Clustering Short Text and Its Evaluation

Prajol Shrestha, Christine Jacquin, and Béatrice Daille

Laboratore d'Informatique de Nantes-Atlantique (LINA),
Université de Nantes, 44322 Nantes Cedex 3, France
{Prajol.Shrestha,Christine.Jacquin,Beatrice.Daille}@univ-nantes.fr

Abstract. Recently there has been an increase in interest towards clustering short text because it could be used in many NLP applications. According to the application, a variety of short text could be defined mainly in terms of their length (e.g. sentence, paragraphs) and type (e.g. scientific papers, newspapers). Finding a clustering method that is able to cluster short text in general is difficult. In this paper, we cluster 4 different corpora with different types of text with varying length and evaluate them against the gold standard. Based on these clustering experiments, we show how different similarity measures, clustering algorithms, and cluster evaluation methods effect the resulting clusters. We discuss four existing corpus based similarity methods, Cosine similarity, Latent Semantic Analysis, Short text Vector Space Model, and Kullback-Leibler distance, four well known clustering methods, Complete Link, Single Link, Average Link hierarchical clustering and Spectral clustering, and three evaluation methods, clustering F-measure, adjusted Rand Index, and V. Our experiments show that corpus based similarity measures do not significantly affect the clusters and that the performance of spectral clustering is better than hierarchical clustering. We also show that the values given by the evaluation methods do not always represent the usability of the clusters.

1 Introduction

Clustering short text is an emerging field of research and is useful in many NLP tasks such as summarization, information extraction/retrieval, and text categorization. In general, clustering consists of two main parts, the first part is to find a score for similarity between short text and then cluster them according to these similarity scores. Short text pose a challenge while clustering because they have few words which is used to determine the similarity between short text in contrast to text documents. Existing methods use a portion of the terms empirically from a frequency list and find similarity between short text based on these terms using different text similarity methods [1] [2]. The clusters are then evaluated using mapping based measures (e.g. Purity, clustering F-measure) which have drawbacks. One of them is that these methods may not be able to evaluate the entire membership of a cluster and do not evaluate every cluster [3]. Due to this drawback of the mapping based measures, the usefulness of the existing short text clustering methods cannot be judged [4].

A. Gelbukh (Ed.): CICLing 2012, Part II, LNCS 7182, pp. 169–180, 2012.

Here in this paper, we use four short text corpus, three created from abstracts of scientific papers and the other created from newspaper paragraphs, described in Sect. 3.1, to give an idea of the different variation present in short text and how different clustering methods behave on them. We use Single link (SHC), Complete link (CHC), and Average link (AHC) hierarchical clustering methods. Along these methods, we use spectral clustering (SPEC) which has not been used in the scope of short text. This clustering method has been very successful in the field of machine learning such as image segmentation [5]. Clustering methods depend on similarity values and to see its effect we use four existing similarity measures. The clusters are then evaluated using clustering F-measure (F), adjusted Rand Index (ARI) and V. We demonstrate that none of these measures relate to the usability aspect of the clusters so they are not always able to properly evaluate the quality of clusters. We start by describing the clustering methods.

2 Clustering Methods

Clustering short text is the task of grouping short text together into groups in such a way that short text related to a category are found in a unique group. It consists of two steps: the first step is to find the similarity or dissimilarity matrix and then clustering the short text with the help of this matrix. In this paper we consider dissimilarity between two text to be one minus the similarity between them. We used four different corpus based similarity methods to create the matrix namely cosine similarity (CS) measure using tf-idf weights, Latent Semantic Analysis (LSA) using log(tf)-idf weights [6], Short text Vector Space Model (SVSM) [7], and Kullback-leibler distance (KLD) [8]. These measures are used by each of the clustering methods. In this section, we give a brief description of all the similarity and clustering methods.

2.1 Short Text Similarity Methods

Cosine Similarity Measure (CS) : This measure has been extensively used in NLP to find similarities between text where the text is represented as a weighted vector [9]. Here we use $tf*idf$ weights where tf and idf stands for term frequency and inverse document frequency respectively. For us documents are short text. Given two short text $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$, their cosine similarity is computed with (1).

$$CS(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|} \tag{1}$$

where, $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$ are m-dimensional vectors of short text a and b over the term set of $T = \{t_1, t_2, ...t_m\}$. Each vector dimension represents a term with its weight corresponding to the short text, which is a non-negative value. As a result, the cosine similarity is non-negative and bounded between [0,1] where 1 indicates the two text are identical.

