ISSN 1744-1986

The Open University

Technical Report N^{o} 2006/16

An approach to the automatic grading of imprecise diagrams

Pete Thomas
Neil Smith
Kevin Waugh

1st December 2006

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

http://computing.open.ac.uk

An approach to the automatic grading of imprecise diagrams

Pete Thomas, Neil Smith, Kevin Waugh

Department of Computing, Open University, Milton Keynes, UK, MK7 6AA

***e-mail*:** p.g.thomas@open.ac.uk

Abstract

In this paper we describe our approach to the grading (marking) of graph-based diagrams for instance those produced by students in the area of Entity-Relationship (E-R) diagrams. This is an application of a more general problem and is based on our framework for diagram understanding. We believe that techniques of NLP commonly used in understanding text have equivalents in the understanding of diagrams and this paper shows how these techniques have been incorporated into a grading tool. The accuracy of the marking tool has been measured in two small-scale trials and the positive results from those trials are presented here. Our approach naturally allows the provision of diagrammatic feedback on student answers to E-R diagramming questions which has been incorporated into a tool for practising E-R diagramming.

1 Introduction

People communicate with each other using a variety of modes, such as speech, text, diagrams, gestures, and so on. In the particular domain of assignments in which students are required to answer factual questions there is a tendency for them to use diagrams to help explain technical details. In our experiments on the automatic grading of examinations in Computing (Thomas 2003) we found that, even in questions that expected textual responses (essays), students would include diagrams to support their arguments. Clearly, when grading such work automatically, we need to extract meaning from a diagram and award it a grade.

We believe that there are many similarities between text and diagrams. In both modes, there are basic elements that can be composed into meaningful units at many levels of abstraction such as, for text, words, phrases, sentences and paragraphs and, for diagrams, lines, boxes, and regions. In both modes, meaning is encoded within the meaningful units and their composition. Therefore, we propose that many of the techniques used to interpret text can also be used (suitably modified) to interpret diagrams.

A domain specific diagram, of which an entity-relationship (E-R) diagram is an example, consists of features that obey domain specific rules of structure and content. It is not an arbitrary combination of drawing primitives such as lines, arcs and text: to do so would miss the inherent meaning in the structure. A practitioner working in a specific domain would recognise well-formed diagrams as carrying meaning within that domain. Furthermore they would interpret some imprecise diagrams as similar to well-formed diagrams and thereby infer the diagrams' intended meanings. Similar notations may be used in different domains, but their interpretation may be different.

Information-containing diagrams, like text, are regular entities with their own rules of well-formed-ness. It follows that diagrams, like text, can have their syntactic rules expressed by grammars or grammar-like formalisms (Marriott and Meyer 1998). However, this approach proved limited for text as the grammars were either too limited to allow the full range of natural language or produced too many alternative parses of a piece of language. These problems were addressed by changing to a statistical approach for natural language understanding (Manning and Schutze 2002).

In contrast, most diagrammatic understanding systems are based around developing visual programming languages (Bottoni and Costagliola 2002, Marriott et al.1998) and assessing the status of diagrams in mathematical proofs; see for example (Jamnik, Bundy and Green 2002). We term diagrams

in these domains as precise; the diagrams represent a formal language and grammar-based approaches are adequate.

Much work has been done on the automatic interpretation of speech and text (Jurafsky and Martin 2000), but to the best of our knowledge little has been done on the automatic interpretation of the informal diagrams drawn for every-day communication. We describe such diagrams as *imprecise* because, in general, they are incomplete, items are malformed and there can be extraneous items (Smith, Thomas and Waugh 2004). Such 'real-life' diagrams deviate from the formal grammar that should define them. In such a case, the meaning of the diagram remains clear to human viewers, while grammar-based approaches are not sufficient to parse it.

In this present work, we assume that both precise and imprecise diagrams will be produced within the rules of a specific domain, and that diagrams will be drawn with the intention that they carry meaningful content. That is, the diagrams are intended to be interpretable within the given domain. Thus, the diagrams we process are assumed to have some intended meaning, and that meaning is encoded using rules of diagram composition from that domain and also from general diagram rules.

Our overall approach to interpreting diagrams is a framework consisting of five stages which we have named: segmentation, assimilation, identification, aggregation and interpretation. The first two, segmentation and assimilation, together translate a raster-based input into a set of diagrammatic primitives (such as lines, boxes and text). The identification stage uses domain knowledge to identify low-level, domain specific features which we call minimal meaningful units (MMUs). MMUs are aggregated into higher-level, abstract features. Finally, the diagram is interpreted to produce meaningful results (Waugh, Thomas and Smith 2004).

The resolution of imprecise diagrams depends on general diagram knowledge (knowledge of how generic diagrams are drawn), domain specific diagrammatic knowledge (how this type of diagram is drawn) and domain specific knowledge (what interpretations are valid within this domain). Inference mechanisms using the identifiable content of the diagram and the domain specific and general knowledge should be able to infer either a plausible interpretation or a repair of the diagram. The resolution may not result in either a complete interpretation of the meaning of the diagram, or a complete repair for the diagram but may provide some acceptable information for the application.

In this paper we examine one area of application of this approach: the marking (or, grading) of diagrams and the provision of feedback to the originator of the diagram. This work has its origins in our attempts to provide instant feedback in online examinations (Thomas, Price, Paine and Richards 2002, Thomas 2003, Thomas 2004).

The paper is structured as follows. In Section 2 we review the current state of the general diagram reasoning literature showing where our work fits in. In Section 3 we describe the problem domain by examining a model of teaching that our work is intended to support. Our basic approach to understanding imprecise diagrams and E-R diagrams in particular, is given in Section 4. Sections 5 and 6 describe the development of a marking tool through two small-scale trials, and show how ideas from NLP have been incorporated into the marking algorithm. Results from the trials are also presented there. Section 7 shows how our approach has naturally led to feedback in diagrammatic form. The final substantive section briefly describes a teaching tool for revising the construction of E-R diagrams using the tools and techniques described in the paper. The paper concludes with a description of planned future work and a summary of our findings.

