LIGHT-WEIGHT CLUSTERING TECHNIQUES FOR SHORT TEXT ANSWERS IN HUMAN COMPUTER COLLABORATIVE (HCC) CAA

Mary McGee Wood, Craig Jones,
John Sargeant and Phil Reed

Light-weight Clustering Techniques for Short Text Answers in Human Computer Collaborative (HCC) CAA

Mary McGee Wood, Craig Jones, John Sargeant and Phil Reed
School of Computer Science,
University of Manchester
mary@cs.man.ac.uk
jonesc@cs.man.ac.uk
js@cs.man.ac.uk
preed@cs.man.ac.uk

Abstract

We first explore the paedogogic value, in assessment, of questions which elicit short text answers (as opposed to either multiple choice questions or essays). Related work attempts to develop deeper processing for fully automatic marking. In contrast, we show that light-weight, robust, generic Language Engineering techniques for text clustering in a human-computer collaborative CAA system can contribute significantly to the speed, accuracy, and consistency of human marking. Examples from real summative assessments demonstrate the potential, and the inherent limitations, of this approach. Its value as a framework for formative feedback is also discussed.

Introduction

Assess By Computer (ABC; Sargeant et al 2004), deployed at the University of Manchester since 2003, follows a human-computer collaborative (HCC) approach to assessment. We focus on constructed answers such as text and diagrams rather than answers requiring mere selection between alternatives. The HCC assessment process is an active collaboration between humans and a software system, where the software does the routine work and supports the humans in making the important judgements.

One feature which distinguishes our approach from “traditional” CAA is our classification of question and answer types, which has three parameters. First, we distinguish constructed from selected answers (we strongly deprecate the traditional use of the term “objective” to mean “selected”).

Second, we distinguish “closed” or truly “objective” from “open” or “subjective” questions. For closed questions, the substance of a correct answer can be specified in advance (although its expression can vary wildly and unpredictably: Wood et al 2005). Open questions typically ask for an original example or argument. A marking scheme can only describe meta-level properties of a correct answer, and a “model answer” can only be an example.

Third, we distinguish loosely between long and short text answers. Length does not necessarily correlate with openness /closure: “Describe the causes of haemolytic disease in the newborn” calls for a paragraph of routine book-work while “Give an original example of an exception to default inheritance” requires only a short phrase. Length also does not necessarily correlate with the levels of Bloom’s taxonomy (Bloom et al 1956). Its main significance in ABC is that different Natural Language Engineering techniques are optimised for different lengths of text. To date we have focussed on simple, robust, generic techniques which are best suited to short answers.

Related Work

The use of text clustering in CAA is far from unique; but the other work we are aware of, such as the examples below, limits itself to formative assessment and/or aspires to be fully automatic.

Lütticke (2005) uses “logical inference” to compare student-drawn semantic networks with a model answer and generate formative feedback: the details of the comparison mechanism are unclear.

Weimer-Hastings et al (2005) use Latent Semantic Analysis to compare student answers with expected answers in an Intelligent Tutoring System in research methods in Psychology. Its use is purely formative, and they have attempted to evaluate student learning gain but not the effectiveness of clustering per se (p.c.). Although the technique is generic, its application is question-specific: they refer to it as “expectation-driven processing”.

Carlson & Tanimoto (2005) induce text classification rules from student answer sets. These rules are used “to construct ‘diagnoses’ of misconceptions that teachers can inspect in order to monitor the progress of their students” and to automatically construct formative feedback.

Pulman & Sukkarieh (2005) aim for automatic marking of “short” (“from a few words up to five lines”) free text answers to factual (objective, in our terminology) science questions. They use relatively heavy-weight techniques from traditional computational linguistics, and compare answers with keyword-based “patterns”, for which machine learning techniques have been investigated. They have worked with real student data, and their best results correlate acceptably with human markers’ judgements, but on a very small sample, and it is not obvious that these techniques will scale up sensibly.

The Paedogogic Potential of Short Text Answers

Constructed-answer questions have significant advantages over selected-answer questions for assessing students, even at the “knowledge” and “comprehension” levels of Bloom’s taxonomy. Recalling even a bare phrase like “mean cell volume” is a greater challenge than recognising it, even among cunningly chosen distractors; let alone the possibility of getting it right by luck.

And even short text answers (1-30 words; or comparably simple diagrams) are surprisingly versatile. As the following examples (with genuine, representative, mostly good student answers) show, short text answer questions, set cleverly, can test all levels of the taxonomy.