Latent Semantic Analysis (LSA) : LSA is a method which is used to find similarity between text using singular value decomposition (SVD), which is a form of factor analysis and is well-known in linear algebra [10]. SVD decomposes the rectangular term-by-short text matrix, $\mathbf{M}$, into three other matrices $\mathbf{M} = \mathbf{U}\Sigma_k\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are column-orthogonal matrices and Σ_k is a diagonal $k \times k$ matrix which contains k singular values of $\mathbf{M}$ such that the singular values are in the descending order, $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k$. We choose $k' \ll k$ and multiply the three matrices to get $\mathbf{M} \simeq \mathbf{U}\Sigma_{k'}\mathbf{V}^T$ which is a re-composed matrix of the original matrix $\mathbf{M}$. The similarity between short text is then computed using the cosine similarity measure between the columns of the new matrix $\mathbf{M}$.

Kullback-Leibler Distance (KLD) : KLD is used in [8] to cluster narrow domain abstracts and is based on Kullback-Leibler (KL) divergence which is used to give a value to the difference between two distributions. For two distributions P and Q the KL divergence on a finite set X is shown in (2).

$$D_{KL}(P\|K) = \sum_{x \in X} P(x) log \frac{P(x)}{Q(x)} \tag{2}$$

This measure is not symmetric but there exists symmetric versions. In [8] they have used some and shown that there is not much difference between them. We implemented the max of the KL distance as in (3).

$$D_{KLD} = max(D_{KL}(P\|K), D_{KL}(K\|P)) \tag{3}$$

To use D_{KLD} as a distance measure for short text we compute the probabilities as shown in (4) and they are based on the distribution of the terms in the vocabulary, V.

$$P(t_k, d_i) = \begin{cases} \beta * P(t_k|d_i), & \text{if term } t_k \text{ occurs in the document } d_i \\ \epsilon, & \text{otherwise} \end{cases} \tag{4}$$

where,

$$P(t_k|d_i) = \frac{tf(t_k, d_j)}{\sum\limits_{t_k \in d_i} tf(t_k, d_i)}$$

and

$$\beta = 1 - \sum_{t_k \in V, t_k \notin d_i} \epsilon \quad \text{such that,} \quad \sum_{t_k \in d_i} \beta * P(t_k|d_i) + \sum_{t_k \in V, t_k \notin d_i} \epsilon = 1$$

In [11] [8], KLD relies only on terms of a certain portion of the vocabulary. This selection of the terms were done using three methods and among them we choose *Document Frequency* (DF) technique because no parameter has to be estimated and it gives a stable result. The concept of this term selection is that, the lower frequency terms in the collection of text do not play a role in predicting the class for the text. In order to implement KLD we select the top 70% of the vocabulary which was sorted in a descending order according to the term frequency.

Short Text Vector Space Model (SVSM) : This method is used in [7] to find similarity between short text. For each text, a text vector is created from term vectors. Given a corpus C of n short text and m unique terms, the term vector, $\overrightarrow{t_j}$, for term t_j is a vector created with n number of possible dimensions where each dimension represents a unique short text. The presence of the term in a short text is indicated by its sentence id and the term's inverse document frequency, idf, here a document is a short text, as shown below:

$$\overrightarrow{t_j} = [(S_1, idf_j), (S_5, idf_j), ..., (S_i, idf_j)]$$

where S_i is the short text id where t_j is present, $i \in 1, .., n$ and idf_j is the idf value of term t_j. This term vector is a reduced vector space representation where short text that do not contain the term is absent which saves space. The dimension of the matrix formed by term vectors can be further reduced using LSA [12] or Principle Component Analysis [13] but are not used here. Once we have the term vectors we can create a short text vector by adding the term vectors of the terms present in that short text. For a short text consisting of terms $t_1, t_2, .., t_k$, the dimension, d_i, of the sentence vector corresponding to the short text S_i, will be $d_i = \Sigma^k_{j=1;t_j \in S_i} idf_j$, where idf_j is the idf value of the term j and $i \in 1, .., n$. The similarity between short text is calculated using the cosine similarity between the text vectors.

