Automatic Generation and Ranking of Questions for Critical Review

Ming Liu[1, 2], Rafael A. Calvo[2] and Vasile Rus[3]

[1]Faculty of Computer and Information Science, Southwest University, Chongqing, China,[2]School of Electrical and Information Engineering, University of Sydney, Australia// [3]Department of Computer Science, University of Memphis, USA // Ming.Liu@sydney.edu.au // Rafael.Calvo@sydney.edu.au // vrus@memphis.edu

Abstract. Critical review skill is one important aspect of academic writing. Generic trigger questions have been widely used to support this activity. When students have a concrete topic in mind, trigger questions are less effective if they are too general. This article presents a learning-to-rank based system which automatically generates specific trigger questions from citations for critical review support. The performance of the proposed question ranking models was evaluated and the quality of generated questions is reported. Experimental results showed an accuracy of 75.8% on the top 25% ranked questions. These top ranked questions are as useful for self-reflection as questions generated by human tutors and supervisors. A qualitative analysis was also conducted using an information seeking question taxonomy in order to further analyze the questions generated by humans. The analysis revealed that explanation and association questions are the most frequent question types and that the explanation questions are considered the most valuables by student writers.

Keywords

Automatic Question Generation, Writing Support, Natural Language Processing, Learning to Rank

Introduction

Academic Writing is one of the most challenging academic activities for most university students. Typical writing skills are not sufficient for academic writing which involves intense argumentation for supporting one's work addressing a major research question. Students must identify relevant information, critically read and analyze research literature, and finally synthesize previous and related work in order to build the case for their answer the research question (Graswell, 2008). Thus, critically evaluating the literature and understanding its key concepts is an important aspect of academic writing. However, many students experience difficulties with critically analyzing the literature. Afolabi (1992), for example, identified some of the most common problems that students have when writing a literature review, including *not being sufficiently critical, lacking synthesis, and not discriminating between relevant and irrelevant materials.*

Trigger questions can be an effective tool to help with writing a critical review. A limited number of generic trigger questions are commonly used to guide students during academic writing (Taylor & Procter, 2008). For example, a research supervisor usually asks trigger questions such as: (1) *Have you critically analyzed the literature you use? Instead of just listing and summarizing items, have you assessed them, discussing strengths and weaknesses? (2) Have you discussed how your project will contribute to that discipline or field?* However, such questions are too general and not likely to provide meaningful feedback for many students writing about a specific topic. This paper focuses on

investigating the performance of an automated system to generate specific trigger questions that offer contextualized feedback relevant to the target topic.

Automated Question Generation (AQG) is a challenging task which involves natural language understanding and generation (Rus & Graesser, 2009). It typically involves three major steps: target content selection (what to ask about), question type selection (the type of question, e.g. Who, Why, Yes/No), and question construction (how to ask question). Current AQG approaches (Mostow & Chen, 2009) focus on generating mostly factual questions for supporting reading comprehension or vocabulary assessment. In contrast, the present work focuses on generating trigger questions for supporting critical thinking in academic writing.

In order to generate such trigger questions, two main AQG approaches were proposed: 1) question generation based on key phrases(Liu, Calvo, Aditomo, & Pizzato, 2012), and 2) question generation based on citations(Liu, Calvo, & Rus, 2012). The focus of this paper is on improving the system performance of the second approach. In this approach, sentences containing citations are chosen as target sources for question generation because such sentences are more relevant to the task of writing research papers. They normally express an author's *Opinion* or the description of an *Application*. For example, the following citation sentence expresses an *Opinion* of the cited author Cannon: *Cannon (1927) challenged this view mentioning that physiological changes were not sufficient to discriminate emotions.* Some examples of trigger questions that could be generated from the above citation sentence are: *Why did Cannon challenge the 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?*

Trigger questions are important during writing as they draw student writers' attention and reflection to other researchers' views and evidences. In our previous approach (Liu, Calvo, & Rus, 2010; Liu, Calvo, & Rus, 2012), rule-based methods were used for all major steps: citation sentence extraction, sentence simplification, classification of simplified sentences, and syntactic transformation. However, in this processing pipeline errors could occur at any stage, such as citation sentence extraction, voice transformation, and classification. To overcome these challenges, an overgeneration-and-ranking approach was adopted. Compared with the previous study (Liu, Calvo, & Rus, 2012), the present study describes the improved question generation system based on the overgeneration-and-ranking approach and its evaluation in an educational context. Specifically, our contributions are as in the followings:

- A novel AQG system is proposed for academic writing support based on an overgeneration-and-ranking approach. Two different ranking algorithms were evaluated: RankSVM (Pairwise Approach) and Logistic Regression (Pointwise Approach).
- An evaluation method for the AQG system is described: the performance of a question ranking model and the quality of system-generated trigger questions.
- A deep analysis of human supervisor and tutor trigger questions based on a question classification proposed by Trabasso et.al. (1996).

The remainder of the article is organized as follows. Background section provides a brief literature review related to question generation and ranking. The System Design and Architecture section

describes the proposed AQG framework and the ranking model. The Evaluation and Result section describes the experiments and results.

Background

This section first reviews the literature of how the questions are used for critical review. Then, it presents current AQG and the learning-to-ranking approaches.

