Improving Clustering of Noisy Documents through Automatic Summarisation

Seemab Latif[1], Mary McGee Wood[2], and Goran Nenadic[1]

[1] School of Computer Science
University of Manchester, UK
latifs@cs.man.ac.uk, G.Nenadic@cs.man.ac.uk
[2] Assessment21, Cooper Buildings
Sheffield, UK
mary@cs.man.ac.uk, mmw@assessment21.com

Abstract. In this paper we discuss clustering of students' textual answers in examinations to provide grouping that will help with their marking. Since such answers may contain noisy sentences, automatic summarisation has been applied as a pre-processing technique. These summarised answers are then clustered based on their similarity, using k-means and agglomerative clustering algorithms. We have evaluated the quality of document clustering results when applied to full-texts and summarized texts. The analyses show that the automatic summarization has filtered out noisy sentences from the documents, which has made the resulting clusters more homogeneous, complete and coherent.

1 Introduction

Document clustering is a generic problem with widespread applications. The motivation behind clustering a set of documents is to find inherent structure and relationship between the documents. Document clustering has been applied successfully in the field of Information Retrieval, web applications to assist in the organization of the information on the web (Maarek et al., 2000), to cluster biomedical literature (Illhoi et al., 2006) and to improve document understanding by using document clustering and multiple document summarisation where summarisation is used to summarise the documents in resultant clusters (Wang et al., 2008).

The quality of document clustering is dependent on the length of the documents and the "noise" present within documents. The main features of a document are the words that it contains. Unimportant words contribute to noise and thus may lead to faulty results from automatic processing. Assess By Computer (ABC) (Sargeant et al., 2004) is a suit of software tools for assessment of students' exams. It uses document clustering to group students' textual answers to support semi-automated marking and assessment in examinations (Wood et al., 2006). One important feature of ABC is abstraction: instead of referring to a large number of students' answers one at a time, ABC refers to them as groups of similar answers at one time. Students' answers are written under time pressure in an examination environment. Therefore, there is always a high chance of students making spelling mistakes and writing sentences or words that are not relevant

Special issue: Natural Language Processing and its Applications.
Research in Computing Science 46, 2010, pp. 253-264

Received 02/12/09
Accepted 16/01/10
Final version 12/03/10

to the question being asked and thus could be considered as noise, in particular, if we want to assess students' understanding of the concepts in questions. These sentences or words may affect the performance of the clustering process. To avoid the deterioration of the clustering results, we hypothesized that Automatic Text Summarisation could be one of the solutions to remove noisy data from the answers efficiently and effectively.

Previous extensive work on document clustering has focused on issues such as initial number of clusters, similarity measures and document representation, and has to some extent ignored the issue of document pre-processing. The principal pre-processing techniques applied to documents have been stopword removal, stemming and case folding. However, document clustering algorithms are highly dependent on the length of the document (Hsi-Cheng and Chiun-Chieh, 2005). Therefore, we hypothesize that pre-processing texts using automatic summarization before document clustering may improve the performance of clustering algorithms in terms of both efficiency and quality of the clustering, and summary based clusters will be more homogeneous, complete and coherent than fulltext clusters.

2 Integrating Text Summarisation into Document Clustering

In this paper, we integrate text summarisation with document clustering to utilize the mutual influence of both techniques to help in grouping of students' free text answers. We apply automatic text summarisation as a pre-processing method for students' answers before clustering. The main aim behind using automatic text summarisation is to extract the information content of the answers, and to improve document clustering by reducing the noise and the length of the answers.

2.1 Automatic Summarisation

In this paper, we have used Keyword-dependent Summarizer (KdS) that summarises a document using keywords given by the user. In case of summarising students' answers, these keywords are given by the marker. This summarisation method follows the shallow approach for summarization, i.e. it employs text extraction techniques to generate the summary (Neto et al., 2002). Abstractive or language generation approach for summarisation was not used, as there is no guarantee that generated the text will be useful (Neto et al., 2000). In order to perform extractive summarisation, the first question arises is that on what granularity segments will be extracted i.e. the "size" of the textual units that are copied from the original text and pasted into the summary. This could be a paragraph, a sentence, a phrase or a clause. We have performed summarisation on sentence level using a structural analysis technique. This technique uses the semantic structure of the text (such as keyword frequency, position of the sentence in the document and paragraph, cue phrases and sentence length) to rank sentences for inclusion in a final summary.