2.2 Clustering Algorithms

The hierarchical agglomerative clustering (HC) and SPEC clustering methods are described in this section. HC are bottom up algorithms in which elements are merged together to form dendrograms and are used extensively in the field of NLP. Different HC algorithms are present but have the same underlying approach and can be formally written as these steps:

1. Compute the dissimilarity matrix with one of the approach given in Sect. 2.1
2. Start with each short text in one cluster and repeat the following steps until a single cluster is formed :
   (a) Merge the closest two clusters.
   (b) Update the dissimilarity matrix to reflect the dissimilarities between the new cluster and the original clusters.
3. Cut the dendrogram in a way we find the required number of clusters.

The three hierarchical clustering, SHC, CHC and AHC used here differ in step 2a where the closest clusters are determined. Below, we state how the closeness are determined for each algorithm.

Single Link HC (SHC) : This clustering method considers two clusters to be close in terms of the minimum dissimilarities between any two elements in the two clusters.

Complete Link HC (CHC) : This clustering method considers two clusters to be close in terms of the maximum dissimilarities between any two elements in the two clusters.

Average Link HC (CHC) : This clustering methods considers two clusters to be close in terms of the average pairwise dissimilarities of all the pairs of elements in the two clusters.

Spectral Clustering (SPEC) : Along with the HC algorithms we also use Spectral Clustering which has been recently used in the community of machine learning [5]. K-means clustering algorithm is the underlying clustering algorithm of SPEC which is applied on the normalized eigenvectors of the similarity matrix. The algorithm for spectral clustering is given below from [14] :

1. Given a set of short text, $S = \{s_1, ..., s_n\}$, the similarity matrix, $M \in \mathbb{R}^{n \times n}$, is generated using some similarity measures mentioned in Sect. 2.1.
2. Create the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by the Gaussian Similarity function, $A_{ij} = exp(-\|r_i - r_j\|^2/2\sigma^2)$ with $\sigma = 0.5$, if $i \neq j$, and $A_{ii} = 0$, where $r_i, ..., r_j$ are rows of M.
3. Construct the normalized graph Laplacian matrix $L = D^{-1/2}AD^{-1/2}$ where, D is a diagonal matrix whose (i, i)-element is the sum of A's i-th row.
4. Compute the eigenvectors of L and select the k largest eigenvectors and stacking them in columns to form $X = [x_1, x_2, ..., x_k] \in \mathbb{R}^{n \times k}$.
5. Normalize the row's of X to have unit length to form the matrix Y (i.e. $Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2}$).
6. Using K-means, cluster the rows of matrix Y into k clusters by treating the row of Y as points in $\mathbb{R}^k$.

3 Experiments and Results

In this section, we analyse the behaviour of the clustering methods, the effect of similarity measures on clustering methods, and the evaluation methods. We start by describing the corpus and the evaluation methods so that we can explain the results of the experiments.

3.1 Corpus

We use 4 different types of corpora with regards to the size, type, and the distribution of short text among the clusters. These corpora consist of paragraphs of text from newspapers as well as narrow domain abstracts which make them representative of short text that are normally dealt in the field of written NLP. We use three corpora namely CICLing-2002, hep-ex, and KnCr, created from scientific abstracts, which have been used previously for short text clustering [8] and will serve as a reference corpus. We also use a new short text corpus collected from newspapers. Here, we give a short description of each of these corpora.

The LDC Corpus : This corpus is a collection of 12 newspaper articles concerning the *Death of Diana*. The articles were taken from the Linguistic Data Consortium's (LDC) North American News Text Corpus[1]. We consider

[1] LDC Catalog number: LDC95T21

each paragraph a short text and each paragraph was manually annotated[2] with one of the 13 categories it is related to. The annotations were done by two annotators independently and the reliability of agreement on the annotation of these categories according to Fleiss' kappa [15] is 0.91. Which is an almost perfect agreement. The small error that arose was due to the fact that some paragraphs could be related to more than one categories but were assigned to one category. The disagreements were resolved between the annotators by discussing the main idea of the paragraph. Table 1 gives the distribution of the paragraphs according to the categories and some other properties of the corpus.

Table 1. Properties of the LDC corpus

(a) Number of paragraphs in each category

Categories	Paragraphs
Diana's life before accident	22
Driver's life before accident	5
Other's life before accident	9
Just before accident	18
Accident	10
Just after accident	22
Accident aftermath	8
Expression of grief	32
Funeral	46
Accusations	13
Cause	17
Investigation	20
Media	20

(b) Features

Feature	Value
Number of categories	13
Number of paragraphs	242
Total number of terms	5,351
Vocabulary size (terms)	1,761
Term average per paragraph	22.45

The CICLing-2002 Corpus : This is a small corpus consisting of 48 abstracts in the domain of computational linguistics collected from the CICLing 2002 conference. This corpus has 4 classes of 48 abstracts and the abstracts are evenly distributed among the 4 classes which is as follows : {11, 15, 11, 11}.

