An Automatic Question Generation Tool for Supporting Sourcing and Integration in Students' Essays

Ming Liu
School of Elec. & Inf. Engineering
University of Sydney
NSW
liuming@ee.usyd.edu.au

Rafael A. Calvo
School of Elec. & Inf. Engineering
University of Sydney
NSW
rafa@ee.usyd.edu.au

Abstract *This paper presents a domain independent Automatic Question Generation (AQG) tool that generates questions which can be used as a form of support for students to revise their essay. The focus here is on generating questions based on semantic and syntactic information acquired from citations. The semantic information includes the author's name, the citation type (describing the aim of the cited study, its results or an opinion), the author's expressed sentiment, and the syntactic information of the citation. Pedagogically, the question templates are designed using Bloom's learning taxonomy where the questions reach the Analysis Level. We used 40 undergraduate students essays for our experiment and the Name Entity Recognition component is trained on 20 essays. The result of our experiment shows that the question coverage is 96% and accuracy of generated questions can reach 78%. This AQG tool will be integrated into our peer review system to scaffold feedback from peers.*

Keywords Question Generation, Electronic Feedback System for Sourcing and Integration in Students'Essay

1 Introduction

Progress made in question answering systems has motivated a recent growth in automatic question generation systems. Two types of question generation tasks are normally considered. The first is text-to-question, where a document is provided to an AQG system that generates a question for which the answer is contained in the text. The second type is as a component of an Intelligent Tutoring System where a dialogue between the student and the ITS, and a set of propositions, is used as the input to the AQG component. In this case the question is aimed at helping the student elicit an answer containing the propositions.

The former AQG systems can support reading comprehension tasks, automatically suggesting questions that tutors can use in their teaching. Similar systems can be used to generate questions in the medical or security domain, where a system suggest questions to a practitioner based on a the case file. The second type of AQG systems is useful in a growing number of tutoring systems that have natural dialogue capabilities (e.g. Autotutor discussed later).

In this study we are concerned with building an AQG component for a third type of pedagogical applications: supporting students in their academic writing. In this context the common way of addressing the AQG problem is substantially changed:

- The driver for the technology is pedagogical so the questions should be framed in a pedagogical theoretical framework.
- The domain may be very general and a corpora for background knowledge might not be available.
- The questions must be generated from a single document, instead of a whole corpora
- The target audience of the questions is the same author of the document. The author should know the answers, so the goal here is to trigger reflection or get the student to expand on a topic.

Most different genres of academic writing contain citations of third party work on which the student is expected to comment (as in a literature review) or which is being used as evidence in an argument. When writing an essay or literature review, students are expected to learn and reason from multiple documents which require the skill of *sourcing* (i.e., citing sources as evidence to support their arguments) and *Information Integration* (i.e., presenting the evidences in a cohesive and persuasive way).

The development of student's sourcing and integration skills can be supported by using trigger questions such as *Does the essay provide evidence for the claims it makes?* or *Does the conclusion follow from the argument?* But such questions are too general and not likely to provide strong support in the process of writing on a specific topic. More specific questions need to be asked.

Most of the current AQG systems rely on shallow semantic parsing with entity recognizers. For example, Name Entity Recognizer,Verbnet [14] and Framenet [1]

Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

can only ‘understand’ the semantic role of the entities such as agent, time, location and object in a sentence and generate factual questions. To generate deep questions related to a student’s essay, AQG systems depend on some type of domain knowledge. AutoTutor [8] can generate deep questions, using domain specific knowledge in Computer Literacy or Physics.

This paper describes a new AQG system that includes a name entity recognizer for citation extraction, a pattern-matching based classifier for citation type classification and a sentiment analysis component for detecting the author’s opinion polarity. These pieces of information are used to generate template-based questions during student’s academic writing activities and targeting specific levels of Bloom’s learning objectives taxonomy. Section 2 provides a brief review of the extensive literature focusing on approaches and systems that support learning experiences with sourcing and integration as learning goals. Section 3 describes the system’s architecture while Section 4 its evaluation, including coverage and correctness. Section 5 concludes.

2 Related Work

Natural Language Processing techniques have been used to develop a number of tutoring and feedback systems. Section 2.1 reviews some of the projects developing writing support tools, and Section 2.2 systems that generate questions automatically.

