ISSN 1744-1986

Technical Report N° 2004/27

Comparing machine graded diagrams with human markers: some observations

Pete Thomas

16th December 2004

Department of Computing
Faculty of Mathematics and Computing
The Open University
Walton Hall,
Milton Keynes
MK7 6AA
United Kingdom

http://computing.open.ac.uk

The Open University

Comparing machine graded diagrams with human markers: some observations

Pete Thomas

Department of Computing

Open University

Abstract

In this paper we examine the performance of an automatic (machine) grading algorithm for entity-relationship (E-R) diagrams by comparing it with human generated marks for a set of student answers to an assignment question. Using a variety of statistical tests it is shown that the performance of the automatic marker is very close to that of the human markers: the Pearson correlation coefficient is 0.964 (significant at the 0,01 level, 2-tailed, N=26) and the Kendall tau-b correlation coefficient is 0.919 (significant at the 0.01 level, 2-tailed, N=26). The investigation revealed deficiencies in both the machine and human markers. There is prima-facie evidence that the orientation (shape) of a diagram may influence humans to award lower marks than they should.

Introduction

As part of our ongoing research into machine understanding of imprecise diagrams [9], we have been investigating the particular problem of automatically grading student answers to assignment questions that require Entity-Relationship (E-R) diagrams to be drawn [13]. We have developed an automatic E-R diagram marker that is based on the results earlier work on the automatic grading of textual answers to assignments [11]. The diagram marking tool conforms to a 5-stage architecture described in [9 and 10].

The effectiveness of the automatic marker has been judged against the criterion of how well the automatically generated grades for a set of student drawings compare with marks generated by experts in the field. In this paper we report on some of the issues that these comparisons have raised about the nature of grading, both human and machine-based.

In our most recent experiments we have looked at two examples taken from the assessment of a database course. The first experiment was performed on student answers to an assignment early on in the course where the question was tightly specified and where we expected the majority of students to perform well. In this scenario we expected the automatic marker to perform well. In the second experiment, we took student answers to a question posed in the final assignment of the course which was much more open-ended. Here we expected there to be a much wider diversity of answers and consequently a much poorer response from the marking tool. It turned out that our expectations about the performance of the marking tool were not met: the results for the second experiment were better than for the first. This unexpected result caused us to look in depth at the behaviour of both the automatic marker and the human markers, and the way in which we evaluated the effectiveness of the automatic marker.

The paper is structured as follows. The next section discusses how tutors approach the marking of diagrams in our educational context and compares this with the approach used in the automatic marker. The third section compares the initial set of marks produced by the human and machine markers and identifies where the major discrepancies occurred. The fourth section looks in detail at the discrepancies and shows how a closer match between human and machine generated marks was obtained. The paper concludes with a discussion of the findings and sets out the direction for future work.

The marking processes

In this section we shall describe, briefly, how the marking of the E-R diagrams was performed (a) by the human markers and (b) by the marking tool.

Human marking

In our environment (distance education), we typically have large numbers of students (in excess of one thousand) on each presentation of the database course. Student assignments are marked and commented

upon by a team of tutors – experts in the database field with distance teaching experience. Over 40 tutors are employed on this course.

To ensure consistency of performance between tutors, two quality assurance procedures are in place. First, each tutor is provided with a set of 'Tutor Notes' containing both a sample solution and a comprehensive marking guide which explains how the marking scheme is to be applied. If a tutor is faced with a student answer which does not match the sample solution, they are expected to use their professional judgement and to assign marks within the guidelines set out in the Tutor Notes.

Second, a process known as monitoring is invoked in which the work of a tutor is examined by another expert, the monitor, whose role is to check both the marking accuracy and the usefulness of the tutor's feedback to the student. Problems identified by the monitor in the marks awarded by a tutor can result either in an immediate re-grading of the student answer or a request for the answer to be re-marked.

In our experiments we monitored the marking of all tutors and adjusted the marks for those answers where discrepancies were found between the tutor's and monitor's marks. These adjusted marks were then used as the definitive measure of correctness of the students' answers. It is an interesting aside to note that in all but one case where an adjustment was made, the adjustment was of a single mark. However, in the remaining case, the adjustment was 5 marks (the maximum mark for the question was 25); we shall return to this later.