The hep-ex Corpus of CERN : This corpus contains 2,922 abstracts collected by the University of Jaén, Spain on the domain of Physics from the data server of the CERN. These abstracts are related to 9 categories. The distribution of the abstracts among the 9 classes is highly uneven and is as follows: {2623, 271, 18, 3, 1, 1, 1, 1, 1 }

The KnCr Corpus of MEDLINE : This corpus contains abstracts from the cancer domain of the medical field and collected from the MEDLINE documents [1]. It contains 900 abstracts and they are related to 16 categories. The abstracts are distributed among the 16 classes as follows :{169, 160, 119, 99, 66, 64, 51, 31, 30, 29, 22, 20, 14, 12, 8, 6}

[2] Annotations at URL: `http://www.projet-depart.org/public/LINA-PCL-1.0.tar.gz`

3.2 Evaluation Methods

For the purpose of evaluating the quality of the clusters, we use 3 existing measures. These measures will determine which clustering methods produce the best clusters. The details of these measures are based on the initial setting where S number of short text are naturally grouped into classes denoted by $C = \{c_1, c_2, ..., c_n\}$ and are clustered by the clustering algorithms into groups denoted by $K = \{k_1, k_2, ..., k_3\}$.

Clustering F-measure (F) : F is a mapping based measure where evaluation is done by mapping each cluster to a class [16] and is based on precision and recall as follows:

$$F(C) = \sum_{C_i \in C} \frac{|C_i|}{S} max_{K_j \in K}\{F(C_i, K_j)\} \tag{5}$$

where,

$$Recall(C_i, K_j) = \frac{n_{ij}}{|C_i|} \quad Precision(C_i, K_j) = \frac{n_{ij}}{|K_j|}$$

and

$$F(C_i, K_j) = \frac{2 \times Recall(C_i, K_j) * Precision(C_i, K_j)}{Recall(C_i, K_j) + Precision(C_i, K_j)}$$

where n_{ij} is the number of short text of class C_i present in clusters K_j. The F value will be in the range of [0,1], where 1 being the best score. A slight variation of this method has also been used in clustering short text [8] which computes the F according to the clusters rather than the class and is computed as $F(K) = \sum_{K_j \in K} \frac{|K_j|}{S} max_{C_i \in C}\{F(C_i, K_j)\}$ which we do not use in this paper.

Adjusted Rand Index (ARI) : This measure is an improvement of the Rand Index [17] which is based on counting pairs of elements that are clustered similarly in the classes and clusters. With the initial setting the ARI can be computed as below:

$$\frac{\sum_{i,j} \binom{n_{ij}}{2} - [\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}]/\binom{S}{2}}{1/2[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}] - [\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}]/\binom{S}{2}} \tag{6}$$

where n_{ij} is the number of short text of class C_i present in cluster K_j, n_i is the number of short text in class C_i, n_j is the number of short text in the cluster K_j. The upper bound of this measure is 1 and corresponds to the best score and the expected value of this measure is zero.

V : V is based on information theory and uses entropy and conditional entropy to evaluate the cluster [18]. The value of V are computed as in (7).

$$V = \frac{2hc}{h+c} \quad \text{where,} \quad h = \begin{cases} 1 & H(C) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & \text{else} \end{cases} \quad c = \begin{cases} 1 & H(K) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & \text{else} \end{cases} \tag{7}$$

with,

$$H(C) = -\sum_{c=1}^{|C|} \frac{\sum_{K=1}^{|K|} a_{ck}}{S} log \frac{\sum_{K=1}^{|K|} a_{ck}}{S}$$
$$H(K) = -\sum_{k=1}^{|K|} \frac{\sum_{c=1}^{|C|} a_{ck}}{S} log \frac{\sum_{c=1}^{|C|} a_{ck}}{S}$$
$$H(C|K) = -\sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{a_{ck}}{S} log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}}$$
$$H(K|C) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{a_{ck}}{S} log \frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}}$$

where, a_{ck} is the number of short text in C_i which is present in K_j. V gives an evaluation score in a range of [0,1], 1 being the best score.

