Dialogue and Discourse 3(2) (2012) 101–124 doi: 0.5087/dad.2012.205

G-Asks: An Intelligent Automatic Question Generation System for Academic Writing Support

Ming Liu MING.LIU@SYDNEY.EDU.AU
Rafael A. Calvo RAFAEL.CALVO@SYDNEY.EDU.AU

School of Electrical and Information Engineering
University of Sydney
Sydney NSW 2006
Australia

Vasile Rus VRUS@MEMPHIS.EDU
Department of Computer Science
University of Memphis,
Memphis TN38152
USA

Editors: Paul Piwek and Kristy Elizabeth Boyer

Abstract

Many electronic feedback systems have been proposed for writing support. However, most of these systems only aim at supporting writing to communicate instead of writing to learn, as in the case of literature review writing. Trigger questions are potentially forms of support for writing to learn, but current automatic question generation approaches focus on factual question generation for reading comprehension or vocabulary assessment. This article presents a novel Automatic Question Generation (AQG) system, called G-Asks, which generates specific trigger questions as a form of support for students' learning through writing. We conducted a large-scale case study, including 24 human supervisors and 33 research students, in an Engineering Research Method course and compared questions generated by G-Asks with human generated questions. The results indicate that G-Asks can generate questions as *useful* as human supervisors ('useful' is one of five question quality measures) while significantly outperforming Human Peer and Generic Questions in most quality measures after filtering out questions with grammatical and semantic errors. Furthermore, we identified the most frequent question types, derived from the human supervisors' questions and discussed how the human supervisors generate such questions from the source text.

General Terms: Automatic Question Generation, Natural Language Processing, Academic Writing Support

1 Introduction

When students are asked to write a literature review or an essay, the purpose is often not only to develop disciplinary communication skills, but also to learn and reason from multiple documents. This involves skills such as sourcing (i.e., citing sources as evidence to support arguments) and information integration (i.e., presenting the evidence in a cohesive and persuasive way). In learning through writing, students need to consider trigger questions and monitor their understanding. Most students, however, fall short in these metacognitive skills (Graesser & Person 1994). Afolabi (1992) identified some of the most common problems that students have when writing a literature review, including not being sufficiently critical, lacking synthesis, and

 Submitted 2/11; Accepted 1/12; Published online 3/12

not discriminating between relevant and irrelevant materials. Simple generic questions have been used to address these problems. These are questions such as "*Have you clearly identified the contributions of the literature reviewed?*" and "*Did you connect the literature with the research topic by identifying its relevance?*" Reynolds and Bonk (Reynolds & Bonk 1996) showed that students given generic trigger questions in a writing activity perform better than those who received no trigger questions. However, generic questions may not always support the process of writing on specific topics. More content-related questions need to be asked, and most academics would ask such questions in the process of providing feedback to students.

In the field of Automatic Question Generation (AQG), most systems (Heilman & Smith 2009; Rus, Cai, & Graesser, 2007; Wolfe, 1976) focus on the text-to-question task where a set of content-related questions are generated based on a given text. Usually, the answers to the generated questions are contained in the text. For example, Heilman and Smith presented an AQG system to generate factual questions with an 'overgenerating and ranking' strategy based on Natural Language Processing (NLP) techniques, such as Name Entity Recognizer and Wh-movement Rules, and a statistical ranking component for scoring questions based on features. The target applications of such systems are reading comprehension and vocabulary assessment. These are significantly different from academic writing, which is our target application.

This is the first project, to our knowledge, that contributes a system for generating content-specific questions that support writing. Targeting this type of student-generated content has significant challenges distinct from those found in the literature for generating questions to support reading skills, where the content is made of textbooks and expertly written material. We contribute a description of how questions can be automatically generated from students' texts that is grounded on taxonomies developed in the writing research literature. We use the following citation categories - based on work by Lehnert et al. (1990) - relevant to literature review papers: *Opinion*, *Result*, *Aim of Study*, *System*, *Method*, and *Application*. Table 1 shows examples of automatically generated questions based on the citation category. For example, if a student (citer) cites an opinion in his academic writing thus: "*Cannon (1927) challenged this view mentioning that physiological changes were not sufficient to discriminate emotions*", G-Asks will generate trigger questions asking the student for evidence regarding the opinion. Examples of these are shown in the row corresponding to the *Opinion* category in Table 1. A goal of this study is to investigate how human experts generate their trigger questions from the source text, as a form of feedback, to support academic writing, what types of trigger questions are commonly used by human experts in writing, and how useful these questions are.

In addition, this project contributes a novel question generation technique based on sentence classification, particularly citation sentences which are common and informative elements in academic writing. This technique is used to develop and evaluate a tutoring system that scaffolds students' reflections on their academic writing with content-related trigger questions automatically generated from citations using NLP techniques.

Our basic approach to generate trigger questions is to first automatically extract citations from students' compositions together with key content elements. Then, the citations are classified using a machine learning approach, and questions are generated based on a set of templates and content elements. Experiments based on a Bystander Turing Test, a version of the original Turing Test (Person & Graesser 2002), revealed that human evaluators have moderate difficulties distinguishing questions generated by the system from questions produced by humans (Liu et al. 2010). The Bystander Turing Test (Person & Graesser 2002) asks participants to rate if particular dialog moves, such as hint and prompt, in tutoring transcripts are generated by the computer system or human tutors. It also measures the quality of generated dialog moves by asking the participants to give a score.

Category	Question	Source Sentence
Opinion	*Why did Cannon challenge this view mentioning that physiological changes were not sufficient to discriminate emotions? (What evidence is provided by Cannon to prove the opinion?) Does any other scholar agree or disagree with Cannon?*	*Cannon (1927) challenged this view mentioning that physiological changes were not sufficient to discriminate emotions.*
Result	*Does Davis objectively show that this classification accuracy gets higher from about 70 % up to 98 % while actors express emotions and computers perform the...? (How accurate and valid are the measurements?) How does it relate to your research question?*	*This classification accuracy gets higher from about 70% up to 98% while actors express emotions and computers almost perform the same on classifying 5-7 emotions (Davis, 2001).*
System	*In the study of Macdonald, why does workbench tool provide feedback on spelling, style and diction by analyzing English prose and suggesting possible improvements? What are the strength and limitations of the system? Does it relate to your research question?*	*The Writer's Workbench tool provides feedback on spelling, style and diction by analysing English prose and suggesting possible improvements (Macdonal et al, 1982).*
Application	*Why did Hunter use FBG arrays as tunable elements for high-speed signal correlating of grating-based processors? Could the problem have been approached more effectively from another perspective? Does it relate to your research question?*	*Hunter (2003) used FBG arrays as tunable elements for high-speed signal correlating of grating-based processors.*
Method	*Why did Ghosh develop an on-line algorithm capability of covering a complete range of faults from benign fault to faults of...? What are the strengths and limitations of this approach?*	*To achieve this, Ghosh (2004) developed an on-line algorithm capability of covering a complete range of faults from benign fault to faults of an unrestricted nature.*
Aim	*Why does Gawlik conduct this study to investigate biomass conversion in water at pressure and temperature ranges of 30-50 Mpa and 330-410 C? (What is the research question formulated by Gawlik? What is Gawlik's contribution to our understanding of the problem under study?)*	*Gawlik (2003) investigated biomass conversion in water at pressure and temperature ranges of 30-50 Mpa and 330-410°C.*

Table 1: An example of content-related trigger questions produced by G-Asks

The approached presented in this article improves on a former approach we developed in a previous study (Liu et al. 2010). In the previous approach, we proposed a combined Tregex expression rule-based approach with sentiment analysis to classify the citation sentences, and evaluated it with a pilot study on a small group of subjects. Tregex (Levy & Andrew 2006) is a powerful pattern matching technique which can denote the relations between syntactical tree nodes, such as Noun Phrase (NP) or Verb Phrase (VP), from the syntactic tree derived from a sentence by using a sentence parser, such as the Stanford Parser (Klein & Manning 2003). For example, Tregex pattern NP < NN $ VP is denoted as an NP immediately dominate an NN and sister VP. Six conceptual citation categories were used in that study: Opinion, Aim, Result, Method, System and Other. The study's result shows that the accuracy on citation extraction reaches 60% in 145 citation sentences. One of the biggest advantages of the Tregex expression pattern-matching rule is that it can match deep syntactic features of a sentence, such as predicate verb, subject and object. For example, if the predicate verb matches 'argue,' 'challenge,' or 'claim,' then the conceptual category of this citation sentence is 'Opinion'. We also used SENTIWORDNET (Esuli & Sebastiani 2006) to check if a sentence contains sentiment words. If it contains sentiment words, then it also is considered as 'Opinion'. However, creating rules is labor intensive, and this approach is not scalable. Here, we describe machine learning techniques that have significantly improved the citation classifier's performance. Furthermore, we have conducted a full-scale study, in a real course, with writing activities that involved 57 subjects, including 24 supervisors and 33 postgraduate students.