This summarization method has been evaluated qualitatively and quantitatively by both human and automatic evaluations. Both human and ROUGE ((Lin and Hovy, 2003)) evaluations have achieved high positive correlation with the scores assigned to the essays by the human marker.

2.2 Clustering process with Summarisation: Framework

Clustering students' free text answers is a difficult task as there are a number of problems posed by natural language for automated processing of these answers. In this paper, each student answer is referred to as a document. Documents are represented as a vector of words in the Vector Space Model (VSM). The performance of the VSM is highly dependent on the number of dimensions in the model and it can be improved by removing terms that carry less semantic meaning and are not helpful while calculating the similarity between documents (stopword removal), converting terms to their stems (stemming), spelling corrections and term weighting (the process of modifying the values associated with each term to reflect the information content of that term). The similarity between document vectors is calculated using the cosine of the angle between them.

The overall framework of the integration of text summarisation with document clustering for clustering students' answers is given in Figure 1 and it follows the following steps:

- **Step 1: Extracting Answers**
  In this step, students' answer strings for a question, which are to be clustered, are extracted from the XML file provided by the ABC marking tool.

- **Step 2: Summarising Answers**
  Each answer string is then summarised using KdS. Summarisation extracts the information content of the answers and filters out the noise. This reduces the number of terms in the answer string.

- **Step 3: Document Pre-processing**
  The summarised answers are then pre-processed using spelling correction, stopword removal, stemming and term weighting. First three pre-processing steps and summarisation aim to reduce the total number of terms used in clustering while term weighting provides a way to apply different weights to the terms to reflect their importance.

- **Step 4: Creating the Term-By-Document Matrix**
  The term-by-document matrix on full-text answer strings could potentially be a sparse matrix containing over 500 terms (dimensions). Many terms in the matrix are noisy and useless, giving no semantic information about the answer. To improve the performance of the term-by-document matrix, we have created vectors on the summarised answer strings instead of full-text answer strings. The matrix created on summarised answer strings has fewer dimensions as compared to the

matrix crated on full-text answer strings.