3.3 Clustering

We used all three hierarchical clustering and the spectral clustering methods to cluster the short text present in the corpora. The clustering was performed in R[3] which is an environment for statistical computing and graphics. After clustering, the distribution of text among clusters shows that each clustering method has its own characteristics that defines the type of clusters that are created in terms of the distribution of text. Table 2 shows the distributions of the elements in the clusters created by all four clustering methods, which used cosine similarity, on the Cicling-2002 corpus and the hep-ex corpus.

Table 2. Distribution of the short text among the clusters created by SHC, CHC, AHC, and SPEC which uses cosine similarity

(a) Cicling-2002 corpus

	Cluster Index			
Clustering	1	2	3	4
SHC	45	1	1	1
CHC	11	24	7	6
AHC	33	12	1	2
SPEC	13	4	9	22

(b) hep-ex corpus

	Cluster Index								
Cluster	1	2	3	4	5	6	7	8	9
SHC	2912	1	1	1	1	1	1	1	1
CHC	2879	5	11	5	2	4	5	5	4
AHC	2879	13	11	1	5	3	5	2	1
SPEC	298	248	396	337	243	328	371	303	396

From Table 2, we can see that SHC creates many clusters with only one element in it which indicates that for our purpose, single link hierarchical clustering may not be a good choice. The characteristics of SPEC shows that it distributes the text evenly throughout the clusters. CHC and AHC have similar characteristics which lie between SHC and SPEC. These characteristics of the clustering method remain the same irrespective of the corpora but this characteristic alone cannot be used to decide upon the appropriate method for clustering.

There are evaluation methods that give scores on the quality of the clusters and based on these scores we tend to decide on the appropriate clustering

[3] `http://www.r-project.org/`

method. Different evaluation methods have different properties [3], so before we decide on the clustering method we first have to decide on which evaluation method would be appropriate.

We compare the evaluation methods using a direct method where we assign each cluster generated by the clustering method to a unique class in such a way that the average F-score (AF) for each pair of cluster and class is maximized. F-score is defined as $F(C_i, K_j)$ in (5). As we maximize the AF, the resulting pairs of cluster and class could be considered as the best practical solution. Tables 3(a) 3(b) 3(c) 3(d) show the F-score confusion matrix of class against clusters generated by 4 clustering methods, using CS, on the Cicling-2002 corpus. The bold-faced values in each matrix makes the AF maximum. This optimal assignment is done automatically using the Hungarian Algorithm [19]. Table 3(e) shows the scores given to each clustering method by the 3 evaluation methods and maximum AF (MAF). We consider an evaluation method to be good if it resembles the MAF scores because a high value for MAF generally indicates a high level of agreement between the classes and the clusters.

Table 3. In (a),(b),(c), and (d) the F-score confusion matrices for SHC, CHC, AHC, and SPEC applied on the CICLing-2002 corpus are shown and the elements which make the MAF are bold-faced. The classes and clusters are represented by the rows and columns respectively. In (e) the clusters generated by the clustering methods are evaluated using F, ARI, V, and MAF.

(a) SHC

0.17	0	0.36	0
0	0	0.5	**0**
0	0	**0.39**	0
0	**0.17**	0.32	0.17

(b) CHC

0.11	0.29	0.36	**0.12**
0	**0.56**	0.15	0.19
0	0.29	**0.45**	0.12
0.67	0.17	0	0.24

(c) AHC

0	0.17	0.36	**0.15**
0	0.15	**0.54**	0
0	0	0.5	0
0.17	**0.70**	0.05	0.15

(d) SPEC

0.1	**0.27**	0.25	0.3
0	0.11	0.14	**0.65**
0.80	0	0	0.18
0	0.13	**0.67**	0.12

(e) *Cicling-2002*

	F	ARI	V	MAF
SHC	0.40	0.01	0.11	0.21
CHC	0.52	0.10	0.21	0.45
AHC	0.53	0.17	0.29	0.35
SPEC	**0.61**	**0.25**	**0.34**	**0.60**

Table 3(e) does not help us find the best evaluation method because no evaluation method represents the MAF value, but it certainly gives an insight on the performance of the clustering methods. All of the evaluation methods do point towards spectral clustering to be the best clustering method for our case. Table 4 gives the complete results of the experiments. It shows that for all the corpus, excluding hep-ex corpus, spectral clustering performs better than the rest. In the case of hep-ex, the short text are unevenly distributed among the clusters as shown in Sect. 3.1 and as the characteristics of the spectral clustering tends to make evenly distributed clusters the performance decreases.