2 Diagrammatic reasoning

A fundamental question for diagrammatic reasoning is to define what a diagram is, and how it is different from other forms of representations. (Chandrarasekaran 2002) asserts that diagrams are inherently non-propositional. For instance, propositions can be stated regarding the existence of diagrammatic elements at different levels of abstraction, their absolute and relative positions, colour, shape (at many levels of detail), and so forth. This fundamentally non-propositional nature of diagrams forces diagrammatic reasoning systems (which are fundamentally propositional) to pre-select the propositions they will extract from a diagram. This implies that diagrammatic reasoning cannot be truly

generic, but must start with a description of the diagrammatic domain which guides the extraction of propositions.

The set of propositions extracted from a concrete diagram forms a domain-specific abstract diagram, normally represented as a graph of entities and relations (Andries; Engels and Rekers, 1998, Howse, Molina, Shin and Taylor 2002). A typical use of this technique is to represent a diagram as a graph by reifing the entities in the diagram and their geometric relations, and expressing them as nodes and arcs in the graph that is the abstract diagram (Sowa1984, Bottoni and Costagliola 2002, Fish and McCartney 2002, Flower and Howse 2002, Swoboda 2002, Donlon, and Forbus, 1999). This is essentially the process of identifying the base syntactic units and minimal meaningful units in the diagram.

This abstraction becomes easier when the underlying concrete diagram is itself the depiction of a graph embedded in a plane. Many of the diagrams used in computing, such as Entity-Relationship diagrams, UML class diagrams, and Petri nets, are of these sorts. In these cases, the mapping from concrete diagram to abstract diagram is greatly simplified, as the geometric information in the diagram can mostly be ignored. However, the geometry of the entities in the diagram may be significant when the existence of malformed MMUs must be inferred.

3 Teaching Model

One of our aims is to incorporate automatic marking into a computer-based learning system as illustrated by the learning model shown in Figure 1.

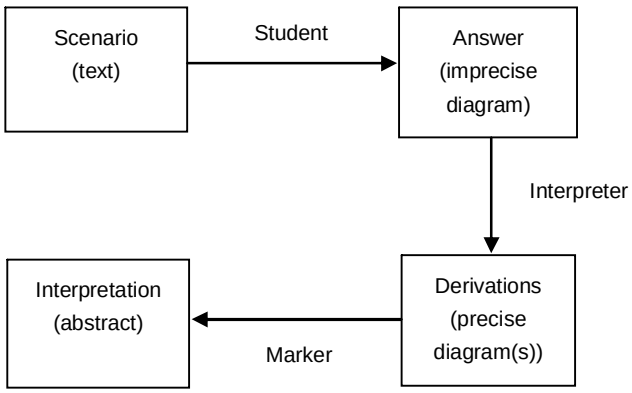

Fig. 1. A computer-based learning model

A typical question is a description of a scenario from which the student is required to produce a diagram. In general, such a diagram will be imprecise. The *Interpreter* analyses the imprecise diagram to provide possibly many derivations (that is representations as precise diagrams). The *Marker* selects one of the derivations and computes a grade and provides feedback in the form of one or more annotated precise diagrams.

For example, suppose that a scenario based on the activities of clients at a fitness centre were posed, with a sample solution shown as the E-R diagram in Figure 2.

The sample solution has six entity types (represented by boxes) and six relationships (represented by straight lines decorated with ‘crows feet’ and small circles). The entity types and relationships have names composed from terms appearing in the scenario.

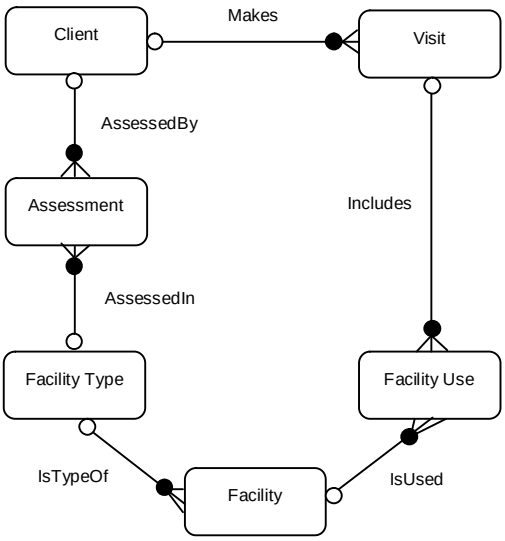

Fig. 2. A sample E-R diagram

Figure 3 illustrates output from the marker which shows a typical student diagram overlaying the specimen solution (the decorations on the relationships have been suppressed for clarity). Solid lines indicate where the student's answer matches the specimen solution. The dashed lines indicate those parts of the specimen solution that have not been matched by the analyser. (In this environment, matching need not be exact; the analyser finds a match that is 'good enough'.)

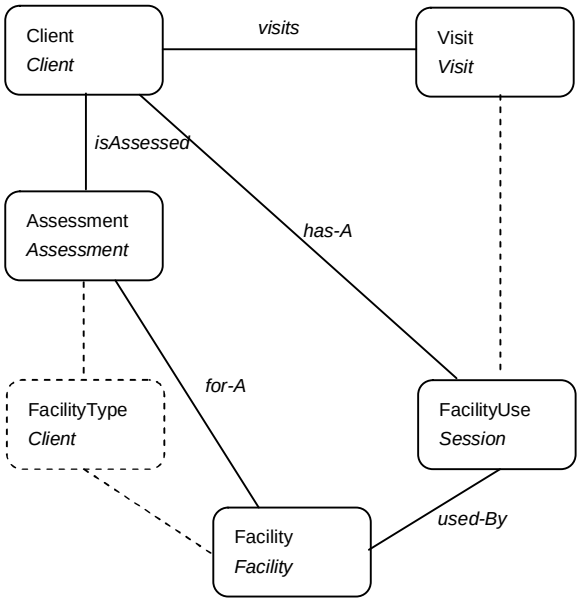

Fig. 3. A student diagram overlaying the sample diagram

Thus, in this example, the student has failed to identify (a) an entity type (*FacilityType*) and its associated relationships, and (b) the relationship between *Visit* and *FacilityUse*. The student has also included an additional relationship (*has-A*). Figure 3 also shows that the student tended to use names that were quite different to those used in the specimen solution. For example, the student used the name *Session* which the analyser has matched with the entity type name *FacilityUse* on the sample diagram.

