Using NLP to Support Scalable Assessment of Short Free Text Responses

Alistair Willis
Department of Computing and Communications
The Open University
Milton Keynes, UK
alistair.willis@open.ac.uk

Abstract

Marking student responses to short answer questions raises particular issues for human markers, as well as for automatic marking systems. In this paper we present the Amati system, which aims to help human markers improve the speed and accuracy of their marking. Amati supports an educator in incrementally developing a set of automatic marking rules, which can then be applied to larger question sets or used for automatic marking. We show that using this system allows markers to develop mark schemes which closely match the judgements of a human expert, with the benefits of consistency, scalability and traceability afforded by an automated marking system. We also consider some difficult cases for automatic marking, and look at some of the computational and linguistic properties of these cases.

1 Introduction

In developing systems for automatic marking, Mitchell et al. (2002) observed that assessment based on short answer, free text input from students demands very different skills from assessment based upon multiple-choice questions. Free text questions require a student to present the appropriate information in their own words, and without the cues sometimes provided by multiple choice questions (described respectively as improved verbalisation and recall (Gay, 1980)). Work by Jordan and Mitchell (2009) has demonstrated that automatic, online marking of student responses is both feasible (in that marking rules can be developed which mark at least as accurately as a human marker), and helpful to students, who find the online questions a valuable and enjoyable part of the assessment process. Such automatic marking is also an increasingly important part of assessment in Massive Open Online Courses (MOOCs) (Balfour, 2013; Kay et al., 2013).

However, the process of creating marking rules is known to be difficult and time consuming (Sukkarieh and Pulman, 2005; Pérez-Marín et al., 2009). The rules should usually be hand-crafted by a tutor who is a domain expert, as small differences in the way an answer is expressed can be significant in determining whether responses are correct or incorrect. Curating sets of answers to build mark schemes can prove to be a highly labour-intensive process. Given this requirement, and the current lack of availability of training data, a valuable progression from existing work in automatic assessment may be to investigate whether NLP techniques can be used to support the manual creation of such marking rules.

In this paper, we present the Amati system, which supports educators in creating mark schemes for automatic assessment of short answer questions. Amati uses information extraction-style templates to enable a human marker to rapidly develop automatic marking rules, and inductive logic programming to propose new rules to the marker. Having been developed, the rules can be used either for marking further unseen student responses, or for online assessment.

Automatic marking also brings with it further advantages. Because rules are applied automatically, it improves the *consistency* of marking; Williamson et al. (2012) have noted the potential of automated

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 243–253,
Denver, Colorado, June 4, 2015.

marking to improve the reliability of test scores. In addition, because Amati uses symbolic/logical rules rather than stochastic rules, it improves the *traceability* of the marks (that is, the marker can give an explanation of *why* a mark was awarded, or not), and increases the *maintainability* of the mark scheme, because the educator can modify the rules in the context of better understanding of student responses. The explanatory nature of symbolic mark schemes also support issues of auditing marks awarded in assessment. Bodies such as the UK's Quality Assurance Agency[1] require that assessment be fully open for the purposes of external examination. Techniques which can show exactly why a particular mark was awarded (or not) for a given response fit well with existing quality assurance requirements.

All experiments in this paper were carried out using student responses collected from a first year introductory science module.

2 Mark Scheme Authoring

Burrows et al. (2015) have identified several different eras of automatic marking of free text responses. One era they have identified has treated automatic marking as essentially a form of information extraction. The many different ways that a student can correctly answer a question can make it difficult to award correct marks[2]. For example:

> *A snowflake falls vertically with a constant speed. What can you say about the forces acting on the snowflake?*

Three student responses to this question were:

(1) *there is no net force*

(2) *gravitational force is in equilibrium with air resistance*

(3) *no force balanced with gravity*

The question author considered both responses (1) and (2) correct. However, they share no common words (except *force* which already appears in the question, and *is*). And while *balance* and *equilibrium* have closely related meanings, response (3) was not considered a correct answer to the question[3]. These examples suggest that bag of words techniques are unlikely to be adequate for the task of short answer assessment. Without considering word order, it would be very hard to write a mark scheme that gave the correct mark to responses (1)-(3), particularly when these occur in the context of several hundred other responses, all using similar terms.