The remainder of the paper is organized as follows. Section 2 provides a review of the literature with a focus on writing support and AQG systems. It also describes several question classification schemas relevant to our work. Section 3 presents the major steps of our approach while section 4 details a case study we conducted to assess the quality of the generated questions using the Bystander Turing Test. Section 5 discusses the obtained results and suggests lines of future work.

2 Related Work

NLP techniques have been used to develop a number of tutoring and feedback systems for academic writing support. Section 2.1 reviews some of the writing support systems. Section 2.2 focuses on systems that generate questions automatically. Section 2.3 summarizes question classification schemas while section 2.4 presents automatic citation classification work.

2.1 Automated Feedback Systems for Writing Support

Computational approaches to writing support have focused primarily on assessment and less on providing automatic feedback on writing (Shermis & Burstein 2002; Williams & Dreher 2004). Despite a variety of initiatives to improve the quality of automatic feedback, the effectiveness of proposed systems remains to be proven and further research is needed. Meanwhile, providing timely and appropriate feedback at key stages of the writing process remains a manual task and therefore a serious challenge for university lecturers.

Some of the early systems include Writers Workshop (Anderson 2005), developed at Bell Laboratories, and Editor (Thiesmeyer & Theismeyer 1990), developed at Rochester Institute of Technology. Both systems focus on grammar and style. Studies of the impact of Editor (Beals 1998) concluded that the pedagogical benefits of grammar and style checking are limited. It could also be argued that these systems only aim at supporting writing to communicate as opposed to writing to learn.

SaK, a writing tutoring system (Wiemer-Hastings & Graesser 2000) developed at the University of Memphis, is based on the notion of voices that speak to the writer during the process of composition. SaK uses avatars to associate each voice with a face and personality. Each avatar provides feedback on a different aspect of the composition, pointing out the strong

and weak parts of the text but without correcting it. SaK uses Latent Semantic Analysis (LSA) to calculate the average distance between consecutive sentences and provide feedback on the overall coherence of the text. LSA is a technique used to measure the semantic similarity of texts (Landauer et al. 2007). SaK can also analyze the purpose of a sentence, identifying clusters of topics amongst student's writings so that when the topic of a new composition is not identified students can be asked for an explanation or reformulation.

Sourcer's Apprentice Intelligent Feedback system (SAIF) (Britt et al. 2004) is an automated feedback tool for writing essays which can be used to detect plagiarism, uncited quotations, lack of citations, and limited content integration problems. Once a problem is detected, SAIF can give helpful feedback to the student as shown in Table 2. SAIF also uses LSA techniques for plagiarism detection, computing the similarity between each essay sentence and the source sentences in the LSA space. For finding citations, SAIF uses a Regular Expression Pattern Matching technique to detect the explicit citations by recognizing phrases containing author name (e.g. *According to*, *As stated in*, *State*). Evaluations showed that SAIF provides feedback that encourages more explicit citations in students' essays. However, SAIF only addresses some basic problems for sourcing and integration. Moreover, it requires a large number of source documents to build the LSA semantic space and a large number of predefined pattern matching rules.

Problem	**Feedback prompts student to:**
1a. Unsourced copied material (plagiarism)	Reword plagiarism and model proper format.
1b. Unsourced copied material (quotation)	Explicitly credit source and model proper format.
2. Explicit citations	Explicitly make a minimum of 3 citations.
3. Distinct sources mentioned	Cite at least 2 different sources.
4. Excessive quoting	Paraphrase more instead of relying on quotations too heavily.
5.Integration from multiple sources	Include a more complete coverage of the documents in set.

Table 2: Types of Problems SAIF addresses and the intended goal of feedback

Glosser is an automated feedback system that provides academic writing support for college students (Villalon et al. 2008; Calvo & Ellis 2010). It uses textual data mining and computational linguistics algorithms to analyse various features of texts based on which feedback is provided to student writers. The feedback is in the form of generic trigger questions (adapted to each course) and document features that relate to each set of questions. For example, by analysing the words in each paragraph, Glosser can measure how related two adjoining paragraphs are. If the paragraphs are too unrelated, this can indicate lack of lexical cohesiveness which Glosser will flag. Glosser (1.0) provides feedback on four aspects of the writing: structure, coherence, topics, and concept visualisation. Glosser does not address sourcing directly, but four trigger questions are provided: *(1) Are the ideas used in the essay relevant to the question? (2) Are the ideas developed correctly? (3) Does this essay simply present the academic references as facts, or does it analyze their importance and critically discuss their usefulness? (4) Does this essay simply present ideas or facts, or does it analyze their importance?*

2.2 Computational Approaches to Natural Question Generation

One of the first automatic question generation systems proposed for supporting learning activities was AUTOQUEST (Wolfe 1976). In this case, as in most of the current research, questions are generated from external sources that students read (as opposed to write). The purpose of these questions is to help novices to learn English. The approach used in AUTOQUEST is similar to that of Kunichika et. al. (2001) who proposed an AQG approach based on both the syntactic and semantic information extracted from the original text. Their educational context was the assessment of grammar and reading comprehension around a story. The extracted syntactic features include subject, predicate verb, object, voice, tense, and sub-clause. The semantic information contains three semantic categories, noun, verb and preposition, which are used to determine the interrogative pronoun for the generated question. For example, in the noun category, several noun entities can be recognized including person, time, location, organization, country, city, and furniture. In the verb category, bodily actions, emotional verbs, thought verbs, and transfer verbs can be identified. It also extracts semantic relations related to the time, location, and other semantic categories, when an event occurs. Because this technique extracts substantial syntactic and time/space semantic information from sentences, the generated questions can be quite sophisticated, leading to very good writing support. Evaluations showed that 80% of the questions were considered by experts as appropriate for novices learning English and 93% of the questions were semantically correct (Kunichika et al. 2001).

For vocabulary assessment, there are recent attempts to automatically generate multiple-choice closed questions. In theory, they take reading materials and generate questions by removing some words from a source sentence. The two major issues in automatic multiple-choice question generation are: 1) to determine which words to remove from the source sentence and 2) to choose the wrong alternatives or distracters. Coniam (1997) determined the words to remove by selecting every nth-word in the text to be a test item and distracters are produced by choosing the same Part of Speech (e.g. noun, verb or adjective) and similar word frequency in a tagged corpus. Mitkov and Ha (2003) determined the words to remove by choosing the *Key Terms*, which are noun phrases with a frequency over a certain threshold. The distracters (e.g. hypernyms and hyponyms of the term) were selected by consulting WordNet (Fellbaum 1998), which is a lexical database that groups English nouns, verbs, adjectives and adverbs into synonym sets or synsets. A synset is linked to other synsets with various relations, including synonym, antonym, hypernym, hyponym and other semantic relations. They demonstrated that automatic generation and manual correction of questions can be more time-efficient than manual question creation alone.

AutoTutor, developed by Graesser et al. (Person & Graesser 2002) at the University of Memphis, is an intelligent tutoring system (ITS) that improves students' knowledge in computer literacy and Newtonian physics through an animated agent asking a series of deep reasoning questions that follow the Graesser-Person taxonomy (Graesser & Person 1994). In each of these subjects a set of topics have been identified. Each topic contains a focal question, a set of good answers, and a set of anticipated bad answers (misconceptions). The system initiates a session by asking a focal question about a topic and the student is expected to write an answer containing 5-10 sentences. Initially, the system used a set of predefined hints or prompts to elicit the correct and complete answer. More recently, the hints and prompts are automatically generated (Rus, Cai, & Graesser, 2007) and then human experts validate them instead of students evaluating the quality of hints and prompts as in Person and Graesser's study. The authors showed that AutoTutor's questioning approach had a positive impact on learning with an effect size on a pretest - post-test study of approximately 0.8 standard deviation units in the areas of computer literacy and Newtonian physics. However, the tutor system is domain dependent and requires a large number of human resources to predefine the content of each topic.

