Approaches to the computerized assessment of free text responses

Approaches to the Computerized Assessment of Free Text Responses

Dave Whittington
Department of Electronics and Electrical Engineering
University of Glasgow

Helen Hunt
Department of Computer Science
University of Strathclyde

Abstract

The automated assessment of student's essays is regarded by many as the Holy Grail of computer aided assessment. Whilst a few people search for the grail, many more deny its existence. This paper describes the various approaches that have been taken over the last 40 years in an attempt to solve the problems involved with the computerized assessment of free text.

The earliest approaches were founded in simple style analysis. Systems, such as Project Essay Grade (PEG), were developed upon the idea that certain surface features of an essay could be manipulated in such a way as to predict the grade that a human examiner would assign to an essay. Other methods, such as Latent Semantic Analysis (LSA), also take a statistical approach to marking, but focus on actual textual content, analyzing groupings and context. The Educational Testing Service (ETS) originally attempted to tackle the problem from a classification point of view whilst more recent work by ETS bears similarities to PEG in its statistical approach. Most of the methods currently being developed have been shown to be capable of generating essay scores that correlate with a human grader's scores at least as well as two human graders correlate with each other..

A novel approach being adopted by the authors to allow the comparison of students' essays against a model answer involves the use of theories developed for inter-lingual machine translation. A few different methods for knowledge representation, and their current uses in machine translation are presented.

Panlingua, an idea for knowledge representation developed by Chaumont Devin, is based on semantic network research. Using a system of nodes over four layers it attempts to model how the brain might translate from sensory patterns it sees or hears at the top level, through syntactic and semantic levels, to a representation of understanding at the deepest level. Another approach to machine translation involves work done by Bonnie Dorr at the University of Maryland based on Lexical Conceptual Structure (LCS) theory. This allows the knowledge represented in a text to be translated into a language independent data structure.

These theories provide a way of representing knowledge that is not reliant on the surface syntax of the text representing the knowledge. This will hopefully allow 'fuzzy matching' of sentences which have different syntactic structures but similar semantic

meaning. The authors will be looking at the possibility of grading essays via a comparison of these data structures.

Introduction

For many years researchers have been working on computerized methods for assessing the quality of a students' free text response to a question. The automated marking of objective tests, such as multiple choice questions, is commonplace but they have been criticized for only being able to assess lower order cognitive skills. For this reason, objective tests are often only used within an overall assessment strategy that would include the manual marking of essays and other types of free text.

Critics of automated free text assessment rightly claim that assessing the quality of an essay is a complex and fundamentally subjective process. On the other hand, researchers claim that the subjective nature of essay assessment leads to variation in grades awarded by different human assessors, which in turn is a source of unfairness. A system of automated assessment would at least be consistent in the way it assessed essays and if the system can be shown to grade essays within the range of those awarded by human assessors then enormous cost and time savings could be made. Critics continue to argue that automated systems can never model complexities of human grading. Meanwhile, researchers continue to develop systems that assess essays with relatively little effort and produce grades that generally emulate the grades of human assessors well.

This paper examines some historical systems and some current research. The algorithms adopted by the various systems are explained and their results are presented. The paper goes on to outline a new approach which is being developed by the authors based on ideas used in the area of machine translation.

Project Essay Grade

Ellis Batten Page of Duke University in the USA developed Project Essay Grade (PEG) (Page, 1968; Page, 1994; Page, 1995) since the mid 60's. Page uses, what he terms, *proxes*, which are computer approximations or measures of *trins*, in*trins*ic variables of interest within the essay (i.e. what a human grader would look for but the computer can't directly measure), to simulate human rater grading. *Proxes* include: essay length (as the amount of words) to represent the *trin* of fluency; counts of prepositions, relative pronouns and other parts of speech, as an indicator of complexity of sentence structure; and variation in word length to indicate diction (because less common words are often longer). *Proxes* are calculated from a set of training essays and are then transformed and used in a standard multiple regression along with the given human grades for the training essays to calculate the regression coefficients. These regression coefficients are the weighting that best simulates the given human grades when used with the calculated *proxes*. They are then used with *proxes* calculated from the unmarked essays to produce expected grades. Page's latest experiments (Page, 1995) have achieved results reaching a multiple regression correlation as high as 0.87, which is more reliable than a 6-judge panel, i.e. the

computer is predicting the scores that judges will assign to essays better than the judges are predicting each other.