Trigger Questions and Critical Review

In higher education, students need to be able to critically review literature related to their research topics. According to Steward (2004), a critical review should be "Comprehensive, Fully referenced, Relevant, A synthesis of key themes and ideas, Balanced between different ideas and opinion, Critical in its appraisal of the literature, and Analytically developing new ideas from the evidence" (p. 496). However, critical review is not an easy task for most college students.

Questions are central aspects in the theories of learning, cognition, and education (Graesser & Person, 1994). It can engage students in learning activities, such as critical review, by helping them to recognize their knowledge deficiencies and reflect on their writing activity. But, many studies have shown that students have problems recognizing their own knowledge deficits (Hacker, Dunlosky, & Graesser, 1998) and ask very few questions (Graesser & Person, 1994). Thus, support in the form of generic trigger questions (asked by a computer or human tutor or supervisor) are often used to help self-reflection. Reynolds and Bonk (1996) showed that a group of students given generic trigger questions performed better than those students who received no trigger questions to support revision in a writing activity. As mentioned in Introduction, the aim of the AQG system is to support students' writing activities by generating trigger questions that helps them self-reflect to important aspects of the critical review writing task.

Question Generation

Recently proposed question generation approaches (Heilman & Smith, 2010; Yao, 2010) focused on generating factual questions for reading compression and vocabulary assessment. These approaches are more related to our work since the question generation objectives are similar: specific shallow/deep questions generation from natural text.

Yao (2010) classified question generation systems into three categories: Template-based, Syntax-based, and Semantic-based. The Syntax-based approach is the most popular approach to automatically generating factual questions. The key idea of this approach is to transform the declarative target sentence into an interrogative one by manipulating the derived syntactic tree typically parsed by using a Context Free Grammar parser. Tregex (Levy & Andrew, 2006) is a powerful syntactic tree search language for identifying syntactic elements (e.g. main verbs of sentences) and which has been used by researchers (Heilman & Smith, 2010) to define wh-movement rules. Compared to the syntax-based approach, the semantic approaches (Yao, 2010) rely on a semantic parse—a deeper level of linguistic analysis than the syntactic parse—aimed at transforming the declarative semantic parse

into questions. Both syntax-based and semantic-based approaches can generate questions having high specificity because these questions contain more information about the answer. But, the generated questions are factual questions and generally not very deep.

Unlike syntax-based and semantic-based approaches, the template-based approach (Mostow & Chen, 2009) does not require complex question transformation rules that convert the parser output into questions. Instead, it focuses on generating deep questions by extracting knowledge from the text, make inferences when possible, and filling empty slots in question templates. This approach can generate deep questions. But, these questions are less specific.

In order to generate specific deeper questions, which are generated from original sentences with predefined deep question templates, a combined syntax-based and template-based question generation approach, as the one described in this article, is recommended.

Learning-to-Rank

Learning-to-rank received increasing attention in both Information Retrieval and Machine Learning research during the past decade. Most of approaches to learning-to-rank are designed as supervised machine learning approaches. All instances are given a (binary or ordinal) score or label indicating their relevance as decided by an independent judgment process. In the training phase, a ranking function is learned based on a set of feature vectors together with their true labels. In the testing phase, the ranking function is used to rank a new set of instances and generate a ranked order of these instances.

Based on how they treat sets of ratings and loss functions, Cao et al. (2007) classified learning-to-rank approaches into 3 categories: 1) Pointwise Approach: learning to classify the question or instance according to their label individually (e.g. positive or negative category), 2) Pairwise Approach): classifying pairs of rated questions into two categories (correctly ranked or incorrectly ranked), and 3) Listwise Approach: optimizing the loss function for ordering the questions related to a single academic paper instance. In the information retrieval literature, the Pointwise approach is viewed as the weakest of the three learning-to-rank approaches because it ignores the cluster of answer instances per query. Machine learning techniques that can be used in conjunction with the Pointwise approach are classifiers (e.g. Naïve Bayes) and regression(e.g. Logistic Regression). The Pairwise approaches are considered more effective than Pointwise approaches because pairs of answer instances are considered. The algorithms used in the Pairwise approaches include RankSVM (Joachims, 2006). Listwise approaches are more recent developments. Liu (2009) shows that the Listwise techniques reached scores similar to or better than Pairwise techniques.

However, Listwise approaches are less suitable than Pointwise and Pairwise approaches in computational linguistics because the labels used in the training set normally have very few categories, such as relevancy or irrelevancy. These few categories would cause Listwise approaches to have problems of accurately obtaining an ordered ranking sequence in the training set. Collins and Koo (2005) trained a logistic regression model, a Pointwise approach for ranking syntactic parses. Duh (2008) proposed an automatic machine translation evaluation based on learning-to-rank. He compared

RankSVM to RankBoost and found that RankSVM performed best when ranking-specific features are considered.

System Design and Architecture

This section presents an overview of the system's pipeline architecture (see Figure 1), briefly describing each step and emphasizing the question ranker.