Machine marking

The algorithm embodied in the automatic marker compares a student diagram with the sample solution and derives a measure of similarity (a value in the range 0 to 1). To derive the similarity measure, both the sample solution and the student diagram are decomposed into their constituent relationships and the 'best' match between the two sets of relationships is determined. This process matches pairs of relationships, one from the student answer and one from the sample solution, and assigns a similarity measure to each pair. The similarity between the two diagrams is based on an aggregation of the similarities of the relationship-pairs.

Finally, the mark scheme is applied (effectively, 6 marks were available for each correct relationship and 1 mark for the correct identification of the entities – this exactly mirrored the instructions given to the human markers). This can be viewed as a shallow approach to determining similarities.

Initial comparison of marks

The marks for the automated tool were compared with the moderated human marks as follows. On the first experiment there were 26 student answers in the marking sample (all were from student volunteers). The first comparison used simple descriptive statistics and the results are shown in Table 1.

N=26	Mean	St. Dev	Range
Human	21.27	3.436	13 – 25
Machine	22.08	2.497	15 – 25

Table 1 Descriptive statistical tests

The descriptive statistics show that the machine marker is the more lenient marker by one mark per student, on average. There is a major discrepancy in the standard deviation with the spread of human marks being much greater than that of the machine marker, a result confirmed by the range of marks awarded. This was not an unexpected result because our experiments with the automatic marking of text have consistently shown the machine marker to have a narrower spread than the human markers.

The next test looks at correlations. Table 2 shows the results with three tests of correlation.

	Correlation	Significance Level
Pearson	0.939	0.01, 2-tailed
Spearman	0.953	0.01, 2-tailed
Kendall	0.889	0.01, 2-tailed

Table 2 Correlation tests

The Pearson correlation coefficient is a (parametric) measure of the closeness of the two sets of marks, whereas Spearman's rho coefficient is a non-parametric test which measures how closely the two sets of

marks rank the students. In both cases, the results are extremely good, showing very close correlation. Kendall's tau-b statistic is another measure of rank ordering which corrects for ties (which there are in this data), and again shows good correlation.

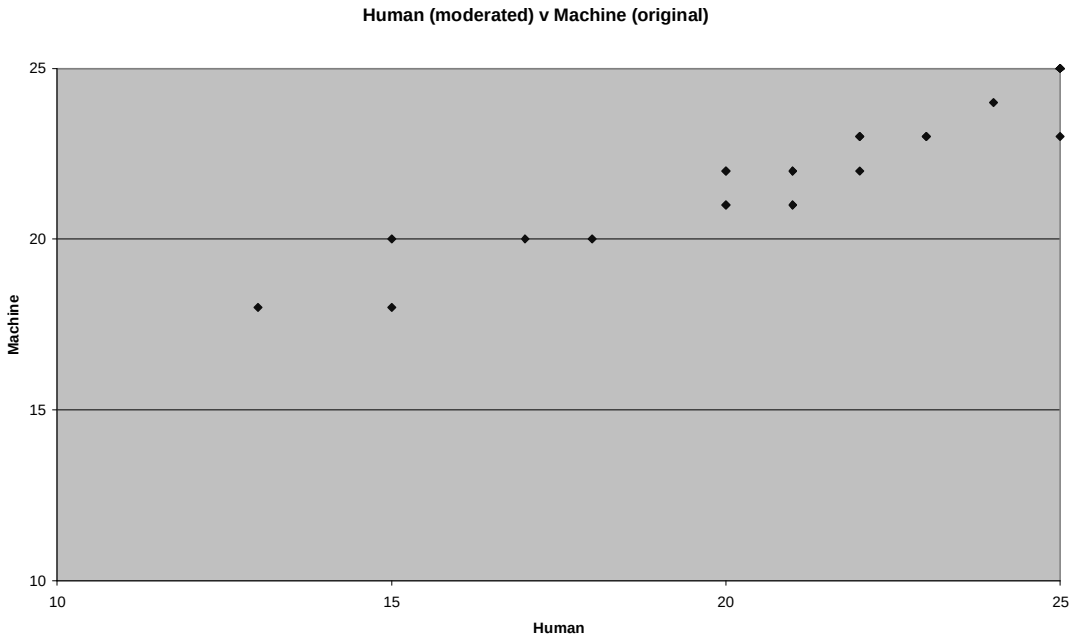


Figure 1 Scatter plot of human and machine marks

Figure 1 is a scatter plot of the two sets of marks. It clearly shows the linearity of the data, with greater variability at the lower end. More revealing is the slope of the regression line, 0.683. If there were an exact match between the machine and human marks, the slope of this line should be 1.

