Saarland University

Department of Computational Linguistics

Seminar: **Intelligent Tutorial Systems**

Automatic Question Generation for Literature Review Writing Support

Author:
Madhumita
Matriculation: 2551615

Supervisor:
Dr. Helmut Horacek

11th September 2014

Abstract

This paper surveys the existing approaches for automatic content related question generation for literature review writing support.

The first approach that has been discussed extracts citations from papers and classified them using a rule based approach. Thereby questions are generated using template-matching and they as used as feedback to students. A study using Bystander Turing Test shows that these questions are found to be more useful than generic questions. Students find it difficult to distinguish between questions generated by the system and the humans (F-score = 0.43).

The second approach discussed uses a similar set up as the previous approach. The difference lies in the classification of citations - machine learning approaches are used as opposed to rule based approaches. In a study conducted using this approach, similar trends were observed as the previous study. Although, higher citation classification accuracy was obtained.

Generating questions only based on citation put a limit on the concepts using which the questions which can be generated. A citation independent approach is discussed in the next paper that has been surveyed. Most frequent question types have been identified. Based on these question types, proposal for an AQG (Automatic Question Generation) framework has been put forward.

In the final paper, an approach to generate concept related questions have been discussed based on previous findings. Key phrases are first extracted from the literature reviews. Matching Wikipedia articles are found. The key phrases are classified into one of the concept categories among Research Field, Technology, System, Term and Other. A conceptual graph structure representation is constructed for each key phrases and related sections in corresponding Wikipedia article. This structure is used to generate questions through a template. A study similar to Bystander Turing Test was conducted, which showed that students found it difficult to distinguish between questions generated by supervisor and the system. The generated questions were found to be as useful as supervisor generated questions and more useful than generic questions. These questions were more useful for the students new to a concept than to those who had more knowledge about the concerned topic.

Contents

1 Introduction

Literature reviews are frequently written in the academia to improve the understanding of existing relevant literature. At the same time, it helps one to reflect about the topics involved and contribute to a critical analysis of the concepts put forward. One may also contribute to the existing discussions by suggesting an improvement over the existing methods. Writing a literature review helps one identify which fields have been well explored and which still need to be explored. A good review, according to Steward (2004) is defined as a Comprehensive, Relevant, A synthesis of key themes and ideas, Critical in its appraisal of the literature, and Analytical developing new ideas from the evidence.

According to Liu et al. (2010), writing a literature review involves two main skills: Gathering information from different sources and integrating them in a coherent manner according to ones understanding. However, one needs to be certain that he has correctly understood the content and has clearly identified the contributions by different researchers. He should understand how certain research ideas are different from each other and how closely they relate to one another. He should be able to provide a critical analysis of different ideas put forward. In order to be able to do this, one must reflect about the content he has read and written down.

Question-answering is believed to be quite critical in finding out how well has one understood any given topic. Classroom lectures and tutoring sessions are often accompanied with a question-answering session where every student is provided with an opportunity to find out their own knowledge deficits and fill the gaps with inputs from lecturers and tutors. Questions help one to reflect about the content, and in the process, improve their own understanding. By training students to ask good questions, improvements in comprehension, learning and memory of technical material can be achieved (see Davey and McBride, 1986; Gavelek and Raphael, 1996; Singer and Donlan, 1982). Literature review writing succeeded by a basic question-answering helps to reflect on the comprehensiveness of the current review and to know how well has one understood the content of his review. According to Reynolds and Bonk (1996), students who are given a generic trigger questions perform better at a writing activity than those who do not receive any trigger questions at all.

Tutors and lecturers often prepare content related questions to provide feedback to students. This procedure is quite slow and involves a great deal of time and energy on their behalf. If the process of generating content related questions could be automated, a great deal of effort could be saved and processes could be fast.

In this paper, different approaches for automatic question generation for literature review writing support have been discussed. The remainder of the paper is structured as follows:

Section 2 describes the approach for automatic question generation for literature review writing support using citations. The next section, section 3 describes the approach for AQG using automatic citation classification through machine learning approaches. Section 4 discusses an approach to generate content-specific questions for literature review writing support without limiting to citations. Section 5 utilizes the previously mentioned approach to generate questions through key concepts in a literature review and relations extracted through Wikipedia articles. The last section, section 6 discusses all these approaches and draws a conclusion.