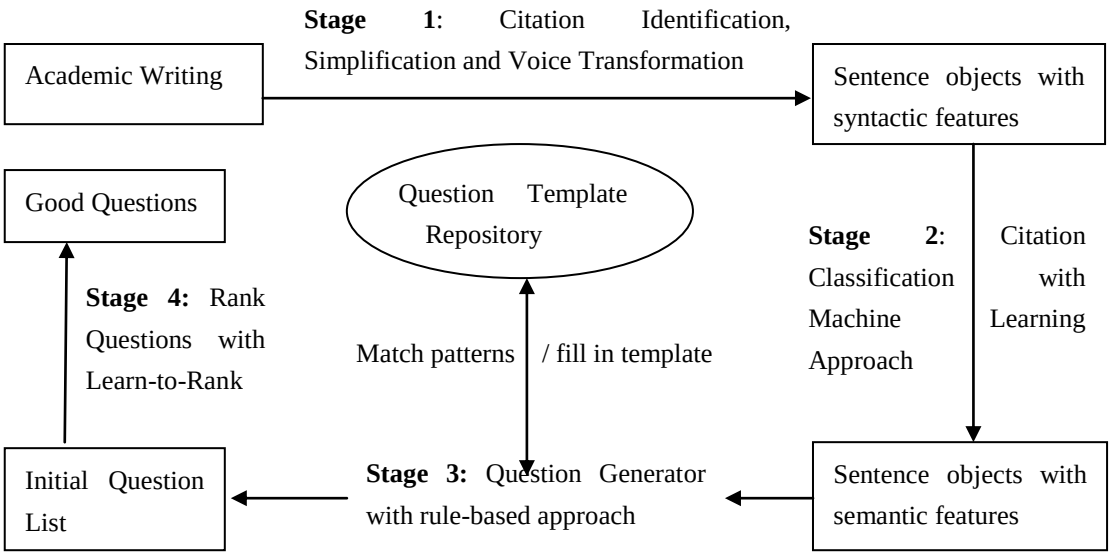

Figure 1: System Architecture: Multiple Stages Question Generation Process

Stage 1 Pre-processing

In stage 1, citation sentences from academic papers are extracted, simplified, and then transformed in active voice. The citation sentence extraction was based on the citation styles defined by Powley and Dale (2007). After the citations are extracted, sentence simplification is performed. This involves splitting compound and complex sentences and also removing phrase types such as appositives, non-restrictive relative clauses, and participial modifiers. Tregex (Levy & Andrew, 2006) were used for this sentence simplification and passive-to-active voice transformation.

Passive-to-active voice transformation is frequently applied since Hyland (1994) identified that the passive construction is one of the common grammatical styles used in citation sentences. The transformation algorithm is to set the new sentence in active form, which is needed for next steps in our processing pipeline, according to the passive construction type. Table 1 shows three common passive construction types used in academic writing (Hyland, 1994). In the examples shown in the table, the parenthesis indicates a syntactic element given by the term in the bracket. The Tregex Expression rules are used to detect the passive construction type and extract these syntactic elements. Finally, we reconstruct the passive sentences into their active form by using these elements.

Table 1:Passive to Active Transformation with Tregex Rules

Passive Construction	Tregex Rules	Active Form
Type 1: It+is+[main verb]+ [that..]. e.g., It is {argued [Main Verb]} {that… [Clause]}	SBAR=Clause $- (VBN < "+Main Verb+")	[Subject]+[Main Verb]+ [That Clause] e.g., Peter argued that…
Type 2: The main verb followed by a particle. e.g., {The noncatalytic decomposition of cellulose in subcritical water [Object]} was {carried [Main Verb]} {out [Particle]} by Sakaki et al (2005).	PRT=Particle $- (VBN < " + Main Verb + ")	[Subject]+[Main verb]+[Particle] + [Object] e.g., Sakai carried out the noncatalytic decomposition of cellulose in subcritical water.
Type 3: The main verb is followed by an infinitive clause or a prepositional phrase. {MDS [Object]} can be {used [Main Verb]} {to describe the evolution of atomic structure during irradiation [Clause]}.	PP|S=Clause $- (VBN < " + Main Verb + ")	[Subject]+[Main Verb]+ [Object]+ [Clause]. e.g., Peter used MDS to describe the evolution of atomic structure during irradiation.

Stage 2. Citation Classification

The goal of this stage is to identify the citation category for each citation sentence extracted in Stage 1. The citation categories shown in the first column in Table 2 were used: 'Opinion', 'Result', 'Aim of Study', 'System', 'Method', and 'Application'. These were chosen based on a taxonomy of conceptual citation categories proposed by Lehnert et al. (1990). To automatically detect the citation category, a Naive Bayes classifier is trained, and 17 generic features were proposed, including Cue Phrases, Sentiment Features, Negation Feature, and Syntactic Feature (Liu, Calvo, & Rus, 2012).

Table 2: Citation Types and Examples

Aim	present the aim of an author' study, e.g. *Bunescu et al. focused on extracting named entities from natural language documents*.
Opinion	express the opinion of an author, e.g. *Reiter and Dale (1997) state that template-based systems are more difficult to maintain and update.*
Result	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.*
Method	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).*
System	describe a system, e.g. *AutoSlog (Riloff, 1996) is a dictionary construction system that creates extraction patterns automatically using heuristic rules.*
Application	apply a method/system to a field, e.g. *Kappa statistics K will be used to measure the reliability (Siegel and Castellan, 1988).*

Stage 3. Question Generation

This is the third stage in our approach to automatically generating trigger questions. Once the citation category and syntactic features were extracted from a citation sentence, a set of predefined patterns are used to generate the corresponding questions. Table 3 shows six rules and their questions templates defined in our Repository of Templates. For example, the following citation sentence is assumed to be extracted in Stage 1:

While it is shown that AES has inter-assessor correlations comparable to that of human assessors (Dikli , 2006) , many scholars are still highly critical of the validity and robustness of the approach (Britt , et al. , 2004) .