2.1 Electronic Feedback System for Sourcing and Integration

Numerous projects have used computational approaches to assessing and providing automatic feedback on writing, most of the focus being on the assessment [15]. Despite a variety of initiatives to improve the quality of automatic feedback the efficacy of the systems remains to be proven and more research is needed. Meanwhile providing timely and appropriate feedback at key stages of the writing process remains a manual task, and a serious challenge for university lecturers.

Some of the early systems include Writers Workshop a system developed by Bell Laboratories, and Editor [16] both focused on grammar and style. Studies on the impact of Editor [2] concluded that the pedagogical benefits of grammar and style checking are limited. It could also be argued that these systems only aimed at supporting writing to communicate and did not address the issue of supporting writing to learn, important in today’s curriculum design.

SaK, a writing tutoring system developed at the University of Memphis [18] is based on the notion of voices that speak to the writer during the process of composition. SaK uses avatars to give the impression of giving each voice a face and a personality [18]. Each avatar provides feedback on a different aspect of the composition, saying what is good or bad about 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 between texts and has been described thoroughly elsewhere [11]. SaK can also analyze the purpose of a sentence, identifying clusters of topics amongst the students so when the topic of a new composition is not identified the student can be asked for an explanation or reformulation.

Sourcer’s Apprentice Intelligent Feedback (SAIF) [3] is an automated feedback tool for writing essays which can be used to detect plagiarism, uncited quotation, 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 1.

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 1: Types of Problems SAIF addresses and the intended goal of feedback

SAIF also uses Latent Semantic Analysis (LSA) techniques for plagiarism detection, computing the similarity between each essay sentence and the source sentences in LSA semantic space. For finding the explicit citations, SAIF uses a Regular Expression Pattern Matching technique to detect the explicit citations by recognizing phrases containing the author’s name (e.g. According to, As stated in, State). Evaluations showed [3] that SAIF provides helpful feedback for students to use more explicit citations in their essays. However, this tool only addressed some basic problems for sourcing and integration. Moreover, it required a large number of source documents to build the LSA semantic space and a large number of pattern matching rules had to be predefined.

Glosser is an automated feedback system for student’s writing [17]. It uses textual data mining and computational linguistics algorithms to quantify features of the text, and produce feedback for the student. This 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 analyzing the words contained in each paragraph, it can measure how close two adjoining paragraphs are. If the paragraphs are too far this can be a sign of what is called lexical cohesiveness and Glosser flags a small warning sign. Glosser (1.0) provides feedback on four aspects of the writing: structure, coherence, topics, and concept visualization.

Glosser does not address sourcing directly, but four trigger questions (and the text features above) 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 analyse their importance and critically discuss their usefulness?
4. Does this essay simply present ideas or facts, or does it analyse their importance?

The AQG algorithms described here are designed to be integrated into Glosser and provide support for sourcing an integration of citation sentences. The students upload a composition and Glosser provides the different forms of feedback. Other approaches for including the automatically generated questions include embedding them within an email, or using them as part of a peer-review process.

2.2 Question Generation

One of the first automatic question generation systems proposed for supporting learning activities was AUTOQUEST [19]. In this case, as in most of the current research questions are generated from external sources that the student *reads* (as opposed to writes).

The approach used here is similar to that of Kunichika et. al. [10] who proposed an AQG approach based on both the syntactic and semantic information extracted from the original text based on DCG (Definite Clause Grammar). 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, used to determine the interrogative pronoun for the generated question. For example, in the noun category, several noun entities can be recognized including the Person, Time, Location, Organization, Country, City, Furniture. In the verb category, the bodily actions, emotional verbs, thought verbs and transfer verbs can be identified. It also builds the semantic links among 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 more sophisticated and provide better support. The empirical result shows that 80% questions were considered by experts as appropriate for novices learning English and 93% of the questions were semantically correct.

AutoTutor, developed by the Graesser et al [8] at the University of Memphis, is an ITS that improves student's knowledge in computer literacy and Newtonian physics through an animated agent asking a series of deep reasoning questions that follow Graesser-Person taxonomy [7]. In each of these themes a set of topics have been identified. Each topic contains a focal question, a set of good answer aspects, a set of hints, prompts or elaborations which used to elicit each good answer aspect, a set of anticipated bad answers and so on. The system initiates a session by asking a focal question about a topic and the student are expected to write an answer containing 5-10 sentences. The system can generate hints or prompts for the student to elicit the correct and complete answer. 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 system is domain dependent and requires a large number of human resources to predefine the content of each topic.