Thus, as expected, the results indicate that the machine marker works well at the upper range of marks, but is less accurate at the lower end. However, the rank correlations indicate that the machine marker compares well with the human markers in ordering the students' performances on this question.

Investigating low-end performance

In an attempt to discover whether it would be possible to improve the low-end performance of the machine marker we examined the three scripts on which the machine marker performed least well. Table 4 shows the human and machine marks for these three scripts.

Student	A	B	C
Human mark	13	15	17
Machine mark	18	18	20

Table 4 The three most poorly correlated marks

The answer from student 'A' showed that there was rule in the marking scheme used by the human markers that had not been incorporated into the automatic marker. Figure 2 shows the sample solution in which it can be seen that one of the relationships, Introduces, is recursive (it relates the entity Member to itself).

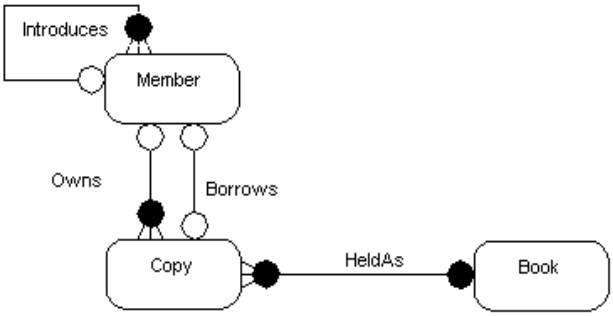


Figure 2 The sample solution

The rule can be paraphrased as “if there is no recursive relationship on the Member entity type, do not award any marks for this part of the diagram”. The machine marker tries to maximise the mark it awards by always trying to find a match in a student’s diagram for each relationship in the sample solution. When the machine marker was amended to incorporate this new rule, the machine generated mark for Student A was 14, which compares favourably with the human mark of 13, although we felt that the human mark was somewhat severe.

At first sight, student A’s answer, shown in Figure 3, appears to have a different structure to the sample solution shown in Figure 2.

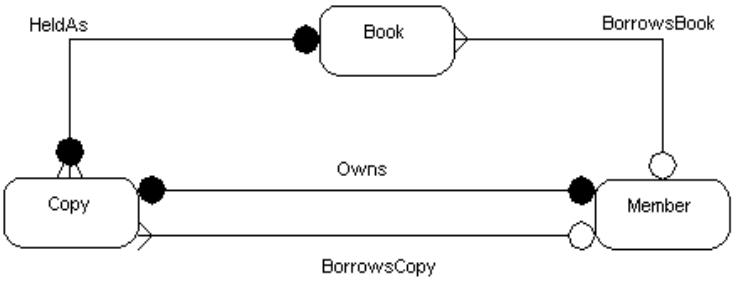


Figure 3 The answer from Student A

We wondered, therefore, whether the way in which the student had drawn the diagram might have influenced the human markers (both the tutor and the monitor). That is, the markers might have been expecting a particular ‘shape’ of diagram and, when faced with a different shape, were inclined to award fewer marks than the diagram deserved.

To test this hypothesis we looked at the answers provided by students B and C: their diagrams are shown in Figures 4 and 5.