The sentence simplifier will split the complex sentence into two simpler sentences: *It is shown that AES has inter-assessor correlations comparable to that of human assessors (Dikli , 2006).* and *Britt (2007) criticized the validity and robustness of the approach*. Because the first sentence is passive, it is transformed into an active sentence: *Dikli showed that AES has inter-assessor correlations comparable to that of human assessors*.

In Stage 2, the classifier categorizes the first sentence as a 'Result' citation and the second sentence as belonging to the 'Opinion' category. Stage 3 applies rule 1 to generate the following trigger questions, which are intended to prompt the student's reflection by asking for evidence or other authors' opinions: *Why did Britt criticize the validity...? Does any other scholar agree or disagree with Britt?* Similarly, rule 3 is applied to generate the trigger questions below for evaluating the evidence. *Did Dikli objectively show that AES has… Is the analysis of the data accurate and relevant to the research question?* Each of these questions is then scored according to features of the source sentence. The following section will describe the scoring process in more detail.

Table 3: Six Rules and Their Question Templates

Rule	Category	Question Template
1	Opinion	Why +subject_auxiliary_inversion ()? What evidence is provided by +subject+ to prove the opinion? Does any other scholars agree or disagree with +subject+?
2	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?
3	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?
4	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?
5	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?
6	Application	Why+Subject_Verb_Inversion()? Could the problem have been approached more effectively from another perspective? Does it relate to your research question?

Question Ranker

The previous stages generate questions that vary in their quality, from syntactic to semantics to importance. This is unavoidable and happens for different reasons, such as errors in sentence parsing, voice transformation, and citation classification. To address this problem, ranking the large pool of questions according to their quality is needed. Stage 4 implements a learning-to-rank algorithm to meet this challenge.

Ranking Model

In the ranking model, two common learning-to-rank approaches were used: Pointwise approach (Logistic Regression; (Collins & Koo, 2005) and Pairwise approach (RankSVM; (Joachims, 2006)). The logistic regression model is learned by fitting training data to a logit function by using the predictor binary variable which indicates whether a question is acceptable or not. The RankSVM model is learned using a Pairwise approach which can naturally specify questions that are of an equivalent rank. The Support Vector Machines (SVM) algorithm has been used previously for preference ranking in the context of Information Retrieval. The same framework was adopted for ranking questions in this case. In this model, given a collection of questions ranked according to preferences between two questions represented by feature vector q_i and q_j, respectively, and a linear learning function f,

$$q_i \succ q_i \Leftrightarrow f(q_i) > f(q_j) \quad (1)$$

Where $\succ$ indicates that q_i is preferred over q_j. The function f is defined as f(q)=w·q, where

$$f(q_i) > f(q_j) \Leftrightarrow w \cdot q_i > w \cdot q_i \quad (2)$$

In the context of SVM, these weight vectors or support vectors (w) are identified by minimizing the function using slack variables ξ_{ij} :

$$\min_{w,\xi_{ij}} \frac{1}{2}\|w\|^2 + C\sum_{ij}\xi_{ij} \quad (3)$$

Subject to the constraints:

$$\forall(q_i, q_j): w \cdot q_i > w \cdot q_i + 1 - \xi_{ij}$$

$$\forall(i,j): \xi_{ij} \geq 0$$

Finding the support vectors and the generalization of the Ranking SVM is done differently (Joachims, 2006). If the data are linearly separable, the ξ_{ij} are all equal to 0. In this case, the ranking function is considered as projecting the data points onto the separating hyperplane and the support vectors as the two points q_i and q_j whose projections are nearest each other on the hyperplane. The generation is

accomplished by calculating w to maximize the distance between these closest points. The distance or margin between these two points is formulated as $\frac{w(q_i - q_j)}{\|w\|}$. Like the classification SVM algorithm , the margin is maximized when $\|w\|$ is minimized.

Feature Definition

The features used in the ranking models were developed by an in-depth analysis of questions generated manually by human experts for the training set which contained 504 citations. This is the same set used for training the citation classifier. Because our current ranking models focus on acceptability of a question in terms of grammatical and semantic correctness, these features should indicate the likelihood of generating an acceptable or unacceptable question in terms of the complexity of source sentences (Num.of NamedEntities, NameAppearInBoth, Num.ofClauses and Length), the transformation performed during the processing (IsPronounResovled, PredicationConfidence and IsPassiveVoice), and a citation sentence in reporting form (ReportingVerb and NameAppearInSubject). For example, if the source sentence is very long and complex (many clauses), the parser is more likely to generate errors. Moreover, if some syntactic transformation was performed, like passive-to-active, the transformation stage might generate errors. However, if the source sentence is in reporting form using reporting verbs, it most likely generates a question without syntactic errors because this type of sentence is normally simple and transformation rules for this type of source sentence are well defined.

Some NLP tools are used or developed to extract these features. A state of art Named Entity Tagger, LBJ (Ratinov & Roth, 2009), was used to extract name entities in a sentence. A simple Pronoun Resolver, finding the nearest Name Entity appearing before the pronoun, was implemented to identify citations with pronominals. Tregex rules are developed to detect number of clauses and passive voice detection. The reporting verb list was obtained from academic writing tutorial websites (Centre for Academic Success, 2011).

