Bridging Gaps in Computerised Assessment of Texts

David Callear
Dept. of Info. Systems
University of Portsmouth
United Kingdom
David.Callear@port.ac.uk

Jennifer Jerrams-Smith
Dept. of Info. Systems
University of Portsmouth
United Kingdom
Jenny.Jerrams-Smith@port.ac.uk

Victor Soh
Dept. of Info. Systems
University of Portsmouth
United Kingdom
Victor.Soh@port.ac.uk

Abstract

A survey of major systems for the automated assessment of free text answers is presented. This includes the Project Essay Grade (PEG), Intelligent Essay Assessor (IEA) which employs Latent Semantic Analysis (LSA), and Electronic Essay Rater (e-rater). All these systems have the same weakness in that they are unable to perform any assessment of text content. The word order is not taken into account. In an effort to bridge the gaps in knowledge about this research problem, an introduction to a novel Automated Text Marker (ATM) prototype is given in this paper.

1. Introduction

A number of computerised free text assessment systems have been developed in the USA. Some based on the Latent Semantic Analysis (LSA) technique [6] have been independently tested in France [3] and there is at least one system being studied in the UK. None of the systems satisfactorily assesses text content.

This paper consists of a survey of the various major systems and an introduction to the authors' *novel* Automated Text Marker (ATM) prototype, primarily designed to assess text content.

2. Project Essay Grade (PEG)

Research in computerised assessment of students' essays dates back to the mid nineteen sixties, when Ellis Batten Page of Duke University in the USA developed the Project Essay Grade (PEG) [8].

Text content and word order are not taken into account, although PEG produced high correlations of around 80% between the computer-predicted and the human-assessed essay grades. This approach is based on the superficial surface features of an essay (counts of commas, semicolons, average word count, etc.) as indicators of its quality.

Whether PEG is suitable for assessing creative writing skills is debatable, although it is an indication that the assessment of essay style has been successfully automated. It is, however, unsuitable for assessing factual disciplines, in which text content is very important.

3. Intelligent Essay Assessor (IEA)

The Intelligent Essay Assessor (IEA) [5,6] was developed in the late 90s. It utilised the Latent Semantic Analysis (LSA) technique [6] which was originally meant for indexing documents and text retrieval in the late 80s. LSA was developed by Thomas K. Landauer of the University of Colorado, Boulder and Peter W. Foltz of the New Mexico State University in the USA.

LSA is not suitable to assess short answer questions and factual disciplines. The grammar and word order are not taken into account. LSA does not distinguish sentences such as, *"Country A bombed country B"* and *"Country B bombed country A"*, from each other.

4. Electronic essay rater (*e-rater*)

Jill Burstein of the Educational Testing Service (ETS) in the USA and others developed the Electronic Essay Rater (e-rater) [2,4] in 1998. The ETS is an organisation which conducts a number of world-wide standardised tests for the purpose of admissions to universities.

E-rater uses the Microsoft Natural Language Processing (MSNLP) tool [7] for parsing all sentences in the essays. In syntactic structure analysis, features identified include the numbers of complement clauses, subordinate clauses, infinitive clauses, relative clauses, subjunctive modal auxiliary verbs and others.

A grade prediction accuracy is determined by comparing human and e-rater grades across 15 test questions. The empirical results range from 87% to 94%. The system is similar to PEG, but the final linear regression model

0-7695-1013-2/01 $10.00 © 2001 IEEE

incorporates syntactic, rhetorical and some topical features. E-rater does not assess text content beyond spotting weighted keywords. The empirical results obtained are not from essays on factual disciplines.

5. Automated Text Marker (ATM)

The Automated Text Marker (ATM) prototype is designed for the purpose of automating the assessment of text content in detail. This adds up to a final summative score which reflects the detailed assessment of text content incorporating word order.

An infection is the invasion and multiplication of microorganisms on body tissue that produce signs and symptoms as well as an immunologic response.

Figure 1: Example Sentence

The two main components of ATM are the syntax and semantics analysers. ATM is written in Prolog. A model answer or an expertly written examiner answer to a closed-ended topic is automatically broken down into its basic concepts and dependencies, and the same is done with each student answer, then the two are compared.

```
DEPENDENCY GROUP 1
group(1,([infection] ➔ [is] ➔ [invasion])).
group(1,([infection] ➔ [is] ➔ [multiplication])).
group(1,([invasion] ➔ [of] ➔ [microorganism])).
group(1,([multiplication] ➔ [of] ➔ [microorganism])).

DEPENDENCY GROUP 2
group(2,([group(1)] ➔ [in] ➔ [body tissue])).

DEPENDENCY GROUP 3
group(3,([group(2)] ➔ [produce] ➔ [sign])).
group(3,([group(2)] ➔ [produce] ➔ [symptom])).
group(3,([group(2)] ➔ [produce] ➔ [immune response])).
```

Figure 2: CD Form of Example Sentence

Syntax analysers of varying complexity can be written in Prolog relatively simply and efficiently. The grammar can be augmented to include a wide-coverage, context-free and formalised grammatical description such as the Generalised Phrase Structure Grammar (GPSG) formalism [1].

A simple example of a sentence to be analysed is shown in Figure 1. For clarity, its conceptual dependency groups (CD) are shown in Figure 2 in output form, not the Prolog internal representation within the program.

Each fragment of concept is either totally independent (a dependency group by itself) or falls under a major dependency group, and is automatically given a numerical tag (number). Each numbered dependency group represents the context within which fragments of concept must be reclustered and segregated. These major dependency groups can be further related to each other so that successively larger dependency groups are generated and numbered automatically.

6. Conclusion

The difficulties in automating the assessment of text content incorporating word order are addressed in this paper. A solution is provided by ATM. Text passages are automatically broken down into their smallest viable units of concepts, and compared to an examiner's model answer.

7. References

[1] Bennett, P., *A Course in Generalised Phrase Structure Grammar*, UCL Press, London, 1995.

[2] J. Burstein, K. Kukich, S. Wolff, Chi Lu, M. Chodorow, L. Braden-Harder and Mary Dee Harris, "Automated Scoring Using a Hybrid Feature Identification Technique", *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics*, Montreal, Canada, 1998.

[3] P. Dessus, B. Lemaire and A. Vernier, "Free Text Assessment in a Virtual Campus", *Proceedings of the 3rd International Conference on Human System Learning*, Europia, Paris, 2000, pp. 61-75.

[4] Educational Testing Service (ETS), *E-rater*, website: http://www.ets.org/research/erater.html/, Princeton NJ.

[5] P. W. Foltz, D. Laham, T. K. Landauer, "Automated Essay Scoring : Applications to Educational Technology", *Proceedings of ED-MEDIA '99 Conference*, AACE, Charlottesville, 1999.

[6] T. K. Landauer, P. W. Foltz, D. Laham, "Introduction to Latent Semantic Analysis", *Discourse Processes*, 1998, vol. 25, pp. 259-284.

[7] Microsoft Corporation, *MSNLP,* website: http://research.microsoft.com/nlp/, USA.

[8] E. B. Page, "New Computer Grading of Student Prose : Using Modern Concepts and Software", *Journal of Experimental Education*, vol. 62, no. 2, pp. 127-142, 1994.