3 System Design and Architecture

The AQG tool described here is designed to generate questions from a student's essay and a set of templates designed by the instructor. The system was evaluated using a corpus of student essays discussed in Section 4. Sentences from that corpus are used here as examples on how the questions are generated. The corpus contains essays on the topic "English as a Global language".

In this section we provide an overview of the system's architecture shown in Figure 1 and describe each step in a pipeline process. The input to the system is an essay and the output is the generated questions.

Table 2 shows an example of questions generated by the AQG tool and their mapping to cognitive levels in Bloom's Taxonomy. In this example, the questions are generated from the raw sentence written by a student as part of an essay.

The question generation process follows 3 steps shown in Figure 1:

Step 1. Pre-processing. This includes citation extraction, filtering 'noisy' segments, splitting complex sentences and sentence transformation if it uses a noun or passive voice to refer to resources. There are two major components to perform these tasks: 1 *Sentence Extractor*, performs citation sentence extraction using the combination of trained Stanford Name Entity Recognizer [5], and a Pronoun Resolver, which is implemented by finding the nearest Name Entity appeared before the pronoun, and 2 *Filter* performs the rest of tasks which involved to clean up "noisy" segment, split complex sentences, transform other types of citation form to reporting verb type by using Tregex Pattern Match Techniques[12].

Examples of students' compositions include:

1. **According to Crystal,** *more people in the world speak Chinese than any other language.*

Level	Description	Example
Recognition	Ability to identify the specific content.	1.Who is Graddol? 2.What does Graddol point to in his study?(Sourcing)
Recall	Ability to retrieve the specific content from memory.	The same to Recognition
Comprehension	Ability to understand the learning material in terms of generation inferences, interpretation information, explanation and summarization information.	Why would Graddol point to the social and economic inequality that the dominance of English could lead to? (What evidence does Graddol provide to prove that?) (Sourcing)
Application	Ability to apply the knowledge from the learning material to a problem or situation.	How did you present Graddol opinion as evidence to confirm the thesis in your essay?(Integration)
Analysis	Ability to disassemble the elements and find the relationship between elements.	1. Is Crystal against Graddol's opinion? 2. Since you say Crystal's opinion is against Graddol, can you find the contradictive evidence provided by Crystal? (Integration)

Table 2: An example of questions generated from the sentence *"Graddol on the other hand points to the social and economic inequality that the dominance of English could lead to"*.

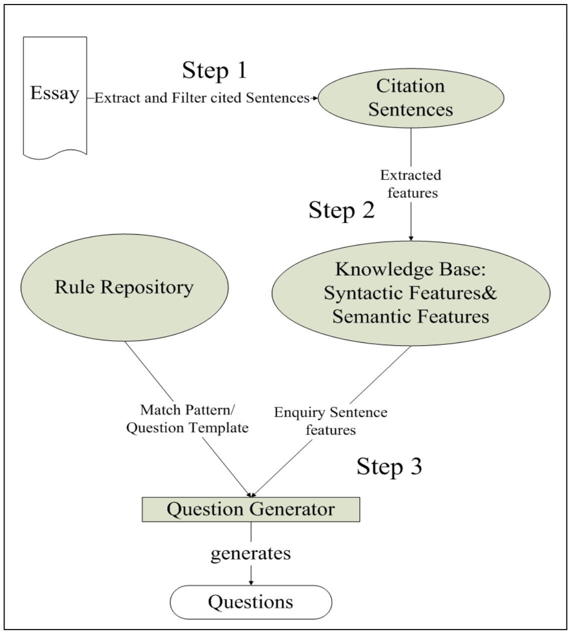
Essay
Step 1
Extract and Filter cited Sentences
Citation Sentences
Extracted features
Step 2
Rule Repository
Knowledge Base: Syntactic Features& Semantic Features
Match Pattern/ Question Template
Enquiry Sentence features
Step 3
Question Generator
generates
Questions

Figure 1: System Architecture

2. Although *Crystal and Graddol use many statistical evidence to discuss the spread of English as a Global language and the resulting consequences of this*,**Wallraff actually challenges the notion that English has the global status most people believe it to have.**

3. *Wallraff's* **opinion** *is that there is a rate of growth of other languages in the USA which is higher than the rate of growth of English.*

In sentence 1 the noisy segment is shown in Bold. Sentence 2 is a complex sentence divided into two simple sentences shown in Bold and Italics respectively. Sentence 3 uses the noun 'opinion' to refer to the reference and the system will convert it into a reporting verb type (explained later). The new reported verb type version for Sentence 3 is:

Wallraff states that there is a rate of growth of other languages in the USA which is higher than the rate of growth of English.