2.3 Question Taxonomy

Question taxonomies are developed based on analysis of human questions in tutoring situations, classroom teaching, or even technical manuals. The types of question are often related to Bloom's Taxonomy (Bloom 1984), which is a framework to assess cognitive processes. Deeper questions are highly correlated to higher-level processes in Bloom's Taxonomy. Question taxonomies have been proposed according to different application domains, including computational modeling of question answering as a cognitive process (Lehnert 1978), analyzing students' questions in a dialog between a human and intelligent tutoring agent (Acker et al. 1991), analyzing tutors' questions in human tutorial dialog (Graesser & Person 1994; Nielsen et al. 2008; Boyer et al. 2009). The most well known question taxonomy was one proposed by Graesser and Person (1994) based on their two studies about human tutors and students' questions during tutoring sessions in a college research method course and middle school algebra course. Six trained human judges coded the questions in the transcipts, obtained from the tutoring sessions, on four dimensions: Question Identification, Degree Specification (e.g High Degree means questions contain more words that refer to the elements of desired information), Question-content Category, and Question Generation mechanism (the reasons for generating questions include *knowledge deficit* in the learner own knowledge base, *common ground* between dialogue participants, *social actions* among dialogue participants, and *conversation control*). They defined following 18 question categories according to the content of information sought rather than on the interrogative words (i.e. why, how, where, etc).

1. Verification: invites a yes or no answer.
2. Disjunctive: Is X, Y, or Z the case?
3. Concept completion: Who? What? When? Where?
4. Example: What is an example of X?
5. Feature specification: What are the properties of X?
6. Quantification: How much? How many?
7. Definition: What does X mean?
8. Comparison: How is X similar to Y?
9. Interpretation: What does X mean?
10. Causal antecedent: Why/how did X occur?
11. Causal consequence: What next? What if?
12. Goal orientation: Why did an agent do X?
13. Instrumental/procedural: How did an agent do X?
14. Enablement: What enabled X to occur?
15. Expectation: Why didn't X occur?
16. Judgmental: What do you think of X
17. Assertion:
18. Request/Directive

After analyzing 5,117 questions in the research methods and 3,174 questions in the algebra sample, they found four frequent question categories: verification, instrumental-procedural, concept completion, and quantification questions. They stated that some questions could belong to more question categories. For example, the question "Did the drug dosage decrease the anxiety?" is ambiguous since it belongs to verification and antecedent questions. But, this type of question should not be construed as a weakness in the classification scheme because polythetic classification schemes can be used. In our study, we adapted the question taxonomy proposed by Graesser and Person to analyze questions generated by human experts. Section 4.6 describes it in more detail.

2.4 Citation Classification and Extraction

Citations are commonly used in research documents. Similar to question taxonomies, different citation classification schemas and methods have been proposed motivated by different purposes. Lehnert et al. (1990) presented a citation taxonomy based on a corpus of machine learning research papers, defining 18 conceptual reference categories including System, *Method, Concept, Result, Fact, Criticism, Example, Application, Proposal, Problem and Argument,* and 3 relationships between these categories including, *Similarity, Difference and Flagship*. The purpose of this research project was to summarize a scientific research paper in terms of underlying research trend. They used a rule-based sentence parser, called CIRCUS, to automatically fill out predefined frames containing reference categories and relationships slots. It was reported that the system correctly classified 75% of citations sentences. However, their evaluation is not quite convincing since they only used 28 citations sentences from 2 research papers. This rule-based approach does not generalize well to unseen data, and designing the parsing rules is very time-consuming. Because our work is similar to this study, we adapted some reference categories from Lehnert et al., but we used a an approach with greater generalization power based on a machine learning method proposed by Tefuel (2006).

Comparing to a rule-based approach proposed by Lehnert et al., Tefuel proposed a supervised machine learning approach to automatically classify the citation types in order to improve the impact factor calculations and citation indexer. They defined 12 mutually exclusive categories including *Weak (Weakness of Previous Researches), Contrast, Base (Current work is based on other research), Use (Current work uses other method), Similarity (Current work is similar to other work) and Neutral (Neutral description of cited work)*. The feature set contains 12 features that record the presence of 892 cue phrases identified by annotators, which include agent type (the authors of the paper or everybody else), action type (e.g. aim of study), location, verb tense and voice. They tested the approach using 2,829 citations from 116 articles, randomly selected from ACL (Association for Computational Linguistics) conferences. To report the classification results, Teufel and colleagues (2006) used a macro-averaging F1-Score (0.57) and the Ibk (k=3) algorithm, which is an alternative version of the k-nearest neighbour. There are several drawbacks of this approach and methodology: the distribution of citation categories is skewed and the proposed cue-phrases feature is too coarse-grained.

Since different automated citation classification methods were proposed, citation extraction and analysis have become a great interest to researchers. Powley and Dale (2007) introduced terminologies to describe the variety of citations styles, such as *Textual Citation (it uses author-year pair to refer to an entry in the reference list) and Index Citation (it uses numbers)*. They further divided the *Textual Citation* styles into 4 categories. Examples of textual citations are provided below in **bold** face:

1. Textual Syntactic: citations form a syntactic part of the sentence. E.g. ***Levin (1993)*** *provides a classification of over 3000 verbs.*
2. Textual Parenthetical: citations are enclosed in parentheses. E.g. *Two current approaches to English verb classifications are WordNet* ***(Miller et al., 1990)*** *...*
3. Prosaic: people name is used to refer to an earlier citation. E.g. ***Levin*** *groups verbs based on an analysis of their syntactic properties . . .*
4. Pronominal: pronoun is used to refer to an earlier citation. E.g. ***Her*** *approach reflects the assumption that the syntactic behavior of a verb is determined.*

Finally they used regular-expression-based heuristics to automatically extract citations from a research paper. The citation extraction recall was reported as 0.99 based on a collection of papers from 2000-2005 containing 294 citations. In our study, we will focus on *Textual Citation* extraction by using regular-expression-based heuristics.

3 System Design and Architecture

G-Asks has been integrated into our iWrite Web Application (Calvo et al. 2010), which allows students to write and submit their assignments, and provides them with a complete solution for supporting the write-review-feedback cycle of a writing activity. iWrite used automatic feedback tools for students to do revision, such as Glosser described in the Related Works section. These tools are based on TML (http://sourceforge.net/projects/tml-java/), a multipurpose text mining library that provides functionalities for the pre-processing of documents, such as tokenization, stemming, stop-word removal, sentence segmentation and building latent semantic space. It maintains three corpora, adding each new document, at the sentence, paragraph, and document level to an Apache Lucene database, which is commonly used in information retrieval and text mining tasks. Compared to a traditional database, such as MySQL, the speed of Lucene fulltext search and indexing is much faster.

Figure 1 shows the G-Asks system architecture and its integration to iWrite. The input to G-Asks is a literature review paper stored at the sentence level after the preprocessing in iWrite and the output is a set of generated questions used by iWrite, which delivers the questions to the student author.

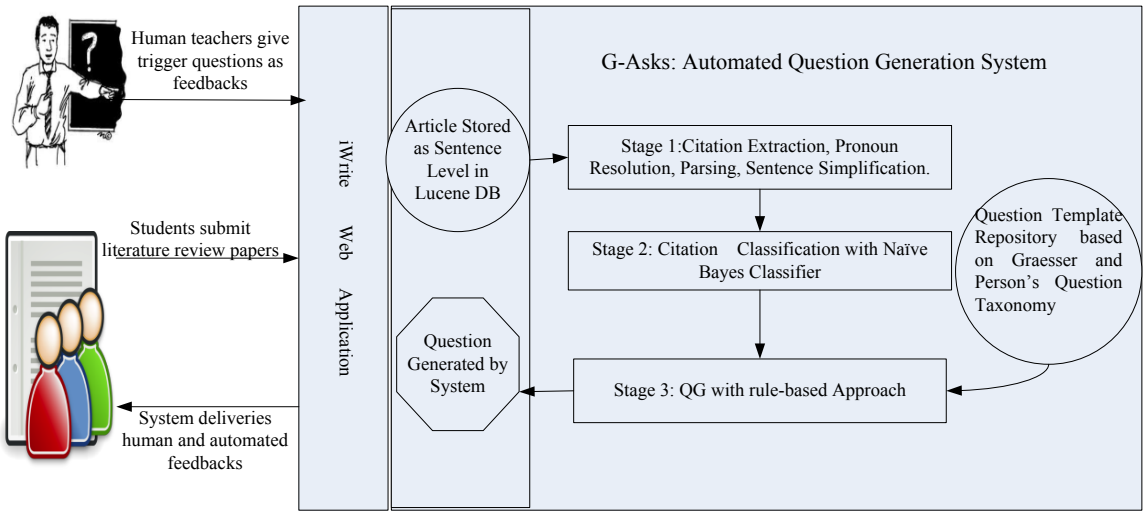

Figure 1: The G-Asks System Architecture integrated in iWrite Web Application

The question generation process follows 3 major stages.

