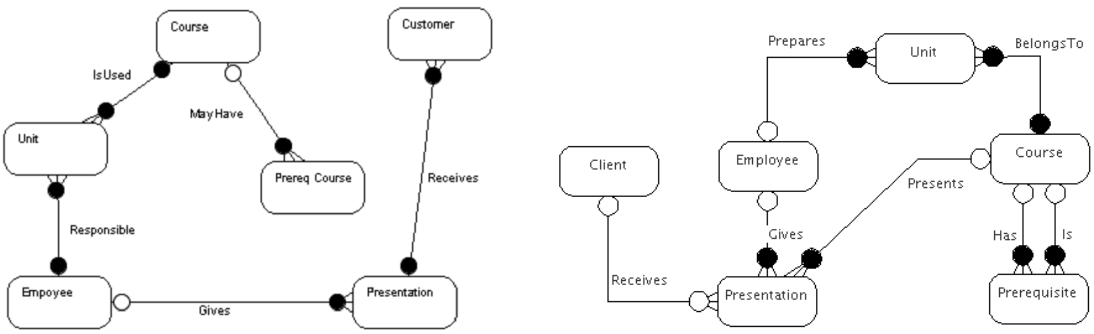
The role of labels in the automatic assessment of graph-based diagrams

Introduction

The ability to draw diagrams in free-form is rarely found in e-assessment systems. This paper examines one crucial area which needs to be well understood if automatic marking of diagrams is to be feasible: the analysis of labels.

Graph-based diagrams include geometrical elements of differing shapes connected by lines to carry the semantics of the domain being modelled. Some diagrams also use the relative location of elements to express information. Examples are entity-relationship diagrams (ERDs), illustrated in Figure 1, sequence diagrams (from the UML), biological flow diagrams and electrical circuit diagrams.

Labels play a central role in providing meaning both as identifiers, distinguishing between geometric elements, and the values of attributes of elements. Such information is central to the process of comparing two diagrams as is done when marking student-drawn diagrams.



(a) Student-drawn diagram (b) Model solution diagram

Figure 1 A student drawn ERD and the model solution diagram

In Figure 1, it is clear that there are many points of similarity, both in the entities (rounded rectangles) and the adornments on the relationships (circles and 'crowsfeet'), but there are several places where the diagrams differ. The student diagram (a) uses different, yet similar, labels and there are spelling errors. There are many ways in which a student diagram can deviate from a model solution yet still be considered to be a good match. Elements can be missing and additional, extraneous, elements can be included. We say that student diagrams are likely to be *imprecise*. The aim, therefore, is, in the presence of imprecision, to match a student diagram with a model solution and award a mark which is close to that which an expert human would give.

Our approach to marking decomposes the diagrams into sub-parts having certain properties that we have named *meaningful units* (MUs). The best match between the MUs of two diagrams is characterised by a set of similarity measures (values in the range [0..1]). This set of similarities is combined with a marking scheme – equivalent to the marking scheme that would be used by a human marker – to determine a mark for the student's diagram. Fuller descriptions of the theory behind our method and the overall approach can be found elsewhere (Smith et al., 2004). Here we shall concentrate on our approach to determining the similarity of labels.

Other researchers have begun to look at how diagrams can be marked (Hoggarth & Lockyer, 1998, Higgins & Bligh, 2006, Batmaz & Hinde, 2007, Tselonis & Sargeant, 2007) and there is one paper that looks at the specific problem of label comparison. Jayal and Shepperd (2009) have shown that the number of synonyms for labels can increase significantly as the number of student answers to a question increases thereby highlighting one of the difficult aspects of this work – recognising synonyms in student diagrams for labels occurring in model solutions. In this paper we shall explain how we compare labels in the presence of imprecision and how we cope with the synonym problem.

Matching labels

Our aim is to determine a measure of the similarity of two labels, one in a student-drawn diagram the other in a model solution. If two labels are sufficiently similar (above some chosen threshold) we assume that the student label is equivalent to (or matches) the label in the model solution. By equivalent we mean that the two labels are synonymous in the context of the problem being solved. However, because of imprecision, the student label might contain several defects. For example, it may be misspelt, be an abbreviation, or have a different lexical structure (e.g. embedded punctuation) to the model solution label. In most cases, such defects would be recognised by an expert human marker who would readily identify when two labels are equivalent, but in automatic marking the situation is not so simple.

In general, labels consist of one or more words and often those words are concatenated. Therefore, the first step is to determine word boundaries and this consists primarily of looking for punctuation and changes of case ('camel-case' is a common way of constructing labels from multiple words, identifying individual words with an initial capital letter). Of course, with imprecision, not all student labels conform to the rules and ad-hoc analyses have to be applied. For example, hyphenation is often used inappropriately and, as a result, can indicate a word boundary. However, if used correctly, hyphenation means the conjunction of two words which ought not to be split. Word boundaries can also be determined by searching for words which occur in the model solution.