PEG relies purely on a statistical approach that assumes that the quality of the essay, the *trins*, is reflected in the measurable *proxes*. No natural language processing is used and lexical content is not taken into account. PEG also requires training, in the form of assessing a number of previously manually marked essays for *proxes*, in order to calculate the regression coefficients, which enables the marking of new essays.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) was developed, in the main, by Thomas K. Landauer of the University of Colorado, Boulder and Peter W. Foltz of New Mexico State University (Landauer, 1997a; Landauer 1998, see also the LSA Website). LSA was not initially developed for use in automated essay grading but has been applied to this. It was first used for indexing documents and information retrieval and, therefore, much work on the essay grading aspect remains unpublished (for example, Landauer, Laham & Foltz (1998), 'Computer-based Grading of the Conceptual Content of Essays').

The technique, which is a method of representing contextual usage of words, proceeds along a number of steps. Firstly the essay is transformed into a matrix representation whereby each row represents a unique word and each column is a 'context', such as a sentence or a paragraph. Each cell then contains the frequency of the word appearing in that context. For example:

The sentences,
‘The man likes going on holiday’ (Context A)
and, ‘Spain is a popular holiday destination’ (Context B)
would give the matrix,

	Context A	**Context B**
man	1	0
holiday	1	1
Spain	0	1

Even though not all of the words from the original text would be represented in the matrix (certain *stopwords*, such as the, and, if etc. would be removed and morphological differences allowed for) a real essay would obviously create a very large matrix.

The initial matrix is then transformed. Each word occurrence is weighted as an estimate of it's importance in the passage, and inversely with the degree to which knowing that a word occurs provides information about which passage it appeared in. This means that if a word appears frequently in one context, but rarely in another, then that word is an important keyword for that particular context.

This first transformation of the initial matrix is similar to inverse document frequency (IDF) weighting (van Rijsbergen, 1979) that is often used in indexing and information retrieval. The weighting gives more importance to index terms that are more specific, i.e. those that occur less, because indexing specificity is inversely proportional to the number of documents an index term occurs in. IDF weighting is also used in the latest work by ETS (Burstein, 1998a; Burstein, 1998b; Burstein, 1998c).

Singular Value Decomposition (SVD), a form of factor analysis, is then applied is to the weighted matrix. SVD involves decomposing a rectangular matrix into the product of 3 other orthogonal (i.e. AAT=I, the original matrix multiplied by the its transpose is equal to the identity matrix) matrices. This gives you two rectangular/square matrices (depending on the dimensions of the original matrix) and one diagonal matrix consisting of the singular values of the original matrix.

For example, if we have the matrix, $A = \begin{bmatrix} 96 & 172 \\ 228 & 96 \end{bmatrix}$

Its' singular value decomposition would be,

$$\begin{bmatrix} 6 & -8 \\ 8 & 6 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 8 & 6 \\ 6 & -8 \end{bmatrix}$$

Next the dimension of the diagonal matrix is reduced thereby reducing the rank (the number of linearly independent rows and columns) of the original matrix. Finding the optimal dimension for the matrix is important so that correct induction of underlying relations between words and contexts can come through.

A new weighted version of the original matrix is now reconstructed by multiplying these three component matrices together. This dimension reduction increases the dependency of the data on each other, increasing the links between words and contexts, theoretically meaningfully.

Cosine correlation (also used in ETS's latest work) is used to measure the similarity of the reduced dimensional space constructed from a 'model answer', such as an instructional text taken from a course text or an essay prepared by the class tutor, against a student essay.

LSA has produced some impressive results. To use their own words, their approach has produced grades that correlate "approximately as well with experts' assigned scores as such scores correlate with each other, sometimes slightly less well, on average slightly better" (Landauer, 1998).

LSA makes no use of word order although they claim (Landauer, 1997b) that it is not the most important factor in collecting the sense of a passage. It also requires large amounts of data in order to be able to construct a suitable matrix representation of word use/occurrence and due to the size of the matrices involved any mathematical calculations will be, computationally, very expensive.