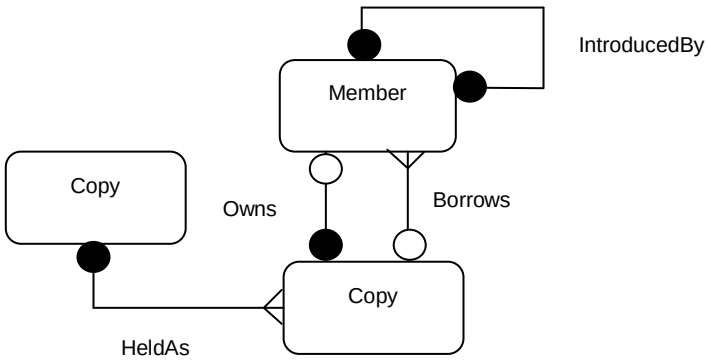


Figure 4 The answer from student B

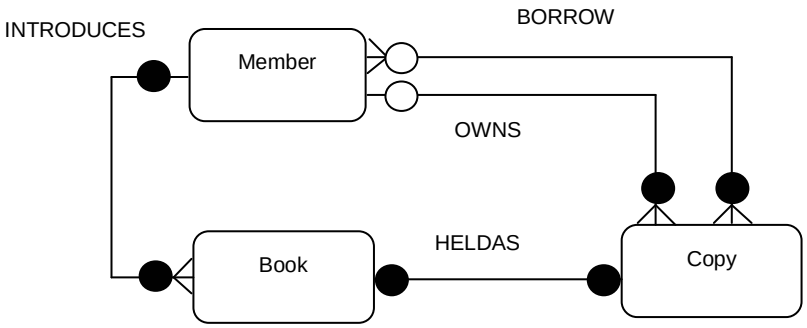


Figure 5 The answer from student C

It certainly seems that these answers do have a different shape to the sample solution. It is worth noting that the diagram provided by student A was drawn using a software tool. However, student B's diagram was hand drawn and hence more difficult to interpret than our machine drawn version shown in Figure 4. Student C's diagram was a mixture of machine and hand drawing!

To test the theory that the shape of a diagram might have influenced the human markers, we redrew the three diagrams in a form that more closely corresponded to the shape of the sample solution and asked a human marker to mark the result. For example, the redrawn Student A's diagram is shown in Figure 6.

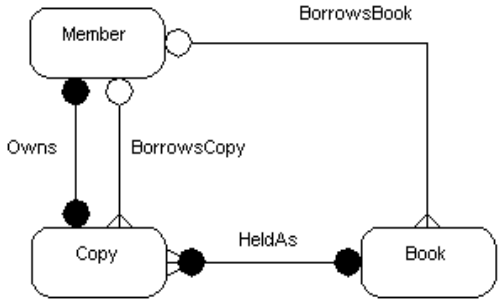


Figure 6 Student A's diagram redrawn

The resulting human marks for students A, B and C (and the revised machine mark for student A) are shown in Table 5.

Student	A	B	C
Human mark	15	18	20
Machine mark	14	18	20

Table 5 Revised human marks

The correspondence between the human and machine mark has certainly improved for these examples, providing prima-facie evidence that humans are influenced by the shape of the diagrams.

Putting all this together, the statistical comparison between the revised human marks and the revised machine marks are as follows.

N=26	Mean	St. Dev	Range
Human	21.38	3.008	15 – 25
Machine	22.00	2.953	14 – 25

Table 6 Descriptive statistical tests for revised marks

The means are still of the same magnitude but are closer with the machine being the less severe marker. The standard deviations are now much closer with the human marks showing the wider spread.

	Correlation	Significance Level
Pearson	0.964	0.01, 2-tailed
Spearman	0.969	0.01, 2-tailed
Kendall	0.919	0.01, 2-tailed

Table 7 Correlation tests for revised marks

The three correlation coefficients have all improved, particularly Kendall's tau-b. The scatter plot for the revised marks is shown in Figure 7.

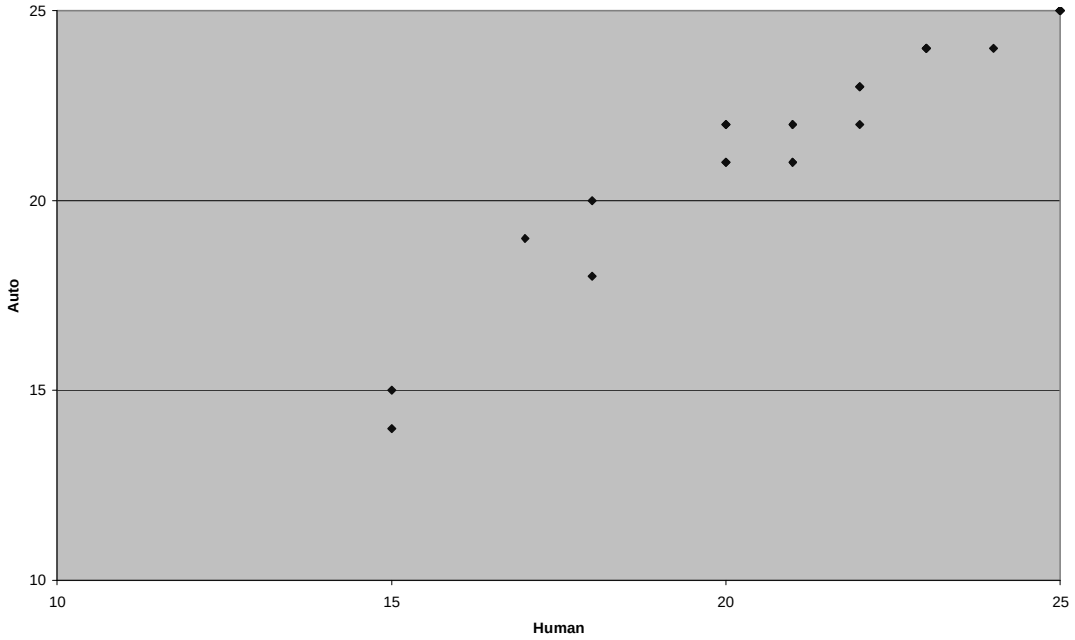


Figure 7 Scatter plot of revised human and machine marks

The slope of the revised regression line is 0.946259 which confirms that the results are well-correlated across the range of marks.