Once a lexical division has been obtained, a check on the appropriateness of the division is made which recombines adjacent 'words' to see, for example, whether a prefix has been inadvertently separated from its stem.

This lexical process produces a 'phrase' – a sequence of words. In the kinds of diagram that we are interested in, it is possible to distinguish between labels that are used to label objects and those that label relationships between those objects. When a single word is used, the former are essentially nouns and the latter are verbs. Therefore, we model objects as noun phrases and relationships as verb phrases. The aim now becomes the identification of the noun in the noun phrase and the verb in the verb phrase.

A noun phrase can be viewed a noun accompanied by a set of modifiers (such as adjectives). Therefore, when comparing two noun phrases we determine the similarity of the two nouns and, separately, the similarity of the two sets of modifiers. The similarity of the two phrases is the weighted sum of the noun and modifier similarities. When comparing two verb phrases, the aim is to isolate the verbs and determine their

similarity. This implies discarding auxiliary verbs and prepositions which are the most common verb modifiers.

Words in a student-constructed label may occur in a different form to the expected words in the model solution (for example, where a different verb tense is used). Therefore, words are stemmed (we use the Porter stemming algorithm). Finally, words are compared using a similarity measure based on their edit-distance. This means that we can tolerate misspelling by accepting similarities above a given threshold.

The most important issue concerns synonyms. The labels occurring in a model solution are only one manifestation of the labels that can legitimately appear. In our current approach we include commonly occurring synonyms explicitly in the mark scheme (noting that specialist synonyms can exist in the domain of the question). The multiple synonym problem reported by Jayal and Shepperd (2009) is ameliorated by the use of techniques such as stemming (to reduce words to a canonical form) and similarity measures (to estimate the closeness of words). The synonym lists serve to pick up synonyms with different stems. In addition, the scenarios used in our questions are designed to limit the growth in synonyms (the need to design scenarios in this area has been investigated by Batmaz & Hinde (2008)).

Analysis and Results

We have examined around two-thousand student-drawn diagrams and derived a set of rules for comparing two labels. To determine how well the label matching process works we examined the behaviour of our automatic marker on a corpus of 591 diagrams obtained from answers to an examination question.

Table 1 shows the performance of the automatic marker when compared with the marks awarded by human markers on 394 of the diagrams (we used 197 diagrams for development work). The table shows that 91% of diagrams were awarded a mark by the automatic marker that was no more than half a mark different from the mark awarded by the human markers. More details of the overall marking process and an analysis of the results can be found in (Thomas et al., 2008a).

Table 1. The difference between automatic and human marking (N = 394)

Difference	0	0.5	1	1.5	2	2.5	3.0	3.5
Number	270	89	29	4	1	0	0	1
%	68.53	22.59	7.36	1.02	0	0	0	0.25
Cum. %	68.53	91.12	98.48	99.49	99.75	99.75	99.75	100

The single outlier which differed by 3.5 marks from the human mark was due to the existence of two entities with the same label and automatic marker choosing the wrong match. A change in the algorithm to examine the context of an entity has resolved this problem.

We then looked at which entity label in the model solution best matched each entity in a student-drawn diagram. Table 2 shows that the majority of student labels match with a label in a specimen solution with a high similarity value.

Table 2. Matching entity labels (N = 591)

	Similarity						
Label in model solution	≥0.9	≥0.8	≥0.7	≥0.6	≥0.5	<0.5	Total
Client	575	2	0	0	1	15	593
Employee	586	1	0	0	1	2	590
Unit	584	0	0	0	2	11	597
Course	699	3	0	1	0	0	703
Presentation	545	1	1	3	3	1	554
Prerequisite	167	3	11	2	3	0	186

Care needs to be taken in interpreting Table 2. First, more than one student label can match with a single model solution label. Second, not all student diagrams included labels that correspond to those in the model solutions and not all student diagrams had the same number of entities as the model solution. Most student diagrams included labels that correspond to 'Client', 'Employee', 'Unit' and 'Presentation' (a small number of diagrams included two entities with the same label, particularly 'Unit').

The relatively low number of matches with 'Prerequisite' is because this label only occurred in one of the two model solutions which corresponded to around a fifth of student answers.

A large number of student diagrams contained two or more labels that best matched with 'Course' (or one of its synonyms) reflecting, to some extent, the fact that labels were being reduced to a canonical form. This form of ambiguity does not necessarily mean that the automatic marker has difficulty in determining a 'global' match of entity labels in a given student diagram. The process determines a match for each individual label that maximises the overall similarity of labels in a diagram.

The most important message to be taken from Table 2 is that there are very few instances where the similarity measure is not decisive. The situation with relationship labels is much less clear as indicated by Table 3.