Educational Testing Service I

Jill Burstein and Randy Kaplan of the Educational Testing Service (ETS) spent some of the early 1990s working on a method for scoring short answer free text (Burstein, 1996). Their system does take account of actual content but works only on a sentence or a sentence fragment of between 15 and 20 words.

They present scoring of answers as a classification problem, scoring is achieved by correctly classifying responses by content. They claim that the system should also be able to determine when responses have duplicate meaning, i.e. when they paraphrase each other. The system works but only if the paraphrase, or appropriate metonym or synonym, has occurred in the training data. It uses the training data to tell the computer what to look for, but won't match against any other syntactic structure that might occur, although a thesaurus is used too.

They've been using the system to mark a 'formulating hypotheses' question type (F-H item), where students are given a situation and have to hypothesize as to why the situation has occurred.

A sample question might be:

‘Average students’ examination grades have been rising over recent years, explain why this might be.’

With possible answers:

‘Better teacher training’

‘Better facilities in schools’

‘Students can use CAL packages in their own time to revise’

The technique uses, what they term, lexical-semantic techniques, to build a scoring system, based on small data sets. It uses a domain-specific, concept-based lexicon and a concept grammar, both built from training data (200 essays from 378 in this run which is not really such a small training set) with some manual intervention which will be discussed later.

The training essays are parsed by Microsoft's Natural Language Processing Tool (MSNLP), any suffixes are removed by hand, and a list of *stopwords* is also removed. This produces a lexicon made up of any one, two or three word terms in the training data, modelled on the layered lexicon developed by Bergler (Bergler, 1995). This means the list of words and terms in the lexicon remain constant whilst the features associated with each entry are modular, so can be replaced as necessary. Some manual classification is needed at this point, such as the specification of some words as metonyms of each other and so on.

Grammar rules are then constructed, again manually, for each category of answer (each category should contain all the paraphrases for that possible answer) using syntactic parses of sentences from the training data along with the lexicon.

New essays are then parsed by the 'phrasal node extraction program' which outputs the sentences' noun, verb, adjective, adverb & preposition phrases etc. The system does not make use of specific parts of speech at this stage so they are collapsed into a generic XP phrase type and the sentence, or what is left of it now, is checked for

matches against the grammar and the lexicon. The XP phrase type is taken from X-bar syntax, which attempts to model common properties between the different syntactic components of noun, verb, adjective and preposition phrases etc. Instead of building a different grammar rule for each different type of phrase, X-bar syntax generalizes out to a single rule that applies universally to all (see Haegeman, 1991 for more information on X-bar syntax).

ETS claim 80% accuracy when marking the test set of essays and 90% when marking both the training and test set, i.e. using the training set to train the system and then including it in the test set for the marking part as well. 40% of errors are caused by gaps in the lexicon, where words that hadn't been manually identified as metonyms because they hadn't occurred in the training data and weren't synonyms in any thesaurus they were using. The second run, therefore, used an augmented lexicon. Constructing this new lexicon involved examining the test set as well as the training set to manually place metonyms (this was only done on the training set for the first run). This achieved 93% accuracy when marking on the test set and 96% when marking on both the training and test sets.

This high accuracy rate could be attributed to the question type only having a right or wrong answer. Grading is not actually done on a scale, or as a percentage, so there is a higher chance of the system matching the human assessor's mark. Other errors were caused by: concept structure problems, where a response could not be classified because it's concept-structure did not match any of the grammar rules for any of the categories; categorical cross-classification, where there was significant enough similarity between two categories for misclassification to occur, meaning students could lose marks because the system would think they had given two answers from the same category where in fact they had given two different categorical answers.

The system involves lots of pre-processing and much of it is manual, although ETS argue that the cost, in time, is still worth the saving. The F-H item is only a pilot type for the Graduate Record Examination (GRE) so they are only guessing that the questions wouldn't be marked faster than the 40 hours they say it takes to build the lexicon and the grammar (although with 28,000 students able to give up to 15 responses for each question, they're probably right). They have also said that they were planning to automate the generation of the grammar, which should cut the pre-processing time in half, but there has been no further word of this.