Knowledge: *What single measurement would you make to confirm that an individual is anaemic?*

Student answer: *haemoglobin concentration*

Comprehension: *A blood sample was taken from a patient and he was found to have a high white cell count. On further investigation the patient was found to have a neutrophil count of 22 x 10^9/L. Give two examples of what this could be indicative of.*

Student answer: *A recent or present bacterial infection. Or an allergic reaction.*

Application: *What is the value at the root of this minimax tree?*

Student answer: *42*

Analysis: … *What general significant problem with the size of search spaces does this illustrate?*

Student answer*: There are too many to calculate. This problem illustrates the number of possible choices AI problems have to deal with; it is a combinatorial explosion.*

Synthesis: *Rewrite the following replacing the underlined part with the appropriate pronoun: Ho regalato I quaderni a Paolo.*

Student answer: *Glieli ho regalati*.

Evaluation: *For each of the following pairs of classes, state whether or not it would be appropriate to relate them by inheritance, and why. If not, what other sort of relationship would be appropriate? – Car and Wheel*

Student answer: *This one may be better as a composition instead. A car as an association with wheel, but a wheel can exist on its own without the car class.*

Text Clustering

Clustering is the process of grouping similar objects together. A measurement of similarity, or distance, is used to assign objects within a set into subsets or clusters. Clustering is used in other fields such as Bioinformatics (Heyer et al 1999), finding nearest neighbours of a document (Buckley & Lewitt 1985), and for the organisation of search engine results (Zamir et al 1997).

Clustering offers a number of benefits in HCC assessment. The examples used here are free text student responses to assessment questions. Similar work at Manchester using the ABC system is looking at diagram responses (Tselonis et al 2005). Clustering similar answers together can help the human marker, as it provides a review mechanism to check that marking is consistent, and potentially offers a basis for rapid formative feedback.

The simplest form of text clustering is based on keywords, which may be specified in advance or (according to the HCC approach) expanded during the marking process. This has proved useful in some cases (as shown below), but is not a general solution. In this paper we concentrate mainly on the consequences of clustering the complete texts of short answers.

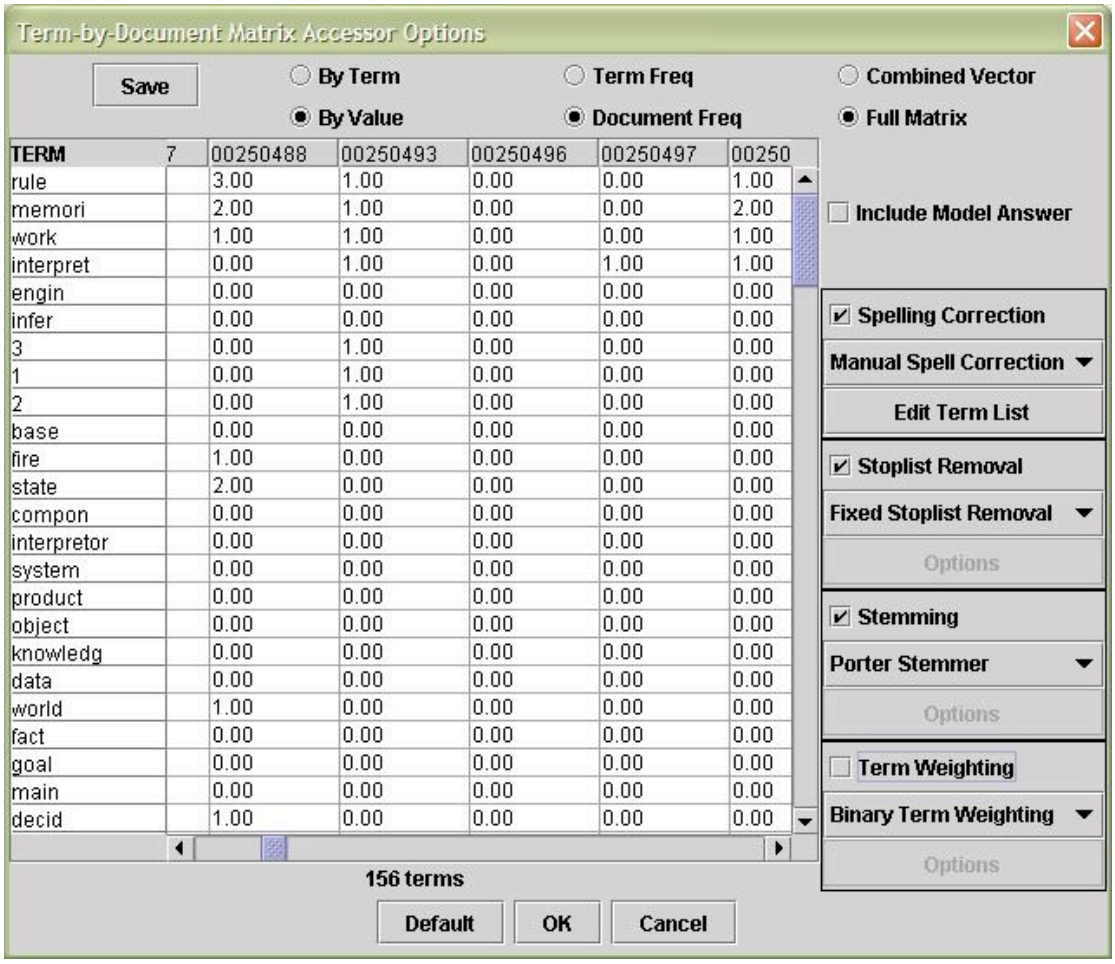
Clustering offers a tradeoff: the larger the clusters, the more fewer there are to process, but the less similarity there is between answers within a cluster. For formative applications we may be able to live with some inaccuracy in order to be able to give rapid feedback per cluster. In the summative case very high standards of accuracy are required if the students are to have confidence in the assessment software and procedures.