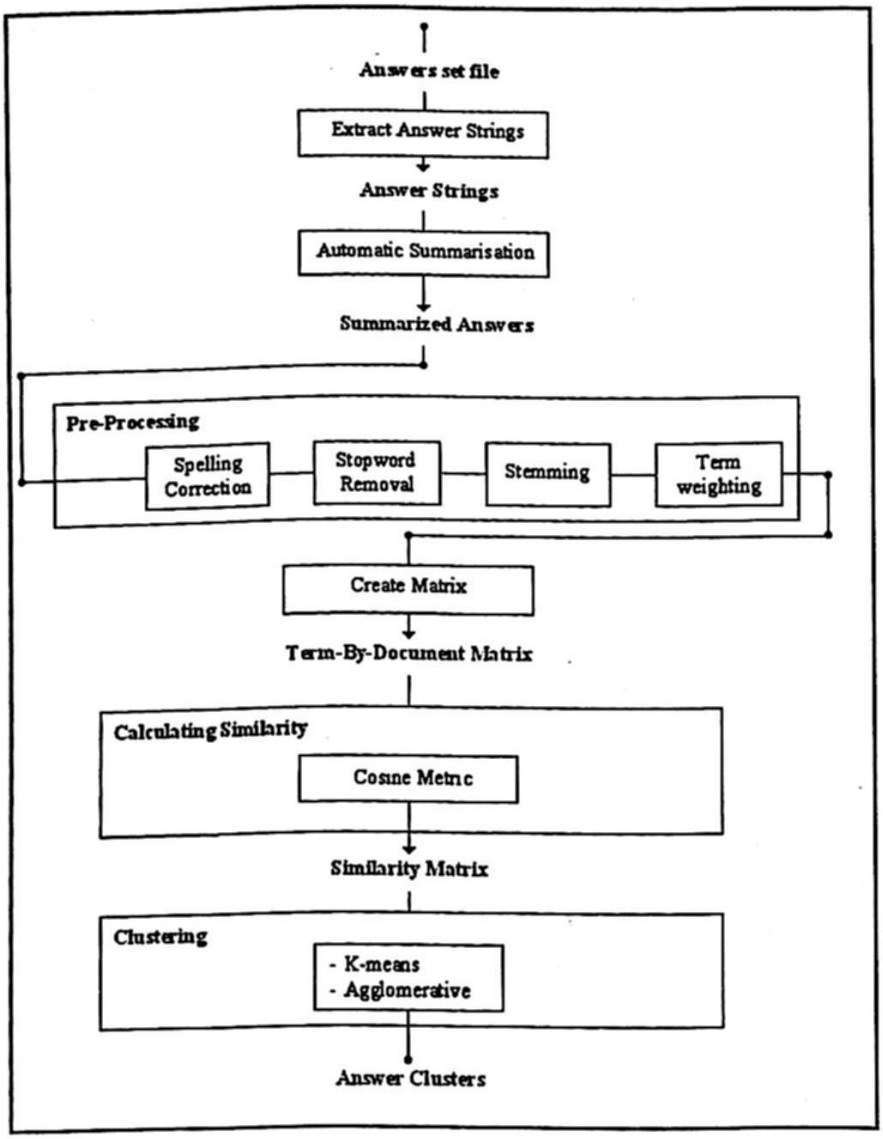

Fig. 1. Framework for Clustering Students' Answers

– **Step 5: Creating Clusters**
The last step in this procedure is to create clusters. Two clustering algorithms are used to generate clusters. Clustering algorithms cluster answers based on the terms present within the answer and the similarity between answers.

3 Experimental Setup

In this section, the design and results of an experiment are discussed. Here, automatic summarisation has been applied as a pre-processing step for document clustering to reduce the noise. The full documents and summarised documents are then clustered and clustering results are evaluated using the precision and recall measures (defined below).

The aim of this experiment is to evaluate the clustering results, when carried on full-text and summarised documents, for their quality and accuracy.

Experimental Design: The automatic summarisation pre-processing method was applied for partitional (k-means) and hierarchical (agglomerative) clustering algorithms. We have used KdS and the keywords were taken from the "Model Answer" to the examination questions. The clustering algorithms were run on the full-text documents and on five different levels of summarisation compression. Each document in the dataset was compressed to 50%, 40%, 30%, 20% and 10% of the size of its original length, and both algorithms were run on the documents in each dataset having five different levels of compressed documents and the whole full-text document.

Datasets: The datasets used for this experiment were taken from the ABC marking tool, and include marked exams from Pharmacy, Biology and Computer Science undergraduate courses conducted at the University over the past five years. The reason for taking these datasets is that human assigned marks for these datasets are available and these marks will be used to evaluate the clustering results. These exams were marked by an instructor who was teaching the course.

Table 1 gives the number of answers in each dataset, average number of terms in the answers and the number of clusters generated for each dataset. The number of clusters for each dataset was the total number of distinct marks for that question plus one (for mark 0). For example, if we want to cluster answers to a question worth of 4 marks then the number of clusters will be 5 (one for mark 0). As a sample, one question along with its answer from the Biology and Pharmacy datasets is given in appendix A.

Table 1. Datasets statistics for evaluating summary-based clusters

Dataset	Question ID	Number of Documents	Number of Features	Number of Clusters
CS	CS1	108	880	5
Biology	Biology1	279	960	6
	Biology2	280	708	6
	Biology3	276	909	7
	Biology4	276	996	6
Pharmacy	Phar1	139	1103	5
	Phar2	165	1278	7

3.1 Evaluation Metrics

In general, the categories are not known before hand, but in our case, we have used human-marked data from the ABC marking tool to evaluate the quality of clustering.

The answers in each dataset were grouped according to their marks awarded by the human marker. These marked categories have served the purpose of the "gold standard" categories for the evaluation. For the evaluation of document clustering results, Precision and Recall were defined as follows.

Precision associated with one document represents what fraction of documents in its cluster belongs to its marked category. Precision of a clustering result on a dataset can be either calculated at the document level (micro-precision) or at the cluster level (macro-precision). If the micro-precision is high, then it means that the number of noisy documents or misclassified documents in the clusters is low. If macro-precision is high then it means that the most documents from the same category are grouped together.
Recall associated with one document represents what fraction of documents from its marked category appears in its cluster. Similarly to precision, recall of a clustering result on a dataset can be either calculated on the document level (micro-recall) or on the cluster level (macro-recall). If the micro-recall is high then it means that most of the documents from its model category lie in its cluster. If macro-recall is high then it means clusters are similar to the model categories.