To achieve these, the input sentence is parsed into a Phrase Structure tree, and then the Tregex Expressions are used to detect the syntactic patterns, and finally we use Tsurgeon to perform required opersions.

Tregex, developed by Stanford NLP group, is a powerful pattern matching technique which can match an individual word, regular expression, a POS tag or group of POS tags such as a Noun Phrase. Once the matched node is found by the Tregex, the Tsurgeon tool can perform delete, add, remove the node from the syntactic tree as shown in Figure 2 .

According to a study by Hyland [9], there are mainly three grammatical ways to refer to sources, which use Reporting Verb, Noun and Passive construction. Here, we call this as three grammatical patterns for citation. In our implementation, the citation sentence which is either Noun or Passive construction patterns would be transformed into reporting verb pattern because it would be easier to transform the citation sentence with reporting verb pattern into questions in later stage. Therefore, the Tregex Expression are defined to detect the three grammatical patterns and extract right Subject, Predicate Verb, Predicate, Auxiliary Verb for processing in later stage. The code segment in Figure 2 is used to split the complex sentence 2.

```
find_adv=TregexPattern.compile
          ("ADVP =rb >>,(NP >(S > ROOT)) | > S");
find_clause=TregexPattern.compile
          ("SBAR=sbar<(IN<Although|though)<S");
find_comma=TregexPattern.compile
          ("/,/=comma \$ (NP >( S > ROOT ))");
Tsurgeon.parseOperation("delete rb");
Tsurgeon.parseOperation("delete sbar");
Tsurgeon.parseOperation("delete comma");
```

Figure 2: An example of code segment using TregexPattern and Tsurgeon for splitting a complex sentence

Step 2. Syntactic and Semantic features. The purpose of this step is to extract the Syntactic feature and Semantic feature, such as the citation type (Study Result, Author's Opinion, Aim of Study) and the Author's Opinion Polarity. AQG then inserts the Semantic features as facts into a prolog knowledge base to be used in Step 3. There are two components to perform these tasks: *a Sentence Feature Extractor* which performs Syntactic Feature and Semantic feature extraction, and *a Sentiment Classifier* which detects the Author's Opinion Polarity.

Sentence Feature Extractor uses Tregex Expression on the Syntactic Tree for pattern match to extract syntactic features: Subject, Predicate Verb, Link Verb, Modal Verb and Predicate which are essential elements for question generation. In addition to Syntactic Features Extraction, *Sentence Feature Extractor* also uses predefined Reporting verb to define the Citation Type by matching the predicate verb in a sentence. In our database, Reporting Verb have been classified into three categories which correspond to different citation types.

Sentiment Classifier is used to detect the Author's opinion polarity about a topic. For the sentiment analysis AQG defines three elements: Opinion Holder, Topic and Opinion Polarity. At the moment, AQG only handles one Author appearing in a sentence and the opinion holder is the Author mentioned in the citation sentence. The topic is detected by choosing the most frequent noun or noun phrases among citation sentences expressed as a Sentence-Term matrix containing rows corresponding to the citation sentences and columns corresponding to the terms appeared in the sentence. Because AQG doesn't consider the number of times a word appears, a Binary Weighting schema is used. The topic is chosen by finding the term with maximum value and the Equation 1 is defined below, where $a_{ij} = 1$ if the term j appears in the citation sentence i, n is the number of citation sentences in an essay and the m is the number of terms appearing in these sentences.

$$\max_{\forall j \in m}\{\sum_{i}^{n} a_{ij}\} \tag{1}$$

For example, two citation sentences are extracted from an essay.