In fact, techniques such as Latent Semantic Analysis (LSA) have been shown to be accurate in grading longer essays (Landauer et al., 2003), but this success does not appear to transfer to short answer questions. Haley's (2008) work suggests that LSA performs poorly when applied to short answers, with Thomas et al. (2004) demonstrating that LSA-based marking systems for short answers did not give an acceptable correlation with an equivalent human marker, although they do highlight the small size of their available dataset.

Sukkarieh and Pulman (Sukkarieh and Pulman, 2005) and Mitchell et al. (2002) have demonstrated that hand-crafted rules containing more syntactic structure can be valuable for automatic assessment, but both papers note the manual effort required to develop the set of rules in the first place. To address this, we have started to investigate techniques to develop systems which can support a subject specialist (rather than a computing specialist) in developing a set of marking rules for a given collection of student responses. In addition, because it has been demonstrated (Butcher and Jordan, 2010) that marking rules based on regular expressions can mark accurately, we have also investigated the use of a symbolic learning algorithm to propose further marking rules to the author.

Enabling such markers to develop computational marking rules should yield the subsequent benefits of speed and consistency noted by Williamson et al., and the potential for embedding in an online systems to provide immediate marks for student submissions (Jordan and Mitchell, 2009). This proposal fits with the observation of Burrows et al. (2015), who suggest that rule based systems are desirable for "re-

[1] `http://www.qaa.ac.uk`

[2] Compared with multiple choice questions, which are easy to mark, although constructing suitable questions in the first place is far from straightforward (Mitkov et al., 2006).

[3] As with all examples in this paper, the "correctness" of answers was judged with reference to the students' level of study and provided teaching materials.

term(R, Term, I)	The I^{th} term in R is *Term*
template(R, Template, I)	The I^{th} term in R matches *Template*
precedes(I_i, I_j)	The I_i^{th} term in a response precedes the I_j^{th} term
closely_precedes(I_i, I_j)	The I_i^{th} term in a response precedes the I_j^{th} within a specified window

Figure 1: Mark scheme language

peated assessment" (i.e. where the assessment will be used multiple times), which is more likely to repay the investment in developing the mark scheme. We believe that the framework that we present here shows that rule-based marking can be more tractable than suggested by Burrows.

2.1 The Mark Scheme Language

In this paper, I will describe a set of such marking rules as a "mark scheme", so Amati aims to support a human marker in hand crafting a mark scheme, which is made up of a set of marking rules. In Amati, the mark schemes are constructed from sets of prolog rules, which attempt to classify the responses as either correct or incorrect. The rule syntax closely follows that of Junker et al. (1999), using the set of predicates shown in figure 1.

The main predicate for recognising keywords is *term(R, Term, I)*, which is true when *Term* is the I^{th} term in the response R. Here, we use "term" to mean a word or token in the response, subject to simple spelling correction. This correction is based upon a Damerau-Levenshtein (Damerau, 1964) edit distance of 1, which represents the replacement, addition or deletion of a single character, or a transposition of two adjacent characters. So for example, if R represented the student response:

(4) *no force ballanced with gravity*

then *term(R, balanced, 3)* would be true, as the 3^{rd} token in R is *ballanced*, and at most 1 edit is needed to transform *ballanced* to *balanced*.

The predicate *template* allows a simple form of stemming (Porter, 1980). The statement *template(R, Template, I)* is true if *Template* matches at the beginning of the I^{th} token in R, subject to the same spelling correction as *term*. So for example, the statement:

template(R, balanc, 3)

would match example (4), because *balanc* is a single edit from *ballanc*, which itself matches the beginning of the 3^{rd} token in R. (Note that it would not match as a *term*, because *ballance* is two edits from *balanc*.) Such templates allow rules to be written which match, for example, *balance*, *balanced*, *balancing* and so on.

The predicates *precedes* and *closely_precedes*, and the index terms, which appear as the variables I and J in figure 1, capture a level of linear precedence, which allow the rules to recognise a degree of linguistic structure. As discussed in section 2, techniques which do not capture some level of word order are insufficiently expressive for the task of representing mark schemes. However, a full grammatical analysis also appears to be unnecessary, and in fact can lead to ambiguity. Correct responses to the *Rocks* question (see table 1) required the students to identify that the necessary conditions to form the rock are *high temperature* and *high pressure*. Both *temperature* and *pressure* needed to be modified to earn the mark. Responses such as (5) should be marked correct, with an assumption that the modifier should distribute over the conjunction.

