Improving the Output from Software that Generates Multiple Choice Question (MCQ) Test Items Automatically using Controlled Rhetorical Structure Theory

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

Abstract

A combination of established theories [1].[2].[3].[4] are applied in an attempt to improve the output from a system [5],[6] which automatically generates MCQ (Multiple Choice Question) test items from source documents. The literature observes that NLG (Natural Language Generation) system evaluation is non-trivial [7] and so the method is evaluated using a process suited to the featured domain [8]. The experiment intersperses 38 MCQ test items whose question stems have been generated using Controlled Rhetorical Structure Theory (CRST) with 62 manually created MCQ test items to form an item bank. A usability score is assigned to each item by a domain expert and these scores are used in the evaluation of the effectiveness of the method. The results provide some evidence to support the incorporation of CRST into future versions of the software.

.

Keywords

Controlled Language, Natural Language Generation (NLG), Rhetorical Structure Theory (RST), Multiple Choice Question (MCQ) test item generation, Controlled Rhetorical Structure Theory (CRST)

1. Introduction

Multiple Choice Question (MCQ) test items have been used by the UK Company featured in this study to regularly 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 [8]. However an internal study proved that creating and updating the item bank manually is an expensive process. In response to these results we are investigating various ways to automatically generate MCQ test items, the most promising one being the application of a MCQ test item generator [5], [6]. The creators of this system were the only researchers in the field who expressed an interest in collaborating with us in order to improve their system.

The MCQ test item generator [5], [6] uses the following steps to generate MCQ test items:

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

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. These pre-processing methodologies aim to avoid important clauses being filtered out in step 2 above.

In order to examine the benefits and problems that arise from applying each combination of theories in the pre-processing methodology, a domain specific evaluative measure is used. In the experiment presented in this paper 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 stems in the selections made by the domain expert is used as an evaluative measure of the proposed adaptations of the system.

The rest of this paper is organized as follows: section 2 describes the motivation for the study and provides a description of Controlled Rhetorical Structure Theory (CRST). Section 3 provides an example to illustrate the application of CRST from source document to output MCQ test item stem and Section 4 describes how the experiment was conducted before presenting the table of results. Conclusions and descriptions of proposed applications can be found in Section 5.

2. Context

2.1 Motivation

The Policy Library for the featured UK Company consists of a small number of general policy statements (POLs) and a large number of Standard Techniques (STs). The STs are intended to give precise instructions for the correct methods that the staff must apply when they are carrying out work on behalf of the company. Several of the Standard Techniques contain requirements for staff to complete sequences of MCQ test items (called 'CBT tests').

For example:

ST:OS7D – Relating to Audits of Operational field staff

"3.1 All Senior Authorised and Authorised Persons who hold an authorisation for HV Operational Work (11SW, 33SW, 66SW, 132SW and restricted variations) shall complete an annual CBT test to the satisfaction of an Examining Officer qualified to examine for that authorisation."

ST:HS17A – The Management of Asbestos Found in Operational, Non Operational Buildings and Equipment