The source sentence refers to the citation sentence in the descriptions of the 11 features defined below:

Num. of Named Entities: this numeric feature describes the number of author names in the source sentence.

IsPronounResolved: this Boolean feature detects whether the pronoun resolution has been resolved.

IsPassiveVoice: this Boolean feature detects whether the source sentence is passive voice.

ContainsAnswer: this is a Boolean feature that detects the presence of answer cue phrases, such as due to, in order to, considering, etc.

Negation: this Boolean feature detects whether the source sentence contains negation cue phrases (not, no, never), restrictive adverbs (few, rarely, seldom), negative verbs (fail, deny), or negative adjectives (insufficient).

PredicationConfidence: this numeric feature shows the predication confidence from the citation classifier.

ReportingVerb: this is a Boolean feature that detects whether the source sentence contains reporting verbs, such as show, argue, discuss, and explain.

NameAppearInBoth: this is a Boolean features indicating the presence of name entities in both subject and predicate.

NameAppearInSubject: this is a boolean feature to detect the presence of name entities in subject.

Num. of Clauses: this numeric feature shows the number of clauses in the source sentence. Each clause consists of a noun phrase (NP) followed by a verb phrase (VP).

Length: this feature judges whether the source sentence is too short or too large.

An Application Scenario in a Writing Environment

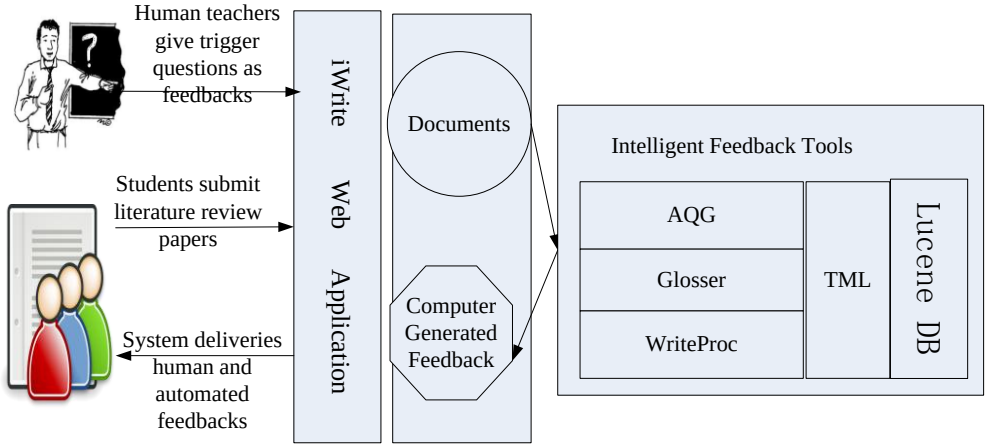

Figure 2: AQG in a writing support environment

This section briefly describes the potential application of the AQG tool which will be integrated into iWrite (Calvo, O'Rourke, Jones, Yacef, & Reimann, 2011). iWrite is a web-based writing support environment 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 uses automatic feedback tools that help students conduct revision, such as Glosser (Villalon, Kearney, Calvo, & Reimann, 2008) . AQG will be integrated into iWrite as one of the feedback tools.

In Figure 2, the input to the iWrite is an academic writing paper, written in natural text by a student using Google Docs. When the assignment deadline is due, the iWrite assignment manager downloads the student's document and passes it to Intelligent Feedback Tools for processing. The output is the textual feedback including trigger questions, generated by human or intelligent feedback tools, such as AQG, which is delivered to the student author.

Evaluation and Result

Data Collection

In order to train our question ranking model, 504 questions were collected and manually annotated as a training set. Those questions were generated from citations included in 45 academic papers from ACL conference and International Joint Conferences on Artificial Intelligence (IJCAI). Our testing set includes 489 questions generated from 33 literature review papers separately written by 33 engineering research students enrolled in a Research Methods course. Two human experts, who also annotated the citations for the classification task, were asked to label each question as acceptable or unacceptable according to two major criteria, (1) grammaticality and (2) semantic correctness. Grammaticality refers to the presence or absence of grammar errors. Semantic correctness refers to the overall meaning of the generated question within the context of the citation sentence that triggered the question, e.g. a generated question may not make any sense due to wrong author name entity extraction or wrong question template used because of errors in the citation classification step. For example, in the following generated question: *Why does MANET conduct this study to consider the effect of velocity of the nodes?*, MANET was wrongly identified as an author. The Cohen' Kappa inter-agreement between human annotators was 0.65 (substantial agreement).

Besides collecting system generated questions from each literature review paper, questions were collected including those from the student's supervisors, peers, and five generic questions. Table 4 shows the five generic questions used by the research method course intended to help students write literature review.

Table 4: Five Generic Questions used in Literature Review Support

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

Like the Bystander Turing Test, our judges, the student writers, rated the quality of each generated question using a Likert scale (1 was "strongly disagree" and 5 "strongly agree") :

QM1: This question is correctly written.

QM2: This question is clear.

QM3: This question is appropriate to the context.

QM4: This question makes me reflect about what I have written.

QM5: This is a useful question.

