# A Genetic-Based HAC Technique for Parallel Clustering of Bilingual Malay-English Corpora

Ng Zhen Wei, Chan Chen Jie, Rayner Alfred
School of Engineering and Information Technology
Universiti Malaysia Sabah
Kota Kinabalu, Sabah, Malaysia
zhenwei_1988@hotmail.com, jie26488@hotmail.com,
ralfred@ums.edu.my

Joe Henry Obit
Labuan School of Informatics Science
Universiti Malaysia Sabah
Labuan, Malaysia
joehenryobit@yahoo.com

*Abstract*—**Multi Multilingual corpora, containing the same documents in a variety of languages, are becoming an essential resource for natural language processing. Clustering multilingual 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. For instance, in this way, the most relevant articles to be translated from language Malay to language English can be selected after studying the clusters of abstracts in language English. In this paper, we report on our work on applying Hierarchical Agglomerative Clustering (HAC) to a large corpus of documents where each appears both in Malay and English. We cluster these documents for each language and compare the results both with respect to the content of clusters produced. On the data available, the results of clustering one language resemble the other, provided the number of clusters required is relatively small. Further, we study the effects of changing the method used to compute the inter-clusters distance that includes single link, complete link and average link distance between clusters. Finally, we describe an experiment employing a genetic algorithm to fine-tune the individual term weights 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 language Malay when applied to the corresponding documents in language English. Other possible applications include training the algorithm on a hand-clustered set of documents, and subsequently applying it to a superset, including unseen documents, incorporating in this way expert knowledge about the domain in the clustering algorithm.**

***Keywords*—bilingual corpora; hierarchical agglomerative clustering; parallel clustering; genetic algorithm; Malay-English Corpora**

## I. INTRODUCTION

In particular, clustering algorithms that build illustrative and meaningful hierarchies 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 research in clustering text documents. However, there are few experiments that 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. Applying clustering algorithm to a set of documents based on a set of fine-tuned terms 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 a set of fine-tuned terms, can increase the quality of clustering results, since the weights of terms are fine-tuned according to a predefined fitness function implemented in the evolutionary algorithm.

The aim of the experiments presented in this paper is to investigate the effect of applying a clustering technique based on a set of fine-tuned terms, to parallel multilingual texts. 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 mappings and the most representative terms extracted for Malay-English clusters.

In this paper, we provide the results of clustering parallel corpora of Malay-English texts based on the set of fine-tuned terms. In addition, we also present the findings obtained in mapping the Malay-English clusters and also the most representative terms extracted for Malay-English clusters.

We will first explain some of the background about the vector space model representation of documents, the hierarchical agglomerative clustering method, genetic algorithm and the semi-supervised clustering technique. Next, we describe the experimental design set-up and the experimental results and conclude this paper with future works.

## II. Background

### A. *Vector Space Model Representation*

In this experiment, we use the vector space model [1], in which a document is represented as a vector in *n*-dimensional space (where *n* is the number of different words in the collection). Here, documents are categorized by the words they contain and their frequency. Before obtaining the weights for all the terms extracted from these documents, stemming and stopword removal is performed. Stopword removal eliminates unwanted terms (e.g., those from the closed vocabulary) and thus reduces the number of dimensions in the term-space. Once these two steps are completed, the frequency of each term across the corpus is counted and weighted using *term frequency – inverse document frequency* (tf-idf) [1], as described in equation (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 [2].

### B. *Hierarchical Agglomerative Clustering*

In this work, we concentrate on hierarchical agglomerative clustering. Unlike partitional clustering algorithms that build a hierarchical solution from top to bottom, repeatedly splitting existing clusters, agglomerative algorithms build 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 [3]. 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 one if the documents are identical, and zero if they share no words. The two clusters to merge at each step are found using either the single link, complete link or average link method. 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. 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 ‘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 language. Clustering in the other language is then carried out and *purity* [4], equation (5), used to compare each of the resulting clusters $C \in C_{ALL}$ to its closest match among all clusters $L_{ALL}$. (*Precision* is the probability of a document in cluster C being labeled L. Purity is the percentage of correctly clustered documents.)

### C. *Genetic Algorithm*

A Genetic Algorithm (GA) is a computational abstraction of biological evolution that can be used to some optimization problems [5]. 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 [6].