We have used only micro-precision and micro-recall measures for clustering result evaluations, which are combined using the standard combining metric, Van Rijsbergen's F-Measure (Rijsbergen, 1974). Both precision and recall of a cluster will be high when the computed cluster is similar to the model cluster. The precision will be 1 and the recall of each document will decrease by $\frac{1}{|\mathrm{ModelCategory}|}$ when each document is placed in independent cluster i.e. each cluster has only one document. The recall will be 1 and the precision of each document in the cluster will decrease by $\frac{n}{n+1}$ when all the documents will be clustered into one cluster. Note that it is not possible to have precision or recall of 0 because at least a cluster and a category share one common document.

There are four mathematical constraints proposed by (Amigo et al., 2009) that should be satisfied by the clustering evaluation metrics. These constraints are Cluster Homogeneity, Cluster Completeness, Rag Bag and Cluster Size versus Quantity. The metrics that we have introduced here satisfy these constraints. Micro-precision satisfies cluster homogeneity and ragbag constraints while micro-recall satisfies cluster completeness and size versus quantity constraints.

4 Results and Analysis

Because of random initialization of the k-means, the algorithm was run 10 times on each dataset and then mean values were calculated for the evaluation. All summarisation-based clustering results have performed better then the full-text clustering results. In tables 2 and 3, the micro-precision values are higher for the summarisation-based clusters than the micro-precision values for the full-text clusters.

Table 2. Micro-precision for k-means clustering

Dataset	Question ID	Fulltext	Compression Level				
			50%	40%	30%	20%	10%
CS	CS1	0.513	0.831	0.837	0.793	0.753	0.728
Biology	Biology1	0.519	0.843	0.880	0.838	0.776	0.711
	Biology2	0.495	0.904	0.899	0.852	0.832	0.780
	Biology3	0.617	0.814	0.823	0.759	0.711	0.677
	Biology4	0.672	0.777	0.786	0.764	0.719	0.679
Pharmacy	Phar1	0.569	0.774	0.800	0.737	0.691	0.642
	Phar2	0.586	0.798	0.811	0.770	0.720	0.663

Table 3. Micro-precision for agglomerative clustering

Dataset	Question ID	Fulltext	Compression Level				
			50%	40%	30%	20%	10%
CS	CS1	0.466	0.840	0.840	0.783	0.750	0.694
Biology	Biology1	0.715	0.855	0.862	0.780	0.805	0.727
	Biology2	0.738	0.900	0.895	0.865	0.849	0.736
	Biology3	0.673	0.809	0.843	0.764	0.718	0.675
	Biology4	0.695	0.786	0.800	0.753	0.707	0.705
Pharmacy	Phar1	0.599	0.775	0.779	0.762	0.718	0.602
	Phar2	0.655	0.772	0.812	0.761	0.721	0.678

According to the definition of micro-precision, it is high when most of the items from a single model category are clustered together in one cluster or when majority of the items have their individual clusters. In our case, the numbers of clusters are fixed for each dataset, which omits the possibility of each document having its own cluster. This means that most of the documents belonging to a single model category are clustered together and that summarisation has filtered out the feature terms that are not useful in distinguishing the documents. Therefore, summary-based clusters are more homogeneous and noise free than full-text clusters.

Micro-recall values for each of the algorithms are given in tables 4 and 5. These values are high for full-text clusters. According to the definition of micro-recall, it is high when the resultant clusters are similar to the model categories or when majority of the items are clustered in one cluster. We analysed the full-text document clustering results manually, which showed that the high micro-recall for full-text clusters is due to the clustering of most of the documents in one cluster.

