joh, negeri, sukan
E1-M2	Bank, parti, pkr, umno, cent, bn, mca, presid, polic, ong	Pos, pam, pow, mesin, khidmat, unit, gerak, matik, serupa, jabat
k = 10		
E2-M1	Eti, solar, tech, batteri, mou, lithium, sirim, technolog, green, system	Parti, bank, umno, bnm anggota, pkr, joh, negeri, atus, sukan
E1-M2	Bank, parti, pkr, umno, cent, bn, mca, presid, polic, ong	Pos, pam, pow, mesin, khidmat, unit, gerak, matik, serupa, jabat
k = 15		
E6-M1	Paint, voc, chemic, soo, eco, odour, hazard, low, fume, opac	Parti, bank, umno, bn, anggota, pkr, joh, negeri, sukan, wang
E1-M2 iproperty,	Paint, voc, chemic, soo, eco, odour, hazard, low, fume, opac	Hartanah, sunway, atus, templer, ieli, country, com, janj, ekar

3.2. Comparison of Extracted Terms

The ten most representative terms that describe the matching English and Malay clusters have a similar meaning as illustrated in **Table 2-4** (k = 5, k = 10 and k = 15), for each different method of measuring the inter-cluster distance (Single link, Complete link and Average link).

The only notable exception is listed in the first two mappings (E3-M1 and E1-M2 (k = 5), E2-M1 and E1-M2 (k = 10) and E6-M1 and E1-M2 (k = 10)) in **Table 2**, where all top English terms are less related to the Malay terms extracted when clustering using the Single link. **Table 3** shows the mappings (E6-M1 and E1-M4 (k = 10) and E5-M1, E4-M2, E3-M5, E7-M6, E6-M9 and E10-M14 (k = 15)) that indicate less related terms extracted between the two sets of documents in different languages (Malay and English)). However, when k = 5, the mappings are well aligned and the terms extracted for the Malay and English clusters are very well related. **Table 4** shows the mappings (E4-M1 and E1-M2 (k = 5), E10-M2 and E5-M3 (k = 10), E10-M2, E15-M3, E8-M4, E2-M8 (k = 15)) that indicate less related terms extracted between the two sets of documents in different languages. **Table 5** shows the percentage of mappings with less related terms extracted from the mappings of Malay and English clusters. The lowest percentage of mappings with less related terms extracted occurs when the complete link distance measure is used to cluster bilingual Malay-English documents. However, mapping Malay-English clusters, with k = 10, will produce better results on average.

Table 3. Terms extracted for complete link based clustering

Mapping	English cluster	Malay cluster
k = 5		
E1-M1	Polic, rubber, sailor, embassy, finance, banana, million, risda, develop, compani	Ancong, umno, Labuan, daftar, tronas, polis, wang, kawas, taman, air
E2-M2	Bank, umno, cent, honei, hsbc, property, custom, eon, internet, syndrom	Bank, atus, madu, internet, getah, jualan, daun, udang, hsbc, khidmat
E3-M3	Pkr, parti, mca, bn, ong, presid, elect, tm, pbb, mp	Parti, pkr, bn, anggota, mca, anwar, presiden, ayar, parlimen, pbb
E4-M4	Athlet, boxer, gold, medal, category, ronoh, fuad, swim, Terengganu, ironman	Pemain, beregu, law, wei, joh, filem, jarring, minit, chong, buka
E5-M5	Race, minut, win, goal, chong, team, cup, wei, titl, singl	Tm, tinju, sukan, lumba, inju, kategor, pingat, engganu, sukma, atlet
k = 10		
E6-M1	Tourism, park, penang, hot, spring, tawau, dengu, tourist, venu, seberang	Ancong, tronas, taman, ng, wang, kawas, miri, unjung, panas, najib
E1-M4	Sailor, embassy, banana, leaf, thaipusam, petrona, film, innov, rice, lubric	Ayar, down, anak, sindrom, unta, india, perahu, denggi, pakist, latih
k = 15		
E5-M1	Miri, najib, visit, project, plaza, facebook, mainten, muhyiddin, contractor	Ancong, tronas, taman, ng, wang, kawas, miri, unjung, panas, najib
E4-M2	Rubber, risda, replant, smallhold, choi, nurin, hectar, itrc, ik, summon	Madu, getah, udang, galah, benih, hartanah, risda, inovas, tualang, atus
E3-M5	Finance, port, asli, devic, orang, cent, change, market, prudenti, company	Atus, fdi, bas, laluan, rapid, prudential, change, perty, equities, suku
E7-M6	Honei, syndrome, paint, children, tualang, nose, fama, rhinitis, language, kdsf	Ayar, down, anak, sindrom, unta, india, perahu, denggi, pakist, latih
E6-M9	Seedstock, camel, chef, prawn, antique, pastri, jefri, cake, academi, ng	Chef, kedai, pastri, poh, kek, antic, akadem, lanz, jefri, keris
E10-M14	Tm, pbb, taekwondo, Sarawak, spdp, secretary, elect, bn, parti, baling	Pkr, anwar, fairus, parti, parlimen, mohammad, anggota, bangkang, long, rakyat