Conclusions

We conclude that the (revised) automatic marker gives marks that correlate very well with the (revised) human marks across the range of marks. It is particularly pleasing to note the improvement at the lower end of the marks range where we found evidence that the original discrepancy was not simply a function of the machine algorithm, but that there were inaccuracies in the human marks.

We have found some evidence that the inaccuracies in the human marks at the low end might be due to the orientation (shape) of the student diagrams.

Furthermore, the closeness of the two sets of marks poses the question of whether the human markers were employing a similarly shallow approach to marking as the machine algorithm.

Clearly, this investigation needs to look at larger datasets, but the results so far are encouraging. The results suggest that building an accurate automatic marker in this domain is possible. An initial use for such a marker, given its consistent approach, is to act as a 'second marker' identifying potential discrepancies in human generated marks and thereby identifying student answers that should be remarked.

The machine marking algorithm reported upon here has been incorporated into a 'revision tool' that contains a number of questions of the same type as the assignments on the database course. The tool enables the

student to draw E-R diagrams as solutions to the questions and have the tool provide feedback on their efforts based on the findings of the automatic marker.

Future work

We have already performed a second experiment on data obtained from student answers to a more open ended question. The results support the conclusions drawn here, but we need to complete the statistical analysis. This second experiment has identified two directions in which the automatic marker might be improved.

First, in areas where students have freedom to choose their own identifiers, the identification of synonyms becomes a significant problem. We have implemented an approach to this problem which needs to be rigorously tested.

Second, the automatic marking algorithm is too lenient, always trying to maximise the score. This leads to matches between relationships that are implausible in the sense that a human marker would disregard any suggestion that the relationships were similar. Therefore, we are looking for mechanisms that will identify 'plausible' matches.

In the area of E-R diagrams there are circumstances where two quite different (usually small) diagrams can be regarded as equivalent. For example, in data analysis, we often replace a many-to-may relationship by an equivalent pair of one-to-many relationships. We believe that we can use this idea to identify missing relationships in a student diagram. There are several such well-known patterns (we refer to them as clichés) that can be exploited here. However, this requires the automatic marker to have a synthesis phase in which individual relationships are built up into more complex structures. Work on identifying clichés in a diagram is underway.

As we said above, we need to work with larger datasets, and this will be done once we have collected the necessary data from the next presentation of the course. Whilst we have a drawing tool that helps with the construction of E-R diagrams, the default is for students to hand draw diagrams which poses another problem – accurately transforming hand drawn diagrams into machine readable diagrams.

References

[1] Anderson, M., McCartney, R. (2003) Diagram processing: Computing with diagrams, Artificial Intelligence 145 (1-2) 181-226.

[2] Blackwell, A, Marriott, K., Shimojima, A (eds) (2004) *Diagrammatic Representation and Inference*. Lecture Notes in Artificial Intelligence (**2980**). Springer.

[3] Bowers, D.S., (200n)

[4] Burstein, J., Leacock, C., Swartz, R. (2001). Automated Evaluation of Essays and Short Answers. Fifth International Computer Assisted Assessment Conference, Loughborough University, UK.

[5] EAP (2002). Electronic Assessment Project. http://mcs.open.ac.uk/eap

[6] Jamnik, M (1998) Automating Diagrammatic Proofs of Arithmetic Arguments, PhD thesis, University of Edinburgh.

[7] Manning, C.D., Schutze, H. (1999) Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

[8] Marriott, K. and Meyer, B. (eds) (1998) Visual Language Theory. Springer-Verlag, New York.

[9] Smith, N, Thomas, P.G. and Waugh, K. (2004) Interpreting Imprecise Diagrams. *Diagrammatic Representation and Inference: Third International Conference, Diagrams 2004, Cambridge, UK, March 22-24, 2004. Proceedings.* K. M. Alan Blackwell, Atsushi Shimojima. Heidelberg, Springer-Verlag. **2980:** 239 - 241Thomas, P. (2003) Grading Diagrams Automatically, Technical Report, Computing Department, Open University, UK, TR2003/01.