When the student's diagram was marked by a tutor, a mark of 7 (out of 12) was awarded (later confirmed as an accurate mark by a monitor). The analyser was slightly more generous and awarded 8 marks.

4 Automatic Marking of E-R diagrams

4.1 The basic approach

The basis of the automatic marking algorithm is a search for the 'best' match between the MMUs in a student's answer and MMUs in the sample solution. The 'best' match is based on maximising the sum of the similarities between pairs of similar MMUs.

In E-R diagrams we have two types of MMU: relationships and entity types. A relationship involves two entity types and a set of attributes. The attributes take the form of a name and a set of adornments – crows feet, representing the degree of the relationship, and circles, representing the participation of entity types in the relationship.

In our first attempt, the similarity between two relationships is a function of *all* the attributes of the relationships which we have termed *full similarity*. The algorithm attempts to match every relationship in a student answer with a relationship in the sample solution. If there are more relationships in the student answer than the sample solution, a subset of the answer relationships, equal in size to the set of relationships in the sample solution, is matched and a correspondence which maximises the mark is chosen. If there are fewer relationships in the student answer than the sample solution, a correspondence between relationships is again chosen to maximise the mark.

The similarity between the names of entity types (and names of relationship) assumes that only names with a simple structure will be encountered and therefore uses stemming and edit-distance to determine a suitable measure of similarity.

4.2 Evaluating the marking tool

The marking tool has been tested in two trials based on student drawings from the assessment of a final-year database course. The first trial was performed with student answers to part of an assignment presented early in the course where the question was tightly specified. We expected the majority of students to perform well on this question and for the automatic marker to perform well also. The second trial was based on a question in the final course assignment which was considerably more open-ended than the first. Here we expected a wider diversity of student answers and a much poorer response from the automatic grader.

In our distance education environment, we typically have large numbers of students on each presentation of the database course (circa one thousand). There are several assignments throughout the course and student answers are graded and commented upon by a team of *tutors*, experts in the database field with distance teaching experience. Around 50 tutors are employed on this course and each one is responsible for a group of approximately 20 students.

To ensure consistency of performance between tutors, we use two main quality assurance procedures. First, each tutor is provided with a copy of a set of *Tutor Notes* containing both a sample solution and a comprehensive marking guide for each question on each assignment. The marking guide is a comprehensive set of instructions on how to apply the given marking scheme. When tutors are faced with student answers which do not match the sample solution, they are expected to use their professional judgement to assign marks within the guidelines set out in the Tutor Notes.

The second procedure, which we call *monitoring*, is where the work of a tutor is reviewed by another independent expert (a monitor) who checks both the grading accuracy and the appropriateness of the tutor's feedback to the student. Problems identified by a monitor can result in either an immediate re-grading of the student answer or a request for the answer to be completely re-marked by the tutor.

In these trials, we compared the grading of the automatic marker to the grades awarded by the tutors (after monitoring). The adjusted marks were used as the definitive measure of correctness of the students' answers.

The marks available for a correct answer were 25 in the first trial and 12 in the second trial. The difference in the number of marks available in the two trials is not significant; it simply represents the relative importance of the task within a complete assignment.

5 The First Trial

5.1 Results from the first trial

The grades from the automated tool were compared with the moderated human marks in three ways as follows. In the first trial there were 26 student answers in the marking sample (we were restricted to student volunteers). The first comparison used simple descriptive statistics and the results are shown in Table 1.

Table 1 Descriptive statistics (max mark 25)

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

The descriptive statistics show that the automatic (machine) grader is the more lenient marker by one mark per student, on average. There is a major difference 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 (Thomas 2003) have consistently shown the automatic grader to provide a narrower spread of grades than the human markers.

The next test looked at correlations. Table 2 shows the results with two common tests of correlation.

Table 2 Correlation tests

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

The Pearson correlation coefficient is a (parametric) measure of the closeness of the two sets of marks, and shows very close correlation. Kendall's tau-b statistic is a measure of rank ordering (preferred to the Spearman rho statistic because it corrects for ties in the data, which there are in this data set), and again shows good correlation.

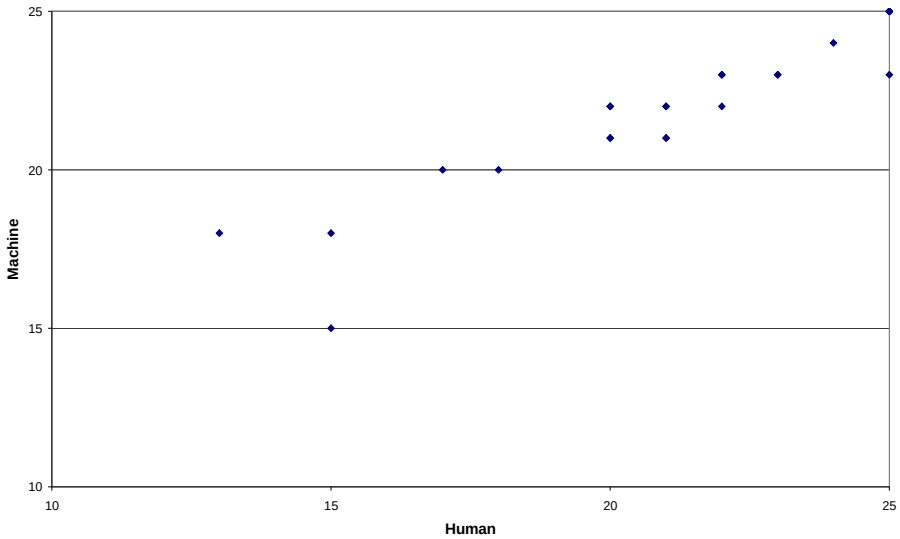

Fig. 4. Scatter-plot of human and machine marks

The third test is shown in Figure 4 where a scatter-plot of the two sets of marks is shown. The linearity of the data is quite apparent, with greater variability at the lower end. More revealing, however, is the slope of the regression line, equal to 0.683. If there were an exact match between the machine and human grades, the slope of this line would be 1. This simple test shows that the 'low-end' performance of the automatic grader was poor.