Lightweight Clustering Techniques

A commonly used measure of similarity from the field of Information Retrieval is the Vector Space Model (Salton 1971). Documents are expressed as vectors within a multi-dimensional space. The similarity between two documents is calculated as the distance between their respective vectors.

This clustering process can be broken down into a number of distinct steps, which have been implemented within a prototype extension of the ABC marking tool. The first step is the creation of a *term-by-document matrix*, a list of terms (words) and a count of the number of times they appear in each answer (see Figure 1). Each column is a vector representing the term frequency counts of an individual answer. Several pre-processing steps can be performed on the matrix to improve performance. These include spelling correction, removal of stop words (commonly occurring words of little interest such as *"the"*), stemming (removal of affixes from a word to leave a common stem. e.g. *"interpreter"* is converted to *"interpret"* - Porter 1980) and applying different weights to terms, in our case binary.

The next step is to calculate the similarities between vectors. The simplest way is to take the Euclidean distance between vectors. However this does not normalise vectors for length, and so the measure commonly used is the cosine of the angle between two vectors. This gives a range between 0.0 and 1.0, where a value of 0.0 indicates two answers that share nothing in common, and a value of 1.0 indicates two answers that are identical after pre-processing. This similarity measure can then be used to cluster the answers.



TERM	7	00250488	00250493	00250496	00250497	00250
rule		3.00	1.00	0.00	0.00	1.00
memori		2.00	1.00	0.00	0.00	2.00
work		1.00	1.00	0.00	0.00	1.00
interpret		0.00	1.00	0.00	1.00	1.00
engin		0.00	0.00	0.00	0.00	0.00
infer		0.00	0.00	0.00	0.00	0.00
3		0.00	1.00	0.00	0.00	0.00
1		0.00	1.00	0.00	0.00	0.00
2		0.00	1.00	0.00	0.00	0.00
base		0.00	0.00	0.00	0.00	0.00
fire		1.00	0.00	0.00	0.00	0.00
state		2.00	0.00	0.00	0.00	0.00
compon		0.00	0.00	0.00	0.00	0.00
interpretor		0.00	0.00	0.00	0.00	0.00
system		0.00	0.00	0.00	0.00	0.00
product		0.00	0.00	0.00	0.00	0.00
object		0.00	0.00	0.00	0.00	0.00
knowledg		0.00	0.00	0.00	0.00	0.00
data		0.00	0.00	0.00	0.00	0.00
world		1.00	0.00	0.00	0.00	0.00
fact		0.00	0.00	0.00	0.00	0.00
goal		0.00	0.00	0.00	0.00	0.00
main		0.00	0.00	0.00	0.00	0.00
decid		1.00	0.00	0.00	0.00	0.00

Figure 1: A Term by Document Matrix

Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering (Jain et al 1999) starts with each object forming a separate cluster. The process then follows these steps.

1. Find the two most similar clusters, A and B.
2. Combine A and B into one cluster.
3. Repeat until a designated stop point.

The ultimate end point is a single cluster that contains all answers. This is uninformative. One of our most interesting questions is how to determine the most effective stop point for a given question for a given purpose, bearing in mind the speed / accuracy trade-off discussed above.