Table 4. Terms extracted for average link based clustering

Mapping	English cluster	Malay cluster
k = 5		
E4-M1	Grave, skelet. Rebury, graveyard, kin, reloc, remain, pusara, burial, tier	Parti, bank, umno, bn, anggota, pkr, atus, joh, negeri, sukan
E1-M2	Bank, parti, pkr, umno, cent, bn, mca, polic, presid, ong	Abuh, westports, araf, pinang, denggi, eti, tech, solar, teu, bateri
k = 10		
E10-M2	Miri, muhyiddin, tamu, bintulu, visit, muhibbah, kedayan, educt, bakam, arriv	Parti, umno, bn, pkr, anggota, tm, mca, negeri, presiden, rakyat
E5-M3	Grave, skelet, rebury, graveyard, kin, reloc, remain, pusara, burial, tier	Polis, long, kes, saman, singapura, yeludup, sawat, nurin, gunasegar, yelamat
k = 15		
E10-M2	Antique, jefri, shop, kri, stone, bundl, collect, nut, slicer, coin	Getah, ancong, anak, Bandar, filem, down, sindrom, gram, wilayah, risda
E15-M3	Miri, muhyiddin, tamu, bintulu, visit, muhibbah, kedayan, educt, bakam, arriv	Parti, umno, bn, pkr, anggota, mca, parlimen, presiden, anwar, rakyat
E8-M4	Grave, skelet, rebury, graveyard, kin, reloc, remain, pusara, burial, tier	Polis, long, kes, saman, singapura, yeludup, sawat, nurin, gunasegar, yelamat
E2-M8	Syndrome, develop, asli, innov, citi, tourism, devic, orang, najib, park	Eti, tech, solar, bateri, amam, mou, litium, sirim, etera, teknolog

Table 5. Percentage of mappings using less related terms

Inter-Cluster Distance	Percentage of mappings with less related terms extracted			
	k = 5	k = 10	k = 15	Average
Single link	50.0	33.3	16.7	33.3
Complete link	00.0	22.2	46.2	22.8
Average link	50.0	25.0	30.8	35.3
Average	33.3	26.8	31.2	30.5

Table 6. DBI values for different number of clusters

Inter-cluster distance	DBI		
	k = 5	k = 10	k = 15
Complete link (Without GA)	2.11	1.81	1.74
Complete link (With GA)	1.18	1.64	0.58

Table 7. Percentage of Malay-English clusters alignment with weights adjustment using a genetic algorithm

Inter-cluster distance	Cluster alignment (%)			
	k = 5	k = 10	k = 15	Average
Single link	80.0	60.0	80.0	73.3
Complete link	100.0	80.0	86.7	88.9
Average link	100.0	80.0	93.3	91.1
Average	93.3	73.3	86.7	84.4

3.3. Fine-Tuning Weights of Terms using a Genetic Algorithm

Table 6 indicates that DBI values are improved (decreased) when the weights of all terms are optimized using the genetic algorithm. This results show that a better clusters structure is obtained by the clustering text documents using the fine-tuned tf-idf weights.

4. DISCUSSION

Based on **Table 1**, the percentage of aligned clusters, between the two sets of clusters is 80% when k = 5 and 15. However, when k = 10, there are more clusters that cannot be aligned in the clusters mapping. When using a Single link distance measure, two clusters are combined, when there are two points, one from each cluster, that have the smallest distance between them. It is suspected that the clusters produced may not be as compact as possible. As a result, the clusters produced may not be well separated among themselves. When a Complete link is used to cluster the text documents, the percentage of cluster alignment is 100% when k=5 and this percentage decreases as the number of clusters increases to k = 10 and k = 15. This is probably because when using the complete link distance measure for two different clusters in order to cluster text documents, two highly dense clusters are more likely to be combined because the distance between two clusters is measured based on two points from two different clusters that are separated the farthest. Thus, this causes a highly dense clusters produced when the final clusters are produced. When a more dense set of clusters is produced for both English and Malay, more clusters can be aligned as the clusters produced are more compact and related to each other. In contrast, the percentage of aligned clusters increases as the number of clusters increases from k = 5 and k = 10 to k=15, when using the Average link distance measure to cluster text documents. When using the Average link distance measure to cluster documents, the average distance between two different centers is considered in clustering these documents. The results obtained are not encouraging. This is probably due to the fact that Malay documents have a greater number of distinct terms due to the complex and rich morphology of Malay language. As the Malay language has more word forms to describe English phrases, this may affect the computation of weights for terms in finding centers of each cluster.

