Improve the output from a MCQ test item generator using Statistical NLP

Robert Michael Foster

Research Institute in Information and Language Processing ,
University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1LY, United Kingdom
R.M.Foster@wlv.ac.uk

Abstract—**In this study I use statistical Natural Language Processing and adapted Controlled Language methods to pre-process individual documents before they are used as source documents for a system which automatically generates MCQ (Multiple Choice Question) test items. The literature observes that Natural Language Generation system evaluation is non-trivial and so the success of the featured methods is evaluated using a process suited to the featured domain. Generated MCQ test items are combined with items that have been created using traditional methods and then a routine is selected by a domain expert. The results provide some evidence to support the incorporation of some of the featured methods into future versions of the software.**

Keywords: e-learning, Natural Language Processing, Multiple Choice Question (MCQ) test item generation.

I. INTRODUCTION

The UK Company featured in this study regularly uses Multiple Choice Question (MCQ) test items to confirm staff knowledge of documents from the company's Policy Library. The MCQ test items are delivered in the form of pre and post tests associated with training courses and field audits.

The stored responses from these tests allow the company to demonstrate that training has been received by staff in accordance with requirements stated in UK Legislation. However an internal study proved that creating and updating the item bank manually is an expensive process. I am therefore investigating various ways to automatically generate MCQ test items, the most promising one being the application of a MCQ test item generator [4].[5]. The creators of this system were the only researchers in the field who expressed an interest in collaborating in order to make their system more usable for our domain. The MCQ test item generator [4].[5] uses the following steps to generate MCQ test items:

1. Identify significant terms within the source
2. Apply a clause filtering module
3. Transform the filtered clauses into questions
4. Use semantic similarity to select distractors.

During initial experiments with a particular policy document, most of its clauses were filtered out and so the number of usable MCQ test items produced was very small. In order to improve upon this performance various experiments are planned in which a variety of theories about language and learning are applied in pre-processing the source documents. The pre-processing featured in this paper aims to improve the accuracy of the selection of significant terms in step 1 above, reduce the number of sentences filtered out in step 2 and perhaps also improve the chances of locating believable distracters in step 4 above.

In order to examine the benefits and problems that arise from applying each theory within the pre-processing methodology, a domain specific evaluative measure is used. The evaluation is done by analyzing the selections made by a domain expert from a bank of MCQ test items. The relative proportion of automatically generated to manually created MCQ test items in the selections made by the domain expert is used as an evaluative measure of the proposed adaptations of the system.

This paper describes various attempts to improve system performance by incorporating statistical language processing techniques into the pre-processing of source documents. The evaluation method chosen must demonstrate that the generated MCQ test items are as close as possible to manually created ones

II. TECHNIQUES

A. Identifying suitable MCQ test item content

A count of selected features of each phrase within the source document leads to the identification of significant content for MCQ test items. This includes processing of previous versions of the relevant documents within the policy library.

B. Planting candidate MCQ components

Lexicalisation Theories in the literature seek to explain language change [8]. Language change theory is applied in this study by accelerating naturally occurring language change processes using Cataphoric Lexicalisation [12].

C. Restructuring source text

The domain is well defined by the company's policy document library. However, some problems have been encountered when applying natural language processing techniques to these documents in the past, and so an adapted form [3] of a popular theory about discourse structure [11] is applied to the source documents as a final stage in the pre-processing before the MCQ test item generator software [4].[5] is applied.

III. EXPERIMENT

The hypothesis is that pre-processing source documents in this way will assist the MCQ test item generator's significant term selection process. Another possible benefit is that the deliberate application of some Lexicalisation processes to source documents might produce some believable distracters.

The language processing methods described above were applied in the source document pre-processing stage of MCQ test item generation covering a subset of the topics taught in a training course. The domain expert who ran the training course had asked for a MCQ knowledge check from the researcher. The domain expert had drawn the material for the course from a specified document from the policy library. The theories were applied within a simulation as opposed to a reprogramming of the question generator [4], [5] in order to ensure careful and thorough application of the theories.

The sentences resulting from the pre-processing were inserted into the source documents in place of the extracted sentences and then the amended documents were used as source documents during the manual simulation of the operation of the MCQ test item generator[4], [5].