(5) *high temperature and pressure*

While the *precedence* predicate is adequate to capture this behaviour, using a full parser creates an ambiguity between the analyses (6) and (7).

(6) *(high (pressure)) and temperature* $\times$

(7) *(high (pressure and temperature))* $\surd$

The example suggest that high accuracy can be difficult to achieve by systems which commit to an early, single interpretation of the ambiguous text.

So a full example of a matching rule might be:

term(R, oil, I) ∧
term(R, less, J) ∧
template(R, dens, K)
precedes(I, J) → *correct(R)*

which would award the correct marks to responses (8) and (9):

(8) *oil is less dense than water* √

(9) *water is less dense than oil* ×

The use of a template also ensures the correct mark is awarded to the common response (10), which should also be marked as correct:

(10) *oil has less density than water*

2.2 Incremental rule authoring

The Amati system is based on a bootstrapping scheme, which allows an author to construct a rule-set by marking student responses in increments of 50 responses at a time, while constructing marking rules which reflect the marker's own judgements. As the marker develops the mark scheme, he or she can correct the marks awarded by the existing mark scheme, and then edit the mark scheme to more accurately reflect the intended marks.

The support for these operations are illustrated in figures 2 and 3. To make the system more usable by non-specialists (that is, non-specialists in computing, rather than non-specialists in the subject being taught), the authors are not expected to work directly with prolog. Rather, rules are presented to the user via online forms, as shown in figure 2. As each rule is developed, the system displays to the user the responses which the rule marks as correct.

As increasingly large subsets of the student responses are marked, the system displays the set of imported responses, the mark that the current mark scheme awards, and which rule(s) match agains each response (figure 3). This allows the mark scheme author to add or amend rules as necessary.

2.3 Rule Induction

As the marker constructs an increasingly large collection of marked responses, it can be useful to use the marked responses to induce further rules automatically. Methods for learning relational rules to perform information extraction are now well established (Califf and Mooney, 1997; Soderland, 1999), with Inductive Logic Programming (ILP) (Quinlan, 1990; Lavrač and Džeroski, 1994) often proving a suitable learning learning technique (Aitken, 2002; Ramakrishnan et al., 2008). ILP is a supervised learning algorithm which attempts to generate a logical description of a set of facts in the style of a prolog program. Amati embeds the ILP system Aleph (Srinivasan, 2004) as the rule learner, which itself implements the Progol learning algorithm (Muggleton, 1995), a bottom up, greedy coverage algorithm. This allows an author to mark the current set of questions (typically the first or second block of 50 responses), before using the ILP engine to generate a rule set which he or she can then modify. We return to the question of editing rule sets in section 4.1.

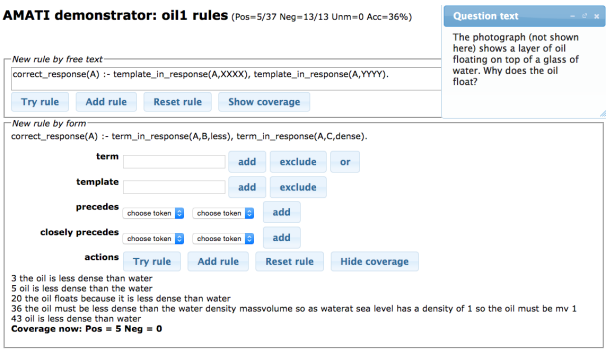
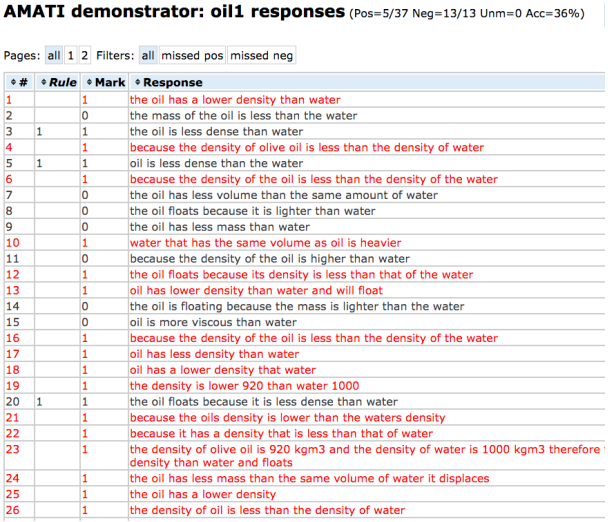