In this paper, we examine the clustering algorithm that minimizes some objective function applied to *k*-cluster centers. In our case, we consider the *cluster dispersion*.

$$\text{tf-idf} = \text{tf(t,d)} \cdot \text{idf(t)} \tag{1}$$

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

$$\text{sim}(\text{d}_i, \text{d}_j) = \frac{\left(\text{d}_i \text{d}_j\right)}{\left(\|\text{d}_i\| \cdot \|\text{dj}\|\right)} \tag{3}$$

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

$$\text{Purity} = \sum_{C \in C_{ALL}} \frac{|C|}{|D|} \cdot P(C, L) \tag{5}$$

$$\text{Precision (EMM)} = \frac{C(E) \cap C(M)}{C(E)} \tag{6}$$

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

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 dispersion metric* is used, such as the Davies-Bouldin Index (DBI) [7]. 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 $k \leq K$ 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 cluster the text documents based on the minimum (single link) and maximum (complete link) distances between clusters.

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 to evaluate a particular segmentation of data.

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

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

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

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

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

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 is produced. More specifically, given $N$ points and $K$-clusters, select the weight of each terms in that minimize the objective function defined in (12).

## III. Experimental Design

There are two main stages in this experiment. (I). In the first stage, we perform the task of clustering parallel corpora of Malay-English texts. We look at the similarities and differences of two main areas: Malay-English cluster mappings and the extracted most representative terms for English-Malay clusters. (II). Next, in the second stage, we apply the genetic algorithm to optimize the weights of terms considered in clustering parallel corpora of Malay-English texts.

### A. *Clustering Parallel Corpora*

In the first stage of the experiment, there are two set of parallel corpora in two different languages, Malay and English. 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, where the impact of stemming is potentially greater, the process of building an accurate algorithm becomes a more challenging task [8,9]. In this experiment, the Malay texts are stemmed by using the Rules Frequency Order (RFO) stemmer [9]. Figure 1 illustrates the experimental design set up for the first stage of the experiment. The 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 the cluster and the cluster members. The next section contains a detailed comparison of the results for the two languages looking at each of these elements.

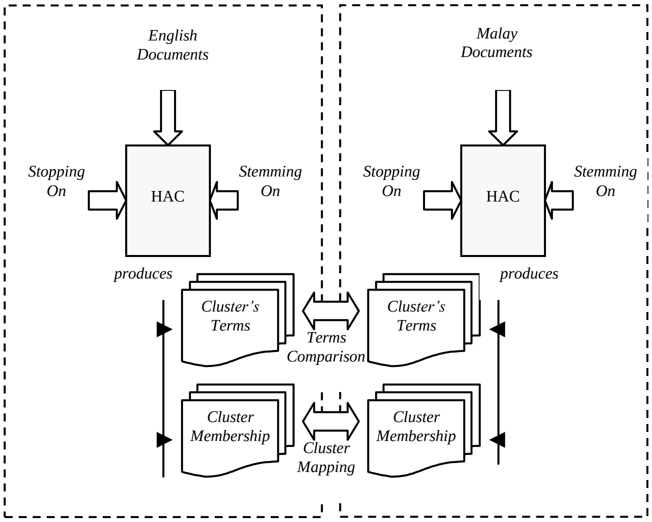

Figure 1. Experimental set up for parallel clustering task

### B. *Clustering Parallel Corpora Based on a Set of Fine-Tuned Weights of Terms using a Genetic Algorithm (GA)*

The second stage of the experiment is clustering the documents based on genetic algorithm that optimizes the weight of the terms to best cluster the documents according to the fitness function of the GA. Here, we describe the representation of the problem in the Genetic Algorithm setting.

Population Initialization Step: A population of *X* strings of length *m* is randomly generated, where *m* is the number of terms (e.g. cardinality of terms). *X* strings are generated with continuous numbers (0.5, 1.0 and 1.5) representing the weight of terms.