1. *"The increasing use of English is also negative in respect to the advantage gained by its native-speakers, not to mention the "threat to the identity of nations" through the inevitable increase of use of minority languages (Crystal, 1992)."*

2. *"Graddol on the other hand points to the social and economic inequality that the dominance of English could lead to."*

The word *'English'* has been chosen as Topic because it has the largest value 2 according to Equation 1. After the Opinion Holder and a Topic are detected, AQG detects the Opinion Polarity about the topic. The Opinion Polarity is decided by the Sentiment Region containing sentiment words in a sentence. The size of Sentiment Region is very important and AQG defines it as the set of nearest sentiment words around the topic in a sentence, and use the SENTIWORDNET [13] to determine the sentiment of a word. The SENTIWORDNET, a publicly available lexical resource for opinion mining, is an extension of WORDNET2 [4] and has defined three categories for a word sentiment with some magnitude: positive, negative and neutral.

Sentence	Opinion Holder	Topic	Polarity	Sentiment words list
S1	Crystal	English	Negative	(negative=-1.0), gain=0.5, increase=0.5
S2	Graddol	English	Negative	Inequality=-1.0

Table 3: an example of Author's Sentiment Classification

Table 3 shows the result of Sentiment Classification from the two citation sentences in the above example. *Crystal* is the Opinion Holder for Sentence 1, the *English* is chose as the Topic and the Opinion Polarity is *Negative* because AQG calculates the sum of the two nearest sentiment words: *Negative=-1.0* and *increase=0.5* which is negative. It is similar to sentence 2. Once finishing the sentiment analysis AQG will insert the extracted facts including Opinion Holder,Topic and Opinion Polarity into our prolog knowledge base showed in Figure 3 which will be used to infer if the Author's opinion is against/support each other.

```
#Facts
author(crystal).
author(graddol).
against(graddol,english).
against(crystal,english).
opinion(english).
#Inference rules
support(person1,noun1).
against(person2,noun2).
ally(X,Y):-support(X,Z),support(Y,Z),opinion(Z),X\=Y.
ally(X,Y):-against(X,Z),against(Y,Z),opinion(Z),X\=Y.
enemy(X,Y):-support(X,Z),against(Y,Z),opinion(Z),X\=Y.
enemy(X,Y):-against(X,Z),support(Y,Z),opinion(Z),X\=Y.
```

Figure 3: An example of Author's Opinion Polarity in Prolog knowledge base

Step 3. Generation This is the final step to generate template-based questions where the *Question Generator* uses the extracted syntactic features and the knowledge base, and then matches the predefined patterns in our Rule Repository, and finally generates template-based questions. In our current implementation, we have defined 5 rules and each rule defines the pattern for matching and 5 question templates. Each citation sentence would be applied by only one of the five rules. If a citation sentence matches both reporting verb and sentiment words, we would consider the rule for reporting verb because sentiment words have higher error rate to determine the citation type. In the future, we will use Machine Learning techniques to train a citation type classifier which will use the weight of selected features (reporting verb, sentiment words, numbers and etc) rather than current fixed pattern matching technique. Table 4 shows that the five rules are defined in our Rule Repository.

The Pattern Matching is based on the Reporting Verb and Word Sentiment in the citation sentence. In

Rules	Pattern	Citation Type	The Purpose of Generated Question
Rule 1	Reporting Verb	Opinion	Ask the student to provide evidence which support the Opinion (Sourcing), to provide other Author's contradictive opinion or result about the topic(Integration) if applicable
Rule 2	Reporting Verb	Aim of Study	Ask the student to identify the motivation for this Author's study and the outcome of the study (Sourcing).
Rule 3	Reporting Verb	Result	Ask the student to identify if the Author's Result is objective and what opinion does the result support (Sourcing)
Rule 4	Sentiment Word	Opinion	The same to Rule 1
Rule 5	Sentiment Word	Result	The same to Rule 3

Table 4: The Rule Definition for Patterns and Templates

our database, the reporting verb has been classified under one of three citation types and matches the predicate verb extracted from Step 2. If they are not matched, the sentiment words is used to detect the citation type. In our Rule repository, the question templates are designed according to the citation type. For example,

Graddol on the other hand points to the social and economic inequality that the dominance of English could lead to.

The predicate verb is *point to* and it matches a reporting verb under Opinion Type in our repository, then we apply Rule 1 shown in Table 4 to generate the template-based questions. Table 5 gives an example of question templates defined in Rule 1 and Table 2 shows an example of generated template questions defined in Rule 1. As you noticed, the following questions are generated by using prolog inference engine described in Step 2.