Figure 2: Form for entering marking rules



#	Rule	Mark	Response
1		1	the oil has a lower density than water
2		0	the mass of the oil is less than the water
3	1	1	the oil is less dense than water
4		1	because the density of olive oil is less than the density of water
5	1	1	oil is less dense than the water
6		1	because the density of the oil is less than the density of the water
7		0	the oil has less volume than the same amount of water
8		0	the oil floats because it is lighter than water
9		0	the oil has less mass than water
10		1	water that has the same volume as oil is heavier
11		0	because the density of the oil is higher than water
12		1	the oil floats because its density is less than that of the water
13		1	oil has lower density than water and will float
14		0	the oil is floating because the mass is lighter than the water
15		0	oil is more viscous than water
16		1	because the density of the oil is less than the density of the water
17		1	oil has less density than water
18		1	oil has a lower density that water
19		1	the density is lower 920 than water 1000
20	1	1	the oil floats because it is less dense than water
21		1	because the oils density is lower than the waters density
22		1	because it has a density that is less than that of water
23		1	the density of olive oil is 920 kgm3 and the density of water is 1000 kgm3 therefore density than water and floats
24		1	the oil has less mass than the same volume of water it displaces
25		1	the oil has a lower density
26		1	the density of oil is less than the density of water

Figure 3: Application of rule to the *Oil* response set

Our use of ILP to support markers in developing rules has several parallels with the Powergrading project (Basu et al., 2013). Both our work and that of Basu et al. focus on using NLP techniques primarily to support the work of a human marker, and reduce marker effort. Basu et al. take an approach whereby student responses are clustered using statistical topic detection, with the marker then able to allocate marks and feedback at the cluster level, rather than at the individual response level. Similarly, the aim of Amati is that markers should be able to award marks by identifying, via generated rules, commonly occurring phrases. The use of such phrases can then be analysed at the cohort level (or at least, incrementally at the cohort level), rather than at the individual response level.

In practice, we found that markers were likely to use the predicted rules as a "first cut" solution, to gain an idea of the overall structure of the final mark scheme. The marker could then concentrate on developing more fine-grained rules to improve the mark scheme accuracy. This usage appears to reflect that found by Basu et al., of using the machine learning techniques to automatically identify groups of similar groups of responses. This allows the marker to highlight common themes and frequent misunderstandings.

Short name	Question text
Sandstone	A sandstone observed in the field contains well-sorted, well rounded, finely pitted and reddened grains. What does this tell you about the origins of this rock?
Snowflake	A snowflake falls vertically with a constant speed. What can you say about the forces acting on the snowflake?
Charge	If the distance between two electrically charged particles is doubled, what happens to the electric force between them?
Rocks	Metamorphic rocks are existing rocks that have "changed form" (metamorphosed) in a solid state. What conditions are necessary in order for this change to take place?
Sentence	What is wrong with the following sentence? *A good idea.*
Oil	The photograph *(not shown here)* shows a layer of oil floating on top of a glass of water. Why does the oil float?

Table 1: The questions used

3 Evaluation

The aim of the evaluation was to determine whether a ruleset built using Amati could achieve performance comparable with human markers. As such, there were two main aims. First, to determine whether the proposed language was sufficiently expressive to build successful mark schemes, and second, to determine how well a mark scheme developed using the Amati system would compare against a human marker.

3.1 Training and test set construction

A training set and a test set of student responses were built from eight questions taken from an entry-level science module, shown in table 1. Each student response was to be marked as either correct or incorrect. Two sets of responses were used, which were built from two subsequent presentations of the same module. Amati was used to build a mark scheme using a training set of responses from the 2008 student cohort, and then that scheme was applied to an unseen test set constructed from the responses to the same questions from the 2009 student cohort.