Fitness Computation: The computation of the objective or fitness function is based on the Cluster Dispersion. In order to get cluster of better quality, we need to minimize the DBI, defined in (11). Since in GA, we want to maximize the objective and fitness function, the objective fitness function (OFF) that we want to maximize will be as follows (13).

$$\begin{aligned} OFF &= 1/\text{Cluster Dispersion} \\ OFF &= 1/DBI \end{aligned} \tag{13}$$

Selection Process: For the selection process, a rouleete wheel with slots sized according to the fitness is used. The construction of such a roulette wheel is as follows;

- Calculate the fitness value for each chromosome, $f_i$ and $i \leq X$, and get the total overall fitness for $X$ strings of chromosome, $T_{Fitness}$.
- Calculate the probability of a selection $p_i$ for each chromosome, $i \leq X$, $p_i = f_i/T_{Fitness}$.
- Calculate the cumulative probability $q_i$ for each chromosome, $q_i = \sum_{j=1}^{i} p_j$.

The selection process is based on spinning the roulette wheel, $X$ times; each time we select a single chromosome for a new population in the following way:

- Generate a random number $r$ from the range of [0..1].
- Select the $i$-th chromosome such that $q_{i-1} < r \leq q_i$

Crossover: 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 $p_c \cdot X$ of chromosomes, which undergo the crossover operation. We proceed by

- Generating a random number of $r$ from the range [0..1].
- Performing 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.

Mutation: 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

- Generate a random number of $r$ from the range [0..1].
- Do mutation of each bit if $r < p_m$.

Following selection, crossover and mutation, the new population is ready for its next generation. This evaluation is used to build the probability distribution for a construction of a roulette wheel with slots sized according to current fitness values. The rest of the evolution is just a cyclic repetition of selection, crossover and mutation until a number of specified generations or specific threshold has been achieved.

## IV. Experimental Results

### A. Mapping of Malay-English Clusters Alignment

In a 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, Politic, 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).

Figures 2–10 show the precisions for the EMM and MEM for the cluster pairings obtained with varying numbers of clusters, k (k = 5, 10, 15), 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 ID 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, in Figure 2, 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%.