Thus, the results indicate that the machine grader 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. Further details can be found in (Thomas 2004b).

5.2 Investigating low-end performance

The results of the first trial caused us to investigate more closely the reasons for the poor low-end performance of the automatic grader. We examined the three student answers on which the automatic marker performed least well. Table 3 shows the human and machine grades for these answers.

Table 3 The three most poorly correlated marks

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

The answer from student *A* quickly revealed that there was rule associated with recursive relationships in the marking scheme used by the human markers that had not been incorporated into the automatic marker (the relationship *Introduces*, shown in Figure 5, is recursive). The inclusion of the additional rule removed the grading discrepancy.

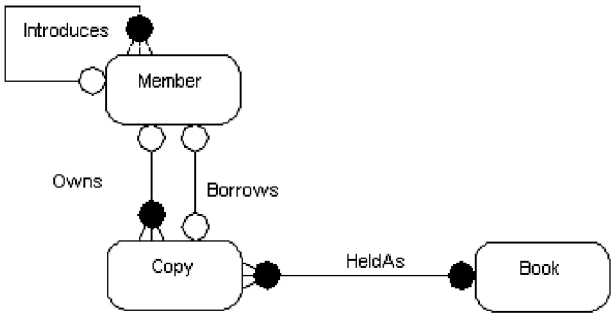

Fig. 5. The sample solution

The investigation of the answers from students *B* and *C*, led us to a conjecture. We felt that the grades awarded by the automatic marker were quite reasonable and that the tutors (and the monitors) had under-valued the answers. For example, the drawing in Figures 5 appears to have different overall structure, or 'shape', to the sample solution in Figure 5, even though, at the individual relationship level, they have much in common – it is the way that they are laid out that differs.

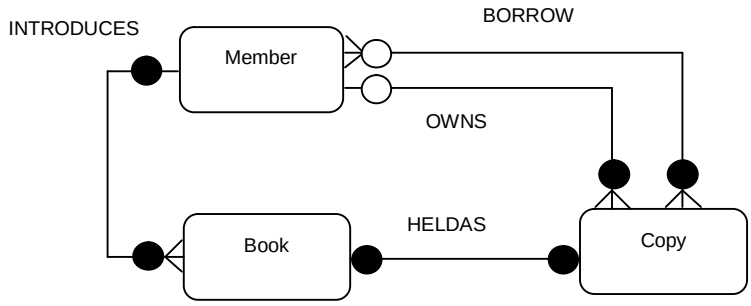

Fig.6. The answer from student *C*

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 humans might have been expecting a particular layout of the diagram – as presented in the specimen solution – so that, when faced with a different layout, they were inclined to award fewer marks than the diagram deserved. In E-R diagrams, the positioning of elements is irrelevant, it is the connections between entity types that matters. It is worth noting that the diagram provided by student *B* was completely hand drawn and Student *C*'s diagram was a mixture of machine and hand drawing which may have made the diagrams even more difficult to interpret than our machine drawn copy shown in Figures 6!

To test this hypothesis that the layout of a diagram might have influenced the human markers, we redrew the three diagrams (using a software tool) in a form that more closely corresponded to the layout of the sample solution and asked an independent human marker to mark the results. The outcome was that the human grades matched those of the automatic marker for students *B* and *C* identically. There was no change to the mark for student *A*.

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

Taking the revised marks into account, the statistical comparison between the revised human marks and the revised machine marks are as shown in Figures 4 and 5.

Table 4 Descriptive statistical tests for revised marks

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

The means are still of the same magnitude but they are closer with the machine continuing to be the less severe marker. The standard deviations are now much closer but the human marks still show a wider spread.

Table 5 Correlation tests for revised marks

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

Both correlation coefficients have improved particularly Kendall's tau-b. The slope of the revised regression line was 0.946259 which confirms that the results are well-correlated across the range of marks.

We concluded that the (revised) automatic marker gave marks that correlated very well with the (revised) human marks across the range of marks. It was 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. Furthermore, the closeness of the two sets of marks poses the intriguing question of whether the human markers were employing a similarly shallow approach to marking as the machine algorithm.

6 The second trial

The second trial, based on the data from the later assignment, included a number of different investigations. First, we looked at the correspondences between the relationships in a student answer with those in the sample solution as determined by the automatic marker, and compared these to the correspondences that a human marker had determined. Once again, there was good agreement at the upper end of the marking scale, but poor agreement at the lower end. That is, the use of full similarity resulted in some correspondences that were not considered reasonable by the human markers. We wanted another measure of 'best' correspondence which was more plausible to humans. We use the term *plausible* because, in low scoring answers, some of the relationships are incorrect or incomplete in

one way or another but a human will look for a match that is not unreasonable, that is, plausible. This led to an additional requirement for the automatic marker: not only must its marks be close to human generated marks, but also its relationship correspondence must also be close to that determined by the human markers.

Therefore, as a first attempt to obtain better relationship correspondences, we took a relationship similarity measure that was based solely on the names used in the relationships. We call this the *name similarity* and refer to the correspondence based on name similarity as *plausible matching.*

The dependence on names alone for determining correspondences brings into sharp focus the assumption made in the first trial that names would be simple and obtained from the provided scenario. However, in the second assignment, students were at liberty to choose their own names for relationship and entity types, so we expected considerable variation from one student to the next. Indeed, we expected names to be chosen that were very different to those in the sample solution although we did expect names that were related to the terms used in the question scenario.

In the second assignment, students were given a scenario based on a fitness club that had various types of fitness machines (called facilities in the scenario) that its members could use: see Figure 2. Members of the club had to be assessed before being authorised to use a particular facility. Furthermore, each member would have a membership card that would give entry to the club. The sample solution used names that were taken directly from the scenario, but students were not similarly constrained. Hence, the automatic marker needed to recognise synonyms that were specific to this scenario. For example, the notion of a fitness facility was on several occasions indicated in student answers by the use of the term *equipment* whereas the term *facility* was used in the sample solution. The terms *facility* and *equipment* are synonymous in normal usage and were detected as such using a thesaurus. However, the entity type *assessment* used in the sample solution appeared as *authorized* in several student solutions. Clearly, these words are not synonymous in normal usage, and it would appear unlikely that they could be automatically recognised as such in advance, using a thesaurus based approach.