Table 4. F,ARI, V, and MAF values for four clustering methods SHC, CHC, AHC and SPEC on four corpus KnCr, hep-ex,Cicling-2002, and LDC. The best score achieved by each evaluation method on every corpus are bold-faced.

Corpus		*KnCr*				*Cicling-2002*			
Cluster	Similarity	F	ARI	V	MAF	F	ARI	V	MAF
SHC	Cosine	0.20	0.00	0.03	0.04	0.40	0.01	0.11	0.21
	KLD	0.20	0.00	0.03	0.04	0.40	0.01	0.11	0.21
	LSA	0.21	0.00	0.04	0.05	0.40	0.00	0.11	0.17
	SVSM	0.20	0.00	0.03	0.04	0.40	0.01	0.11	0.21
CHC	Cosine	0.21	0.01	0.12	0.14	0.52	0.10	0.21	0.45
	KLD	0.20	-0.01	0.11	0.16	0.45	0.06	0.18	0.33
	LSA	0.21	0.03	0.12	0.09	0.52	0.11	0.23	0.52
	SVSM	0.22	0.01	0.09	0.10	0.46	0.07	0.19	0.40
AHC	Cosine	0.25	0.04	0.12	0.13	0.53	0.17	0.29	0.35
	KLD	0.21	0.00	0.04	0.05	0.40	0.02	0.15	0.25
	LSA	0.21	0.00	0.06	0.07	0.40	0.00	0.10	0.21
	SVSM	0.20	0.00	0.04	0.04	0.40	0.02	0.15	0.25
SPEC	Cosine	**0.30**	**0.09**	**0.19**	**0.19**	0.61	0.25	**0.34**	0.60
	KLD	0.23	0.04	0.11	0.14	0.51	0.15	0.26	0.51
	LSA	0.24	0.04	0.15	0.17	0.55	0.19	0.27	0.52
	SVSM	0.22	0.03	0.13	0.16	**0.64**	**0.26**	**0.34**	**0.64**

Corpus		*hep-ex*				*LDC*			
Cluster	Similarity	F	ARI	V	MAF	F	ARI	V	MAF
SHC	Cosine	**0.86**	0.01	0.01	0.10	0.19	0.00	0.09	0.08
	KLD	**0.86**	-0.02	0.01	0.10	0.19	0.00	0.09	0.07
	LSA	**0.86**	0.01	0.01	0.11	0.19	0.00	0.09	0.07
	SVSM	**0.86**	0.01	0.01	0.11	0.19	0.00	0.09	0.08
CHC	Cosine	**0.86**	-0.01	0.01	0.11	0.21	0.10	0.15	0.13
	KLD	0.81	-0.02	0.00	0.10	0.29	0.02	0.26	0.21
	LSA	0.41	0.01	0.02	0.08	0.29	0.08	0.28	0.25
	SVSM	0.56	0.03	0.07	0.11	0.41	0.24	0.42	0.24
AHC	Cosine	**0.86**	0.03	0.01	0.11	0.38	0.18	0.38	0.21
	KLD	**0.86**	-0.01	0.00	0.10	0.35	0.14	0.36	0.18
	LSA	**0.86**	**0.10**	0.05	**0.13**	0.43	0.22	0.42	0.28
	SVSM	**0.86**	0.00	0.01	**0.13**	0.31	0.14	0.32	0.14
SPEC	Cosine	0.28	0.01	**0.08**	0.09	0.50	**0.29**	**0.50**	0.41
	KLD	0.47	0.00	0.03	0.08	0.26	0.05	0.24	0.21
	LSA	0.28	0.01	**0.08**	0.09	**0.51**	0.27	0.49	**0.43**
	SVSM	0.29	0.01	**0.08**	0.09	0.45	0.23	0.45	0.36

For the hep-ex corpus, F evaluation method gives a good result for SHC even though the distribution of the short text in the clusters are clearly undesirable for other clusters as seen in Table 2. This is due to the drawback of F as it may not take into account the membership of the clusters and may not evaluate the

clusters. From this table we can also see that none of the evaluation measure resembles the MAF values. But if required, we would select V as the best out of the three evaluation methods. The reason behind this selection is that, V resembles the variation in the range of MAF more than the other evaluation measures. Among the 16 possible range of MAF, present in each box in Table 4, V resembles MAF 9 times where as ARI 7 times.