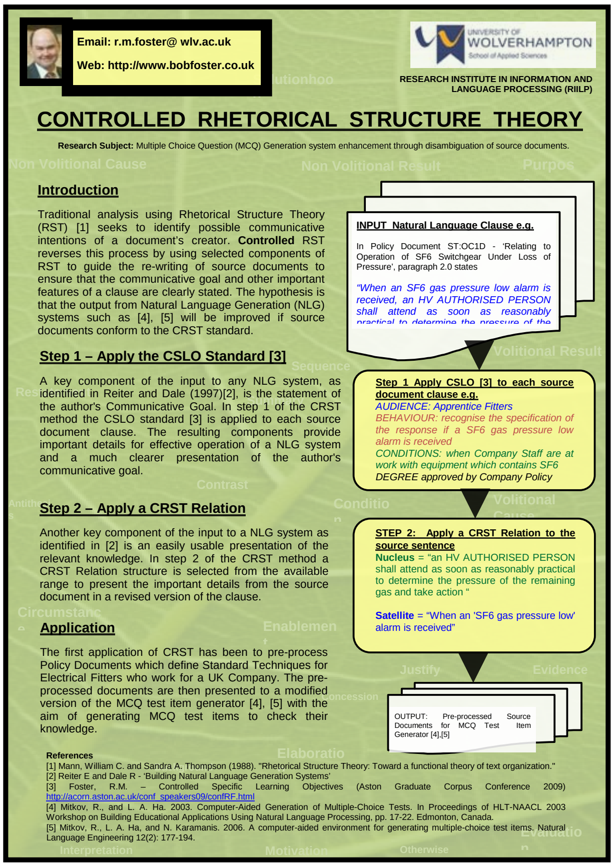
"3.2 After completing the above awareness course all staff shall, on an annual basis, complete the CBT Asbestos Knowledge Refresher which can be accessed via the Safety and Training Resources Catalogue."

In 2007 a review was carried out of the costs of producing and maintaining an item bank of 130 MCQ test items that were first created in 1991. The study demonstrated that item production and maintenance is particularly time consuming. A follow up research project has therefore been set up to analyse and improve the process for creating and maintaining the MCQ assessment tests used by this company.

The most promising approach identified so far has been the use of a MCQ test item generator [5], [6] to generate MCQ test items and post edit them to form the item bank. It has been reported in [5] and [6] that generating MCQ items using the generator can speed up the process by 4 times without compromising the quality of the output. However preliminary experiments applying the system [5], [6] to the policy library from the featured company delivered no usable MCQ test items so significant improvements in performance are necessary before the system could be adopted. This paper describes one attempt to improve system performance by incorporating CRST into both the pre-processing of source documents and the generation of the MCQ test items. An effective method for evaluating the output from this method must be identified and used consistently if the automatic generation of MCQ test items is going to be accepted. The evaluation method chosen must also demonstrate that the generated MCQ test items are as close to manually created ones as possible.

2.2 Controlled Rhetorical Structure Theory

Rhetorical Structure Theory [1] (RST) defines some widely used tools for Natural Language Discourse Processing. Controlled RST (CRST) adapts some of these tools to guide the controlled construction of discourse elements within a well specified domain.


Non Volitional Cause
Non Volitional Result
Purpos
Volitional Result
Sequence
Summary
Contrast
Antithesis
Conditio n
Volitional Cause
Circumstanc e
Enablemen t
Justify
Evidence
Concession
Elaboratio n
Background
Evaluatio n
Interpretation
Motivation
Otherwise

Email: r.m.foster@ wlv.ac.uk

Web: http://www.bobfoster.co.uk

UNIVERSITY OF WOLVERHAMPTON
School of Applied Sciences

RESEARCH INSTITUTE IN INFORMATION AND LANGUAGE PROCESSING (RIILP)

CONTROLLED RHETORICAL STRUCTURE THEORY

Research Subject: Multiple Choice Question (MCQ) Generation system enhancement through disambiguation of source documents.

Introduction

Traditional analysis using Rhetorical Structure Theory (RST) [1] seeks to identify possible communicative intentions of a document's creator. **Controlled** RST reverses this process by using selected components of RST to guide the re-writing of source documents to ensure that the communicative goal and other important features of a clause are clearly stated. The hypothesis is that the output from Natural Language Generation (NLG) systems such as [4], [5] will be improved if source documents conform to the CRST standard.

INPUT Natural Language Clause e.g.