In short, as the number of clusters increases, there are more clusters that can be aligned in the clusters mapping. This is probably due to the fact that Malay documents have a greater number of distinct terms. As the Malay language has more word forms to express the same concepts as the English phrases, this may affect the computation of weights for the terms during the clustering process. Besides that, from **Table 1**, the Complete link has the highest value of percentage of aligned clusters which is 92.3%. This may be due to the fact that the Complete link is less susceptible to noise and outliers.

Compared to **Table 1 and 7** shows that when we applied the genetic algorithm to fine-tune the weights of terms considered in clustering bilingual corpus, the result actually shows an increase in the percentage of aligned clusters. However, the percentage of aligned clusters when k = 10 is lower when the GA is applied to fine-tune the weights of terms. This is probably because the tf-idf weights are overly adjusted and hence this results in a big change on the structure of the clusters formed during the clustering process and hence affects the clusters mapping.

5. CONCLUSION

This study has presented the idea of using hierarchical agglomerative clustering on a bilingual parallel corpus. The aim has been to illustrate this technique and provide mathematical measures, which can be utilized to quantify the similarity between the clusters in each language. The differences of all the clusters were compared, based on the terms extracted.

We can conclude that with a smaller number of clusters, k = 5, all of the clusters from English texts can be mapped into the clusters of Malay texts, by using the Complete link distance measure in clustering a bilingual parallel corpus. In contrast, with a larger number of clusters, fewer clusters from English texts can be mapped into the clusters of Malay texts.

To summarize, here we compared the results of clustering of documents in each of two languages with quite different morphological properties: English, which has a very modest range of inflections, as opposed to Malay with its wealth of verbal, adjectival and nominal word forms. The clusters produced and the top 10 most representative terms for each language and cluster listed. In the study, we also have clustered a bilingual Malay-English corpus based on a set of fine-tuned weights of terms using a GA in the clustering process. When we applied the genetic algorithm to fine-tune weights of terms considered in clustering bilingual corpus, the result actually showed an increase in the percentage of aligned clusters.

6. REFERENCES

Abdullah, M.T., F. Ahmad, R. Mahmod and T.M.T. Sembok, 2009. Rules frequency order stemmer for malay language Int. J. Comput. Sci. Netw. Security, 9: 433-438.

Alfred, R., 2009. A parallel hierarchical agglomerative clustering technique for billingual corpora based on reduced terms with automatic weight optimization. Proceedings of the 5th International Conference on Advanced Data Mining and Applications, Aug. 17-19, Springer Berlin Heidelberg, Beijing, China, pp: 19-30. DOI: 10.1007/978-3-642-03348-3_6

Davies, D.L. and D.W. Bouldin, 1979. A cluster separation measure. IEEE Trans. Patt. Anal. Mach. Intell., PAMI-1: 224-227. DOI: 10.1109/TPAMI.1979.4766909

Filho, J.L.R., P.C. Treleaven and C. Alippi, 1994. Genetic-algorithm programming environments. Computer, 27: 28-43. DOI: 10.1109/2.294850

Holland, J.H., 1992. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. 1st Edn., MIT Press, Cambridge, MA., ISBN-10: 0262581116, pp: 211.

Khalilian, M. and N. Mustapha, 2010. Data stream clustering: Challenges and issues. Proceedings of the International Multi Conference of Engineers and Computer Scientists, Mar. 17-19, Hong Kong, pp: 978-988.

Pantel, P. and D. Lin, 2002. Document clustering with committees. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 11-15, ACM Press, Tampere, Finland, pp: 199-206. DOI: 10.1145/564376.564412

Rijsbergen, C.J.V., 1979. Information Retrieval. 2nd Edn., Butterworths, London, ISBN-10: 0408709294, pp: 208.

Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM., 18: 613-620. DOI: 10.1145/361219.361220

Sembok, T.M.T. and Z.A. Bakar, 2011. Characteristics and retrieval effectiveness of n-gram string similarity matching on Malay documents. Proceedings of the 10th WSEAS International Conference on Applied Computer and Applied Computational Science, (ACACS' 11), ACM Press, Stevens Point, Wisconsin, USA., pp: 165-170.

Torres, G.J., R.B. Basnet, A.H. Sung, S. Mukkamala and B.M. Ribeiro, 2009. A similarity measure for clustering and its applications. Int. J. Elect. Comput. Syst. Eng., 3: 164-164.

Zhao, Y., G. Karypis and U. Fayyad, 2005. Hierarchical clustering algorithms for document datasets. Data Min. Knowl. Discov., 10: 141-168. DOI: 10.1007/s10618-005-0361-3