[10] Thomas, P. (2004) Drawing Diagrams in Online Examinations. In Proceedings of the 8th Annual International Conference on CAA, Loughborough University, Loughborough, UK.

[11] Thomas, P.G., Price, B., Paine, C. and Richards, M. (2002) Remote Electronic Examinations: an architecture for their production, presentation and grading. British Journal of Educational Technology33 (5).

[12] Tsintsifas A., (2002), A Framework for the Computer Based Assessment of Diagram-Based Coursework' Ph.D. Thesis, Computer Science Department, University of Nottingham, UK

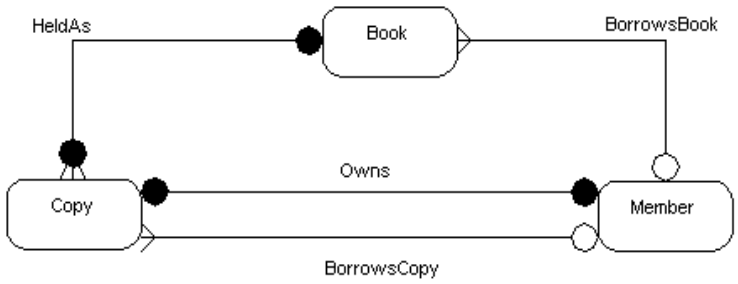
[13] Waugh, K.G., Smith, N. and Thomas, P.G. (2004) Toward the automated assessment of entity-relationship diagrams. In Teaching, Learning and assessment in Databases, LTSN-ICS.

[14] Whittington, D. and H. Hunt (1999). Approaches to the Computerised Assessment of Free Text Responses. 3rd International Conference on Computer Assisted Assessment, Loughborough University, Loughborough, UK

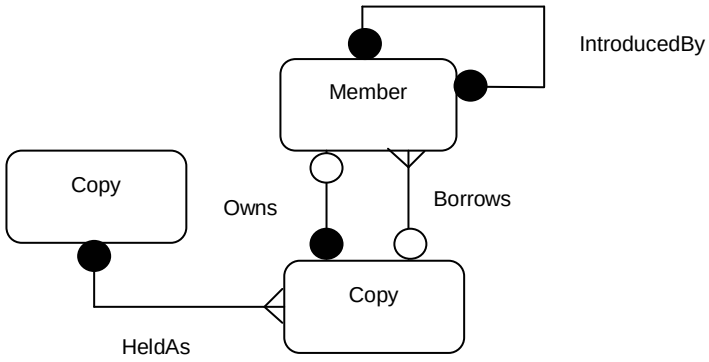
Appendix (Not for inclusion in the final version)

The student answers – as drawn in submitted document

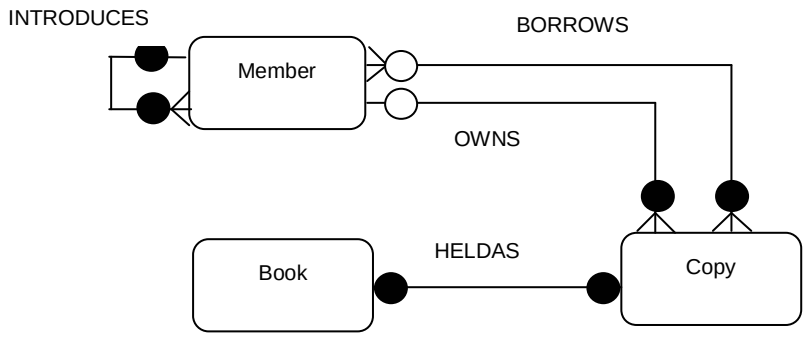
Student A



Student B

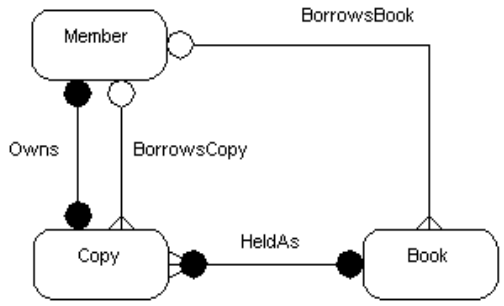


Student C



Redrawn diagrams

Student A



Student B