Existing manually created MCQ test items already provided assessment coverage of the source document. Limited time availability for item review by the subject expert within the commercial environment led to the decision to work with only 6 of the identified Noun Phrases and to apply 3 different Lexicalisation processes to those 6 paragraphs. The MTF (Multiple True/False) template for MCQ test items was selected as the default MCQ test item presentation format for this study following a review of relevant MCQ test item creation guidelines from the literature [9]

The simulated run of the MCQ test item generator upon the pre-processed source document therefore generated MTF test items for 3 language change processes x 6 source document changes = 18 new MTF items. These 18 new items were added to the 6 original items covering the same topics alongside the other 46 manually created MCQ test items.

The final selection by the domain expert was to consist of about 50 items with the aim of confirming apprentices' ability to recognise and recall facts following their attendance at his training sessions. He had no involvement in the creation of either the manually or automatically generated items and had no prior knowledge of which were generated MCQ items, therefore these factors could not have any bearing upon his decision about which items to include in his MCQ test item routine. The following usability scores were used to record the domain expert's assessments of the items:A= Use the item stem unchanged, B= Make minor changes and then use the item, C= Do not use the item.

Once the usability categories had been assigned for each of the items in the item bank, it could be seen that 2 of the 6 manually created MCQ test items had been replaced by MTF test items that had been generated following application of the method. Both of the selected MTF items had been built from the version of the pre-processed document containing the 'acronym' Language change.

IV. CONCLUSIONS AND FUTURE WORK

This has been an extremely small experiment. A larger number of MCQ test items and a more varied range of statistical NLP processes needs to be investigated before any confident conclusions can be drawn about the techniques used. Nevertheless, the fact that some of the generated items were chosen in preference to the original manually created items goes some way towards meeting the criteria specified in the introduction, whereby generated items need to be indistinguishable from manually created items when viewed by a domain expert.

In this study, the statistical methods have been applied manually and a pragmatic, domain specific evaluation method of output from the system has been used. The development effort likely to be involved in creating the software to implement these methods for the featured domain is less than might be imagined thanks to the well defined policy document library protected by a well organized change management system.

In future work, other relevant theories of human learning, controlled language [2] and computational linguistics [7] will be applied within modified versions of the pre processing methodology as we continue to seek to improve the quality of the output from the MCQ test item generator software [4],[5] in the domain featured in this paper.

REFERENCES

[1] Manning, C.D, and Schutze, H., 2007 *"Foundations of Statistical Natural Language Processing"* The MIT Press

[2] Goyvaerts, P. 1996 *"Controlled English, curse or blessing? A user perspective"* Proceedings of the 1st International workshop on Controlled Language Applications (CLAW '96) (Leuven). 137-42

[3] Foster, R.M. 2009 *"Improving the Output from Software that Generates Multiple Choice Question (MCQ) Test Items Automatically using Controlled Rhetorical Structure Theory"* In proceedings of RANLP 2009, Borovets – Student Conference

[4] Mitkov, R., and L. A. Ha. 2003. *"Computer-Aided Generation of Multiple-Choice Tests."* In Proceedings of HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing, pp. 17-22. Edmonton, Canada.

[5] Mitkov, R., L. A. Ha, and N. Karamanis. 2006. *"A computer-aided environment for generating multiple-choice test items."* Natural Language Engineering 12(2): 177-194.

[6] Reiter and R. Dale. 2000. *"Building natural language generation systems."* Cambridge University Press.

[7] Mitkov, R., 2004. *"The Oxford Handbook of Computational Linguistics "* Oxford University press

[8] Brinton, L.L., Traugott, C.G, 2005 *"Lexicalisation and Language Change"* - Cambridge University Press

[9] Haladyna, T.M., Downing, S.M., Rodriguez, M.C., 2002 *"A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment"* Applied measurement in education 15(3), 309–334

[10] Grishman, R. And Kitteridge R. (eds) 1986 *"Analysing Language in restricted domains: Sublanguage description and processing"* Hillsdale NJ: Lawrence Erlbaum Associates

[11] Mann, William C. and Sandra A. Thompson, 1988. *"Rhetorical Structure Theory: Toward a functional theory of text organization."* Text 8 (3): 243-281.

[12] Foster, R.M. 2010 *"Cataphoric Lexicalisation"* In proceedings for the ICEIS 2010 Conference, Funchal, Madeira