Having examined the student answers, it became clear that in several cases students were using their own identifiers consistently and correctly. That is, the structure of (parts of) their diagrams corresponded well with the structure of the sample solution; only the identifiers differed. If, within a given context or environment, a student's identifier were replaced by the identifier in the same context in the sample solution, acceptable agreement would be obtained. Therefore, we implemented a rule to the effect that, if an entity type identifier occurred in a student answer in two or more relationships which corresponded closely with solution relationships within the same context (apart from this entity type identifier) probable synonyms had been found.

However, a constraint is placed on the use of this rule. If the newly found potential synonym, that is the identifier in the solution entity type, occurred elsewhere in the student answer, the names are *not* taken as synonyms for clearly the student answer contains two entity types with different names and because the student has distinguished between them they cannot be synonymous in this context.

A second synonym identification rule came from the observation that some students introduced entity types that could be traced to the given scenario but which were not required in the solution. Some of these entity types formed a one-to-one, mandatory relationship with expected entity types and therefore could be viewed as synonymous. If one of the entity type identifiers in such a one-to-one relationship occurs in the solution but the other does not, the entity type identifiers are considered synonyms. Provided that the entity which does not occur in the sample solution occurs in precisely one answer relationship, the two entity types in the answer are effectively interchanged (further investigation may show that it is possible to be less restrictive than this constraint).

Table 6 shows a selection of synonyms found by these rules (half the student answers had at least one occurrence of these synonyms and a few contained two examples). Note how a single name in the solution, such as *FacilityUse*, was represented by several, very different names in student answers.

Table 6 Synonyms identified by the automatic marker

Name in solution	Name in answer
Assessment	Authorised
Visit	ClientVisits
FacilityUse	Usage
FacilityUse	ActivationSlot
FacilityUse	Session
FacilityUse	Workout
FacilityType	Card
FacilityType	Staff

In most cases, this scheme worked well, identifying synonyms that were felt to be accurate. However, in the case of the name *FacilityType*, associating it with *Card* and *Staff* is incorrect. In fact, this latter example indicates a serious flaw in student thinking.

In many cases, entity names were constructed from the concatenation of two or three identifiers as in *ClientVisits*, *TypeOfFacility* and *IsTypeOf*. In some cases, as with *ClientVisits*, the synonym rules discussed above identified the correspondence. The remaining problem names were dealt with by viewing them as simple noun phrases – a noun qualified by one or more adjectives and determiners. In the similarity measure, the noun was weighted more than the adjectives and determiners. In addition, to recognize the similarity between, for example, *FacilityType* and *TypeOfFacility*, the latter was re-written as the phrase *Facility Type* by dropping the term *Of* and reversing the order of the remaining two terms. It is worth noting that this rule should be applied to entity type names but not to association names which ought to be modelled as verb phrases (if the rules for constructing E-R diagrams are adhered to).

With the additional rules switched on, the statistics relating to the performance of the automatic marker applied to the second assignment are given in Table 7.

Table 7 Correlations with all rules switched on

N=14	Human Marker	Automatic Marker
Mean	8.071	8.0
St. Dev	2.336	2.353
Pearson		0.938**
Kendall		0.837**
Slope		0.945

** Significant at the 0.01 level, 2-tailed, N=14

The effect of each of these rules on the overall statistics was quite small. However, their effect on individual student answers, particularly at the low-end of marks was more noticeable, with the correspondence between relationships closer to the human choices. The effectiveness of this approach encouraged us to examine the problem of providing feedback on the student answers as we describe in the next section.

7 Feedback

The primary feedback provided to a student is a grade. However, in our environment we expect tutors to provide substantial written feedback to indicate why a particular grade was awarded, to enable the student to discover why mistakes had been made (if appropriate) and to indicate how the student might correct their misunderstandings (if any). In the case of ER-diagrams, tutors often write over a student's work indicating erroneous relationships and drawing in missing or corrected relationships. Such corrections are normally accompanied by a textual comment (see Figure 7 for an example). However,

in cases where substantial corrections are applied, the result can be an almost illegible drawing. Nevertheless, the idea that the communication from tutor to student should be, partially at least, in diagrammatic form seems appropriate. Therefore, we began to investigate how the marking tool might provide feedback in diagrammatic form.