However, many of the feature terms in full-text documents have low distinctive power and are not useful in clustering. This suggests that if initial documents are misclassified then these documents "attract" other documents with high similarity and a large number of common feature terms. This will make the cluster noisy with the doc-

Table 4. Micro-recall for k-means clustering

Dataset	Question ID	Fulltext	Compression Level				
			50%	40%	30%	20%	10%
CS	CS 1	0.795	0.715	0.769	0.678	0.592	0.549
Biology	Biology1	0.688	0.724	0.730	0.838	0.616	0.545
	Biology2	0.769	0.668	0.715	0.643	0.588	0.518
	Biology3	0.674	0.760	0.769	0.706	0.655	0.618
	Biology4	0.697	0.772	0.772	0.750	0.702	0.650
Pharmacy	Phar1	0.628	0.756	0.764	0.695	0.653	0.590
	Phar2	0.638	0.754	0.771	0.725	0.667	0.606

Table 5. Micro-recall for agglomerative clustering

Dataset	Question ID	Fulltext	Compression Level				
			50%	40%	30%	20%	10%
CS	CS1	0.863	0.727	0.768	0.712	0.556	0.564
Biology	Biology1	0.574	0.733	0.733	0.722	0.620	0.599
	Biology2	0.496	0.645	0.680	0.622	0.586	0.530
	Biology3	0.613	0.74	0.777	0.727	0.675	0.636
	Biolog 4	0.667	0.781	0.779	0.739	0.684	0.664
Pharmacy	Phar1	0.557	0.713	0.728	0.683	0.690	0.558
	Phar2	0.598	0.764	0.761	0.713	0.676	0.617

uments that are not part of it.

Tables 6 and 7 and figures 2 and 3 give the micro-F-measure values for the k-means and agglomerative clusterings on the three datasets.

Analyses of the results show that both clustering algorithms results have achieved optimal clustering at the compression level of 40% summarised documents. At this level, micro-F-measure is $\simeq$ to 80% for both algorithms, averaged over all the datasets. The datasets with large number of documents have a smooth curve with peak at 40% summary based clustering. For most of the datasets even 10% summarised text clustering has performed better than the full-text clustering.

5 Conclusion

In this paper, the experiments have been discussed which evaluate the method of improving clustering results by performing automatic summarisation of students' textual answers. Automatic summarisation has reduced the length of the documents and hence the number of feature terms that were potentially noisy for clustering. The results suggest that automatic summarisation has filtered out the noise from the documents and

Table 6. Micro-F-measure for k-means clustering

Dataset	Question ID	Fulltext	Compression Level				
			50%	40%	30%	20%	10%
CS	CS1	0.624	0.768	0.802	0.731	0.662	0.626
Biology	Biology1	0.591	0.778	0.797	0.741	0.685	0.616
	Biology2	0.602	0.767	0.796	0.733	0.689	0.622
	Biology3	0.644	0.786	0.795	0.731	0.682	0.646
	Biology4	0.684	0.774	0.779	0.757	0.690	0.664
Pharmacy	Phar1	0.597	0.764	0.782	0.715	0.672	0.614
	Phar2	0.611	0.775	0.790	0.746	0.693	0.633
			Average	0.791			