Table 3. Matching relationship labels (N = 591)

	Similarity						
Label in model solution	≥0.9	≥0.8	≥0.7	≥0.6	≥0.5	<0.5	Total
receives	913	5	24	7	17	275	1241
gives	30	1	0	0	0	1	32
prepares	749	5	6	4	14	78	856
belongsTo	246	3	8	10	20	16	303
has	0	0	0	6	1	3	10
is	55	0	35	1	1	0	92
presents	14	0	0	0	25	25	64

Table 3 illustrates that there is much more variability in the students' choice of relationship labels with two or more student labels best matching the same model solution label. Fortunately, this ambiguity is not as significant as it may appear because we place more emphasis on entity labels than relationship labels when matching diagram elements. Relationship labels are of more importance in distinguishing between

multiple relationships between two entities. However, it is much more difficult to determine an appropriate similarity threshold for relationship labels because there is a large number of matches around the similarity of 0.5 and it can be the case that label matches that a human would accept can be missed.

A more revealing analysis of labels was obtained when we examined labels in a student's diagram that were not successfully matched with labels in a model solution. An unmatched label can result either because of a mistake on the part of the student or a defect in the recognition process of the automatic marker. Two major problems were identified: one in each category.

The main student mistake was the labelling of relationships using a preposition ('for' and 'of' were the most common). In all 594 diagrams, there were 128 instances where a relationship was named by a single preposition.

We found that a small number of unexpected labels occurred frequently throughout student diagrams. In particular, the labels 'attends' (used 100 times), 'contains' (used 67 times) and 'responsible' (used 28 times) were not in the model solution nor in the list of known synonyms. Their presence can be explained with an example. The model solution relationship labelled 'belongsTo' occurs between two entities labelled 'Unit' and 'Course' and can be interpreted as 'a unit *belongs to* a course'. However, it would be quite acceptable to give a different reading to this relationship as in 'a course *contains* a unit'. These unexpected labels represent a 'converse' labelling of relationships that needs to be taken into account.

Other defects in the automatic label matching process were found, but occurred relatively infrequently. There were 6 instances where word boundaries were incorrectly identified and 7 instances of abbreviation. There were very few (5) occasions where a spelling error was the main reason why the matching process failed. In a very small number of cases, students tried to offer alternatives labels for the same relationship by writing, for example, 'gives/wants'.

Conclusions

While our approach to the matching of labels in diagrams has been successful *in our environment*, there are a number of qualifications. First, the problems which students are asked to solve are quite prescriptive and limit the degree of expressiveness in labels, particularly in entities. Entity labels tend to consist of one or two words which are normally easily derived from the question. Labels on relationships are, however, more complex in structure and students have more difficulty in expressing the required concepts. There is, therefore, a greater need to analyse these labels.

We have incorporated this automatic marking method into several tools (see Figure 2) designed to help students learn diagramming skills in entity-relationship and sequence diagrams. Students are offered a series of graded questions (see left-hand pane) and draw their answers in the upper right-hand pane. The tools mark their attempts and provide feedback (both textual and diagrammatic). A student can examine the model solution (shown in the bottom right-hand pane) to receive yet more feedback on how the model solution is derived.

In this formative use, feedback from students has been very positive (Thomas et al., 2007) and students have not queried inaccuracies in marking or invalid feedback.