The difficulties in attempting to build any corpus in which the annotations are reliable are well documented (Marcus et al.'s (1993) discussion of the Penn Treebank gives a good overview). In this case, we exploited the presence of the original question setter and module chair to provide as close to a "ground truth" as is realistic. Our gold-standard marks were obtained with a multiple-pass annotation process, in which the collections of responses were initially marked by two or more subject-specialist tutors, who mainly worked independently, but who were able to confer when they disagreed on a particular response. The marks were then validated by the module chair, who was also called upon to resolve any disputes which arose as a result of disagreements in the mark scheme. The

cost of constructing a corpus in this way would usually be prohibitive, relying as it does on subject experts both to provide the preliminary marks, and to provide a final judgement in the reconciliation phase. In this case, the initial marks (including the ability to discuss in the case of a dispute) were generated as part of the standard marking process for student assessment in the University[4].

3.2 Effectiveness of authored mark schemes

To investigate the expressive power of the representation language, a set of mark schemes for the eight questions shown in table 1 were developed using the Amati system. The training data was used to build the rule set, with regular comparisons against the gold standard marks. The mark scheme was then applied to the test set, and the marks awarded compared against the test set gold standard marks.

The results are shown in table 2. The table shows the total number of responses per question, and the accuracy of the Amati rule set applied to the unseen data set. So for example, the Amati rule set correctly marked 98.42% of the 1711 responses to the *Sandstone* question. Note that the choice of accuracy as the appropriate measure of success is determined by the particular application. In this case, the important measure is how many responses are marked correctly. That is, it is as important that incorrect answers are marked as incorrect, as it is that correct answers are marked as correct.

To compare the performance of the Amati ruleset against the human expert, we have used Krippendorf's α measure, implemented in the python Natural Language Toolkit library (Bird et al., 2009) following Artstein and Poesio's (2008) presentation. The rightmost column of table 2 shows the α measure between the Amati ruleset and the post-reconciliation marks awarded by the human expert. This column shows a higher level of agreement than was obtained with human markers alone. The agreement achieved by independent human markers ranged from a maximum of $\alpha = 88.2\%$ to a minimum of $\alpha = 71.2\%$, which was the agreement on marks awarded for the *snowflake* question. It is notable that the human marker agreement was worst on the same question that the Amati-authored ruleset performed worst on; we discuss some issues that this question raises in section 4.3.

Question	# responses	accuracy/%	α/%
Sandstone	1711	98.42	97.5
Snowflake	2057	91.0	81.7
Charge	1127	98.89	97.6
Rocks	1429	99.00	89.6
Sentence	1173	98.19	97.5
Oil	817	96.12	91.5

Table 2: Accuracy of the Amati mark schemes on unseen data, and the Krippendorf α rating between the marks awarded by Amati and the gold standard

The marks awarded by the marker supported with Amati therefore aligned more closely with those of the human expert than was achieved between independent markers. This suggests that further development of computer support for markers is likely to improve overall marking consistency, both across the student cohort, and by correspondence with the official marking guidance.

4 Observations on authoring rulesets

It is clear from the performance of the different rule sets that some questions are easier to generate mark schemes for than others. In particular, the mark scheme authored on the responses to the *snowflake* question performed with much lower accuracy than the other questions. This section gives a qualitative overview of some of the issues which were observed while authoring the mark schemes.

4.1 Modification of generated rules

A frequently cited advantage of ILP is that, as a logic program, the output rules are generated in a human-readable form (Lavrač and Džeroski, 1994; Mitchell, 1997). In fact, the inclusion of templates means that several of the rules can be hard to interpret at first glance. For example, a rule proposed to

[4]We have not presented inter-annotator agreement measures here, as these are generally only meaningful when annotators have worked independently. This model of joint annotation with a reconciliation phase is little discussed in the literature, although this is a process used by Farwell et al. (2009). Our annotation process differed in that the reconciliation phase was carried out face to face following each round of annotation, in contrast to Farwell et al.'s, which allowed a second anonymous vote after the first instance.

mark the *Rocks* question was:

template(R, bur,I) $\wedge$
template(R, hea,J) $\rightarrow$ *correct(R)*

As the official marking guidance suggests that *High temperature and pressure* is an acceptable response, *hea* can easily be interpreted as *heat*. However, it is not immediately clear what *bur* represents. In fact, a domain expert would probably recognise this as a shortened form of *buried* (as the high pressure, the second required part of the solution, can result from burial in rock). As the training set does not contain terms with the same first characters as *burial*, such as *burnished*, *Burghundy* or *burlesque*, then this term matches. However, a mark scheme author might prefer to edit the rule slightly into something more readable and so maintainable:

template(R, buri,I) $\wedge$
term(R, heat,J) $\rightarrow$ *correct(R)*

so that either *buried* or *burial* would be matched, and to make the recognition of *heat* more explicit.