In the initial state it is straightforward to calculate similarity between clusters, as they each contain only one item. Similarity between clusters containing multiple answers is more complex. “Average linkage” (the mean of the distance between all elements within cluster A and cluster B) is commonly used as a measure.

Within Cluster Similarity is a measure of how similar answers are to each other within any given cluster. Average Within Cluster Similarity is a measure

of how good the clusters are overall. A value approaching 1.0 indicates answers within each cluster are highly similar to each other.

Experimental Design

The data used here comes mostly from first year undergraduate summative examinations in Artificial Intelligence in the School of Computer Science (although ABC assessments have been run in a variety of subject areas, including Italian, Linguistics, and Pharmacy). All answers shown here have been marked by a human assessor.

Similarity between answers was calculated using the Vector Space Model as outlined above. The clustering algorithm was run to each of three termination points, which we believe (on the basis of experience) can produce useful clusters. *Optimal* termination points will vary among questions and assessment modalities, further reinforcing the tenet of HCC that some control must reside with the human marker.

The first termination point is to take the last clustering step when the Average Within Cluster Similarity value is equal to 1.0, indicating that all answers within each cluster are identical after pre-processing. The second is to cluster to a value of Average Within Cluster Similarity of 0.95. At this point answers within a cluster are not functionally identical to each other, but should still be reasonably similar. The third is to examine clustering from an efficiency aspect, considering how much effort could be saved for the human marker if marking by cluster were to be safe. For this we took a point when the number of clusters is 50% of the initial number of answers.

Examples

Experience in marking reveals three categories of question and answer: those where we can mark fairly consistently by cluster, those where marking by cluster is unsafe but reviewing marks by cluster is valuable, and those where clustering buys us little or nothing.

Answers where we can consistently mark by cluster

This knowledge-level question responds well to clustering:

CS141204 q1.1a. *In the "Hector's World" lab, conflict resolution is handled by "salience". Name two other conflict resolution strategies which can be used in production systems*.

Model answer: *Any two of rule ordering, specificity, recency, random.*
NB priority is not acceptable, as it is a synonym for salience.

Partial analysis at the limit of Average Within Cluster Similarity 1.0:

Cluster 1 (“Specificity”, “Random”): 13 answers, Mark = 2
Cluster 2 (“Specificity”, “Rule Ordering”): 8 answers, Mark = 2
Cluster 3 (“Specificity”, “Source File Ordering”): 6 Answers
    Mark = 2: 5 answers
    Mark = 1: 1 answer (“Specification” – error of stemming)
Cluster 4 (“Source File Ordering”, “Random”): 5 answers, Mark = 2
Cluster 5 (“Specificity”, “Priority”): 5 answers
    Mark = 2: 2 answers
    Mark = 1: 3 answers
…
Outliers = 67 answers

The anomaly in Cluster 5 is due to human error by the marker. The version of the ABC marking tool used for this exam did not yet incorporate clustering: had it done so, this mistake would have been avoided. As shown in column 4 of Table 1 in the Appendix, when clustering is continued to the point where the number of clusters is half the number of answers, 4% of the answers have marks different from the rest of their cluster – an acceptable level of accuracy for some types of assessment, and certainly for formative feedback by cluster..

Further clustering improves efficiency, but at a corresponding cost to accuracy. Answers missing correct terms, or with incorrect terms, are merged with correct answers if the clustering process is taken too far.

The fact that clustering collapses word-order can provide useful generalisations, as it does for this question. For another knowledge level question,

CS141205 q3.1a: *The CS1412 "Hector's World" lab uses the programming environment JESS. What does "JESS" stand for?*

Model answer: *Java Expert System Shell*

some students answered “Java Expert Shell System”. These were clustered with “Java Expert System Shell”, and were marked as correct by the human.[1] And BL181104A Q 1.1 ``What single measurement would you make to confirm that an individual is anaemic?'' returned, as its fourth largest cluster, 13 minor variants on “haemoglobin concentration in the blood”, comprising 11 distinct text strings which had been correctly collapsed by pre-processing (see Figure 2). (We will see below, however, that there are other questions for which word order information about the answers is needed.)

[1] (As they were, compared to such outliers as “Java Encapsulated System Software”, “Java emulator simulator system”, or “Java Expressions Structurated System”.