Educational Testing Service II

More recent work by Jill Burstein and others at the ETS (Burstein, 1998a; Burstein, 1998b; Burstein, 1998c) has focused on the Electronic Essay Rater (e-rater). Like PEG, e-rater uses statistical analysis but does take content into account, though like LSA it doesn't consider word order.

The new technique developed by ETS uses a 'Hybrid Feature Technology' incorporating syntactic, discourse structure and content features to emulate good essay traits, as suggested in the Graduate Management Admissions Test (GMAT) manual scoring guide, to calculate essay grades.

The GMAT manual scoring guide says that syntactic variety is an indicator of good essay writing skills. ETS use counts of the number of complement, subordinate, infinitive, relative clauses & occurrences of modal verbs such as would, could, might etc. to calculate ratios of these syntactic structure types, per essay and per sentence, as possible measures of syntactic variety.

Discourse structure analysis measures how well an argument has been formed, it gives an indication of the organization of an essay. More than 60 different features, similar to PEG's *proxes*, are used, most are discourse related syntactic features. The automated argument partitioning and annotation program (APA) identifies discourse units from surface clue words and non-lexical (i.e. syntactic) clues. For example: in summary and in conclusion denote summarization; perhaps & possibly are belief words so indicate argument development; this & there indicate staying on the same topic; infinitive and complement clauses might characterize the beginning of a new argument.

Annotated output from the APA is then fed into the topical content analysis unit. It has been found that good essays tend to use a more specialized and specific vocabulary, therefore, a good essay can be expected to resemble other good essays in its' choice of words, and similarly with bad essays. The e-rater system compares words in new essays to words found in training essays using two similarity measures, EssayContent and ArgContent.

EssayContent is based on word frequency and computed over the essay as a whole. Firstly, a representative vector for each of the six GMAT score levels is built by taking all the training essays with a particular score level and calculating a total frequency count for the words in the essays. This does not include *stopwords* and takes place after suffix stripping. An EssayContent vector is similarly derived for the new essay and compared against the score vectors using the cosine correlation. The new essay is then assigned the score of the most similar vector.

ArgContent is based on word weight, computed for each argument in each essay. The word frequency score vectors used in EssayContent are converted to weight vectors. Each word frequency in each vector is weighted with respect to how often the word occurs with respect to other words, and with respect to how many essays the word occurs in, in a similar fashion to the weighting technique used in LSA. Each argument, as partitioned by APA, is then evaluated separately (weight vectors having been constructed for each argument in each essay) against the score-level vectors to produce a set of scores for the essay. An adjusted mean, which allows for bias on essays with few or many arguments (few arguments in an essay gives a lower than average score from human raters and similarly, essays with many arguments get slightly higher than average grades), of these values gives the ArgContent.

Stepwise linear regression, as in PEG, is used to predict a human grader's score from computer analyzed features. Optimal weights, i.e. the regression coefficients, are found using a set of training essays. In this experiment 270 essays: 5 score level 0, 15 score level 1, 50 for each score level from 2 through 6, were used to train the system. The weights are then used with calculated features from new essays to predict scores.

Results from e-rater are impressive. EssayContent scores alone correlate with human graders at 69% whilst ArgContent scores alone correlate with human graders at 82%. Using both of these features together, correlations between 87-94% were achieved, so weighting obviously improves the scores.

Again statistical approach requiring training has been used, although it is better than PEG in that content is taken into account. Word order is still not considered, the system is simply spotting individual word occurrences, even though they are weighted for significance.

Machine Translation

As noted by Ellis Page (Page, 1966) much background work done in the area of linguistics, and in particular machine translation, could be put to use in our attempt to solve the problem of analysis of free text. We are able to use their ideas without having to consider many of their problems as we often do not need the same level of actual understanding of the text involved, rather just a means of representation to enable sensible comparison.

The following are theories of knowledge representation, which abstract away from the surface syntax of a language to try and capture the underlying semantics. We hope to employ such a language independent structure to enable us to overcome some of the problems we have highlighted in current essay grading systems. Translating essays into this sort of semantic structure will hopefully allow us to avoid having to have previous domain-specific knowledge, or having to train the system with large amounts of previously marked texts, whilst also avoiding the logical arguments attached to simply applying statistical methods. The task should be reducible to a comparison of students' essays against some model answer, both represented in semantic form, allowing 'fuzzy matching' of sentences which have different syntactic structures but similar semantic meaning. The recognition of matching sentences, a simple example being, 'John copies Jane' and 'Jane is being mimicked by John', is one of the fundamental problems of computerizing the essay grading process.