It is also difficult to comment on the similarity measures because the clusters formed are highly affected by the different characteristics of the corpora which overshadows the effect of the similarity measures. But as we consider spectral clustering to be a good clustering method, according to the number of best evaluation scores achieved shown in Table 4, we analyse the similarity measures based on these best scores. By doing so, we see that in most of the cases the spectral clustering which uses KLD similarity measure produces clusters whose evaluation score have the highest difference with the best evaluation score. This could indicate that the performance of KLD is the least among the other similarity measures. The other three similarity measures do not differ much in most of the cases comparing the evaluation measures.

4 Conclusions

In this paper, we cluster short text from four different corpora containing different type, size, and distribution of short text. This difference in the corpora is important to present a generalized solution for the clustering of short text. Among the corpora, three of them have been used in previous research on clustering abstracts. We present a new annotated corpus containing newspaper paragraphs to analyse the clustering of short text. The cluster list for this corpus can be freely downloaded for further research on this field.

To analyse the clustering of short text, we used three hierarchical clustering algorithms which is famous in the field of NLP and spectral clustering which is based on k-means clustering algorithm to show that the latter seems to be a good choice over hierarchical clustering especially when the text are evenly distributed among the clusters. We also show that the performance of KLD method, which uses term selection, is the least compared to the other three measures and the performance of CS, LSA, SVSM do not differ much from each other.

Using the Hungarian algorithm, we assigned each cluster to a class so that the average F-score, AF, is maximized. The maximized AF method can also be considered as an evaluation method if the number of class is the same as the clusters. This optimized assignment was the basis of choosing the best evaluation method. Unfortunately, none of the evaluation method closely resembled the MAF but taking into account the number of times a method shows resemblance to the MAF measure, V has an upper hand. Existing work of short text clustering evaluate the clusters using mapping based methods such as clustering F-measure or Purity. We show that these measures are not able to evaluate the entire membership of the clusters which is a huge drawback. This implies that results from previous work which use these mapping based evaluation methods have to be analysed carefully.

References

1. Pinto, D., Rosso, P.: Kncr: A short-text narrow-domain sub-corpus of medline. In: Proceedings of the TLH 2006 Conference. Advances in Computer Science, pp. 266–269 (2006)
2. Makagonov, P., Alexandrov, M., Gelbukh, A.: Clustering Abstracts Instead of Full Texts. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 129–135. Springer, Heidelberg (2004)
3. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval 12, 461–486 (2009)
4. Reichart, R., Rappoport, A.: The nvi clustering evaluation measure. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pp. 165–173 (2009)
5. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17, 395–416 (2007)
6. Nakov, P., Popova, A., Mateev, P.: Weight functions impact on lsa performance. In: EuroConference RANLP 2001, Recent Advances in NLP, pp. 187–193 (2001)
7. Shrestha, P., Jacquin, C., Daille, B.: Reduction of search space to annotate monolingual corpora. In: Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011) (2011)
8. Pinto, D., Benedí, J.-M., Rosso, P.: Clustering Narrow-Domain Short Texts by Using the Kullback-Leibler Distance. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 611–622. Springer, Heidelberg (2007)
9. Manning, C.D., Raghavan, P., Schütze, H.: Clustering Narrow-Domain Short Texts by using the Kullback-Leibler Distance. Cambridge University Press (2008)
10. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to latent semantic analysis. In: Discourse Processes (1998)
11. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Clustering Abstracts of Scientific Texts Using the Transition Point Technique. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 536–546. Springer, Heidelberg (2006)
12. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 391–407 (1990)
13. Jolliffe, I.T.: Principal component analysis. Chemometrics and Intelligent Laboratory Systems 2, 37–52 (1986)
14. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, pp. 849–856. MIT Press (2001)
15. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76, 378–382 (1971)
16. Fung, B.C., Wang, K., Ester, M.: Hierarchical document clustering using frequent itemsets. In: Proceedings of SIAM International Conference on Data Mining, SDM 2003 (2003)
17. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification 2, 193–218 (1985)
18. Rosenberg, A., Hirschberg, J.: V-measure: a conditional entropy-based external cluster evaluation measure. In: EMNLP 2007 (2007)
19. Harold, K.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly 2, 83–97 (1955)