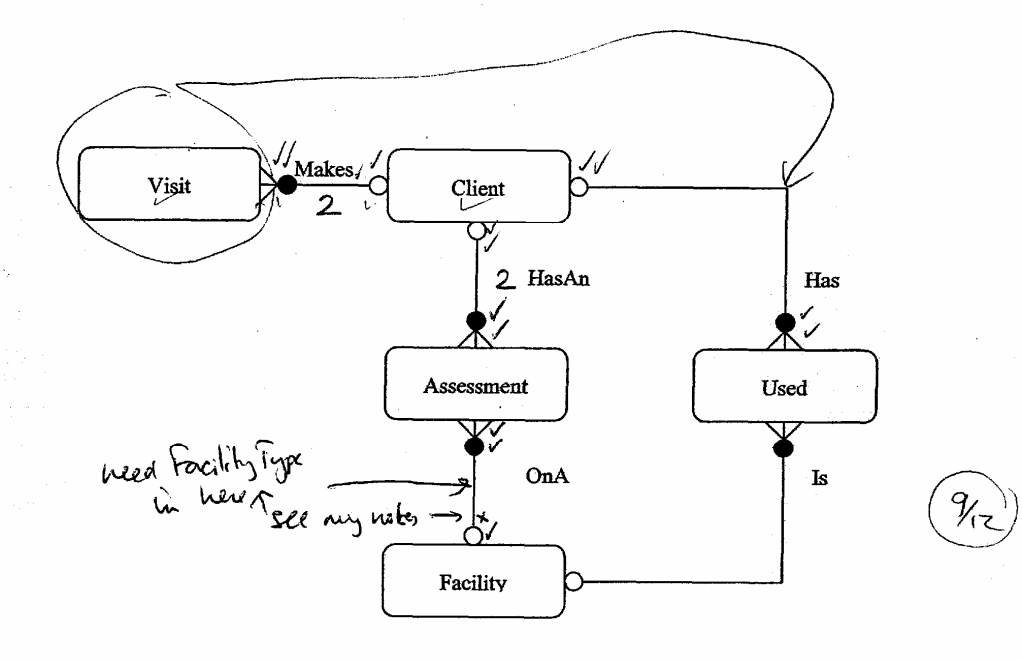

Fig. 7. Tutor comments on a student answer

The first, and most obvious, approach was to inform a student how the tool had analysed their answer diagram into separate relationships and to show the correspondences it had made with the relationships in the sample solution. Figure 8 shows an example of this type of feedback (the student's relationships are on the left and only a subset is shown).

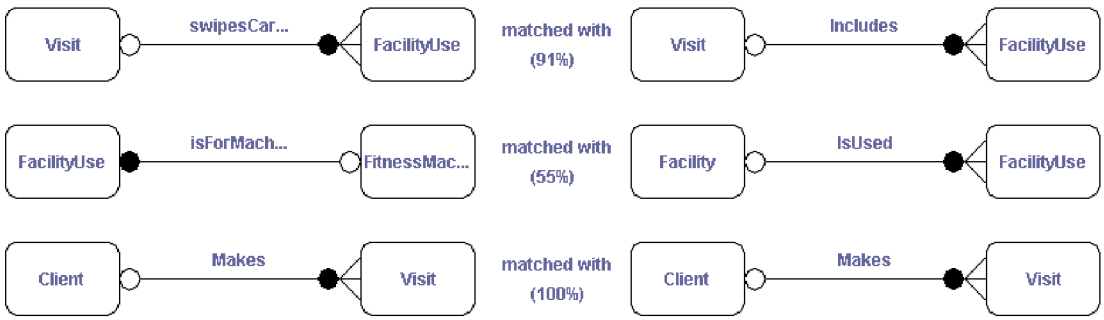

Fig. 8. Comparing relationships in a student's answer with those in the sample solution

In the first pair of matched relationships, the student's relationship named *swipesCardFor* has been matched with a relationship in the sample solution named *Includes*. The only difference between these two relationships is in their association name. The closeness of the match is displayed as a percentage (91%) based on the tool's similarity measure. The third matched pair in Figure 8 shows an exact match.

The example in the middle of Figure 8 illustrates a poorer match (55%) in which differences occur in: the association name, the degree of the relationship (one-to-one instead of many-to-one), and one of the entity type names (*FitnessMachine* compared with *Facility*). In fact, although one cannot tell from the diagram, the tool has deduced that *FitnessMachine* is a synonym for *Facility* which accounts for the slightly higher percentage correlation than, at first sight, might have been expected.

This form of feedback has limited use, particularly as we currently only show those matches with a correlation greater than 45% (the threshold for an 'acceptable' match) – to avoid giving potentially misleading information to students. This is a very localised view and it is not immediately obvious where, in a diagram taken as a whole, which parts of the student's answer are being matched with which parts of the sample solution. That is, there is no global or relative information between relationships. In an attempt to solve this problem, we wondered to what extent we might be able to show how well a student's answer – as a whole – corresponds with the sample solution – as a whole.

We solved this problem by aligning a student's answer with the sample solution and presenting the result as a diagram with the student's answer 'overlaid' on top of the sample solution. Figure 9 shows that, for the same student answer shown incompletely in Figure 8, a very good alignment was obtained.

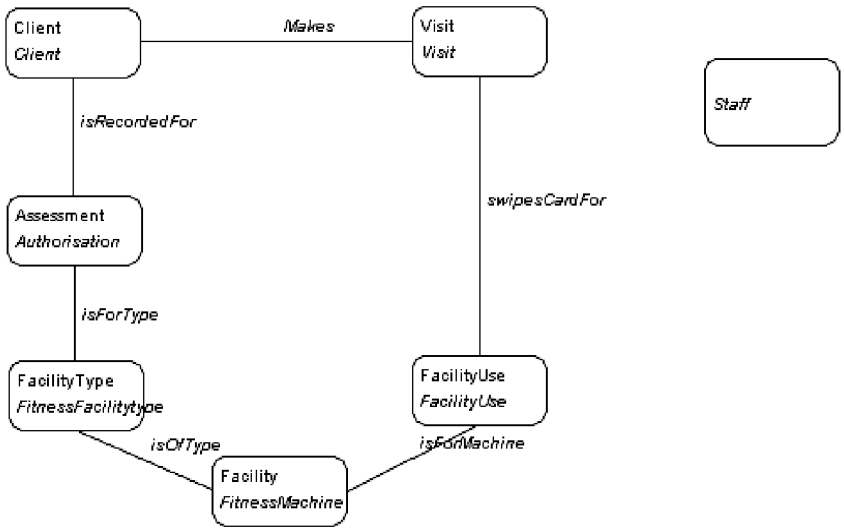

Fig. 9. A student's answer overlaying the sample solution

Each entity type in Figure 9 has two names: the top name is the one that appears in the sample solution and the lower name is the equivalent name in the student answer. The relationships shown as solid lines in Figure 9 are those that appeared in the student answer. The equivalent relationships in the sample solution have been hidden (overlaid) by the student answer relationships. For comparison, the sample solution is shown in Figure 2.

The entity type *Staff* shown on its own to the right of the main diagram in Figure 9 indicates an entity type occurring in the student answer that does not appear in the sample solution.

In this approach, the degree of the relationships and the participation conditions are not shown because of the difficulty of distinguishing those in the answer from those in the solution. They are, of course, shown in Figure 8. In this example, both the tutor and the marking tool awarded 10 (out of 12) marks for the answer.