Panlingua

Panlingua (see the Panlingua Website) is a new idea currently being developed by Chaumont Devin. Panlingua was first mentioned in a small article in PC Week, 1st Dec 1998 and it involves a method of knowledge representation, based on semantic networks. Devin believes that there exists a universal language that he has called Panlingua. Language is independent from how knowledge is represented, i.e. irrelevant of the language in use the method of understanding what has been said, and accessing knowledge, is the same.

Devin attempts to model how the brain might translate from sensory patterns it sees or hears at the top level, through syntactic and semantic levels, to a representation of understanding at the deepest level. The word order, what he terms 'surface syntax', required to say the same thing differs from language to language, but the meaning of what is being said remains the same, i.e. syntax may vary but the semantics remains the same. However, he also says that some words have relationships at a deeper

level, which cannot be determined by the linear order of the words, that there is some sort of sub-surface level syntax that holds words together.

He models the brain as 4 horizontal layers and the links between those layers. The layers are:

- the top layer, the phonological plane - sights, sounds identified as symbols;
- the second layer, the syntactic plane - consists of nodes representing Panlingua atoms, or words;
- the third layer, the semantic plane - consisting of nodes called *semnods* (each atom in the syntactic plane will have a *semnod*);
- the fourth layer, the lower brain - where the stimuli associated with various symbols is handled.

To allow a more clear explanation of the links consider the sentence, 'The zebra was killed by a lion'. The links between the planes are:

- *synlinks* - these fall in the syntactic plane and they link Panlingua atoms, e.g. 'this' links to 'zebra' as a determiner, 'zebra' links to 'killed' as a patient, 'lion' links to 'killed' as an agent;
- *lexlinks* - these are between atoms and *semnods* in the syntactic and semantic planes. Most are of type 'default', i.e. no particular type at all, e.g. 'this', 'zebra' and 'lion'. 'Killed' however is a *lexlink* of type ' past tense declarative transitional'. These *lexlinks* are supposed to carry elements of meaning.

Devin claims there is no thought that cannot be represented in Panlingua, and it can also be used, by implementing semlinks in the semantic plane, as a way of creating an ontology. Semlinks correspond to English auxiliary verbs, such as 'to be' and 'to be able to (can)', therefore, crab could be linked to animal by 'isa' and so on. Ontologies can only represent binary relations though. It could manage 'Roses are red', but not 'John can dance the Jitterbug'. The different semlink types (hypernym, holonym, metonym etc.) allow much to be said about certain words with relatively few entries in the ontology because of the relationships inherent in the link types.

Panlingua requires a lexicon to work. Each entry consist of a word, its' *lexlinks* (of which it can have more than one as a word can have more than one meaning), and it's part of speech (not in the traditional grammar sense, as there is no guarantee that different languages will all have the same parts of speech, but derived from these, meaning the set of *synlinks* allowed for that symbol). Words are entered in the lexicon without regard for morphology, so there must be an entry for each of eat, eats, eating, eaten, ate, eatable etc.

Devin has outlined a possible application of Panlingua for machine translation. First separate lexicon-ontologies for language A and B are created. Then a translation table linking the *semnods* of language A to those of B is created. Ideally these would be bi-directional. Next you need to create the algorithms which map the Panlingua representations of language A to Panlingua representations of language B. Finally, you would need a parser for language A and a text generator for language B.

We have rejected the idea of using Panlingua as our method of representing sentences as all the work to date is completely of a theoretical nature, and in no way tested in practical areas. This is probably due to the fact that, by Devin's own

admission, it is inordinately difficult to construct a successful parser because of the need for successful disambiguation so as to correctly link to the right *semnod*. This would be necessary for our essay grading application, as well as the machine translation task, in order for the correct Panlingua representation to be built.