A more complex instance of the same phenomenon is illustrated by the generated rule:

term(R, high,I) $\wedge$
term(R, temperature,J) $\rightarrow$ *correct(R)*

Although the requirement for the terms *high* and *temperature* is clear enough, there is no part of this rule that requires that the student also mention *high pressure*. This has come about because all the student responses that mention *high temperature* also explicitly mention *pressure*. Because Progol and Aleph use a greedy coverage algorithm, in this case Amati did not need to add an additional rule to capture . Again, the mark scheme author would probably wish to edit this rule to give:

term(R, high,I) $\wedge$
term(R, temperature,J) $\wedge$
term(R, pressure,K) $\wedge$
precedes(I,J) $\rightarrow$ *correct(R)*

which covers the need for *high* to precede *temperature*, and also contain a reference to *pressure*. A similar case, raised by the same question, is the following proposed rule:

term(R, high,I) $\wedge$
term(R, pressure,J) $\wedge$
term(R, and,K) $\wedge$
precedes(I,K) $\rightarrow$ *correct(R)*

which requires a conjunction, but makes no mention of *temperature* (or *heat* or some other equivalent). In this case, the responses (11) and (12):

(11) *(high (pressure and temperature))* $\surd$

(12) *(high (pressure and heat))* $\surd$

are both correct, and both appeared amongst the student responses. However, there were no incorrect responses following a similar syntactic pattern, such as, for example, (13) or (14):

(13) *high pressure and altitude* $\times$

(14) *high pressure and bananas* $\times$

Students who recognised that *high pressure* and *something else* were required, always got the *something else* right. Therefore, the single rule above had greater coverage than rules that looked individually for *high pressure and temperature* or *high pressure and heat*.

This example again illustrates the Amati philosophy that the technology is best used to support human markers. By hand-editing the proposed solutions, the marker ensures that the rules are more intuitive, and so can be more robust, and more maintainable in the longer term. In this case, an author might reasonably rewrite the single rule into two:

term(R, high,I) $\wedge$
term(R, pressure,J) $\wedge$
term(R, temperature,K) $\wedge$
precedes(I,K) $\rightarrow$ *correct(R)*

term(R, high,I) $\wedge$
term(R, pressure,J) $\wedge$
term(R, heat,K) $\wedge$
precedes(I,K) $\rightarrow$ *correct(R)*

removing the unnecessary conjunction, and providing explicit rules for *heat* and *temperature*.

4.2 Spelling correction

It is questionable whether spelling correction is always appropriate. A question used to assess knowledge of organic chemistry might require the term *butane* to appear in the solution. It would not be appropriate to mark a response containing the token *butene* (a different compound) as correct, even though *butene* would be an allowable misspelling of *butane* according to the given rules. On the other hand, a human marker would probably be inclined to mark responses containing *buttane* or *butan* as correct. These are also legitimate misspellings according the table, but are less likely to be misspellings of *butene*.

The particular templates generated reflect the linguistic variation in the specific datasets. A template such as *temp*, intended to cover responses containing *temperature* (for example), would also potentially cover *temporary*, *tempestuous*, *temperamental* and so on. In fact, when applied to large sets of homogenous response-types (such as multiple responses to a single question), the vocabulary used across the complete set of responses turns out to be sufficiently restricted for meaningful templates to be generated. It does not follow that this hypothesis language would continue to be appropriate for datasets with a wider variation in vocabulary.

4.3 Diversity of correct responses

As illustrated in table 2, the *Snowflake* question was very tricky to handle, with lower accuracy than the other questions, and lower agreement with the gold standard. The following are some of the student responses:

(15) *they are balanced*

(16) *the force of gravity is in balance with air resistance*

(17) *friction is balancing the force of gravity*

(18) *only the force of gravity is acting on the hailstone and all forces are balanced*

The module chair considered responses (15), (16) and (17) to be correct, and response (18) to be incorrect.