1. Is Crystal against Graddols opinion? 2. Since you say Crystals opinion is against Graddol, can you find the contradictive evidence provided by Crystal? (Integration)

If the sentence does not contain any reporting verb but some sentiment words, then it is also considered as the Author's Opinion. For example,

The increasing use of English is also negative in respect to the advantage gained by its native-speakers, not to mention the "threat to the identity of nations" through the inevitable increase of use of minority languages (Crystal, 1992).

As the word *Negative* has been detected as a sentiment word, the sentence is consider as expressing Author's Opinion, and then AQG applies Rule 4 to generate questions. Rule 5 is similar to Rule 4 for pattern matching except the sentence does not contain the sentiment words and the citation are expressed as Study Result.

Pattern	The predicate verb matches reporting verb for expressing Authors opinion purpose.
Template	• Who is [Author Name]? • What does [Author Name] [predicate verb Lemma]? • In the [Author Name]s study, do you agree that [Author Name] [Predicate]? Have you evaluated [Author Name]s opinion? • Why would [Author Name] [Predicate]? (What evidence does [Author Name] provide to prove that?) • How did you present [Author]'opinion as evidence to confirm the thesis in your essay? • Is [other Author Name] against [Author Name]'s opinion? Since you say [Other Author Name]s opinion is against [Author Name], can you find the contradictive evidence provided by [Other Author Name]?

Table 5: A Example of Question Template in Rule 1

4 Evaluation

This section describes a preliminary evaluation of the technique focused on two aspects : 1) The Question Coverage. 2) The Semantic Correctness of generated questions. In the last section we comment on planned evaluations that will study the learning impact of such a system, and self (the writer's view) and 3rd person reports on the quality features of the questions generated.

The evaluation was performed using 40 essays written by students at the University of Sydney. Students gave informed consent as approved by the Human Ethics Committee of the University of Sydney.

4.1 Question Coverage

The citation sentence extraction approach is based on the Author Name Recognition. The Expected Number of Questions depends on the total number of citation sentences. Table 6 shows that AQG can reach 96% coverage. This dataset contains 127 citation sentences(127*2=254 questions) and 123 citation sentences (123*2=246 questions) are extracted by AQG. We only evaluate 2 generated questions per citation sentence because some template-based questions only require Author Name, a relatively easy task, the evaluation does not include these questions.In other words, two questions are evaluated per citation sentence. For example, in Rule 1 question 3 and 4 are evaluated which is shown in Table 5. The problem for missing these citation sentence extraction is that some Author Names are not identified by the Name Entity Recognizer which cause these citation sentences can not be detected by AQG.

Expected Number of Questions	Number of Generated Questions	The Question Coverage
254	246	96%

Table 6: Question Coverage

4.2 The Correctness of Generated Questions

123 citation sentences were extracted from the 40 essays. Of these, 5 citation sentences had serious grammatical errors which caused the sentence Parser to fail. Therefore only the 118 remaining sentences were considered for evaluation. Because we only evaluate two questions per rule, the total number of evaluated questions is 236.

Table 7 shows that the semantic correctness of question reach to 78%. One of the main problems is that the rules used are too rough to handle multiple Authors appeared in a sentence. For example, the sentence

"*Wallraff suggests that the number of Spanish speakers in the USA has grown by 50% in the 1980-1990 census, thus refuting Crystal and Graddol's arguements for English being a global language.*"

Another major problem is the misclassification for the citation type: Opinion and Result. For example, the sentence

"*Many Chinese-speakers (four out of five of about 2.4 millions) in America prefer to speak Chinese at home rather than English (Wallraff, 1999).*"

In this case, although it contains *prefer* as a sentiment word with a positive term, the citation sentence should be considered as Study Result.

Rules	Number of Generated Questions	Number of Semantic Correct Questions
Rule 1	82	72
Rule 2	12	12
Rule 3	40	36
Rule 4	64	34
Rule 5	38	30
Total	236	184

Table 7: Question Generation Result

5 Conclusion and Discussion

Sourcing and Integration are important quality features in writing, and are part of the skills that college students must learn to master. The importance of asking questions has been shown to be an important part of teaching and learning experiences, so we designed an implemented a tool for automatically generating questions from an essay.

This domain independent AQG tool supports student's essay writing in the areas of sourcing and integration. Although we have not yet been able to assess the impact on student learning, the system was evaluated using real student essays.