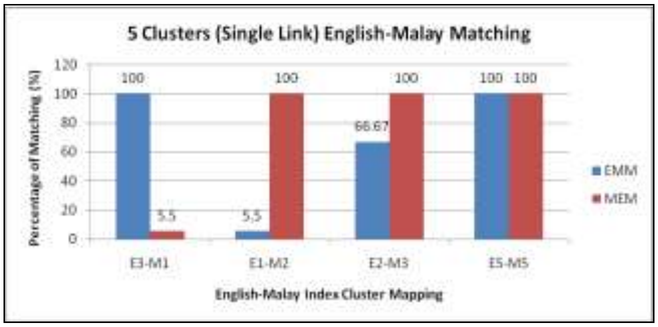

Figure 2. Cluster mapping results for Single link with 5 clusters

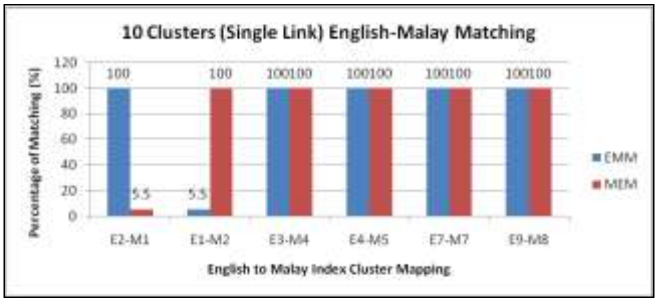

Figure 3. Cluster mapping results for Single link with 10 clusters

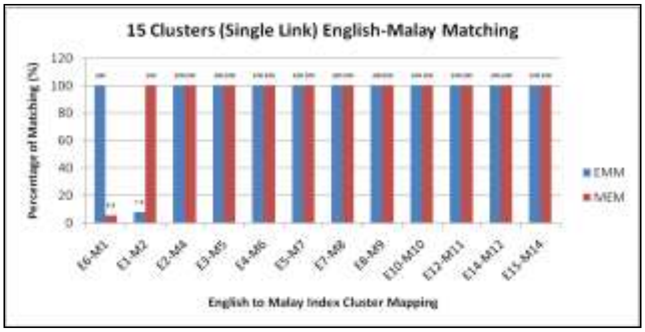

Figure 4. Cluster mapping results for Single link with 15 clusters

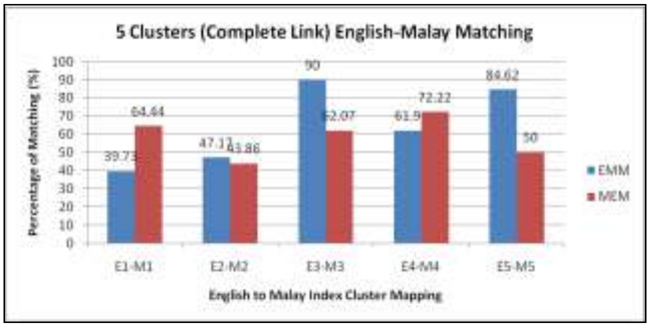

Figure 5. Cluster mapping results for Complete link with 5 clusters

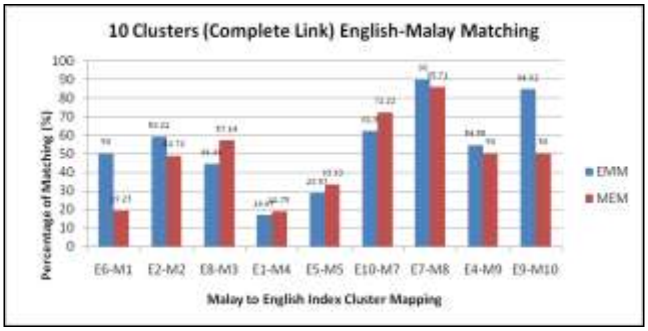

Figure 6. Cluster mapping results for Complete link with 10 clusters

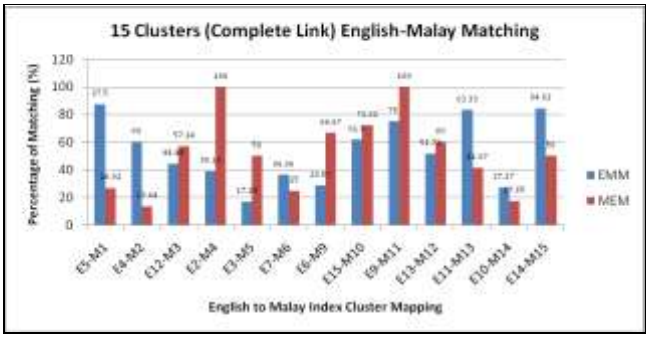

Figure 7. Cluster mapping results for Complete link with 15 clusters

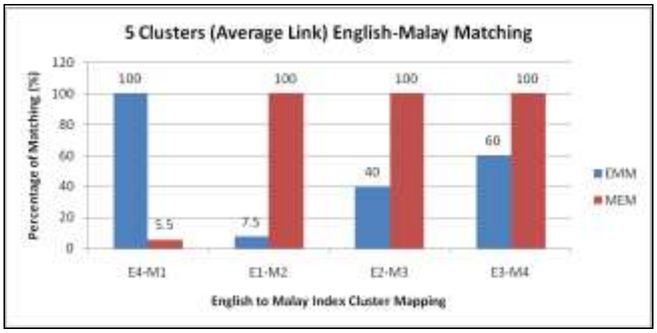

Figure 8. Cluster mapping results for Average link with 5 clusters

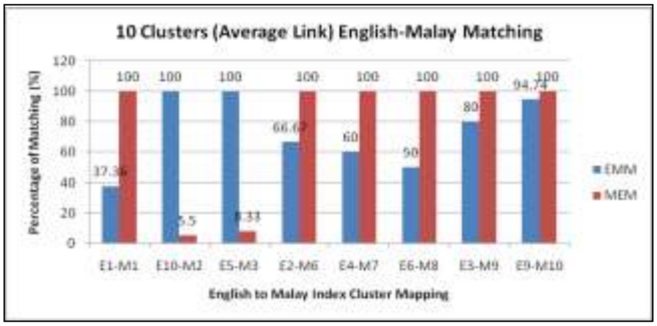

Figure 9. Cluster mapping results for Average link with 10 clusters

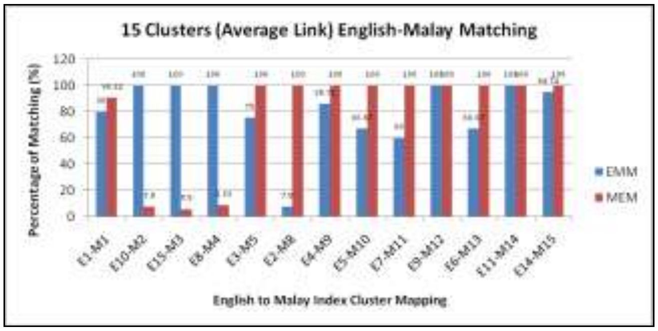

Figure 10. Cluster mapping results for Average link with 15 clusters

TABLE I. PERCENTAGE OF MALAY-ENGLISH CLUSTERS ALIGNMENT

| Inter-Cluster Distance | Percentage of 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 |

Another point of interest is the extent to which the mapping EMM matches MEM. Table 1 shows the alignment percentage for Single link between the two sets of clusters is 80% when k = 5 and 15. However, when k = 10, there are more clusters that are unaligned between the mappings. 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. As a result, the clusters produced may not be as compact as possible. As a result, the clusters produced may not be well separated among themselves. As a result, a less number of clusters can be aligned between the Malay and English clusters since the clusters formed can be based on different centres.