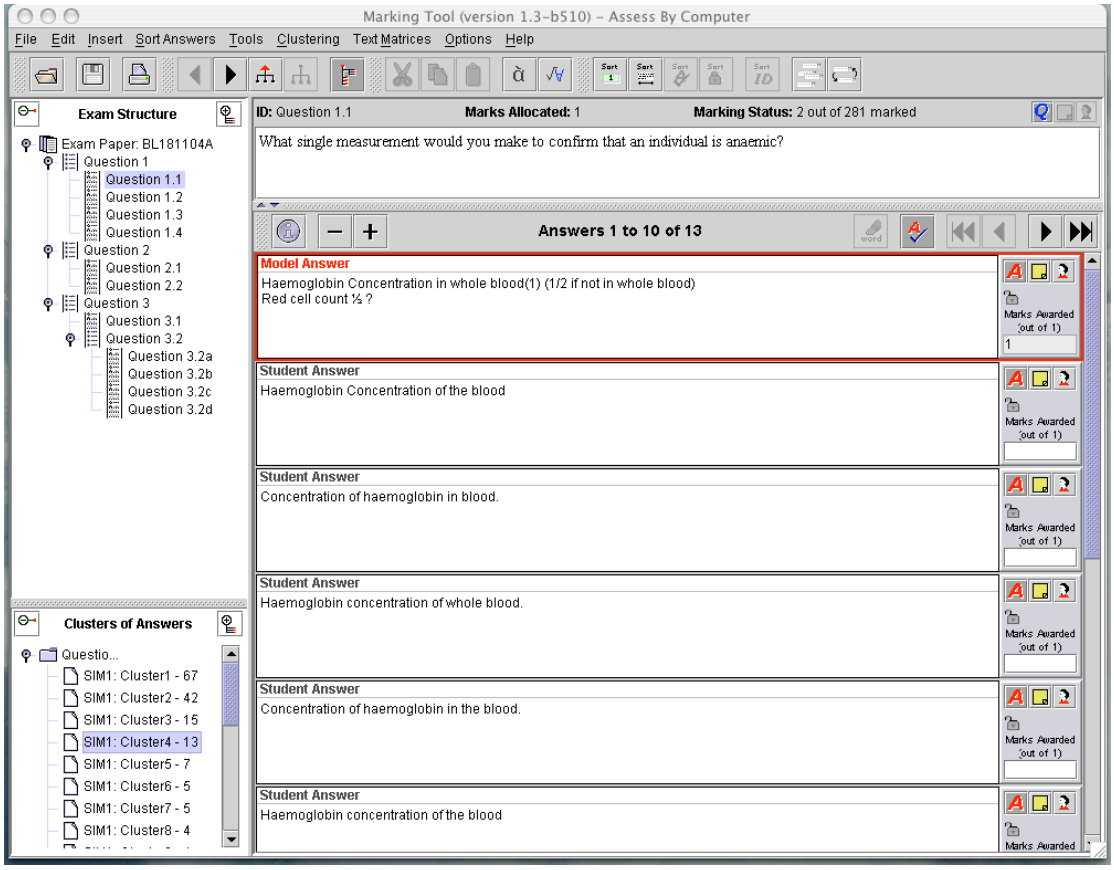


Figure 2: *anaemia*, Agglomerative Hierarchical clustering

Clustering used as a review

The following example shows a type of question which is more difficult to mark:

CS141205 q1.2: *A cricket ball used in a day-night match is white. What problem does this cause for semantic networks? Give another example of the same problem (but* ***not*** *the example featured in the lectures).*

Model answer: *Exceptions to default inheritance. Anything sensible except penguins as non-flying birds, since that was the primary lecture example*.

Answers with a high degree of similarity for the first part of the question might have different responses to the second; or vice versa. The question also asks for an original example. As a result the answers are highly variable. So are the marks awarded to answers within a cluster (see column 6 of Table 1: 22% of answers have marks anomalous for their cluster). Clustering is most useful for the second part, identifying similar examples (especially those students who ignored the question and used penguins as an example).

At Average Within Cluster Similarity 0.95 (91 answers, 73 clusters):

Cluster 1, 4 answers: (non-flying birds, 3 penguins, 1 ostrich)
Cluster 2, 3 answers: (3 wheeled cars)
Cluster 3, 3 answers: (non-flying birds, 2 penguin, 1 chicken)
…
Outliers, 60 answers (also includes 3 wheeled cars, ostriches)

Here clustering by keyword comes into its own (see Figure 3). Answers using the word "multiple" demonstrated a predicted common misunderstanding and were awarded, at most, one mark for a good example.[2] Any answer containing "exception" was awarded full marks unless it also contained the word "penguin".