Stage 1*: Citation Extraction.* In this stage, all the sentences are retrieved from the Lucene database and citation sentences are extracted, parsed and then simplified. In our current approach, we are only interested in *Textual Citations,* defined by Powley and Dale, containing names as this fits best with our goal of generating trigger questions for writing support. Thus, the *Index Citation* style is ignored. A pattern matching technique is used to extract the Textual Syntactic and Textual Parenthetical citation styles. The regular expression code is shown below.

\([a-zA-Z]\s*\d{4}\)|\([p.]+\s*\d{1,4}\)|\([a-zA-Z]+\s*[a-zA-Z]*
\s*[a-zA-Z]*\W*\d{4}|\([^)]*\d{4}\s?\)*

As you can see, this regular expression code would match the *textual syntactic* style by \([a-zA-Z]*\s*\d{4}\)|\([p.]+\s*\d{1,4}\) or match the *textual parenthetical* \([a-zA-Z]+\s*[a-zA-Z]* \s*[a-zA-Z]*\W*\d{4} .

A state of the art Named Entity Tagger (NER), LBJ (Ratinov & Roth 2009), is used to identify citations with Prosaic style, and a simple Pronoun Resolver, finding the nearest Name Entity appearing before the pronoun, was used to identify citations with Pronominal style.

Once citations are extracted using the previous approach, sentence simplification is performed which involves splitting compound and complex sentences and also removing phrase types such as appositives, non-restrictive relative clauses, and participial modifiers. After the sentence simplification is performed, we parse the simple citation sentences to get Syntactic features, including subject, main verb, and auxiliary verb (e.g. be, am, will, have and can) predicate, voice and tense, which are essential to perform subject-auxiliary inversion. We used the Stanford Parser (Klein & Manning 2003) and Tregex (Levy & Andrew 2006) for parsing sentence and sentence simplification separately. The Tregex expressions in Table 3 and 4 are used to simplify a sentence and extract subject, predicate verb and predicate from that sentence.

Description of Transformation	Tregex Expression
A simple sentence, containing one subject and a compound verb, is split into two sentences: 1 *subject + predicate1*; 2 *subject+ predicate2.*	*CC=conjuct $+ VP=predicate1 $- VP=predicate2 > (VP=predicateParent > (S > ROOT) $- NP=subject)*
A compound sentence, containing two or three independent clauses joined by a coordinator, is split into two or three sentences: *s1, s2, s3.*	*CC=conj $+ (S=s1 > (S=smain > ROOT)) $-- (S=s2 > (S > ROOT)) | $- (S=s3 > (S > ROOT))*
A complex sentence, containing an independent clause joined by one dependent clauses, is split into two sentences: *s1, np+vp*	*SBAR < IN < S =s1 $ (NP=np > (S > ROOT) $ VP=vp)*
Remove the apposition by *delete app, lead and trail.*	RRC|PP|SBAR|VP|NP=app $- /,/=lead $+ /,/=trail !$ CC !$ CONJP
Remove non-restrictive clauses by *delete comma and modifier*.	ROOT=root << (VP !< VP < (/,/=comma $+ /[^`].*/=modifier))

Table 3: Examples of Tregex expression rules used in stage 1 to split and compress complex sentences

Syntactic Feature	Tregex Expression
subject	NP > (S > ROOT)
predicate verb	/^VB/ > (VP > (S >ROOT))
predicate	VP > (S > ROOT)

Table 4: Examples of Tregex expression rules used to extract syntactic features.

Stage 2*: Citation Classification.* The goal of this stage is to identify the citation category (described next) for each citation candidate retrieved based on the citation style detection rules described above.

The description of each citation category is shown below:

1. Aim: to present the aim of an author's study, e.g. *Bunescu et al. focused on extracting named entities from natural language documents*.
2. Opinion: to express the opinion of an author, e.g. *Reiter and Dale (1997) state that template-based systems are more difficult to maintain and update.*

3. Result: to report the result of an author's study, e.g. *McCallum and Nigam (1998) show that the multivariate Bernoulli model performs well with small vocabularies.*
4. Method: to describe a method, algorithm, technique, model, or framework proposed by an author, e.g. *Bi-Normal Separation is a relatively new feature selection method introduced by Forman (2003).*
5. System: to describe a system, e.g. *AutoSlog (Riloff, 1996) is a dictionary construction system that creates extraction patterns automatically using heuristic rules.*
6. Application: to apply a method/system to a field, e.g. *Kappa statistics K will be used to measure the reliability (Siegel and Castellan, 1988).*

We implemented a statistical citation classifier using a machine learning approach. We represent each citation as a vector of 17 generic features. As a training set we used 504 citations from 45 academic papers. The features are described in the following:

Cue Phrases. This is similar to Teufel's approach (2006) that defines a verb cluster as a feature for a category. However, our approach provides more information to identify a feature. We call it a Cue Phrase Group that includes a noun cluster, an adjective words cluster and an adverb cluster. According to Hyland's study (1994), reporting verbs are widely used in academic writing and each reporting verb can be used for different reasons (Aim of study, Opinion and Result) in a citation. For example, a reporting verb list in the opinion verb cluster contains *argue, claim, view, reason, explain, emphasize and etc*. The opinion noun verb cluster includes *opinion, view, claim, limitation and etc*. Some verbs or nouns can be used by more than one category and we define these verb or noun clusters in a shared Cue Phrase Group. For instance, both the Opinion and Result category share the Cue Phrase Group containing *suggest, note, point to, observe and etc*. In other words, the Shared Cue Phrase Group Feature gives weight to both Opinion and Result Category. We define 12 Cue Phrase Group Features: 1) Aim Cue Phrase Group (verb, noun and adjective clusters), 2) Opinion Cue Phrase Group (verb and noun clusters), 3) Shared Opinion, 4) Result Cue Phrase Group (verb cluster), 5) Result Cue Phrase Group (result verb and none cluster), 6) Shared Aim & Result Cue Phrase Group (verb cluster), 7) Application Cue Phrase Group (verb and noun cluster), 8) Method Cue Phrase Group (noun cluster), 9) System Cue Phrase Group (noun cluster), 10) Shared System & Method Cue Phrase Group (verb and noun cluster), 11) Own (no cluster) and 12) Other Cue Phrase Group (verb, noun cluster and adjective cluster).

Sentiment Feature. This is a binary feature to check if the citation sentence contains sentiment words with polarity that is either positive or negative. The SENTIWORDNET (Esuli & Sebastiani 2006), a publicly available lexical resource for opinion mining, is an extension of WordNet and has three categories for a word sentiment with some magnitude: positive, negative and neutral. We utilize this resource to identify sentiment words.

Negation Feature. We define the following four cue phrase groups (70 words in total) to detect negation in a citation sentence:

1. Traditional negation words, such as not, no, never, neither, nor, none and not only.
2. Restrictive adverbs, such as few, little, rarely, seldom, hardly, scarcely, barely.
3. Verbs with negative meaning, such as fail, deny, avoid.
4. Adjectives with prefix in-, dis-, un- and non, such as insufficient, imbalance, uncommon, nonassessable and insignificant.

These words are obtained from the frequent academic word lists (Coxhead 2000).

Syntactic Features. We use the voice and tense features from Teufel's study (2006).

Other Features. We use the length of a citation sentence and the numeric features which indicate if the citation sentence contains numeric characters.