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 a Complete link 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 cluster alignment 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 the text documents. When using the Average link distance measure to cluster documents, the smallest distance between two different centers is considered in clustering documents. The results obtained are not encouraging. 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 describe English phrases, this may affect the computation of weights for the terms in finding centers of each cluster during the clustering process.

### B. *Comparison of Terms Extracted*

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

TABLE II. TERMS EXTRACTED FOR SINGLE LINK BASED CLUSTERING WHERE K=5

| Mapping | English Cluster | Malay Cluster |
|---|---|---|
| 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 |
| E2-M3 | Seedstock, camelm prawn, farm, freshwat, haiezack, pond, breeder, breed, aquacultur | Udang, galah, benih, kolam, projek, haiezack, akuakultur, ternak, simen, buah |
| E5-M5 | Honei, tualang, fama, research, usm, process, sugar, nerang, benefit, health | Madu, tualang, fama, usm, yelidik, manfaat, gula, umpul, universit, hadam |

TABLE III. TERMS EXTRACTED FOR SINGLE LINK BASED CLUSTERING WHERE K=10

| Mapping | English Cluster | Malay Cluster |
|---|---|---|
| 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 |
| E3-M4 | Camel, desert, saud, milk, farm, Saudi, imam, falih, male, bread | Unta, susu, saud, haiwan, baka, imam, falih, lading, Saudi, pasir |
| E4-M5 | Seedstock, prawn, freshwat, haiezack, farm, pond, aquaculture, breeder, tank, breed | Udang, galah, benih, kolam, projek, haiezack, akuakultur, ternak, simen, buah |
| E7-M7 | Chef, pastry, cake, recip, academi, lanz, mazlan, aseri, law, brother | Chef, pastri, kek, akadem, lanz, mazlan, recipe, aseri, biras, resepi |
| E9-M8 | Honei, tualang, fama, research, usm, process, sugar, nerang, benefit, health | Madu, tualang, fama, usm, yelidik, manfaat, gula, umpul, university, hadam |