In Policy Document ST:OC1D - 'Relating to Operation of SF6 Switchgear Under Loss of Pressure', paragraph 2.0 states

"When an SF6 gas pressure low alarm is received, an HV AUTHORISED PERSON shall attend as soon as reasonably practical to determine the pressure of the

Step 1 – Apply the CSLO Standard [3]

A key component of the input to any NLG system, as identified in Reiter and Dale (1997)[2], is the statement of the author's Communicative Goal. In step 1 of the CRST method the CSLO standard [3] is applied to each source document clause. The resulting components provide important details for effective operation of a NLG system and a much clearer presentation of the author's communicative goal.

Step 1 Apply CSLO [3] to each source document clause e.g.
AUDIENCE: Apprentice Fitters
BEHAVIOUR: recognise the specification of the response if a SF6 gas pressure low alarm is received
CONDITIONS: when Company Staff are at work with equipment which contains SF6
DEGREE approved by Company Policy

Step 2 – Apply a CRST Relation

Another key component of the input to a NLG system as identified in [2] is an easily usable presentation of the relevant knowledge. In step 2 of the CRST method a CRST Relation structure is selected from the available range to present the important details from the source document in a revised version of the clause.

STEP 2: Apply a CRST Relation to the source sentence
Nucleus = "an HV AUTHORISED PERSON shall attend as soon as reasonably practical to determine the pressure of the remaining gas and take action "

Satellite = "When an 'SF6 gas pressure low' alarm is received"

Application

The first application of CRST has been to pre-process Policy Documents which define Standard Techniques for Electrical Fitters who work for a UK Company. The pre-processed documents are then presented to a modified version of the MCQ test item generator [4], [5] with the aim of generating MCQ test items to check their knowledge.

OUTPUT: Pre-processed Source Documents for MCQ Test Item Generator [4],[5]

References

[1] Mann, William C. and Sandra A. Thompson (1988). "Rhetorical Structure Theory: Toward a functional theory of text organization."
[2] Reiter E and Dale R - 'Building Natural Language Generation Systems'
[3] Foster, R.M. – Controlled Specific Learning Objectives (Aston Graduate Corpus Conference 2009) http://acorn.aston.ac.uk/conf_speakers09/confRF.html
[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.

The poster presentation at RANLP 2009 (provided above) explains how CRST unites standard Rhetorical Structure Theory [1] with the theory of Controlled Specific Learning Objectives (CSLO) [4] which in turn incorporates concepts from AECMA Simplified English [3]. The inherent restriction of these theories to well defined domains does not present a problem in the context of this research since the domain is well defined by the company's policy document library from which the source documents are taken. MCQ test item stem templates are applied to the output from the CRST pre-processing and this paper presents the results when the MCQ test items produced are reviewed by a domain expert.

The first step in applying CRST to the process of creating a MCQ item routine is to use established text analysis methods [2].[3] encapsulated within the CSLO standard [4] to produce an unambiguous statement of the communicative goal of the MCQ test item routine. We use the term ‘communicative goal’ in the sense defined in the NLG methodology presented in [11]

The second step is to use these statements of objectives to select an appropriate sequence of CRST templates that when populated will produce an unambiguous presentation of the facts contained within the source document. In the third step of this application of CRST, a series of MCQ test item templates is populated using the content of the CRST-compliant statements generated in step 2.

The hypothesis of the CRST standard is that the translation of source documents into a sequence of CRST templates provides a presentation of the required facts that can be more easily interpreted. The reader of a CRST-compliant document can identify the writer’s communicative goal through the structure of the document i.e. the choices of CRST templates. Also the content words used within the CRST templates are either single sense words, as defined in the AECMA simplified English lexicon [3], or they are words used as defined in a domain specific lexicon compiled for the featured domain.

Disambiguation of this kind is particularly relevant to automatic MCQ test item generation from source documents because it has been noted that the current generator [5], [6] sometimes picks ‘the wrong clause’ during question stem production. The application of the CRST standard enforces clarity of content and the communicative goals of the source document writer and thereby makes such a mistake less likely.