Stage 3: *Generation*. The final stage of our approach is the actual trigger question generation module. It uses a template-based approach. Once the semantic and syntactic features extracted

from a citation match the predefined patterns in our repository of templates, the corresponding questions are generated. Table 5 shows the 6 rules defined in our Rule Repository. For example, a citation is extracted in Step 1: *Cannon (1927) challenged this view mentioning that, physiological changes were not sufficient to discriminate emotions.* In Step 2, the citation classifier categorizes it as Opinion. Step 3 applies Rule 1 to generate the following question to trigger the student's reflection by asking for the evidence for other person's opinion: *Why did Cannon challenge this view mentioning that physiological changes were not sufficient to discriminate emotions? (What evidence is provided by Cannon to prove the opinion?) Does any other scholar agree or disagree with Cannon?* In order to fill in this question template shown in the Rule 1 of Table 5, the Subject_Auxiliary_inversion operation occurs where the auxiliary precedes a subject. The key to this implementation is to find the auxiliary verb. We used predefined Tregex Patterns to implement this. For example, this Tregex pattern " MD > (VP > (S > ROOT)) " is used to find the modal auxiliary, such as can, could, may, might, need and ought while the pattern "/^VB/ > (VP > (S > ROOT)) < (are|is|am|was|were|has|have|had|do|will|would|should|) is used to find the regular auxiliary, such as be, have/has, do, will, would, shall, should and had. If we couldn't find both auxiliary in the sentence, we will set the auxiliary verb as do, does or did depending on the tense of main verb and singular or plural of the main verb.

Rule	Category	Question Template
1	Opinion	Why +subject_auxiliary_inversion()? What evidence is provided by +subject+ to prove the opinion? Do any other scholars agree or disagree with +subject+?
2	Result	Subject_auxiliary_inversion()? Is the analysis of the data accurate and relevant to the research question? How does it relate to your research question?
3	System	In the study of +subject+, why +subject_auxiliary inversion()? What are the strength and limitations of the system? Does it relate to your research question?
4	Application	Why+Subject_Verb_Inversion()? Could the problem have been approached more effectively from another perspective? Does it relate to your research question?
5	Method	In the study of +subject+, why +subject_auxiliary _inversion()? Which dataset does +subject+ use for this experiment? What are the strengths and limitations of this approach?
6	Aim	Why does +subject+ conduct this study to +predicate+? What is the research question formulated by +subject+? What is +subject+s contribution to our understanding of the problem?

Table 5 Six rules and template-based questions.

4 Evaluation of the Automatic Question Generation System

To evaluate the ability of G-Asks to generate high quality and effective questions for supporting academic writing, we compared questions generated by the system to those produced by humans. Like the Bystander Turing Test conducted by Person and Graesser(2002), in this study judges (student writers) rated the quality of each question according to different measures and were also asked to ascertain whether the question was generated by a human or a system. However, there are some differences between the tests carried out by Person and Graesser and our two evaluations. We used the academic writing task to generate questions while they used a snippet of

a tutorial dialog. Furthermore, our judges were the writers of the content while in Person and Graesser's study judges did not know the content before the experiment.

4.1 Participants and Procedure

We conducted a study with 57 participants (33 PhD students-writers and 24 supervisors). The students were enrolled in a Research Methods course from the Faculty of Engineering at the University of Sydney. Each student submitted a research proposal as part of this subject (and their PhD requirements) to the iWrite web site. Each proposal was read by a peer, who is another PhD student from this course, and the supervisor, who is supervising the student-writer of this proposal, both providing feedback in the form of questions. Having been informed about this experiment, each student was asked to rate the quality of questions generated from his/her literature review paper. These questions were produced by the supervisor, a peer, and by G-Asks, combined with a sample of generic questions shown in Table 6.

	Generic Questions
1	*Did your literature review cover the most important relevant works in your research field?*
2	*Did you clearly identify the contributions of the literature reviewed?*
3	*Did you identify the research methods used in the literature reviewed?*
4	*Did you connect the literature with the research topic by identifying its relevance?*
5	*What were the author's credentials? Were the author's arguments supported by evidence?*

Table 6: Examples of Generic Question Type.

Each question producer (supervisor, peer, G-Asks or generic question) generated a maximum of 5 questions. Therefore, each student evaluated 20 questions at most. Students were then given the following five quality measures, some of which were used in similar studies by Heilman and Smith (Heilman & Smith 2009), to evaluate the questions using a Likert scale where 1 was 'Strongly disagree' and 5 'Strongly agree':

1. *This question is correctly written (QM1).*
2. *This question is clear (QM2).*
3. *This question is appropriate to the context (QM3).*
4. *This question makes me reflect about what I have written (QM4).*
5. *This is a useful question (QM5).*

We received 5 ratings each for 615 questions under each quality measure. Quality measures 1, 2 and 3 focus on 'acceptability' of the generated question (whether it is grammatically correct, not vague, and makes sense according to the context) while the quality measures 4 and 5 focus on the 'usefulness' of the generated question, whether it is helpful to trigger reflection.

4.2 Citation Extraction Evaluation and Result

First, we evaluated G-Asks's performance with respect to citation extraction ability and semantic correctness of generated questions. The training set consisted of 45 papers as described above. The 33 literature review papers used for testing contained 534 citations, out of which 469 were extracted (see Table 7). For example, this citation "*Heesang et al. (2007), based on ant colony behavior, proposed a new path-flow routing algorithm for the backbone network of the next generation networks."* is extracted by using our predefined regular expression for *textual syntactic* style. However, the citation is not valid in our case because the name entity recognizer couldn't identify the *Heesang* as a person name, and the author name is required in our question

generation process. In addition to these 469 extracted citations, 20 non-citations were wrongly extracted as citations because of LBJ NER tagger errors. For example, this citation "*One of the examples is the unbalanced Mach Zehnder interferometer filter which based on the cascade of two couplers."* was wrongly identified as a citation since the name entity recognizer wrongly identified the *Mach Zehnder* as a person name.

Number of Citations	534
Extracted Citations	469
Citation Extraction Rate	88%

Table 7: Citation Extraction Rate.

Within these 469 citations, 21 generated questions that had serious grammatical errors because of the semantic parser and sentence splitter's performance. Therefore, we only evaluated the statistical citation classifier using 448 citations.

4.3 Citation Classification Performance

This experiment examined whether our current method (using machine learning) achieves higher accuracy compared to Liu et al.'s (2010) rule-based approach when running the same conditions. We also evaluated the performance on different learning algorithms. Because in the previous study we found that the application category was frequent, we added this new citation category in our study. The testing dataset contained 448 citations which belong to one of the seven categories. We used balanced F_1-score, precision and recall to measure the classifier's performance. The F_1 of each class is computed by the following formula:

$$F_1 = \frac{2 * precision * recall}{precision + recall} \quad (1)$$

The precision for a class is the number of true positives (i.e. the number of citation sentences correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class, while the recall for a class is the number of true positives divided by the total number of elements that actually belong to the positive class. Table 8 shows the accuracy of the rule-base classifier compared to our method in terms of six common categories.

Model / **Category**	**The Rule-based Classifier**			**Current model (Naïve Bayes)**		
	P	R	F	P	R	F
Opinion	0.49	0.44	0.47	0.76	0.75	0.75
Result	0.90	0.52	0.66	0.76	0.93	0.84
System	0.80	0.32	0.46	0.85	0.41	0.55
Method	0.73	0.28	0.4	0.78	0.75	0.76
Aim	0.85	0.46	0.6	0.71	0.74	0.72
Other	0.68	0.19	0.3	0.49	0.66	0.56
Average	0.74	0.37	0.48	0.73	0.71	0.7

Table 8: The accuracies of Liu et al's rule-based classifier and our method. P stands for Precision, R recall and F F-score.

Running the same conditions, our current machine learning approach achieved higher accuracy than the rule-based classifier based on the Tregex Expression described in the Related Work section. Although the Tregex expression rules are good at extracting the semantic features of a sentence, such as predicate verb, subject and object and matching the well-defined cue phrases, this method still has a problem when dealing with sentences with more complex syntax. We propose that a machine learning approach can handle this very well because only the individual NP or VP are considered.

We also evaluated the citation classifier's performance with different classification algorithms. Table 9 shows the citation classifier's performance and how well each class was predicated by each classifier. We can see that *aim, result, opinion,* and *application* can be identified quite well using defined features, except in some cases where the function of a citation is implicit. For example, this is an implicit citation from the aim category: *There has been much interest recently in devising strong game-theoretic strategies.*

We also observe that the system category is difficult to identify, because it mainly depends on the noun cluster feature containing words such as, *system, tool,* and *agent,* and a verb cluster feature including words such as *devise, present, develop,* and *propose.* When the citation contains a technical term and doesn't contain any verb in the shared verb cluster, it would be difficult to identity. For example, *The CIRCSIM-Tutor [12] teaches cardiovascular physiology by describing…* In this case, it is hard to identify *CIRCSIM-Tutor* as a system name.