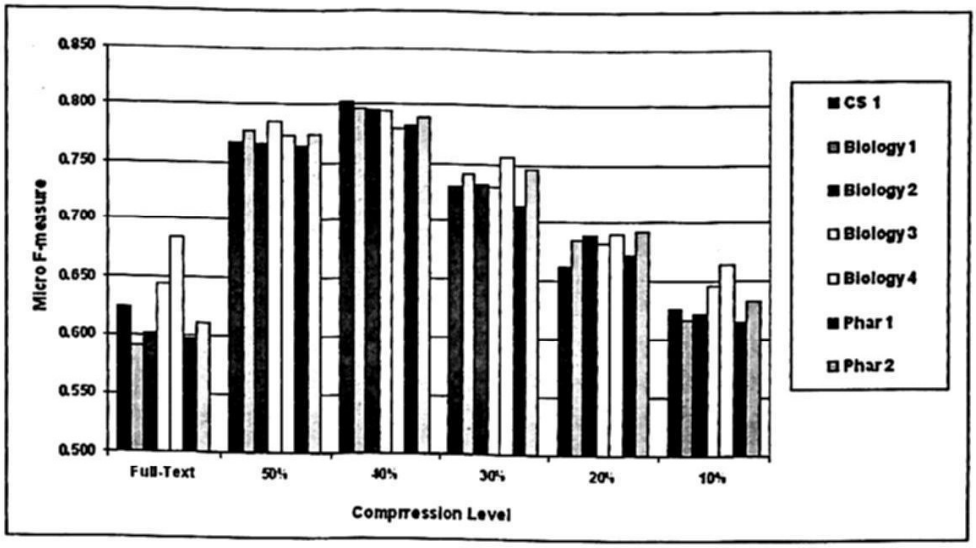

Fig. 2. K-means clustering micro-F-measure

Table 7. Micro-F-measure for agglomerative clustering

Dataset	Question ID	Fulltext	Compression Level				
			50%	40%	30%	20%	10%
CS	CS1	0.605	0.780	0.802	0.746	0.638	0.623
Biology	Biology1	0.637	0.789	0.792	0.750	0.701	0.657
	Biology2	0.593	0.751	0.773	0.723	0.694	0.616
	Biology3	0.641	0.773	0.808	0.745	0.696	0.655
	Biology4	0.681	0.784	0.789	0.746	0.695	0.684
Pharmacy	Phar1	0.577	0.743	0.753	0.720	0.703	0.579
	Phar2	0.625	0.768	0.786	0.736	0.698	0.646
			Average	0.790			

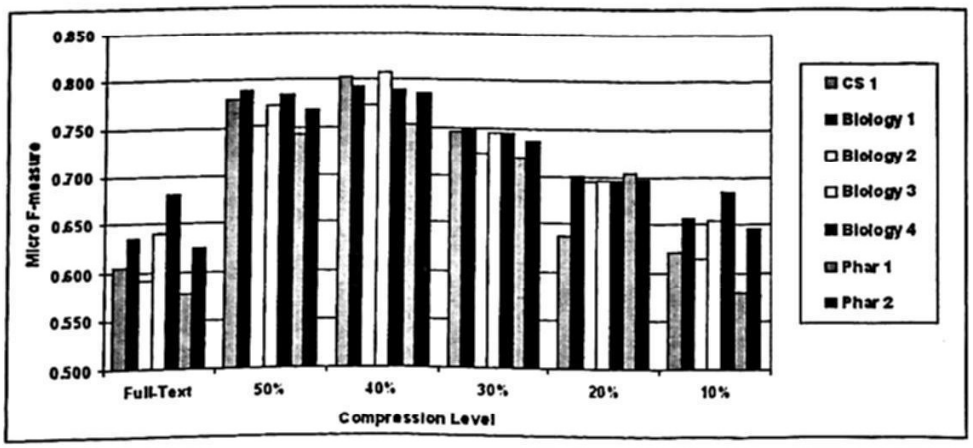

Fig. 3. Agglomerative clustering micro-F-measure

has extracted the relevant information content of the documents.

Due to the noise removal and reduction in document length, the performance of the two document clustering algorithms has improved in terms of quality of the clustering results. While evaluating the clustering results, we came to conclusion that summarisation-based clusters are more homogeneous (as micro-precision of summary-based clusters is higher than the micro-precision of full-text clusters for both the algorithms, see tables 2 and 3) and complete (as micro-recall value of summary-based clusters is higher than the micro-recall for full-text clusters except for 10% summarised-text clusters using k-means, see tables 4 and 5) then full-text clusters. Optimal clustering results were achieved when the documents were summarised to 40% of their original length.

References

1. Amigo, E., Gonzalo, J., Artiles, J. and Verdejo, F. (2009). A Comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints, *Information Retrieval Journal* 12(4): 461– 486.
2. Hsi-Cheng, C. and Chiun-Chieh, H. (2005). Using Topic Keyword Clusters for Automatic Doc- ument Clustering, *Proceedings of the 3rd International Conference on Information Technol- ogy and Applications*, pp. 419–424.
3. Illhoi, Y., Hu, X. and Il-Yeol, S. (2006). A Coherent Biomedical Literature Clustering and Sum- marization Approachthrough Ontology-Enriched Graphical Representations, *Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery*, pp. 374– 383.
4. Lin, C. and Hovy, E. (2003). Automatic Evaluation of Summaries using N-gram Co-occurrence Statistics, *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 71–78.
5. Maarek, Y., Fagin, R., Ben-Shaul, I. and Pelleg, D. (2000). Ephemeral Document Clustering for Web Applications, *Technical Report RJ 10186*, IBM Research Report.
6. Neto, J., Freitas, A. and Kaestner, C. (2002). Automatic Text Summarization using a Machine Learning Approach, *Proceedings of the 16th Brazilian Symposium on Artificial Intelligence, Advances in Artificial Intelligence*, pp. 205–215.