3. Methodology

The examples included in this paper have been chosen to illustrate features of both the proposed source document pre-processing method and the MCQ item stem generation process. They have therefore been taken from the set of generated MCQ test items. However the decision about which items to include as examples was made after the day of the experiment and so their inclusion in no way affected the domain expert’s selections of item stems during the experiment.

CRST is a new application of Rhetorical Structure Theory [1]. This experiment applies CRST both in the source document pre-processing stage and the subsequent MCQ test item generation process. The application of the theory is achieved within a simulation as opposed to a reprogramming of the question generator in order to ensure careful and thorough application of the theory.

The source documents used in the experiment are taken from the policy library of the UK Company that was referred to in the introduction. A description of the method’s application to a particular source document paragraph is presented below in order to illustrate the process that produced the generated stems used in the experiment.

Example 1

In Policy Document ST:OC1D - ‘Relating to Operation of SF6 Switchgear Under Loss of Pressure’, paragraph 2.0 states

“When an SF6 gas pressure low alarm is received, an HV AUTHORISED PERSON shall attend as soon as reasonably practical to determine the pressure of the remaining gas and take action.”

The first step in applying CRST to this paragraph is to define the ‘communicative goal’ of the creator of the MCQ test item routine by applying the CSLO standard [4]. The first draft of this statement in relation to the given paragraph was written as follows:

“Apprentice Fitters recognise the correct description of the response required by Company Policy if a SF6 gas pressure low alarm is received when working with equipment containing SF6”

To ensure this conforms to the CSLO standard, the text is broken up into fragments specifying the Audience, Behaviour, Context and Degree and then re-written to ensure compliance with the controlled Lexicon. This led to several changes but also re-affirmed some of the word choices originally made. This was particularly noticeable within the statement of the required Behaviour where the verb ‘recognise’ was used. This is one of the verbs specified in Bloom’s Taxonomy and is therefore included in the CSLO Controlled Lexicon.

AUDIENCE: Apprentice Fitters

BEHAVIOUR: recognise the specification of the response if a SF6 gas pressure low alarm is received

CONDITIONS: when Company Staff are at work with equipment which contains SF6

DEGREE approved by Company Policy

The full analysis is provided in the table below:

Apprentice Fitters	Domain Specific Lexicon
recognise	CSLO Specific Lexicon
the	Standard Lexicon
specification (changed)	Standard Lexicon
of the	Standard Lexicon
response	CSLO Specific Lexicon
approved (changed)	Standard Lexicon
by	Standard Lexicon
Company Policy	Domain Specific Lexicon
If a	Standard Lexicon

SF6 gas pressure low alarm	Domain Specific Lexicon
Is received when	Standard Lexicon
Company staff (added)	Domain Specific Lexicon
are (added)	Standard Lexicon
at (added)	Standard Lexicon
work (changed)	Domain Specific Lexicon
with equipment	Standard Lexicon
which (added)	Standard Lexicon
contains (changed)	Standard Lexicon
SF6	Domain Specific Lexicon

Table 1 – Analysis and changes made to ensure conformity with the CSLO standard.

This statement clarifies the communicative goal of the creator of the MCQ test item routine, and allows more accurately targeted choices of MCQ test item templates to be made. The objective has been restricted to the checking of apprentices' ability to recognise (level 1 in the Cognitive Domain of Bloom's Taxonomy [9]) the significant facts that they would have been taught during their training. There is no expectation that successful completion of the MCQ test items would provide confirmation of abilities at the other levels within the cognitive domain such as understanding, application, analysis, synthesis or evaluation. Any of these levels which are relevant to the aims of the training and assessment scheme overall are addressed using other approaches, including observation of apprentice work by trainers during training courses and a series of on-site assessments (OSAs).

Having addressed the first step in the application of the CRST method, next comes the choice of an appropriate sequence of CRST templates in order to encapsulate the significant facts. In this case, as was true for many of the sentences processed in this experiment, a single CRST template is sufficient to deliver this content whilst ensuring unambiguous presentation of the communicative goal of the document writer.