Classifier / **Category**	**Naïve Bayes**			**Support Vector Machine**			**J48 Decision Tree**		
	P	R	F	P	R	F	P	R	F
opinion	0.76	0.75	0.75	0.41	0.52	0.46	0.42	0.68	0.52
result	0.76	0.93	0.84	0.79	0.84	0.81	0.76	0.77	0.77
system	0.85	0.41	0.55	0.91	0.1	0.18	0.35	0.8	0.48
application	0.69	0.89	0.78	0.56	0.88	0.68	0.73	0.67	0.7
method	0.78	0.75	0.76	0.39	0.85	0.53	0.57	0.8	0.67
aim	0.71	0.74	0.72	0.56	0.78	0.65	0.65	0.55	0.6
other	0.49	0.66	0.56	0.85	0.01	0.02	0.73	0.19	0.3
average	0.72	0.73	0.71	0.60	0.66	0.48	0.6	0.64	0.58

Table 9: The citation classifier's performance with different learning algorithms. P stands for Precision, R recall and F F-score.

4.4 Question Quality Evaluation and Result

A total of 615 questions were generated based on the 33 literature review papers. Table 10 shows that the supervisors generated 142 questions (107 citation related questions and 35 non-citation related), while the peers generated 151 questions (133 citation related questions and 18 non-citation related). The non-citation related questions addressed other types of writing feedback, such as clarity, organization, and referencing.

Question Producer	Number of questions	Question types
Supervisor	142	Citation related (107)
		Non-Citation related (35)
Peer	151	Citation related (133)
		Non-Citation related (18)
G-Asks	161 (randomly sampled from 469 questions)	161 Citation related
Generic system	161	161 Generic
Total	615	562 questions were evaluated

Table 10. Number of Questions Produced from 33 Literature Review papers

In order to make it comparable, we only used the 562 citation-related questions by supervisors and peers and the 161 questions by the system as well as 161 generic questions. Table 11 shows average scores of the questions according to different quality measures. Supervisors' questions (average score: 4.4) outscored G-Asks questions (average score: 3.8), peer generated questions (average score: 3.77), and generic questions (average score: 3.73).

Quality Measures / Producer	QM1: correctness	QM2: clarity	QM3: context appropriate	QM4: trigger reflection	QM5: usefulness	Average
G-Asks	3.94	3.91	3.76	3.69	3.69	3.80
Supervisor	4.57	4.53	4.45	4.26	4.20	4.40
Peer	3.93	3.96	3.77	3.62	3.56	3.77
Generic	4.04	3.92	3.65	3.49	3.57	3.73

Table 11: Comparisons of Normalized Mean Scores

A one-way ANOVA was conducted to examine whether differences in the average score, as well as in each quality measure, were statistically significant. The ANOVA yielded a significant difference in Average ($F(3,558)=12.15$, $p<0.05$), QM1 ($F(3,558)=9.254$, $p<0.05$), QM2 ($F(3, 558)=8.863$, $p<0.05$), QM3 ($F(3, 558)=12.552$, $p<0.05$), QM4 ($F(3, 558)=11.103$, $p<0.05$), and QM5 ($F(3, 558)=7.913$, $p<0.05$). Follow-up Fishers' least significant difference (LSD) tests with 95% confidence interval were performed to determine which pairs of treatments differed from one another. Table 12 shows the mean differences (MD) and LSD results from which we conclude that the questions from supervisors significantly outscored G-Asks in all the measures. There were no statistically significant differences between questions generated by the peer and G-Asks, or between Generic Question and the G-Asks.

Mean Difference / Criteria	G-Asks Vs Supervisor	G-Asks Vs Peer	G-Asks Vs GQ
QM1	MD =0.632	MD =0.006	MD =0.099
	LSD=0.259	LSD=0.246	LSD=0.236
QM2	MD =0.626	MD =0.056	MD =0.012
	LSD=0.264	LSD=0.251	LSD=0.239
QM3	MD =0.691	MD =0.009	MD =0.112
	LSD=0.270	LSD=0.256	LSD=0.245
QM4	MD =0.572	MD =0.065	MD =0.199
	LSD=0.267	LSD=0.254	LSD=0.243
QM5	MD =0.516	MD =0.126	MD =0.118
	LSD=0.280	LSD=0.266	LSD=0.255
Average	MD =0.607	MD =0.026	MD =0.063
	LSD=0.237	LSD=0.225	LSD=0.215

Table 12: Fisher's least significant difference (LSD) tests with 95% confidence interval.

Because the automatically generated questions were randomly sampled, some had grammatical and semantic errors due to the performance of the Semantic parser, the Citation Classifier and the LBJ Name Entity Tagger. This may have degraded the scores of the G-Asks questions. We performed further analysis on the quality of the AQG system by excluding 35 questions with grammatical and semantic errors. As shown in Table 13, supervisors' questions still got the highest score in each quality measure, but the G-Asks's questions now takes the second place with an average score of 4.10. We performed post-hoc analysis using LSD tests as we did before. Table 14 shows that the questions from the G-Asks significantly outscored generic questions in each quality measure while outperforming peers' questions in quality measures 1, 3, 4, 5 and Average. As expected, the G-Asks outscored Generic Questions because the content-related questions were more helpful than the generic questions. Moreover, the system questions significantly outperformed human peer questions, which might be explained by the following factors. First, peers may not be familiar with the topic of the literature review paper which he or she reviewed so the peer cannot generate very useful questions. Secondly, peers tend to generate conceptual questions which are not deep enough to trigger reflection. For example, this peer's question *What is the beginning process of MIC that shown by Little and Lee?* is only concerned with asking the student-author to identify the concept of *the beginning process of MIC*.

The difference between supervisor generated questions and G-Asks questions for QM5 was not significant. This result indicates that questions produced by the G-Asks system were perceived to be as useful as questions from human supervisors. This positive result might be explained by two factors. First, the system's questions are useful because they are content-related and have semantic meaning. Second, students may have intended to give high scores to questions which they thought were from supervisors, where actually they were from the system. In Table 16, we can see that 21% (34 out of 161) of the G-Asks questions have been wrongly identified as supervisor questions.

Criteria / Question Producer	QM1	QM2	QM3	QM4	QM5	Average
G-Asks	4.26	4.18	4.06	3.99	4.02	4.10
Supervisor	4.57	4.53	4.45	4.26	4.21	4.40
Peer	3.93	3.96	3.77	3.62	3.56	3.77
Generic	4.04	3.92	3.65	3.49	3.57	3.73

Table 13: Comparisons of Normalized Mean Scores after removing questions with grammatical and semantic errors

Mean Difference(MD) / Criteria	G-Asks Vs Supervisor	G-Asks Vs Peer	G-Asks Vs GQ
QM1	MD =0.308	MD =0.330	MD =0.225
	LSD=0.242	LSD=0.230	LSD=0.219
QM2	MD =0.350	MD =0.220	MD =0.263
	LSD=0.248	LSD=0.236	LSD=0.225
QM3	MD =0.385	MD =0.297	MD =0.418
	LSD=0.258	LSD=0.246	LSD=0.235
QM4	MD =0.269	MD =0.368	MD =0.501
	LSD=0.257	LSD=0.245	LSD=0.234
QM5	MD =0.182	MD =0.460	MD =0.452
	LSD=0.269	LSD=0.256	LSD=0.245
Average	MD =0.299	MD =0.335	MD =0.372
	LSD=0.226	LSD=0.215	LSD=0.206

Table 14: Fisher's least significant difference (LSD) tests with 95% confidence interval after removing questions with grammatical and semantic errors.

Rule / Criteria	Aim	Application	Method	Opinion	Result	System
QM1	4.08	4.35	4.43	4.10	4.42	4.10
QM2	4.00	4.30	4.36	3.90	4.27	4.40
QM3	3.88	4.13	4.14	3.81	4.24	4.20
QM4	3.88	4.04	4.07	3.90	4.03	4.10
QM5	3.92	4.13	4.00	3.81	4.09	4.30
Average	3.95	4.19	4.20	3.90	4.21	4.22

Table 15: Comparisons of Scores for Each Rule.

The quality of each generation rule was also evaluated. Table 15 shows the average scores. The rule for System got the highest average score (4.22). Rules for Result, Application, and Method obtained similar scores (above 4.19), while rules for Aim and Opinion obtained lower scores (below 4.0). Because we used a five-point likert scale (3 means 'Neither agree nor disagree' while 4 means 'agree'), these scores for rules (system, result, application and method) indicate that evaluators agreed that those questions are correctly written, clear, appropriate, helpful to reflect and useful. The students' evaluation results about the quality of the Aim and Opinion's question template type were not as good as other types, but it was almost close to 'agree with the good quality of questions'.