Keyword="multiple": 13 answers
Mark=0: 9 answers (misunderstanding)
Mark=1: 3 answers (good example, first part wrong)
Mark=2: 1 answer (see footnote)
Keyword="exception" & NOT "penguin": 30 answers
Mark=1: 3 answers (bad examples, 1 chicken)
Mark=2: 27 answers

While marking by cluster is dangerous for this type of question, clustering is still of some benefit in allowing a human user to review their marking judgments, and may offer a useful basis for per-cluster formative feedback.

[2] With one exception: "A traditional cricket ball is red. If the semantic network has defined a cricket ball it as red, then multiple hierarchy will be needed for a white the ball to define a different type of ball with colour white.

Another example of this would be Manchester United players IsA Footballer, Manchester United players have Skill: High, Intelligence: High. Phil Neville IsA Manchester United player. Skill: Low, Intelligence: Low. It does not fit the normal semantic network for a Manchester United player."

The first part of this answer is wrong; but the human marker (a Stockport County supporter) awarded a bonus mark for the originality of the second.

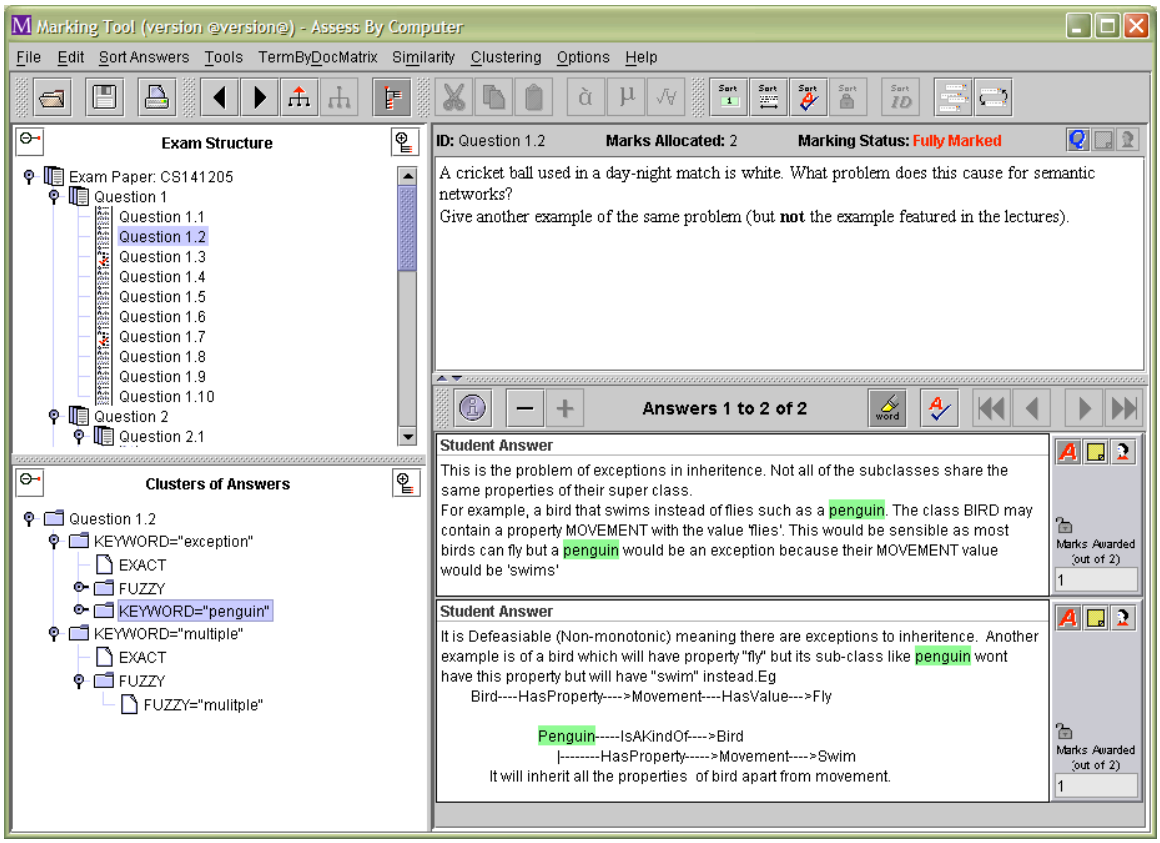


Figure 3: *penguins*, clustering by keyword

Where clustering can't take us

Thus far we have considered question types where correctness was determined by the content of the answer. Any vector-based approach must fail where correctness is a meta-level property of the *structure* of an answer. Consider this example:

CS141205 Q1.1. *A traditional cricket ball is red. Express this fact as a very simple semantic network, in two different ways.*

Model answer: *cricket ball --<has property>-- colour --<value>-- red*
cricket ball --<has colour>-- red

Clustering 93 answers into 63 clusters (with average within-cluster similarity 0.97), we find this, clustered with four correct answers:

Cricket Ball → HAS-COLOUR -→ Red
Cricket Ball:
Colour Red

The clustering is based on the word "has-colour". This answer is wrong because the two "networks" are not sufficiently different from each other (as can be seen by comparison with the model answer). It is inherently impossible

for any clustering technique based purely on word occurrences to detect this. More sophisticated techniques would be more expensive and more fragile.

As with the penguins, the keyword manager can be useful here – all the answers including (variants on) “has-value” received full marks. This reinforces our position that light-weight techniques manipulated by human intelligence offer a viable and valuable strategy for CAA.

A less significant weakness of vector-based clustering approaches is that word order is not taken into account. In some cases this is acceptable or even advantageous, as shown above. However consider the following:

CS141203 q1.1a: *… What conflict resolution strategy would you use to force rule 2 to fire? What strategy would you use to force rule 3 to fire?*

Model answer: *Rule 2 – specificity. Rule 3 - priority*

The answers are short, but clustering shows a much lower correlation with human judgement than for the previously analysed questions. This is largely because the incorrect answers “priority, specificity” were clustered with answers using the same words in the correct order.

In this case, a setter familiar with clustering in marking would have set the question as two separate “leaves”. However, for language translation exercises, word order within a sentence is critical. Thus in a diagnostic test in Italian (IT1200a Q.4.7), the answer “le abbiamo incontrato” received one mark and “abbiamo l’incontrato” none.

Conclusion

Experiments comparing relatively small differences in similarity metrics and clustering algorithms have so far proved inconclusive, yielding only small differences in the correlation of clustering results to human marking judgements. We expect further experiments with a wider range of language engineering techniques to improve performance, especially for slightly longer text answers.

Differences in types of question had much larger effects. Although clustering is most effective on very short answers, this is far from the whole story. Answers where word order is significant, or where original examples are required, for instance, need treating differently from ones where this is not the case.

Clustering is a good tool for thinking about the nature of questions and answers as well as improving speed and consistency of marking in some cases. It clearly has great potential for reducing the workload, and hence improving the timeliness, involved in formative feedback. The examples shown in this paper support our general view that fully automatic summative assessment of constructed answers is generally unsafe in view of What Students Really Say.

Short text answer questions do have paedogogic value if used thoughtfully, and are amenable to light-weight processing in an HCC framework. Analysing answer data (especially marked answer data) can bring some surprising insights into paedogogic aspects of seemingly simple questions.

References

Bloom, B., M. Englehart, E. Furst, W. Hill, & D. Krathwohl (1956) **Taxonomy of educational objectives: the classification of educational goals. Handbook I: Cognitive Domain**. New York: Longmans.

Buckley, C. & A. Lewitt (1985) *Optimization of inverted vector searches*. Proc. 8th Annual SIGIR Conference on Research and Development in Information Retrieval, pp. 97-110.

Carlson, A. & S. Tanimoto (2005) *Text Classification Rule Induction in the Presence of Domain-Specific Expression Forms*. Mixed Language Explanations in Learning Environments (XLANG), AIED (Artificial Intelligence in Education) 2005), Amsterdam.

Heyer, L.J., S. Kruglyak, & S. Yooseph (1999) *Exploring Expression Data: Identification and Analysis of Coexpressed Genes*. **Genome Research** 9:1106-1115.

Jain, A.K., M. N. Murty, & P.J. Flynn (1999) Data *Clustering: A Review*. **ACM Computing Surveys** 31:3

Lütticke, R. (2005) *Graphic and NLP Based Assessment of Knowledge about Semantic Networks*. XLANG.

Porter, M.F. (1980). *An algorithm for suffix stripping.* **Program** 14(3), 130-137

Pulman, S. & J.Z. Sukkarieh (2005) *Automatic Short Answer Marking*. Proceedings of Association for Computational Linguistics.

Salton, G. (1971) *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice Hall, Englewood Cliffs, NJ.

Sargeant J., M.M. Wood & S.M. Anderson (2004) *A human-computer collaborative approach to the marking of free text answers*. 8th International Conference on Computer Aided Assessment, Loughborough University, Loughborough, UK, pp.361-370.