The choice of the <Volitional Cause> CRST template in this case was triggered by the presence of the word 'When' in the source sentence:

> **VC Nucleus** = "an HV AUTHORISED PERSON shall attend as soon as reasonably practical to determine the pressure of the remaining gas and take action "
>
> **VC Satellite** = "When an 'SF6 gas pressure low' alarm is received"

This allowed the third and final step whereby one of the MCQ test item templates that can accept the fields defined in the VC CRST template could be applied. In this case the following MCQ test item template was selected:

> "<VC Satellite> What is the first step required by Policy?"

The resulting Generated stem was as follows:

> ***Generated Stem:*** *When an SF6 gas pressure low alarm is received, what is the first step required by Policy?*

This was accepted by the domain expert for inclusion in his MCQ test item routine without any post-processing

> ***Approved Stem:*** *When an SF6 gas pressure low alarm is received, what is the first step required by Policy?*

There now follows a description of how the experiment was conducted and the results of the selection by the Domain Expert.

4. Experiment and Results

On the day of the experiment a full set of 100 MCQ test items was presented to an expert in the featured domain [8]. Before the decision was taken to apply a logical, repeatable method when generating new MCQ items, 62 items had already been manually created using traditional methods. The specification of the job required a set of 100 items to be presented to the subject expert and so only 38 items were required from the application of Controlled Rhetorical Structure Theory (CRST) as described in section 2. Time constraints within the commercial environment prevented the production of any more items.

The general aim of the domain expert in making selections from the bank of 100 items was to confirm apprentices' ability to recognise and recall facts presented during training following their attendance at a series of training sessions. He had no involvement in the creation of either the manually or automatically generated items and had no prior knowledge of which MCQ item stems had been generated. 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

The items used in this experiment have three or four options and there is only one correct answer for each item in accordance with the recommendations from Haladyna and Downing [10]. For this experiment, only domain expert decisions about question stems are reported. Requirements about changes to the options (correct answer and distracters) within the MCQ test item are not reported.

In the case of the MCQ test item example presented in section 3, the item stem was judged by the domain expert to be accurately targeted towards the stated objectives without requiring post processing and so it was placed in category A. Other generated stems required minor post-processing before the domain expert was prepared to use them and still others were categorized ‘not to be used’.

Examples 2 and 3 provide examples of category B decisions. Example 4 is another example of a Category A decision. All generated items stems that were placed in Category C are unsuitable for inclusion in this paper:

Example 2

In Policy Document: ST:SP2A - Relating to Routine Substation Inspection, paragraph 2.1 states that

> *“A substation inspection is a careful scrutiny without dismantling and is normally done with plant and equipment live.”*
>
> **Generated Stem:** What is the definition of 'a substation inspection'?
>
> **Accepted Stem:** What is the Company Policy definition of 'a substation inspection'?

Example 3

In Policy Document ST:SP2J - Relating to The Routine Maintenance of 11kV and 6.6kV Cable Connected Secondary Switchgear (RMU’s, Switches and Fuse Switches), paragraph 2.7 states

> “*Before lowering tanks or working on equipment, ensure mechanism springs are discharged and all trip, close and power supplies are isolated.”*
>
> **Generated Stem:** *When maintaining 11kv Plant in a substation what must Company staff **ensure** before lowering tanks or working on equipment?*
>
> **Approved: Stem:** *When maintaining 11kv Plant in a substation what must Company staff **check** before lowering tanks or working on equipment?*

Example 4

In Policy Document ST: HS7C - 'Relating to Rescue from Height Techniques and Procedures, paragraph 3.6 states

> *“Before any rescue from height is attempted, an electrical risk assessment shall be carried out to determine the proximity of any adjacent live circuit.”*
>
> **Generated Stem:** What must be done before any rescue from height is attempted?
>
> **Approved Stem:** What must be done before any rescue from height is attempted?