4.5 Human Perception Evaluation and Result

For each of the questions, participants were asked to guess whether it was written by the supervisor, a peer, the G-Asks (an intelligent Computer System), or whether it was a generic question. We use the balanced F-score described in formula 1 to evaluate the classification. Table 16 shows the participants' average performance on the classification, which found that they achieved an F-score of 0.49 on the Supervisor category, 0.46 on the Peer category, 0.34 on the G-Asks category, and 0.60 on the Generic Question category. Generic questions were the easiest to identify, as we expected, while questions produced by G-Asks were the most difficult. Interestingly, 21% (34 out of 161) of the G-Asks questions were wrongly identified as being questions from supervisors, and 40% (64 out of 161) were wrongly identified as questions from peers. There are two major reasons for this result:

1. The questions generated by the system are specific, especially related to the citations.
2. Similar to peers and supervisors, the G-Ask questions used abstract concepts (especially for Application, Result, System and Method citations). Such questions are also quite similar to our questions templates. For example, people often would like to ask student-writers to critically evaluate the method/theory/system and explain why this method is useful to solve a particular problem.

However, the human questions are more concise and correctly written than the system. Some system questions with long length and grammatical errors could be easily identified.

Real / Prediction	Supervisor	Peer	G-Asks	Generic
Supervisor	74(52%)	40(27%)	34(21%)	11(7%)
Peer	41(29%)	82(54%)	64(40%)	16(10%)
G-Asks	14(10%)	23(15%)	51(32%)	51(32%)
Generic	13(9%)	6(4%)	12(7%)	83(51%)
Total	142	151	161	161

Table 16: Human Classification Result on Authorship of Questions. The accuracy is shown in percentage in brackets.

4.6 Question Types Evaluation and Result

Question types are important for the application of automatic question generation, for example it can help us to define the question types which the systems should generate from the source text. There were 142 questions generated by human supervisors from 33 literature review papers written by the students. These questions included 35 surface questions, which concern presentation issues such as formating, spelling, grammatical errors, and some generic questions. Because we are only concerned about specific questions, the remaining 107 questions are analyzed as follows. Two human annotators were asked to independently annotate these 107 questions generated by human supervisors. The annotation was based on 18 frequent question types in Graesser and Pearson's taxonomy (Graesser & Person 1994). Cohen's kappa measures the agreement between two annotators:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (1)$$

Where Pr (a) is the relative observed agreement among annotators and Pr (e) is the overall probability of random agreement. As a result, the Cohen's Kappa coefficient is 0.57 (n=13; N=107; k=2), indicating moderate reliability considering the relatively large number of categories. Annotation of Verification, Causal and Procedural questions were considered to be relatively reliable with more than 74% agreements between annotators, versus only 23% for Judgmental questions, 69% for Concept questions and 50% for Comparison questions. Table 17 shows seven frequent question types in our dataset. The left column gives the definition of each question type while the right column shows an example question generated by human supervisors. Type 1 and 2 questions were classified as simple/shallow, 3 as intermediate, and 4-7 as deep questions in relation to Bloom's Taxonomy of Cognitive Domain (Bloom 1984).

Question Type	Examples
1 Verification: implied yes/no/ answers (shallow)	*Is it possible to reuse some of previous routing techniques, for example those used in cellular networks, in the NGMN?*
2 Concept: Who, When ,What, Where? (shallow)	*Can you give more details about the Generalized Beam Theory?*
3 Comparison: How is X similar to Y? (intermediate)	*In Lim and Nethercot, how well did the numerical results compare with the experimental results?*
4 Causal Antecedent: what event causally led to an event?	*Why network coding in [13] can increase the system throughput?*
5 Causal Consequence: What is the consequence of an event? (deep)	*What is the likely consequence of the nonlinear stress-strain curve on the local-overall interaction buckling behavior of stainless steel structural members?*
6 Procedural: What instrument or plan allows an agent to accomplish a goal? (deep)	*How does the formation of mechanical twins provide corrosion resistance?*
7 Judgmental: What do you think of X?(deep)	*How do you see the Generalized Beam Theory being applied in your project?*

Table **17:** Graesser and Pearson's question taxonomy with examples of questions from academic supervisors

Table 18 shows the frequency of each of question type. As we can see, Concept, Causal, and Procedural questions were more frequent than Judgmental and Verification questions.

Question Category	Frequency
Concept	28
Causal Antecedent/Consequence	28
Procedural	23
Comparison	2
Judgmental	13
Verification	11
Other Types: Feature Specification 1, Goal Orientation 1	

Table 18: Frequency of the Question Type

Table 19 shows the average scores from students' evaluations of the questions. Verification, Concept and Procedural questions obtained slightly higher scores than Causal and Judgmental questions. However, there were no statistically significant differences between scores from these questions types ($F(4,100) = 2.162$, $P > 0.05$).

Question Type	Average Score	Standard Deviation
Concept	4.12	0.993
Causal	3.88	1.021
Procedural	4.34	0.742
Judgmental	3.92	1.145
Verification	4.75	0.430

Table 19: The Average Score of the Question Type

From table 18, we see that human supervisors like to generate simple questions in addition to deep questions. It indicates that conceptual questions are as important as procedural or causal questions, which should be considered when designing the question templates.

In order to investigate how the question was generated from the source text, we classified the source of questions into the following four abstract levels:

Lexicon Level: the question is generated from a key term or concept. E.g., *can you give more details about the Generalized Beam Theory, explain their advantages and disadvantages?* The *Generalized Beam Theory* is the key concept, which is asked to be analyzed critically.

Sentence Level: the question is generated from a single sentence without domain knowledge. E.g., *In Lim and Nethercot, how well did the numerical results compare with the experimental results?* The source sentence was: *Lim and Nethercot compared the numerical tests results with finite element simulation results.* In this case, the source sentence reports comparative tests and the trigger question is about how good the *results* are.

Discourse Level: the question is generated from more than one sentence with inference process and domain knowledge. E.g., *On what basis did Besson et al 1997 compare the solid state and liquid state production of Pyrazine?* The source sentences were *Besson et al. (1997) demonstrated that Pyrazine can be produced by solid state fermentation. They stipulate the concentrations which are much higher than that of liquid state fermentation.* In this case, we should combine two sentences together with some domain knowledge and infer that *Pyrazine can be produced from both solid state and liquid state, and the solid state is better.*

Background Knowledge: the question is generated based on the domain knowledge which is not expressed in the writing. E.g., *what is the power range for each type of the wind turbine?* In this case, we should know that *the power range is one property of wind turbine.*

Source Level	Num. of Questions and Question Type Distribution(only show question type more than 2 times)
Lexicon Level	12 questions include 7 Concept, 5 Judgmental
Sentence Level	49 questions include 20 Causal, 10 Procedural, 10 Concept, 5 Judgmental, 4 Verification.
Discourse Level	14 questions include 4 Procedural, 3 Verification
Background Knowledge	32 questions include 9 Concept, 7 Procedural, 4 Causal, 3 Verification.

Table 20 Frequency of Source Level for Question Generation

Table 20 shows that the number of questions generated at the Lexicon and Sentence levels take 57% (61 out of 107) while the Discourse level takes 13.1% and Background Knowledge level takes 29.9%. The dominant question types were Concept and Judgmental in the Lexicon level; Causal, Procedure, and Concept in the Sentence level; Procedural and Verification in the Discourse level; and Concept and Procedure in the Background level. This indicates promising opportunities for generating questions from Lexicon and Sentence Level by using current NLP technologies.

5 Conclusion and Future Work

This article presented an Intelligent Automatic Question Generation System, as a feedback tool used in the iWrite Web Application, which generates contextualized trigger questions from citations to support literature review writing. The trigger questions are aimed at improving learning during academic writing as opposed to only focusing on writing as a communication skill. In order to evaluate the system's performance and to analyze human expert generated questions, we compared automatically generated questions with human-generated and generic questions using a Bystander Turing test.