Lexical Conceptual Structure (LCS)

Bonnie Dorr, from the Department of Computer Science at the University of Maryland has been working in the machine translation area since the 80's. She works on the principle that in order to be able to get any sort of accuracy from a machine translation system, the system must be capable of capturing language-independent information - such as meaning, and relationships between subjects and objects in sentences; whilst still processing many types of language-specific details, such as syntax and divergence.

Divergence occurs when a translation from one language into another is not literal or word for word. For example, demotional divergence concerns word order -'I like eating' in English translates to the German 'Ich esse gern', literally 'I eat likingly', conflational divergence is where one word in one language translates to more than one in another - 'I stabbed John' in English translates to 'Yo le di punaladas a Juan' in Spanish, literally 'I gave knife-wounds to John', and thematic divergence involves the translation being entirely different - 'I like Mary' in English is 'Me gusta Maria' in Spanish, literally 'Mary pleases me'.

Both forms of knowledge, language specific and language independent are necessary for the machine translation task so Dorr has assumed that it is possible to convert the surface sentence into an internal representation, what they term an 'interlingua', that is common to more than one language. Using this approach, language specific differences between languages are captured by parameters that dictate how the interlingua maps to the syntactic structure (positioning of verbs and their direct objects etc.), whilst language independent differences, such as roles introduced by main verbs, are encoded internally in the interlingua.

Dorr's machine translation system, UNITRAN, (Dorr 1992, Dorr 1993) is capable of translating English, Spanish and German bidirectionally using the idea of an interlingua along with a single, uniform mapping between this representation and the syntactic structures for all three languages. It uses syntactic parameters based on Government-Binding theory, for when word order is different, and lexical-semantic parameters, based on Jackendoff's Lexical Conceptual Structure (LCS), for when the translation is not literal.

To explain briefly how sentences are represented in LCS (see Jackendoff, 1990 and Dorr, 1992 & 1993 for a more in depth discussion): types, such as Event, State, Position, Path, Place etc. are specialized into spatial dimension primitives, such as Go, Stay, Be, Orient; causal dimension primitives, such as Cause, Let; field dimension primitives, which extend spatially oriented primitives to other domains such as Possessional, Temporal, Identificational, Circumstantial, Existential. There are many other types and primitives covering the verb classes.

An example of how a sentence appears in an LCS representation:
The sentence 'John went home' becomes
[Event GO ([Thing JOHN], [Path TO ([Place HOME])])]

This interlingual representation based on LCS "abstracts away from syntax just far enough to enable language-independent encoding, whilst retaining enough structure to be sensitive to the requirements for language translation" (Dorr, 1993). This means that it should, therefore, also be sensitive enough for essay grading which can carried out without any of the trickier problems of machine translation, such as disambiguation.

There are of course problems with using LCS. Dorr identifies one of the major problems with trying to implement an interlingual representation, such as LCS, as its' dependence on the ability to define the primitives in terms of the types allowed. She also goes on to say, however, that "there has been a resurgence of interest in the area of lexical representation (with special reference to verbs) that has initiated an ongoing effort to delimit the classes of lexical knowledge required to process natural language. As a result of this effort, it has become increasingly more feasible to isolate the components of meaning common to verbs participating in particular classes. These components of meaning can then be used to determine the lexical representation of verbs across languages." (Dorr 1993)

She has skirted round this problem by restricting the vocabulary that their translator can operate on. UNITRAN is limited in what it can translate to 150 vocabulary terms in it's lexicon and 20 parameter settings, i.e. it can only translate a finite number of sentences, but enough to be able to see that the idea behind it is a sound one. We could tackle our problem with the same approach, implementing only a few of the types and primitives in order to construct a test set.

An important feature of LCS is that although it can distinguish across verb classes, it cannot make distinctions within classes, i.e. it wouldn't differentiate between roll, hurtle, move etc., they are all just verbs that indicate movement. Verbs can often be distinguished by including Manner in the representation though, which would allow, for example, differentiation between walk and run. Jackendoff states (Jackendoff, 1990) that, "it is not the business of conceptual structure to encode manner" but this could be particularly suitable for our essay marking application. What we have to decide is, exactly how much distinction is necessary between verbs. In some cases an occurrence of a verb class within an answer, as opposed to an actual specific verb, could be satisfactory. For example, if asked 'What did John do?' it might be enough to say 'John went to the shops'. In other cases though you might want to be more specific and say 'John ran to the shops'. The inclusion of Manner in the representation of sentences should perhaps be an optional feature allowing varying degrees of accuracy in answers.