Ranking Model Performance

The following ranking models were derived in order to compare algorithms and feature sets.
RankSVM: this model is implemented by RankSVM algorithm. The RankSVM (ALL) includes all the features described in previous section. According to the characteristic of each feature, the feature set is grouped into two clusters: syntactic group and semantic group. RankSVM(Basic) only contains the basic or syntactic features including isPassive Voice, Negation, Number of Clauses, isProunResolved, and Length. RankSVM(Semantic) includes only the semantic features: Number of Named Entities, Predication Confidence, Reporting Verb, NameAppearInBoth, NameAppearInSubject, ContainAnswer.

Logistic Regression: the model is trained using all the features

Baseline: the expected performance if questions were ranked randomly

Gold Standard: the expected performance if all the acceptable questions were ranked higher than unacceptable

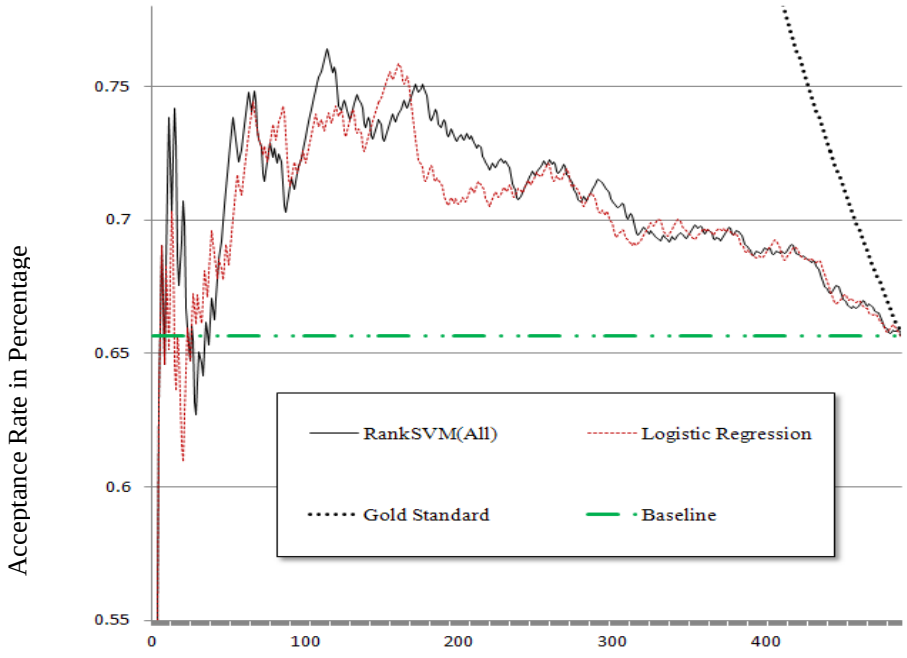

Figure 3: A graph of the percentage of acceptable questions in the top ranked 489 questions using various models

The baseline was 0.657 (65.7% of all test set questions were labeled as acceptable by the human annotator). Figure 3 shows that most of ranking models were unstable in the top 100 questions. The RankSVM (All) and Logistic Regression had very similarly sharp curve, which are higher than baseline

overall. Table 5 shows the ranking results of these models for the top 25% and 50% of the ranked questions. In the top 120 questions (25%), RankSVM (All) got better accuracy (0.758) than Logistic Regression (0.742). RankSVM (Basic) got the lowest accuracy (0.708). Basic features helped as the performance improves from 0.733 for RankSVM (Semantic) to 0.758 for RankSVM (All) with all the features. A one-way ANOVA, at a 95% confidence level, was conducted to examine whether there are statistical differences among these models. The ANOVA indicated a significant difference, $F(3,476)= 3.08$, $p<0.05$. Fishers' least significant difference (LSD) tests at the 95% confidence level were performed to determine whether significant differences occurred between the mean scores for each pair of treatments. Results indicated no statistical difference between RankSVM (All) and Logistic Regression while RankSVM (ALL) significantly outperformed RankSVM (Basic). In the top 224 ranked questions (50%), the RankSVM (All), RankSVM (Semantic), Logistic Regression got equivalent scores, all of which were better than the RankSVM (Basic)

Table 5: The percentage of the top 25% and 50% of ranked questions labeled acceptable with various models

Model	Top 25%	Top 50%
RankSVM(All Feature)	0.758	0.713
RankSVM(Semantic Feature	0.733	0.713
RankSVM(Basic Feature)	0.708	0.692
Logistic Regression	0.742	0.713
Baseline	0.657	0.656

Evaluation of the Top Ranked Questions

The ability of our AQG system to generate quality and effective questions is explored by comparing questions generated by the system to those produced by human supervisor, peer, and five generic questions commonly used in writing literature review. For the evaluation, supervisors generated 107 questions, peers 133, our AQG system top ranked 126, and we also used 161 generic questions (five generic questions have been used repeatedly for each document).

Table 6: Comparisons of Normalized Mean Scores

Assessment / Criteria / Question Producer	QM1	QM2	QM3	QM4	QM5	Ave
AQG	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

A one-way ANOVA and post-hoc analysis using Fisher's least significant difference (LSD) tests were performed.