In contrast to the citation classification with Tregex rule-based approach used in the previous study, for the present study we used a Machine Learning approach based on some useful features, such as cue phrases. The result shows the new approach outperformed the rule-based approach across 5 citation categories. However, from the result of the basic system performance (see section Citation Extraction Rate and Citation Classification Performance), we can see the bottlenecks of the pipeline system were in the NER (Name Entity Recognizer) tagger, the Sentence Parser and the statistical Citation Classifier. The LBJ NER tagger was primarily trained on News Text Corpora, and this might have affected its performance on academic articles. As for the performance of the sentence parser and citation classifier, there is room for improvement. These issues caused the system to generate semantically or syntactically erroneous questions and hence decreased the overall quality of the system. However, these issues can be overcome by a ranking function which would reduce the probability of questions with semantic or syntactic errors being selected.

In order to get more insights on how human experts generate specific trigger question, we analyzed the human supervisors' generated questions for literature review writing support. Six frequent question types based on Graesser and Person's question taxonomy were identified. This can help us design question templates. The results show that 57% of the questions generated at

the Lexicon and Sentence levels were without complex inference processing, which indicates that many potential questions can be exploited by using current NLP techniques.

Future work will focus on using 'the overgenerate-and-rank' approach, which has been applied previously in the Natural Language Generation community (Langkilde & Knight 1998; Walker et al. 2001). As shown in Table 20, Conceptual questions are also important; these questions were rated well by students (achieving an average score of 4.12) and placed second among all question types. We are working to generate Conceptual question types based on key concepts in a single document.

ACKNOWLEDGMENT

The authors would like to thank our colleagues Jorge Villalon and Stephen O'Rourke for the development of the TML java library. This project was partially supported by a University of Sydney TIES grant, an Australian Research Council Discovery Project grant (DP0986873) and Google Research Award for measuring the impact of feedback on the writing process. This research was supported in part by the Institute for Education Sciences (R305A100875) and National Science Foundation (0938239) through grants awarded to Dr. Vasile Rus.

References

Liane Acker, James Lester, Art Souther and Bruce Porter, Eds. (1991). Generating Coherent Explanations to Answer Students' Questions. *Intelligent Tutoring Systems: Evolutions in Design*, Psychology Press, New York.

Jeff Anderson (2005). *Mechanically Inclined:Building Grammar, Usage, and Style into Writer's Workshop*. Stenhouse Publishers Portland, ME.

Timothy J. Beals (1998). Between Teachers and Computers: Does Text-Checking Software Really Improve Student Writing? *The English Journal,* 87(1): 67-72.

Benjamin Bloom (1984). *Taxonomy of educational objectives Book I: cognitive domain*. Addison Wesley, Boston.

Kristy E. Boyer, William Lahti, Robert Phillips, Michael Wallis, Mladen Vouk and James Lester (2009). An Empirically-Derived Question Taxonomy for Task-Oriented Tutorial Dialogue. *In Proceedings of The 2nd Workshop on Question Generation*, pages: 9-16, Brighton.

M. Anne Britt, Peter Wiemer-Hastings, Aaron A. Larson and Charles A. Perfetti (2004). Using Intelligent Feedback to Improve Sourcing and Integration in Students' Essays. *International Journal of Artificial Intelligence in Education,* 14(3): 359-374.

Rafael A. Calvo and Robert A. Ellis (2010). Students' Conceptions of Tutor and Automated Feedback in Professional Writing. *Journal of Engineering Education,* 99(4): 427-438.

Rafael A. Calvo, Stephen T. O'Rourke, Janet Jones, Kalina Yacef and Peter Reimann (2010). Collaborative Writing Support Tools on the Cloud. *IEEE Transactions on Learning Technology,* 4(1): 88-97.

David Coniam (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *The Computer Assisted Language Instruction Consortium Journal,* 14(2): 15-33.

Averil Coxhead (2000). A New Academic Word List. *Teachers of English to Speakers of Other Languages Journal Quarterly,* 34(2): 213-238.

Andrea Esuli and Fabrizio Sebastiani (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *In Proceedings of the 5th Conference on Language Resources and Evaluation*, pages: 417-422, Genoa, Italy.

Christiane Fellbaum (1998). *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Arthur C. Graesser and Natalie K. Person (1994). Question asking during tutoring. *American Educational Research Journal,* 31(1): 104-137.

Michael Heilman and Noah A. Smith (2009). Good question! Statistical ranking for question generation. *In Proceedings of The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages: 609-617, Stroudsburg, PA.

K. Hyland (1994). Academic attribution: citation and the construction of disciplinary knowledge. *Applied Linguistics,* 20(3): 341-367.

Dan Klein and Christopher D. Manning (2003). Fast Exact Inference with a Factored Model for Natural Language Parsing. *In Proceedings of the international conference In Advances in Neural Information Processing Systems*, pages: 3-10, Cambridge, MA.

Hidenobu Kunichika, Tomoki Katayama, Tsukasa Hirashima and Akira Takeuchi: (2001). Automated Question Generation Methods for Intelligent English Learning Systems and its Evaluation. *In Proceedings of The International Conference on Computers in Education*, pages: 1117-1124, Seoul, Korea.

Thomas K. Landauer, Danielle S. McNamara, Simon Dennis and Walter Kintsch (2007). *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum.

Irene Langkilde and Kevin Knight (1998). Generation that exploits corpus-based statistical knowledge. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages: 704-710, Montreal, Quebec.

Wendy Lehnert, Claire Cardie and Ellen Riloff (1990). Analyzing Research Papers Using Citation Sentences. *In Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages: 511-518, Cambridge, MA.

Wendy G. Lehnert (1978). *The Process of Question Answering - A Computer Simulation of Cognition. Hillsdale*. L. Erlbaum Associates, Hillsdale, NJ.

Roger Levy and Galen Andrew (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages: 2231-2234, Genoa, Italy.

Ming Liu, Rafael A. Calvo and Vasile Rus (2010). Automatic Question Generation for Literature Review Writing Support. *In Proceedings of The Tenth Interational Conference on Intelligent Tutoring Systems*, pages: 45-54, Pittsburgh, USA.

Ruslan Mitkov and Le An Ha (2003). Computer-aided generation of multiple-choice tests. *In Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*, pages: 17-22, Morristown, NJ.

Rodney D. Nielsen, Jason Buckingham, Gary Knoll, Ben Marsh and Leysia Palen (2008). A Taxonomy of Questions for Question Generation. *In Proceedings of the 1st Workshop on Question Generation.*, pages: 15-22, Arlington, Virginia.

Natalie K. Person and Arthur C. Graesser (2002). Human or Computer? AutoTutor in a Bystander Turing Test. *In Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, pages: 821--830, London, UK.

Brett Powley and Robert Dale (2007). Evidence-based information extraction for high-accuracy citation extraction and author name recognition. *In Proceedings of the 8th RIAO International Conference on Large-Scale Semantic Access to Content*, pages: 15-22, Pittsburgh, PA.

Lev Ratinov and Dan Roth (2009). Design Challenges and Misconceptions in Named Entity Recognition. *In Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Stroudsburg, PA.

Thomas H. Reynolds and Curtis Jay Bonk (1996). Computerized prompting partners and keystroke recording devices: Two macro driven writing tools. *Educational Technology Research and Development,* 44(3): 83-97.

Mark D. Shermis and Jill C. Burstein (2002). *Automated essay scoring: A cross-disciplinary perspective*. Routledge, Mahwah, NJ.

Simone Teufel, Advaith Siddharthan and Dan Tidhar (2006). Automatic classification of citation function. *In Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages: 103-110, Sydney, Australia.

Elaine C. Thiesmeyer and John E. Theismeyer (1990). *Editor:A System for Checking Usage, Mechanics, Vocabulary, and Structure*. New York: Modern Language Association, Raleigh, NC.

Jorge Villalon, Paul Kearney, Rafael A. Calvo and Peter Reimann (2008). Glosser: Enhanced Feedback for Student Writing Tasks. *In Proceeding of Eighth IEEE International Conference on Advanced Learning Technologies*, pages: 454-458, Santander,Spain.

Marilyn A. Walker, Owen Rambow and Monica Rogati (2001). SPoT: a trainable sentence planner. *In Proceeding of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages: 1-8, Pittsburgh, Pennsylvania.

Peter Wiemer-Hastings and Arthur C. Graesser (2000). Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments,* 8(2): 149--169.

Robert Williams and Heinz Dreher (2004). Automatically grading essays with Markit©. *In Proceedings of Informing Science Conference*, pages: 0693-0700, Rockhampton, Queensland.

John H. Wolfe (1976). Automatic question generation from text - an aid to independent study. *In Proceedings of the ACM SIGCSE-SIGCUE technical symposium on Computer Science and Education*, pages: 104-112, New York.