Both the question coverage and the semantic correctness of generated questions were evaluated. Although the performance of Name Entity Recognizer would be different under different domain, the focus of current work is on interesting question generation. The pattern matching algorithm is based on Hyland's citation study that describes the most common ways of citing third party work. The algorithm captures the major forms of citation and as shown to have excellent accuracy.

Reasoning techniques were implemented in Prolog to detect when two authors are against each other and the generated question can reach to Analysis Level in Bloom's Taxonomy.

The tool can not only detect how many citations the writer has used in their essay but also generate specific or content related questions. Compared to current question generation systems, our tool can generate pedagogically deep questions in a somewhat domain independent form (it still requires templates that may required adaptation). It also presents novel results for using the authors' sentiment to generate questions.

Some limitations of this early work are obvious as the need to handle multiple authors in a sentence and to improve the classification of the citation type.

In future work, we will integrate the AQG tool into Glosser and to our peer review system so it provides extra information to support students' engagement with the writing (or peer-reviewing) process. for example, in a peer-reviewing scenario, the peer could not only evaluate the essay but also the author's answers to these automatically generated questions and provide better feedback. We will also improve the technique by adding ways for extracting multiple authors' arguments in a sentence and use other machine learning techniques to improve the Citation Type classification accuracy.

Acknowledgements

The authors would like to thank Jorge Villalon and Setphen O'Rourke for the development of TML and Glosser. Ming is partially supported by a N.I. Price scholarship. This project was partially supported by an Australian Research Council Discovery Project DP0986873.

References

[1] C. F. Baker, C. J. Fillmore and J. B. Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics*, pages 86–90, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[2] T. J. Beals. Between Teachers and Computers: Does Text-Checking Software Rea lly Improve Student Writing? *English Journal*, pages 67–72, 1998.

[3] M. A. Britt, P. Wiemer-Hastings, A. A. Larson and C. A. Perfetti. Using intelligent feedback to improve sourcing and integration in students' essays. *Int. J. Artif. Intell. Ed.*, Volume 14, Number 3,4, pages 359–374, 2004.

[4] C. Fellbaum and G. Miller. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[5] J. R. Finkel, T. Grenager and M. Christopher. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[6] X. Gong, Y. H.and Liu. Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, New York, NY, USA, 2001. ACM.

[7] A. C. Graesser and N. K. Person. Question asking during tutoring. *American Educational Research Journal*, Volume 31, pages 104–137, 1994.

[8] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan and D. Harter. Intelligent tutoring systems with conversational dialogue. *AI Mag.*, Volume 22, Number 4, pages 39–51, 2001.

[9] K. Hyland. Academic attribution: citation and the construction of disciplinary knowledge. *Applied Linguistics*, Volume 20, pages 341–367, 1994.

[10] H. Kunichika, T. Katayama, T. Hirashima and A. Takeuchi. Automated question generation methods for intelligent english learning systems and its evaluation. pages 1117–1124. Proc. of ICCE01, 2001.

[11] T. K. Landauer, D. S. McNamara, S. Dennis and W. Kintsch. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum, 2007.

[12] R. Levy and A. Galen. Tregex and tsurgeon: tools for querying and manipulating tree data structures. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation.*, 2006.

[13] S. Osinski and D. Weiss. Conceptual clustering using lingo algorithm: Evaluation on open directory project data. In *In IIPWM04*, pages 369–377, 2004.

[14] K. K. Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylyania, 2005.

[15] M. D. Shermis and J. Burstein. *Automated essay scoring: A cross-disciplinary perspective*, Volume 16. The MIT Press, 2003.

[16] E. C. Thiesmeyer and J. E. Thiesmeyer. *Editor:A System for Checking Usage, Mechanics, Vocabulary, and Structure*. New York: Modern Language Association, 1990.

[17] J. Villalon, P. Kearney, R. A. Calvo and P. Reimann. Glosser: Enhanced feedback for student writing tasks. In *Proc. Eighth IEEE International Conference on Advanced Learning Technologies ICALT '08*, pages 454–458, July 1–5, 2008.

[18] P. Wiemer-Hastings and A. C. Graesser. Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, Volume 8, Number 2, pages 149–169, 2000.

[19] J. H. Wolfe. Automatic question generation from text - an aid to independent study. *SIGCUE Outlook*, Volume 10, Number SI, pages 104–112, 1976.