Figure 3, above, shows an example in which there was a poorer correspondence between the student answer and the sample solution. This example illustrates how a missing entity type (*FacilityType*) and a missing relationship (between *Visit* and *FacilityUse*) appear on the overlay – as objects with dashed outlines. The tutor's comments, hand-drawn on the diagram (see Figure 7), consisted of (1) an arrow to show that the *Visit* entity type would have been better placed between *Client* and *Used*, and (2) an annotated arrow to show the position of the missing entity type *FacilityType*. The same information is conveyed by the overlay but with improved clarity. In this example, the tutor awarded 9 marks but the tool awarded only 8 (out of 12).

8 A Revision Tool

The ability to provide the kind of feedback illustrated above has been incorporated into a software 'revision tool' in which students are presented with a collection of typical assessment questions on the construction of E-R diagrams. The revision tool contains a diagramming tool (Thomas 2004a) with

which students draw their answers. The revision tool then marks an answer and provides feedback in terms of a mark and a sequence of relationship diagrams of the form shown in Figure 8. In addition, the tool allows the student to view an interactive version of the sample solution. That is, clicking on a specific part of the solution causes the tool to highlight those parts of the question which relate to the chosen part of the solution. Figure 10 shows the user interface of the revision tool.

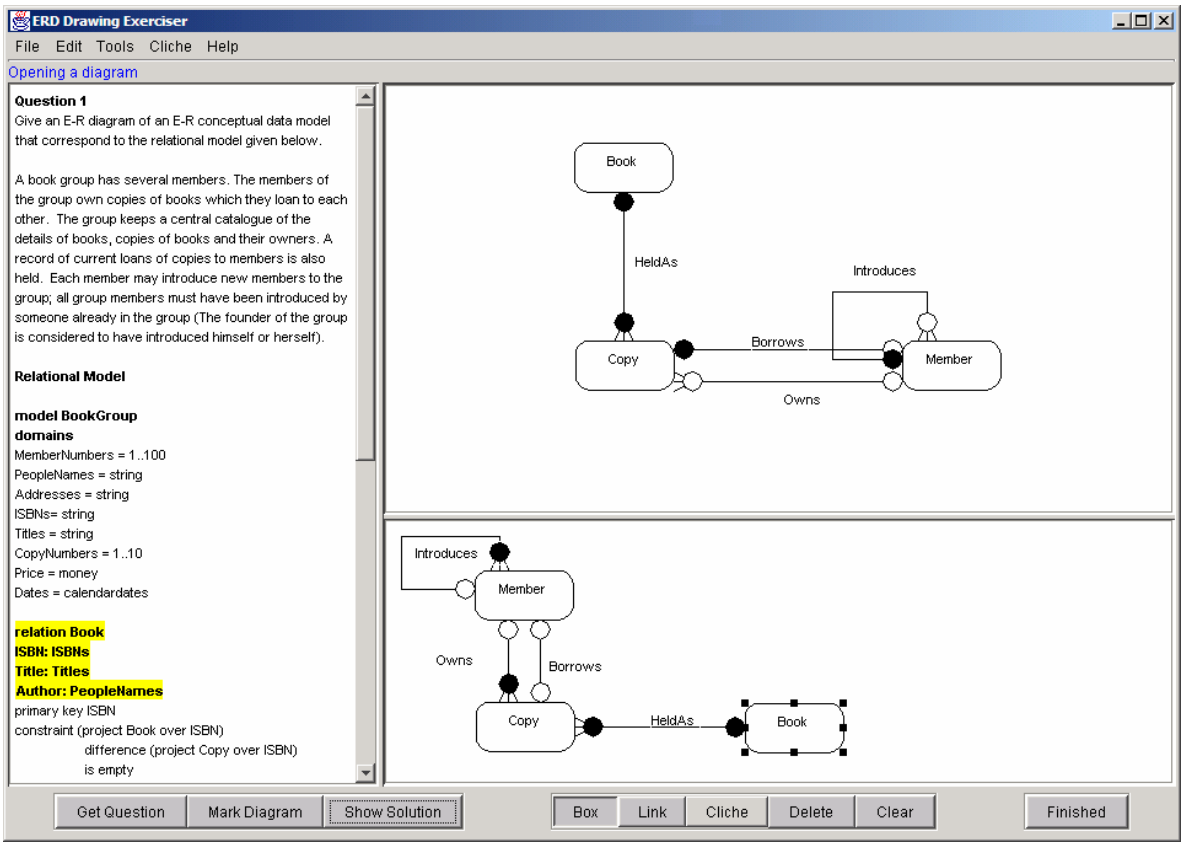

Fig. 10. The revision tool

9 Future Work

There is a clear need to apply the tools and techniques described in this paper to a larger set of student diagrams. Work is already underway to collect samples from all students on our database course from both assignments and examinations. This will provide a marking corpus of around three thousand diagrams from every presentation of the course.

The problem of marking E-R diagrams is just part of the larger problem of diagram understanding. We want to validate our general framework by applying it to a wider variety of interpretation problems. In particular, as well as more complex graph-based diagrams (as found in the UML, say), we wish to look at non-graph based domains.

The approach discussed in this paper depends crucially on the definition and specification of MMUs. We need to develop a theory to support this work, including formal definitions of MMUs and how they can be identified in a particular diagrammatic domain.

We also intend to investigate the remaining transformations of our learning model (the interpreter and comparator shown in Figure 1).

10 Summary

This paper reports on the development of an automatic tool for grading and feeding back on student attempts at drawing Entity-Relationship diagrams to model given scenarios. This work is part of an on-going project concerned with diagram understanding. The automatic marker was designed on the basis that diagrams have features in common with text and that typical text processing tools and techniques

from NLP would be applicable to diagrams. The automatic marker uses shallow matching to compare imprecise student diagrams with a (precise) sample solution.

In two small-scale trials, the automatic marker performed extremely well when compared with human markers. In a few cases where substantial differences where found between the automatic marker and the human markers, it was conjectured that the orientation, or shape, of the student diagram might have influenced the human markers and a simple redrawing of the diagrams in these cases resulted in an amended human mark. Whilst further investigation of this phenomenon is clearly needed, it does suggest that there can be human variation in the marking of diagrams, with automatic markers being more consistent.