The most straightforward form of the answer is along the lines of response (15). In this case, there are no particular forces mentioned; only a general comment about the forces in question. Similar cases were *there are no net forces, all forces balance, the forces are in equilibrium* and so on.

However, responses (16) and (17) illustrate that in many cases, the student will present particular examples to attempt to answer the question. In these cases, both responses identify gravity as one of the acting forces, but describe the counteracting force differently (as *air resistance* and *friction* respectively). A major difficulty in marking this type of question is predicting the (correct) examples that students will use in their responses, as each correct pair needs to be incorporated in the mark schemes. A response suggesting that *air resistance* counteracts *drag* would be marked incorrect. As stated previously, developing robust mark schemes requires that mark scheme authors use large sets of previous student responses, which can provide guidance on the range of possible responses.

Finally, response (18) illustrates a difficult response to mark (for both pattern matchers and linguistic solutions). The response consists of two conjoined clauses, the second of which, *all forces are balanced*, is in itself a correct answer. It is only in the context of the first clause that the response is marked incorrect, containing the error that it is **only** *the force of gravity* which acts.

This question highlights that the ease with which a question can be marked automatically can depend as much on the question being asked as the answers received. Of course, this also applies to questions intended to be marked by a human; some questions lead to easier responses to grade. So a good evaluation of a marking system needs to consider the questions (and the range of responses provided by real students) being asked; the performance of the system is meaningful only in the context of the nature of the questions being assessed, and an understanding of the diversity of correct responses. In this case, it appears that questions which can be correctly answered by using a variety of different examples should be avoided. We anticipate that with increasing maturity of the use of automatic marking systems, examiners would develop skills in setting appropriate questions for the marking system, just as

experienced authors develop skills in setting questions which are appropriate for human markers.

4.4 Anaphora Ambiguity

The examples raise some interesting questions about how anaphora resolution should be dealt with. Two responses to the *oil* question are:

(19) *The oil floats on the water because it is lighter*

(20) *The oil floats on the water because it is heavier*

These two responses appear to have contradictory meanings, but in fact are both marked as correct. This initially surprising result arises from the ambiguity in the possible resolutions of the pronoun *it*:

(21) *[The oil]$_i$ floats on the water because it$_i$ is lighter.*

(22) *The oil floats on [the water]$_j$ because it$_j$ is heavier.*

When marking these responses, the human markers followed a general policy of giving the benefit of the doubt, and, within reasonable limits, will mark a response as correct if any of the *possible* interpretations would be correct relative to the mark scheme.

As with the ambiguous modifier attachment seen in responses (6) and (7), this example illustrates that using a different (possibly better) parser is unlikely to improve the overall system performance. Responses such (21) and (22) are hard for many parsers to handle, because an early commitment to a single interpretation can assume that *it* must refer to *the oil* or *the water*. Again, this example demonstrates that a more sophisticated approach to syntactic ambiguity is necessary if a parsing-based system is to be used. (One possible approach might be to use underspecification techniques (König and Reyle, 1999; van Deemter and Peters, 1996) and attempt to reason with the ambiguous forms.)

5 Discussion and Conclusions

We have presented a system which uses information extraction techniques and machine learning to support human markers in the task of marking free text responses to short answer questions. The results suggest that a system such as Amati can help markers create accurate, reusable mark schemes.

The user interface to Amati was developed in collaboration with experienced markers from the Open University's Computing department and Science department, who both gave input into the requirements for an effective marking system. We intend to carry out more systematic analyses of the value of using such systems for marking, but informally, we have found that a set of around 500-600 responses was enough for an experienced marker to feel satisfied with the performance of her own authored mark scheme, and to be prepared to use it on further unseen cases. (This number was for the *Snowflake* question, which contained approximately half correct responses. For the other questions, the marker typically required fewer responses.)