Table 6 and Table 7 show that questions from the AQG system significantly outscored generic questions in each quality measure (QM) while outperforming peers' questions in quality measures 1, 3,

4, 5 and Average. The difference between supervisor's questions and the system for QM5 is not significant. This result indicates that questions produced by the system are perceived to be as useful as human supervisors. This positive result might be explained by two factors. First, the system's questions are useful because such questions 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.

Table 7: Fisher's Least Significant Difference (LSD) tests with 95% confidence interval and the * represents a significant difference.

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

Human Question Types Evaluation and Result

In order to design an efficient question template, it is very important to investigate the kinds of questions that human reviewers (supervisor and peer in this case) typically generate and how valuable each question type is. The citation questions were classified into 3 categories:

Association questions, which address the entities of the system and their properties, such as who, what, when, or where questions. E.g., *What is a biodegradable polymer?*

Explanation questions, which focus on justifications or explanation for these entities, such as why or how questions. E.g., *Why is it difficult to control the driving forces of natural ventilation?*

Prediction questions, which address the need to foresee consequences, such as what happens next or if-then. E.g., *Do you believe that future progress in microphysics modelling lies with bin or bulk microphysics modelling approaches?*

These categories were proposed by Trabasso et.al. (1996) as a classification of "Information Seeking Questions" (ISQs), i.e. questions that address knowledge deficits when subjects process scientific texts with a particular goal in mind. Since the questions generated by AQG system are "why" related questions, they are mapped to the Explanation category. Table 8 shows that Association and Explanation questions form the majority of human-generated questions, while there are only few Prediction questions.

Table 8: The number of questions generated according to question categories

Source	Type (for citation–related questions only)	Example	Frequency
Supervisor	Association	*Which type of the wind turbine (A to D) has the highest potential electrical conversion efficiency?*	44
	Explanation	*Why coatings and linings are needed to apply to form a monolithic layer in the sewer pipes to inhibit further deterioration in Najafi's paper?*	57
	Prediction	*Do you believe that future progress in microphysics modelling lies with bin or bulk microphysics modelling approaches?*	3
	Association & Explanation	*What are the strength and ductility of materials and how are they related to the energy absorption ability of the materials?*	3
Peer	Association	*What is the beginning process of MIC that shown by Little and Lee?*	89
	Explanation	*Why Carrasco et al. regarded wind energy as a relatively mature technology?*	40
	Prediction	*In your opinion, how NGMN can play a major role in future of Internet?*	1
	Association & Explanation	*What is Emergence? How does it examine agencies interaction?*	3
AQG	Explanation	*In the study of Mulligan, why are organic acids including citric , oxalic , malic , etc extensively used in mineral leaching?*	126

Table 9 shows that Explanation questions are perceived to be the most valuable while Generic questions are the least valuable. A one-way ANOVA test indicated that there are significant differences between Association, Explanation, Generic, and Non-citation related questions in terms of their perceived quality ($F(4,398) = 3.296$, $P<0.05$). Non-citation questions address presentation issues, such as referencing, formatting and numbering. A posthoc test found that Explanation questions including

both system and human generated questions significantly outscore Association and Generic questions (see Table 10).

Table 9: The average score of different question types

Question Type	Average Perceived Quality Score
Association (human)	3.94
Explanation (human)	4.24
Non-Citation related (human) e.g. Reference citing should be checked	3.93
Generic (human) e.g. What is the relationship between your literature and your research?	3.51
Explanation (AQG)	4.10

Table 10: Fisher's least significant difference (LSD) tests with 95% confidence interval and the * represents a significant difference.

Comparison in Pair	Association	Explanation	Non-Citation Related	Generic
Explanation	MD=0.30 * LSD=0.26			
Non-Citation Related	MD=0.01 LSD=0.38	MD=0.31 LSD=0.39		
Generic	MD=0.44 LSD=0.45	MD=0.74 * LSD=0.48	MD=0.43 LSD=0.55	
Explanation(AQG)	MD=0.16 LSD=0.23	MD=0.14 LSD=0.25	MD=0.17 LSD=0.38	MD=0.59 * LSD=0.45

Discussion and Conclusion

Automatic generation of natural questions is a challenging task. This article addressed this challenge and presented a novel AQG system for supporting academic writing by applying overgeneration-and-ranking approaches. The results indicate that this approach is effective since the ranking model improved the acceptability of the top 25% questions by 10%. In particular, RankSVM slightly outperformed Logistic Regression and the experiments revealed that the semantic features are important. However, if the document is poorly written and citation sentences contain grammatical errors, it would cause the system to generate no questions or poor questions with grammatical errors. It is because the question generation is a pipeline process and any error occurs at one stage will influence the following stages. Citations sentence containing grammatical errors could influence the parser to correctly extract predicate verb for sentence classification and transformation. This could cause to misclassify the sentence or incorrectly transform the sentence into a question.

As expected, the top ranked questions generated by the system outperformed the generic questions and are as useful as human generated one if excluding some questions with surface errors. One reason the

system generated questions are as good as the human-generated ones is because the system questions are specific and addressed critical thinking aspects.

Explanation and Association questions are mostly common used by human. Particularly, Explanation questions are more useful than other questions types because they normally trigger deep reflection and invoke critical thinking. This would inspire us to design an effective question template for the AQG system. However, it has been found that association questions are also frequently used by human supervisors and peers. These questions are still valuable to help students to understand key concepts described in the document. Our future work will focus on generating association questions from the key concepts by using information extraction techniques.