References

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

Andries, M.; Engels, G. and Rekers, J. (1998) How to Represent a Visual Specification. In Marriott, K. and Meyer, B. (eds.) *Visual Language Theory*, Chapter 8, 245-259. Springer-Verlag, New York, ISBN 0-378-98367-8.

Bottoni, Paolo and Costagliola, Gannaro (2002) On the definition of visual languages and their editors. In Hegarty, M., Meyer, B. and Narayanan (Eds.) *Diagrams 2002*, LNAI 2317, 305-319, Springer-Verlag Berlin.

J. Burstein, M. Chodorow, and C. Leacock, (2003) Criterion Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence.* Acapulco, Mexico, 2003.

Chandrarasekaran, B (2002) What does it mean for a computer to do diagrammatic reasoning? A functional characterization of diagrammatic reasoning and its implications. In Hegarty, M., Meyer, B. and Narayanan (Eds.) *Diagrams 2002*, LNAI 2317 Springer-Verlag Berlin.

Chok, S.S. and Marriott, K. (1995) Parsing visual languages. In *Proceedings of the Eighteenth Australian Computer Science Conference*, Australian Computer Science Communications, **17**, 90-98.

Donlon, J.J., Forbus, K.D. (1999) Using a geographic information system for qualitative spatial reasoning about traceability. In *Proceedings of the Qualitative Reasoning Workshop*, Loch Awe, Scotland.

Ferguson, R.W. and Forbus, K.D. (2000) GeoRep: A flexible tool for spatial representation of line drawings. In *Proceedings of the 18th National Conference on Artificial Intelligence* (AAAI-2000).

Fish, Dale E., McCartney, Robert (2002) Using Diagrams to Understand Diagrams: A Case-Based Approach to Diagrammatic Reasoning. In Anderson et al (eds.) *Diagrammatic Representation and Reasoning*. Chapter 25, Springer-Verlag Berlin, 447-465.

Flower, Jean and Howse, John (2002) Generating Euler Diagrams. In Hegarty, M., Meyer, B. and Narayanan (Eds.) *Diagrams 2002*, LNAI 2317 Springer-Verlag Berlin, 61-75.

Howse, John, Molina Fernando, Sin, Sun-Joo, Taylor, John (2002) On Diagram Tokens and Types. In Hegarty, M., Meyer, B. and Narayanan (Eds.) *Diagrams 2002*, Springer-Verlag, Berlin, LNAI 2317, 146-160.

Iizuka, K., Tanaka, J. and Shizuki, B (2001) Describing a drawing editor by using constraint multiset grammars. In *Proceedings of the Sixth International Symposium on the Future of Software Technology* (ISFST 2001), Zhengzhou, China. November, 2001.
www.iplab.is.tsukuba.ac.jp/paper/international/iizukia-isfst2001.pdf
(accessed 02/06/04)

Jurafsky, D. and Martin, J. H. (2000) *Speech and Language Processing*. Prentice-Hall, New Jersey, USA. ISBN 0-13-095069-6.

Manning, C.D and Schutze, H. (2002) *Foundations of Statistical Natural language Processing*. MIT Press, Cambridge, Massachusetts, USA. ISBN 0-262-13360-1.

Jamnik, M., Bundy, A. and Green, I. (2002) On Automating Diagrammatic Proofs of Arithmetic Arguments. In Anderson et al (eds.) *Diagrammatic Representation and Reasoning*. Springer-Verlag, Berlin.

Marriott, K. and Meyer, B. (1998) (Eds.) *Visual Language Theory*. Springer-Verlag, Berlin.

Marriott, K., Meyer, B. and Wittenburg, K.B. (1998) A survey of Visual Language Specification and Recognition. In Marriott, K and Meyer, B (eds.) *Visual Language Theory*. Springer-Verlag, New York, 8-85, ISBN 0-378-98367-8.

Shermis, M.D, Burstein, J.C. (2003) (eds.) *Automated Essay Scoring: a cross-disciplinary approach*. Lawrence Erlbaum Associates, Mahwah, NJ, USA. ISBN 0-8058-3973-9.

Smith, N, Thomas, P.G. and Waugh, K. (2004) Interpreting Imprecise Diagrams. In Alan Blackwell, Kim Marriott, Atsushi Shimojima (eds.) *Proceedings of the Third International Conference in the Theory and Application of Diagrams*. March 22-24, Cambridge, UK. Springer Lecture Notes in Computer Science, 2980**,** 239-241. ISBN 3-540-21268-X.

Sowa, J.F. (1984) *Conceptual Structures: Information Processing in Mind and Machine*. Addison Wesley.

Swoboda, Nik (2002) Implementing Euler/Venn Reasoning Systems, In Anderson et al. Theory and *Application of Diagrams*, LNAI 1889, Chapter 21, 371-386 Springer-Verlag, Berlin.

Thomas, P.G., Price, B., Paine, C. Richards, M. (2002) *Remote Electronic examinations: an architecture for their production, presentation and grading*. British Journal of Educational Technology (BJET), **33** (5) 539-552.

Thomas, P.G. (2003) Evaluation of Electronic Marking of Examinations, In *Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education* (ITiCSE 2003), Thesaloniki, Greece, 50-54.

Thomas, P.G (2004a) Drawing Diagrams in an Online Exam, In *Proceedings of the 8th Annual International Conference in Computer Assisted Assessment*. Loughborough University, Loughborough, UK, 403-413.

Thomas, P.G. (2004b) Grading Diagrams Automatically. *Technical Report of the Computing Department*, Open University, UK. TR2004/01.

Thomas, P.G., Waugh, K., Smith, N. (2005) Experiments in the Automatic Marking of ER-Diagrams. To appear in *Proceedings of the 10th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE 05)*, June 27–29, 2005, Lisbon, Portugal.

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

Waugh, K.G., Thomas, P.G., Smith, N. (2004) Toward the Automated Assessment of Entity-Relationship Diagrams. In *Proceedings of the 2nd LTSN-ICS Teaching, Learning and Assessment in Databases Workshop*. Edinburgh.
http://www.ics.ltsn.ac.uk/pub/databases04/index.html
(accessed 26/04/05)