The work described in this paper contrasts with the approach commonly taken in automatic marking, of developing mechanisms which assign marks by comparing student responses to one or more target responses created by the subject specialist (Ziai et al., 2012). Such systems have proven effective where suitable linguistic information is compared, such as the predicate argument structure used by c-rater (Leacock and Chodorow, 2003), or similarity between dependency relationships, as used by AutoMark (now FreeText (Mitchell et al., 2002)) and Mohler et al. (2011). Our own experiments with FreeText found that incorrect marks were often a result of an inappropriate parse by the embedded Stanford parser (Klein and Manning, 2003)), as illustrated by the parses (6) and (7). In practice, we have found that for the short answers we have been considering, pattern based rules tend to be more robust in the face of such ambiguity than a full parser.

A question over this work is how to extend the technique to more linguistically complex responses. The questions used here are all for a single mark, all or nothing. A current direction of our research is looking at how to provide support for more complicated questions which would require the student to mention two or more separate pieces of information, or to reason about causal relationships. A further area of interest is how the symbolic analysis of the students' responses can be used to generate meaningful feedback to support them as part of their learning process.

Acknowledgments

The author would like to thank Sally Jordan for her help in collecting the students' data, for answering questions about marking as they arose and for making the data available for use in this work. Also, David King who developed the Amati system, and Rita Tingle who provided feedback on the usability of the system from a marker's perspective. We would also like to thank the anonymous reviewers for their valuable suggestions on an earlier draft of the paper.

References

James Stuart Aitken. 2002. Learning information extraction rules: An inductive logic programming approach. In *ECAI*, pages 355–359.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Stephen P Balfour. 2013. Assessing writing in MOOCs: Automated essay scoring and calibrated peer review. *Research & Practice in Assessment*, 8(1):40–48.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Steven Burrows, Iryana Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Philip G. Butcher and Sally E. Jordan. 2010. A comparison of human and computer marking of short free-text student responses. *Computers and Education*.

Mary Elaine Califf and Raymond J. Mooney. 1997. Relational learning of pattern-match rules for information extraction. In T.M. Ellison, editor, *Computational Natural Language Learning*, pages 9–15. Association for Computational Linguistics.

F. J. Damerau. 1964. Technique for computer detection and correction of spelling errors. *Communications of the Association of Computing Machinery*, 7(3):171–176.

David Farwell, Bonnie Dorr, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lori Levin, Keith Miller, Teruko Mitamura, Owen Rambow, Florence Reeder, and Advaith Siddharthan. 2009. Interlingual annotation of multilingual text corpora and framenet. In Hans C. Boas, editor, *Multilingual FrameNets in Computational Lexicography : Methods and Applications*. Mouton de Gruyter, Berlin.

Lorraine R Gay. 1980. The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement*, 17(1):45–50.

Debra T. Haley. 2008. *Applying Latent Semantic Analysis to Computer Assisted Assessment in the Computer Science Domain: A Framework, a Tool, and an Evaluation*. Ph.D. thesis, The Open University.

Sally Jordan and Tom Mitchell. 2009. E-assessment for learning? The potential of short free-text questions with tailored feedback. *British Journal of Educational Technology*, 40(2):371–385.

Markus Junker, Michael Sintek, and Matthias Rinck. 1999. Learning for text categorization and information extraction with ILP. In J. Cussens, editor, *Proceedings of the First Workshop Learning Language in Logic*, pages 84–93.

Judy Kay, Peter Reimann, Elliot Diebold, and Bob Kummerfeld. 2013. MOOCs: So Many Learners, So Much Potential.... *IEEE Intelligent Systems*, 3:70–77.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430. Association for Computational Linguistics.

Esther König and Uwe Reyle. 1999. A general reasoning scheme for underspecified representations. In Hans Jürgen Ohlbach and U. Reyle, editors, *Logic, Language and Reasoning. Essays in Honour of Dov Gabbay*, volume 5 of *Trends in Logic*. Kluwer.

T. K. Landauer, D. Laham, and P. W. Foltz. 2003. Automated scoring and annotation of essays with the intelligent essay assessor. In M. Shermis and J. Burstein, editors, *Automated essay scoring: a cross-disciplinary approach*, pages 87–112. Lawrence Erlbaum Associates, Inc.

Nada Lavrač and Sašo Džeroski. 1994. *Inductive Logic Programming*. Ellis Horwood, New York.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards robust computerised marking of free-text responses. In *6th International Computer Aided Assessment Conference*, Loughbrough.