Tselonis, C., J. Sargeant & M.M. Wood (2005) *Diagram matching for human-computer collaborative assessment*. 9th International Conference on CAA, Loughborough, UK. pp. 441-456.

Wiemer-Hastings, P., E. Arnott, & D. Allbritton (2005) *Initial Results and Mixed Directions for Research Methods Tutor*. XLANG.

Wood, M.M., J. Sargeant & C. Jones (2005) *What Students Really Say*. 9th International Conference on CAA, Loughborough, UK. pp. 317-327

Zamir, O., O. Etzioni, O. Madani, & R.M. Karp (1997) *Fast and intuitive clustering of web documents*. KDD-97 pp. 287-290.

Appendix

		Average Linkage, Cluster to 50%					
		CS141203		CS141204		CS141205	
	Q1.1a	Q1.1c	Q1.3d	Q1.1a	Q1.1	Q1.2	Q3.1a
No. of Answers	153	151	137	116	93	91	27
No. of Terms	115	130	386	119	86	510	19
No. of Clusters	76	75	68	58	46	45	13
No. of Outliers	55	70	49	45	26	28	12
Avg Within Cluster Similarity	0.9695	1.0000	0.9182	0.9839	0.9455	0.8625	1.0000
% Marking Reduction	50%	50%	50%	50%	51%	51%	52%
Avg SD of Marks	0.4450	0.0000	0.2100	0.1295	0.3160	0.6340	0.0000
% Anomalous Marks	8%	0%	6%	4%	11%	22%	0%

Table 1. Cluster analysis of answers to questions across three years of the Artificial Intelligence Fundamentals course CS1412. Clusters were created using an Agglomerative Hierarchical algorithm with an Average Linkage metric used to measure distance between Clusters. In each case the algorithm was run to create a number of clusters equal to 50% of the number of answers.

Number of Answers is the total number of answers in the set and *Number of Terms* is the number of terms (words) in the Term-by-Document Matrix. This provides a measure of how variable or diverse the answers are.

Number of Clusters is the total number of clusters at the termination point while the *Number of Outliers* is the number that contain just one answer.

The *Average Within Cluster Similarity* is a measure of how similar answers are within a cluster, i.e. the average number of terms which documents in a cluster share.

% Marking Reduction indicates how much clustering has reduced the number of individual answers a human marker would have to see if they trusted the clustering completely. Whether such trust would be justified is indicated by the

Average SD of marks within Clusters, the overall standard deviation between marks within each cluster, an indication of how well the clustering correlates with the actual human marking.

The Number of Anomalous Marked Answers is another measure of that correlation, the number of answers that were not awarded the same mark as the others within a cluster, while *% Anomalous Marked Answers* gives the same value corrected for the overall number of answers in the cluster.

CS141203 Q1.1a: Here are three rules I might use in deciding how to get to work in the morning:

1. IF weather fine THEN take train
2. IF weather fine AND cold THEN take train and wear woolly hat
3. IF train drivers on strike THEN take bus

What conflict resolution strategy would you use to force rule 2 to fire? What strategy would you use to force rule 3 to fire?

Model answer: Rule 2 - specificity
Rule 3 – priority

CS141203 Q1.1c What are the three components of a production system?

Model answer: Working memory
Rule memory
Interpreter

CS141203 Q1.3d: In artificial intelligence, what is the "Turing test"?

Model answer: A simple test for "intelligence". A tester has to distinguish between communication with a human and with a machine. If they cannot tell the difference, or think the machine is a human, then the machine has passed the test

CS141204 Q.1.1a: In the "Hector's World" lab, conflict resolution is handled by "salience". Name two other conflict resolution strategies which can be used in production systems.

Model answer: Any two of rule ordering, specificity, recency, random.
NB priority is not acceptable, as it is a synonym for salience.

CS141205 Q1.1: A traditional cricket ball is red. Express this fact as a very simple semantic network, in two different ways.

Model answer: cricket ball --<has property>-- colour --<value>-- red
cricket ball --<has colour>-- red

CS141205 Q1.2: A cricket ball used in a day-night match is white. What problem does this cause for semantic networks?
Give another example of the same problem (but **not** the example featured in the lectures).

Model answer: Exceptions to default inheritance.
Anything sensible except penguins as non-flying birds, since that was the primary lecture example

CS141205 Q3.1a: The CS1412 "Hector's World" lab uses the programming environment JESS. What does "JESS" stand for?

Model answer: Java Expert System Shell