TABLE IV. TERMS EXTRACTED FOR SINGLE LINK BASED CLUSTERING WHERE K=15

| Mapping | English Cluster | Malay Cluster |
|---|---|---|
| E6-M1 | Paint, voc, chemic, soo, eco, odour, hazard, low, fume, opac | Parti, bank, umno, bn, anggota, pkr, joh, negeri, sukan, wang |
| E1-M2 | Paint, voc, chemic, soo, eco, odour, hazard, low, fume, opac | Hartanah, sunway, atus, templer, iproperty, ieli, country, com, janj, ekar |
| E2-M4 | Eti, solar, tech, batteri, mou, lithium, sirim, technolog, green, system | Eti, tech, solar, bateri, amam, mou, litium, sirim, etera, teknolog |
| E3-M5 | Prudenti, bumiputera, oropeza, market, agent, penetr, popul, grow, assur, tap | Prudential, bumiputera, oropeza, agen, pasar, assurance, adu, populas, embus, luas |
| E4-M6 | Camel, desert, saud, milk, farm, Saudi, imam, falih, male, bread | Unta, susu, saud, haiwan, baka, imam, falih, lading, Saudi, pasir |
| E5-M7 | Seedstock, prawn, freshwat, haiezack, farm, pond, aquaculture, breeder, tank, breed | Udang, galah, benih, kolam, projek, haiezack, akuakultur, ternak, simen, buah |
| E7-M8 | Banana, leaf, rice, meal, thaipusam, eat, celebr, cultur, Indian, vegetarian | Daun, pisang, nasi, thaipusam, makan, hiding, aya, lapik, ragesh, budaya |
| E8-M9 | Grave, skelet, rebury, graveyard, kin, reloc, remain, pusara, burial, tier | Kubur, pusara, liang, pindah, bumibumi, waris, tapak, indah, arwah, tanah |
| E10-M10 | Chef, pastri, cake, recip, academi, lanz, mazlan, aseri, law, brother | Chef, pastri, kek, akadem, lanz, mazlan, recipe, aseri, biras, resepi |
| E12-M11 | Honei, tualang, fama, research, usm, process, sugar, nerang, benefit, health | Madu, tualang, fama, usm, yelidik, manfaat, gula, umpul, universit, hadam |
| E14-M12 | Norouz, Persian, festiv, Iranian, celebr, neguin, iran, haft, spring, cultur | Norouz, iran, parsi, festival, sambut, haft, seen, neguin, ra, bunga |
| E15-M14 | Embassy, thaipusam, passport, notary, resum, visa, close, festiv, statement, jalan | Duta, thaipusam, tutup, jabat, jumaat, tiausahaha, amerika, mpena, visa |

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 II, III and IV, where all top English terms are less related to the Malay terms extracted when clustering using the Single link.

TABLE V. TERMS EXTRACTED FOR COMPLETE LINK BASED CLUSTERING WHERE K=5

| Mapping | English Cluster | Malay Cluster |
|---|---|---|
| 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 |

TABLE VI. TERMS EXTRACTED FOR COMPLETE LINK BASED CLUSTERING WHERE K=10

| Mapping | English Cluster | Malay Cluster |
|---|---|---|
| E6-M1 | Tourism, park, penang, hot, spring, tawau, dengu, tourist, venu, seberang | Ancong, tronas, taman, ng, wang, kawas, miri, unjung, panas, najib |
| E2-M2 | Bank, cent, honei, hsbc, property, custom, eon, internet, syndrome, debit | Bank, atus, madu, getah, jualan, udang, hsbc, galah, eon, benih |
| E8-M3 | Umno, ghazali, razaleigh, tengku, royalty, voter, wanita, divis, gua, musang | Umno, Labuan, daftar, razaleigh, royalt, tengku, gua, musang, und, lant |
| E1-M4 | Sailor, embassy, banana, leaf, thaipusam, petrona, film, innov, rice, lubric | Ayar, down, anak, sindrom, unta, india, perahu, denggi, pakist, latih |
| E5-M5 | Seedstock, camel, chef, prawn, antique, pastry, jefri, cake, academi, ng | Abuh, pos, westports, araf, chef, pam, kedai, pastry, poh, teu |
| E10-M7 | Race, minut, win, goal, chong, team, cup, wei, titl, singl | Pemain, beregu, law, wei, joh, filem, jarring, minit, chong, buka |
| E7-M8 | Pkr, parti, mca, bn, ong, presid, elect, tm Pbb | Parti, pkr, bn, mca, anggota, anwar, presiden, parlimen, pbb, jawat |
| E4-M9 | Polic, rubber, risda, replant, smallhold, simunjan, Nigerian, gunasegaran, arrest, traffick | Polis, Nigeria, hidung, kes, rhinitis, yeludup, singapura, lapor, agama, masjid |
| E9-M10 | Athlet, boxer, gold, medal, category, ronoh, fuad, swim, Terengganu, ironman | Tm, tinju, sukan, lumba, inju, kategor, pingat, engganu, sukma, atlet |