7. Neto, L., Santos, A., Kaestner, C. A. and Freitas, A. (2000). Document Clustering and Text Summarization, *Proceedings of the 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining*, pp. 41–55.
8. Rijsbergen, C. V. (1974). Foundation of Evaluation, *Journal of Documentation* 30(4): 365–373.
9. Sargeant, J., Wood, M. and Anderson, S. (2004). A Human-Computer Collaborative Approach to the Marking of Free Text Answers, *Proceeding of the 8th International Conference on Computer Assissted Assessment*, pp. 361–370.
10. Wang, D., Zhu, S., Li, T., Chi, Y. and Gong, Y. (2008). Integrating Clustering and Multi-Document Summarization to Improve Document Understanding, *Proceeding of the 17th Conference on Information and Knowledge Management*, pp. 1435–1436.
11. Wood, M., Jones, C., Sargeant, J. and Reed, P. (2006). Light-Weight Clustering Techniques for Short Text Answers in HCC CAA, *Proceedings of the 10th International Conference on Computer Assisted Assessment*, pp. 291–305.

Appendix A

One question along with its answer from Biology and Pharmacy dataset is given in this appendix.

Biology II

Question: What do you understand by the term Haematocrit? Could a person have a normal RBC count but a low Haematocrit? What could be the cause of this?

Answer: The Haematocrit shows the relative proportion of red blood cells to plasma and (white blood cells and other proteins), the figure given being the proportion of 'packed red blood cells'. The ideal proportion of packed cell volume (haematocrit) being, 37-47% in females, and 40-54% in males. If a person was to have a normal red blood cell count (normal amount of cells per micro litre), but smaller cells (microcytic cells) due to a disorder with the peptide chains, for example a missing peptide chain, then the cells would be able to pack closer together, and hence have a low haematocrit. This is classed as microcytic anemia, and can be due to problems in transcribing the two alpha or two beta chains, or perhaps the inability to form the globular protein. Eitherway, the size of the haemoglobin is reduced, hence the volume is reduced, but there is still likely to be the same number of cells. Other reasons for a normal red blood count and a low Haematocrit could be an increase in bodily fluids, i.e. plasma. The red blood cell count would be considered within 'normal' range, however, due to the increase in plasma, the ratio of blood to plasma would be altered in that the Haematocrit value would be lower.

Pharmacy

Question: State the factors to be considered when selecting an antiepileptic regimen for an 18 year old female patient who has been newly diagnosed as suffering from epilepsy by the local Neurologist.

Answer: Firstly, before selecting a treatment regimen, it is important to establish the number of seizures the patient has suffered. This is because treatment is rarely initiated after a single seizure. Usually, a person must suffer from two seizures in twelve months before treatment is given. The type of seizure is also important in selecting a treatment regimen i.e partial seizures are generally treated differently to general seizures. These classes of seizure can be further subdivided and will have specific treatment protocols. Thirdly, it is important to establish whether the patient is taking any other medication or has any other medical conditions which could affect the treatment options. For example, many antiepileptic drugs can interact with and reduce the efficacy of the combined oral contraceptive. It is important to try and give the patient a single agent wherever possible as many antiepileptic patients are successfully controlled with monotherapy. This would avoid the problems asociated with polypharmacy such as drug interactions, increased drug tox-

icity and reduced compliance. The age of the patient also needs to be considered especially if they were very young or elderly as this may limit the treatment options available or the doses may need adjusting. This does not seem to be a problem in this patient as she is 18 years old. However, the fact that she is a female needs to be considered. This is because many of the drugs can cause unacceptable side effects in women i.e. sodium valproate can cause hair loss, phenytoin can cause acne and gingival hyperplasia. This woman is also of child bearing age and it would need to be established if the patient was pregnant or breastfeeding as many of the antiepileptic drugs are teratogenic. It would also need to be established whether this patient was taking the combined oral contraceptive pill as antieplieptic medication can reduce the efficacy of this. This would mean that other precautionary advise on alternative methods of contraception would need to be given.