This sort of system is obviously not intended to be able to mark creative writing assignments, which perhaps some of the other approaches could be applied to. It would only be suitable for use where a model answer of some form or another can be composed. Luckily, most examination questions are of this type, there is generally a correct response, an expected answer which we can use as a standard to compare new responses to.

Conclusions

As has been discussed, several different approaches have been tried to tackle the challenge of computerizing the grading of text responses to questions, and varying degrees of success have been achieved. Most however, require large amounts of training data, and in some cases even manual involvement is necessary, in order for the system to work.

Our approach hopes to reduce the task to a comparison problem using theories taken from the area of machine translation. Parsing answers to be marked into an interlingual representation should allow us to compare these structures against a similarly constructed representation of a model answer.

Our problem of fuzzy matching of sentences bears similarity to what is termed divergence in machine translation, where a translation isn't literal. Hopefully, this similarity will allow us to make use of their technology and apply it in our own area.

References

Bergler, S., (1995), 'From Lexical Semantics to Text Analysis', in P. Saint-Dizier (Ed.), Computational Lexical Semantics, Cambridge University Press

Burstein, J., Kaplan, R., Wolff, S. and Chi Lu, (1996), 'Using Lexical Semantic Techniques to Classify Free-Responses', in Proceedings of SIGLEX 1996 Workshop, Annual Meeting of the Association of Computational Linguistics, University of California, Santa Cruz

Burstein, J., Kukich, K., Wolff, S., Chi Lu and Chodorow, M., (1998a), 'Computer Analysis of Essays', NCME Symposium on Automated Scoring

Burstein, J., Kukich, K., Wolff, S., Chi Lu and Chodorow, M., (1998b), 'Enriching Automated Essay Scoring Using Discourse Marking', in Proceedings of the Workshop on Discourse Relations and Discourse Marking, Annual Meeting of the Association of Computational Linguistics, Montreal, Canada

Burstein, J., Kukich, K., Wolff, S., Chi Lu, Chodorow, M., Braden-Harder, L. and Mary Dee Harris, (1998c), 'Automated Scoring Using a Hybrid Feature Identification Technique', in Proceedings of the Annual Meeting of the Association of Computational Linguistics, Montreal, Canada

Dorr, B.J., (1992), 'The Use of Lexical Semantics in Interlingual Machine Translation', Journal of Machine Translation, 7(3), 135-193

Dorr, B.J., (1993), 'Interlingual Machine Translation: a Parameterized Approach', Artificial Intelligence, 63, 429-492

Haegeman, L., (1991), Introduction to Government and Binding Theory, Basil Blackwell, Oxford

Jackendoff, R.S., (1990), Semantic Structures, MIT Press, Cambridge Massachusetts

LSA Website: http://lsa.colorado.edu

Landauer, T.K., Dumais, S.T., (1997a), 'A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge', Psychological Review, 104, 211-240

Landauer, T.K., Laham, D., Rehder, B., Schreiner, M.E., (1997b), 'How Well can Passage Meaning be Derived Without Using Word Order? A Comparison of Latent Semantic Analysis and Humans', in Proceedings of the 19th Annual Conference of the Cognitive Science Society

Landauer, T.K., Foltz, P.W., Laham, D., (1998), 'Introduction to Latent Semantic Analysis', Discourse Processes, 25, 259-284

Page, E.B., (1968), 'The Use of the Computer in Analyzing Student Essays', International Review of Education, 14, 210-224

Page, E.B., (1966), 'The Imminence of Grading Essays by Computer', Phi Delta Kappan, 47(Jan), 238-243

Page, E.B., (1994), 'Computer Grading of Student Prose: Using Modern Concepts and Software', Journal of Experimental Education, 62(2), 127-142

Page, E.B., (1995), 'The Computer Moves into Essay Grading: Updating the Ancient Test', Phi Delta Kappan, 76(Mar), 561-565

Panlingua Website: http://www.strout.net/info/science/ai/panlingua/

van Rijsbergen, C.J., (1979), Information Retrieval, Butterworths, London