TABLE VII. TERMS EXTRACTED FOR COMPLETE LINK BASED CLUSTERING WHERE K=15

| Mapping | English Cluster | Malay Cluster |
|---|---|---|
| E5-M1 | Miri, najib, visit, project, plaza, facebook, mainten, muhyiddin, contractor, minist | 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 |
| E12-M3 | mno, ghazali, razaleigh, tengku, royalty, voter, wanita, divis, gua, jmusang | Umno, Labuan, daftar, razaleigh, royalt, tengku, gua, musang, und, lant |
| E2-M4 | Bank, cent, hsbc, property, eon, internet, custom, debit, loan, equiti | Bank, hsbc, eon, debit, public, pinjam, sme, kad, biaya, jualan |
| 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 |
| E15-M10 | Race, minut, win, goal, chong, team, cup, wei, titl, singl | Pemain, beregu, law, wei, joh, filem, jarring, minit, chong, buka |
| E9-M11 | Embassy, banana, leaf, thaipusam, rice, celebr, meal, festiv, norouz, passport | Daun, pisang, duta, thaipusam, nasi, makan, hiding, aya, tutup, lapik |
| E13-M12 | Pkr, mca, parti, ong, bn, presid, mp, anwar, deleg, rakyat | Parti, bn, mca, presiden, ong, pbb, puak, sapp, ka, calon |
| E11-M13 | Polic, simunjan, Nigerian, gunasegaran, arrest, traffick, aircraft, crash, mof, singapor | Polis, Nigeria, hidung, kes, rhinitis, yeludup, singapura, lapor, agama, masjid |
| E10-M14 | Tm, pbb, taekwondo, Sarawak, spdp, secretary, elect, bn, parti, baling | Pkr, anwar, fairus, parti, parlimen, mohammad, anggota, bangkang, long, rakyat |
| E14-M15 | Athlet, boxer, gold, medal, category, ronoh, fuad, swim, Terengganu, ironman | Tm, tinju, sukan, lumba, inju, kategor, pingat, engganu, sukma, atlet |

Table V, VI and VII show 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 VIII, IX and X show 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 (Malay and English)).

TABLE VIII. TERMS EXTRACTED FOR AVERAGE LINK BASED CLUSTERING WHERE K=5

| Mapping | English Cluster | Malay Cluster |
|---|---|---|
| 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 |
| E2-M3 | Seedstock, camel, chef, prawn, antique, pastri, jefri, cake, farm, recip | Udang, galah, benih, kolam, projek, haiezack, akuakultur, ternak, simen, buah |
| E3-M4 | Embassy, banana, leaf, thaipusam, rice, celebr, meal, festiv, norouz, passport | Daun, pisang, duta, thaipusam, nasi, makan, hiding, aya, tutup, lapik |

TABLE IX. TERMS EXTRACTED FOR AVERAGE LINK BASED CLUSTERING WHERE K=10

| Mapping | English Cluster | Malay Cluster |
|---|---|---|
| E1-M1 | Bank, cent, tm, develop, finance, service, sale, million, custom, rubber | Bank, atus, jualan, wang, hsbc, eon. Labuan, hartanah, debit, pinjam |
| 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 |
| E2-M6 | Seedstock, camel, prawn, farm, freshwat, haiezack, pond, breeder, breed, aquacultur | Udang, galah, benih, kolam, projek, haiezack, akuakultur, ternak, simen, buah |

| | | |
|---|---|---|
| E4-M7 | Embassy, banana, leaf, thaipusam, rice, celebr, meal, festiv, norouz, passport | Daun, pisang, duta, thaipusam, nasi, makan, hiding, aya, tutup, lapik |
| E6-M8 | Chef, antique, pastri, jefri, cake, recip, academi, lanz, shop, kri | Chef, pastri, kek, akadem, lanz, mazlan, recipe, aseri, biras, resepi |
| E3-M9 | Honei, paint, tualang, nose, fama, rhinitis, voc, health, allergen, allergi | Madu, tualang, hidung, rhinitis, fama, gatal, allergen, usm, alergi, anti |
| E9-M10 | Athlet, category, race, team, boxer, sailor, gold, win, minut, medal | Joh, pemain, minit, beregu, ayar, law, tinju, juara, kategor, acara |