Once the usability categories were assigned for each of the 100 items in the item bank, the following comparison table was produced:

	Generated MCQ stems	**Manually Created MCQ stems**
A=Use the item stem unchanged	**44% (17 stems)**	**43.5% (27 stems)**
B=Make minor changes and then use the item stem	**31% (12 stems)**	**43.5% (27 stems)**
C= Do not use this item stem	**23% (9 stems)**	**13% (8 stems)**

Table 2 - Usability categorization decisions for Generated vs Manually created MCQ test item stems.

5. Conclusions and Future Work

Applying the decision about category according to a clearly observable action allows the same process to be repeated consistently in other MCQ test item generation experiment evaluations. The separation of acceptable item stems between categories A and B also allows the calculation of three ‘alternative’ evaluation measures (A, B and A and B) as performance improves.

The most encouraging outcome from this experiment is the similarity between the proportions of generated to manually created MCQ test items in both categories A and C. This meets the criteria specified in the introduction whereby generated items need to be indistinguishable from manually created items when viewed by a domain expert. The fact that a slightly lower number of generated items required changes compared to the manually created items might have been due to stylistic differences between the domain expert and the writer of the manually created items.

The most serious limitation with this experiment is that other factors apart from the construction of the item stem will have contributed to the category allocation decision for each item that was made by the domain expert. In fact, the reason that none of the items given category C among the items with generated stems could be provided as examples in section 4 of this paper is because they had been categorized ‘not to be used’ for reasons other than poor construction of the stem. However, in defending the applicability of this experiment I return to the motivation laid out in section 2.1 in which it was clearly stated that generated stems must be indistinguishable from manually created items. This experiment provides very strong evidence to support this hypothesis.

The development effort likely to be involved in creating the software to implement the CRST method is not insignificant. However for the featured domain, the task is feasible because the domain is sufficiently well defined by the policy document corpus which has clearly defined boundaries and is protected by a well organized change management system.

This paper has prepared the ground for future experiments seeking improvement in performance of the featured software [5], [6] by pre-processing source documents. CRST has been applied and a pragmatic, domain specific evaluation method of output from the system has been used. It would not be unreasonable to compare the percentage results for categories A B and C arising from this experiment with similarly gathered results for item banks created using other MCQ test item generation methods.

In future work, other relevant theories of human learning, controlled language and cognitive linguistics 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 [5], [6] in the domain featured in this paper. The main obstacle to applying the method in other domains is the requirement for an AECMA Simplified English style domain specific dictionary. Perhaps future work from our team might provide some techniques for achieving the compilation of such a dictionary semi-automatically.

The evaluation method described in this paper will be applied and refined in future experiments for this project. The most significant refinement will be the inclusion of comparisons with the output from other MCQ test item generation systems.

6. References

[1] Mann, William C. and Sandra A. Thompson (1988). "Rhetorical Structure Theory: Toward a functional theory of text organization."

[2] Mager, R. (1975). Preparing Instructional Objectives (2nd Edition). Belmont, CA: Lake Publishing Co.

[3] AECMA Simplified English, http://www.aecma.org/Publications/SEnglish/senglish.htm 2003-04-16

[4] Foster, R.M. – Controlled Specific Learning Objectives (Aston Graduate Corpus Conference 2009) http://acorn.aston.ac.uk/conf_speakers09/confRF.html

[5] 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.

[6] 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.

[7] Reiter, E and Belz, A – Title: A Proposal for Shared-task Evaluation in NLG

[8] UK Legislation Health and Safety at work, etc Act 1974 http://www.hse.gov.uk/legislation/hswa.pdf

[9] Bloom, B. (1956), Taxonomy of Educational Objectives: Book 1 Cognitive Domain, Longman, 1956.

[10] Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? Educ Psychol Meas. 1993;53:999–1009

[11] Reiter E and Dale R - 'Building Natural Language Generation Systems'