However, the question ranking model may not be applied to other question generation approaches since some of the defined features are only related to citations. To generalize this question ranking model, more generic question generation approaches and fine-grained generic features are needed. Despite these shortcomings, we believe that this AQG approach is effective and the evaluation meaningful because real academic writings were used.

Our future work will investigate how the system generated feedback questions would influence student's behaviors and learning performance. For example, a similar writing activity can be conducted, which consists of a draft session and a revise session. In draft session, a student writes a draft proposal and submits to his/her supervisor. After submitting the draft proposal, the student received a score and feedback based on the draft. The feedback could be either human feedback or system feedback. After receiving the feedback, the student revised the proposal in the revise session and finally gets a score for the final proposal. By this way, the system can track how many changes the students would make in the writing after receiving the feedback questions. The system can also found out the score difference between draft proposal and final proposal.

Acknowledgements

Ming is supported by the Young and Well CRC. This work is also supported by Chinese Fundamental Research Funds for the Central Universities under Grant No. XDJK2014A002.

References

Afolabi, M. (1992). The review of related literature in research. *International Journal of Information and Library Research, 4*(1), 59-66.

Cao, Z., Qin, T., Liu, T. Y., Tsai, M. F., & Li, H. (2007). *Learning to rank: from pairwise approach to listwise approach.* Paper presented at the International conference on Machine Learning, June 20-24, New York, USA.

Centre for Academic Success. (2011). Reporting Structure. Retrieved 18 August, 2011

Collins, M., & Koo, K. (2005). Discriminative Reranking for Natural Language Parsing. *Computational Linguistics, 31*(1), 25-70.

Duh, K. (2008). *Ranking vs. Regression in Machine Translation Evaluation.* Paper presented at the The Third Workshop on Statistical Machine Translation, July 19, Stroudsburg, USA.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*(1), 104-137.

Graswell, G. (2008). *Writing for academic success: a postgraduate guide*: SAGE Publications.

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (1998). *Metacognition in educational theory and practice*. Mahwah, US: Mahwah, NJ: Erlbaum.

Heilman, M., & Smith, N. A. (2010). *Good Question! Statistical Ranking for Question Generation.* Paper presented at the Annual Conference of North American Chapter of the Association for Computational Linguistics - Human Language Technologies, June 1-6, Los Angeles, USA.

Joachims, T. (2006). *Training linear SVMs in linear time.* Paper presented at the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, August 20-23, Philadelphia, USA.

Levy, R., & Andrew, G. (2006). *Tregex and Tsurgeon: tools for querying and manipulating tree data structures.* Paper presented at the The Fifth International Conference on Language Resources and Evaluation, May 24-26, Genoa, Italy.

Liu, M., Calvo, R. A., Aditomo, A., & Pizzato, L. A. (2012). Using Wikipedia and Conceptual Graph Structures to Generate Questions for Academic Writing Support. *IEEE Transactions on Learning Technologies, 5*(2), 251-263.

Liu, M., Calvo, R. A., & Rus, V. (2010). *Automatic Question Generation for Literature Review Writing Support.* Paper presented at the 10th International Conference on Intelligent Tutorial Systems, June 14-18, Carnegie Mellon University, USA.

Liu, M., Calvo, R. A., & Rus, V. (2012). G-Asks: An Intelligent Automatic Question Generation System for Academic Writing Support. *Dialogue and Discourse: Special Issue on Question Generation, 3*(2), 101-124.

Liu, T. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval, 3*(3), 7.

Mostow, J., & Chen, W. (2009). *Generating instruction automatically for the reading strategy of self-questioning.* Paper presented at the International conference on Artificial Intelligence in Education, July 6-10, Amsterdam, Holland.

Powley, B., & Dale, R. (2007). *Evidence-based information extraction for high-accuracy citation extraction and author name recognition.* Paper presented at the 8th RIAO International Conference on Large-Scale Semantic Access to Content, May 30 to June 1, Pittsburgh, USA.

Ratinov, L., & Roth, D. (2009). *Design Challenges and Misconceptions in Named Entity Recognition.* Paper presented at the Thirteenth International Conference on Computational Natural Language Learning, August 8-9, Sofia, Bulgaria.

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

Rus, V., & Graesser, A. C. (2009). *The Question Generation Shared Task and Evaluation Challenge*. Paper presented at the 2nd Workshop on Question Generation July 6, Arlington, USA.

Steward, B. (2004). Writing a Literature Review. *The British Journal of Occupational Therapy, 67*, 495-500.

Taylor, D., & Procter, M. (2008). The Literature Review: A Few Tips On Conducting It. . Retrieved 2011, 1rt July, 2011, from http://www.writing.utoronto.ca/advice/specific-types-of-writing/literature-review

Trabasso, T., & Magliano, J. P. (1996). Conscious understanding during comprehension. *Discourse Processes, 21*, 255-287.

Villalon, J., Kearney, P., Calvo, R. A., & Reimann, P. (2008). *Glosser: Enhanced Feedback for Student Writing Tasks.* Paper presented at the The Eighth IEEE International Conference on Advanced Learning Technologies ICALT '08, July 1--5, Santander, Spain.

Yao, X. (2010). *Question Generation with Minimal Recursion Semantics.* Unpublished Master's thesis, Saarland University & University of Groningen Saarbrücken, Germany.