TABLE X. TERMS EXTRACTED FOR AVERAGE LINK BASED CLUSTERING WHERE K=15

| Mapping | English Cluster | Malay Cluster |
|---|---|---|
| E1-M1 | Bank, cent, finance, sale, rubber, hsbc, property, eon, risda, custom | Bank, atus, jualan, hsbc, wang, eon, Labuan, hartanah, debit, pinjam |
| 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 |
| E3-M5 | Penang, port, dengu, Westport, seberang, perai, rapid, rout, marc, teu | Abuh, westports, araf, pinang, denggi, teu, pulau, marc, kontena, minal |
| E2-M8 | Syndrome, develop, asli, innov, citi, tourism, devic, orang, najib, park | Eti, tech, solar, bateri, amam, mou, litium, sirim, etera, teknolog |
| E4-M9 | Tm, internet, taekwondo, cybersecur, Microsoft, club, update, regist, cyber, england | Tm, internet, taekwondo, Microsoft, explorer, kemas, daftar, England, negeri, telekom |
| E5-M10 | Seedstock, camel, prawn, farm, freshwat, haiezack, pond, breeder, breed, aquacultur | Udang, galah, benih, kolam, projek, haiezack, akuakultur, ternak, simen, buah |
| E7-M11 | Embassy, banana, leaf, thaipusam, rice, celebr, meal, festiv, norouz, passport | Daun, pisang, duta, thaipusam, nasi, makan, hiding, aya, tutup, lapik |
| E9-M12 | Chef, pastri, cake, recip, academi, lanz, mazlan, aseri, law, brother | Chef, pastri, kek, akadem, lanz, mazlan, recipe, aseri, biras, resepi |
| E6-M13 | Paint, nose, rhinitis, voc, allergen, allergi, allerg, air, remedy, irrit | Hidung, rhinitis, gatal, allergen, alergi, ong, berair, tekak, iritas, hembus |
| E11-M14 | Honei, tualang, fama, research, usm, process, sugar, nerang, benefit, health | Madu, tualang, fama, usm, yelidik, manfaat, gula, umpul, universit, hadam |
| E14-M15 | Athlet, category, race, team, boxer, sailor, gold, win, minut, medal | Joh, pemain, minit, beregu, ayar, law, tinju, juara, kategor, acara |

TABLE XI. 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 | 0.0 | 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 XI 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 using the Complete link distance measure in order to cluster bilingual documents in Malay and English. However, mapping Malay-English clusters, with k=10, will produce better results on average as shown in Table XI.

## V. CONCLUSION

This paper 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 paper, we also have clustered a bilingual English-Malay corpus based on a set of fine-tuned weights of terms using GA considered in the clustering process. When we applied the genetic algorithm to tune weights of terms considered in clustering bilingual corpus, the result actually showed an increase in the percentage of clusters aligned.

## REFERENCES

[1] Salton, G., and Michael, J. 1986. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill Inc., New York, NY.

[2] van Rijsbergen, C.J. 1979. Information Retrieval, Second edition, London: Butterworths.

[3] Zhao, Y., and Karypis, G., 2005, Hierarchical clustering algorithms for document datasets, Data Mining and Knowledge Discovery, 10(2):141.168.

[4] Pantel, P., and Lin, D., 2002, Document clustering with committees, In Proceeding Of SIGIR'02, Tampere, Finland.

[5] Holland, J., 1975, Adaption in Natural and Artificial Systems, Univeristy of Michigan Press.

[6] Filho, J.L.R., Treleaven, P.C., and Alippi, C. 1994, Genetic algorithm programming environments, IEEE Compu. 27: 28-43.

[7] Davies, D.L., and Bouldin, D.W., 1979, A cluster separation measure, IEEE Transactions and Pattern Analysis and Machine Intelligence, 1(2):224-227.

[8] T. M. T. Sembok and Z. A. Bakar, Effectiveness of Stemming and n-grams String Similarity Matching on Malay Documents, International Journal of Applied Mathematics and Informatics, Issues 3, Volume 5, 2011.

[9] M. T. Abdullah, F. Ahmad, R.Mahmod and M. T. Sembok, 2009, Rules Frequency Order Stemmer for Malay Language, International Journal of Computer Science and Network Security, VOL 9 No. 2.