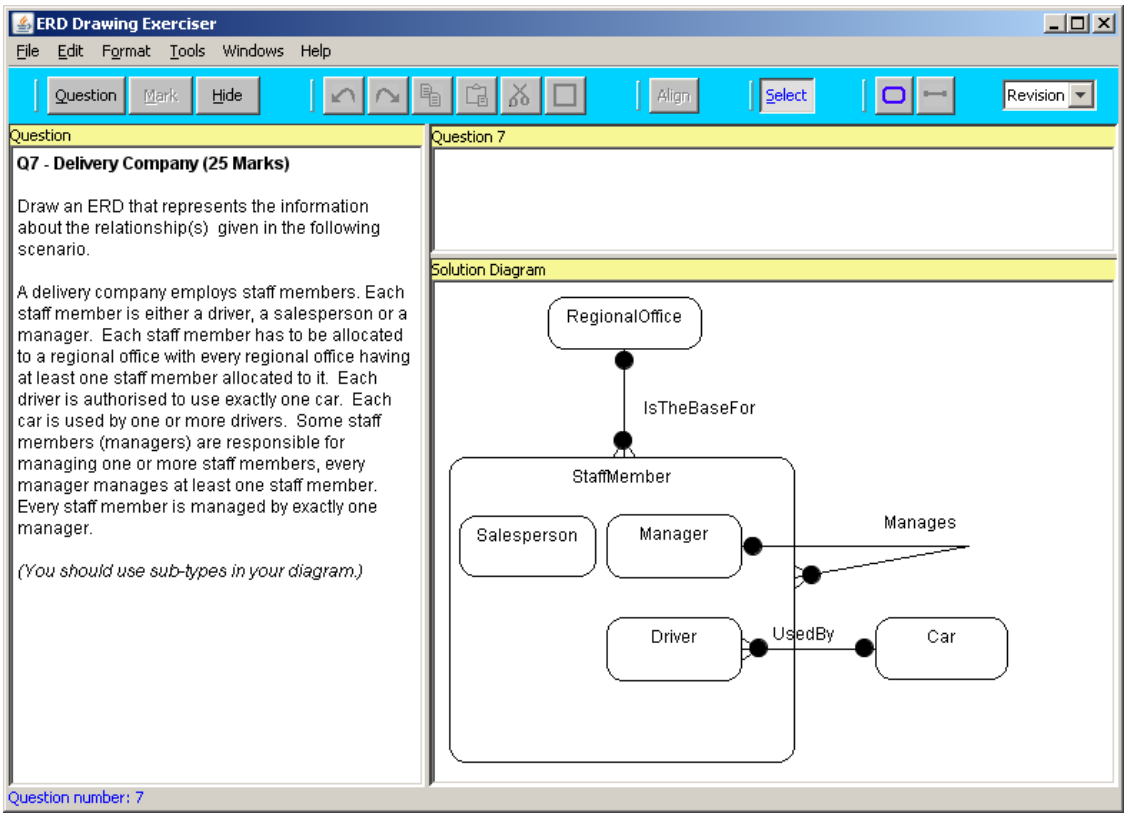


Figure 2. The ERD revision tool

However, if we are to use the method in a summative environment, perhaps initially as a 'second marker', we need to have confidence in the method's effectiveness for the vast majority of student answers. We have investigated the inter-rater reliability of the marker (Thomas et al., 2008a) and found the behaviour of the automatic marker to be similar to that of human markers.

References

Batmaz, F. and Hinde, C.J. 2007 A Web-Based Semi-Automatic Assessment Tool for Conceptual Database Diagram. Proceedings of Sixth IASTED International Conference on Web-Based Education. Chamonix, France, pp 427-432.

Batmaz, F. and Hinde, C.J. 2008 A Method for Controlling the Scenario Writing for the Assessment of Conceptual Database Model. Proceedings of Computers and Advanced Technology in Education (CATE 2008) September 29 – October 1, 2008 Crete, Greece.

Higgins, C. A. and Bligh. B. 2006 Formative Computer Based Assessment in Diagram Based Domains. *Proceedings of the 11th Annual Conference on Innovation and Technology in Computer Science Education*, Bologna, Italy, pp 98-102.

Hoggarth, G. and Lockyer, M. 1998, An automated student diagram assessment system. ACM SIGCSE Bulletin, Vol. 30, No. 3, pp 122-124.

Jayal, A. and Shepperd, M. 2008. The Problem of Labels in e-Assessment of Diagrams, *ACM J. of Educational Resources in Computing*, Vol. 8, No.4, Article 12.

Smith, N, Thomas, P.G., and Waugh, K. 2004 Interpreting Imprecise Diagrams. *Proceedings of the Third International Conference in the Theory and Application of Diagrams*. Cambridge, UK. Springer Lecture Notes in Computer Science, eds: Alan Blackwell, Kim Marriott, Atsushi Shimojima, Vol. 2980 pp 239-241.

Thomas, P.G., Waugh, K., and Smith, N. 2006. Using Patterns in the Automatic Marking of ER-Diagrams. Proceedings of the 11th Annual Conference on Innovation and Technology in Computer Science Education Bologna, Italy, pp 403—413.

Thomas, P.G., Smith, N. and Waugh, K. 2007. Tools for learning and automatically assessing graph-based diagrams. Research Proceedings of ALT-C 2007, Nottingham, pp 61-74.

Thomas, P.G., Smith, N and Waugh, K. 2008a, Automatically assessing graph-based diagrams, *J. Learning, Media & Technology*, Vol. 33, No. 3 pp249-267.

Thomas, P.G., Smith, N and Waugh, K. 2008b. A revision tool for teaching and learning sequence diagrams. Proceedings of ED-MEDIA conference, Austria, pp 5454—5460.

Tselonis, C. et al. 2005 Diagram matching for human-computer collaborative assessment. Proceedings of 9th International Computer Assisted Assessment Conference, Loughborough, UK, available from http://www.caaconference.co.uk/pastConferences/2005/index.asp

Tselonis, C. and Sargeant, J. 2007 Domain specific formative feedback through domain-independent diagram matching. Proceedings of 11th International Computer Assissted Assessment Conference, Loughborough, UK, available from http://www.caaconference.co.uk/pastConferences/2007/proceedings/index.asp

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