2 Automatic Question Generation for Literature Review Writing Support

2.1 Proposed Approach

One of the approaches discussed for automatically generating content related question is generating questions based on citations in a given literature review (see Liu et al., 2010). Citations are used to relate the content of different research papers. Citations together give an overall idea about the content of the literature review. Analyzing the citations to understand why a particular paper has been cited in a specific manner helps us analyze the content and supplement any deficits in our knowledge.

In the proposed system for Automatic Question Generation based on citations in a literature review, first the citations have been extracted from the text. 5 citation styles have been discussed by Powley and Dale (2007). These are: Textual Syntactic, Textual Paranthetical, Prosaic, Pronominal and Numbered. Numbered citations are currently not extracted by Liu et. al. The textual syntactic and textual parenthetical citations have been extracted using pattern matching through the regular expression:

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

Figure 1. Regex for Textual Syntactic or Textual Paranthetical Citation Extraction

To identify and extract the Prosaic citations, and Named Entity Tagger, LBJ has been used. A pronoun resolver is used to identify authors for the Pronominal style citations.

Various syntactic features govern sentence construction in linguistics. Some of these features are the subject of a sentence, the predicate verb, auxiliary verb, tense, voice and predicate of a sentence. Similarly, there are some important semantic features as well. For example, these can be the name of the author and the semantic category of a sentence. The authors exploit these features of the citation sentences to generate questions.

Tregex Levy and Andrew (2006) is a powerful syntactic tree search language for identifying syntactic elements (e.g. main verbs of sentences). Tregex can be used to specify the various relations between the tree nodes. For example, "Node A is the parent of Node B" is denoted as $A < B$ while "B is the rightmost descendant of A" is denoted as $A << -B$. It also supports regular expressions. For example, the expression

$$NP(NounPhrase) < /area|discipline?/$$

matches NP is the parent of areas or areas, discipline or disciplines. The following Tregex is used on Phrase structure trees generated using Stanford parser to identify Method and System tupe citations:

- Method:
  $VP > (S > ROOT) <<, (use|apply) << (NP << -(method|approach|))$
- System:
  $NP > (S > ROOT) << (system|tool)$

In this manner, the subject, the predicate verb and the predicate of the citation sentence have been identified. Tense of a sentence is identified using the verb information. The dependency trees generated by the Stanford parser contain information about the voice of the sentence.

Once the citations have been extracted, the questions are generated based on template matching. Different question templates are prepared for different categories of citation. The syntactic and semantic features are combined with these templates to produce content specific questions.

2.2 Evaluation

To analyze how effective the questions generated by the system were, 6 literature reviews were collected from 6 participants. 5 questions each were generated by a tutor for the topic, an expert lecturer, and the system. Along with this, 5 generic questions were used. The students had to evaluate the quality of the questions generated from their own literature review paper along the following quality measures:

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

Quality rating was done on a scale of 1 to 5, 1 being strongly disagree and 5 being strongly agree.

The recall for retrieving citations from the literature reviews was 0.66 on average. Also, on an average, the precision for generating semantically correct questions was 0.60. Significant improvements in these numbers are required to generate a useful system.

The quality assessment of the questions showed that the proposed approach significantly outperformed the generic questions for overall question quality and was considered more useful than the generic questions. This was an expected result, because review specific questions allow more reflection about the content as opposed to generic questions.

One interesting observation was that there was no significant difference between the questions generated by the lecturer vs. the system, though the questions generated by the tutor were considered to be much better according to multiple evaluation criteria. However, the questions generated by the tutor and the system were also considered to be equally useful. It was also observed that when the questions were generated by the AQG system, it was difficult for the participants to determine who of the three system, lecturer or tutor had generated the question. This shows that the system generated questions were quite similar to the other two.

2.3 Limitations and Discussion

There are several limitations of the approach. First and foremost, the questions are generated only on the basis of citations in a literature review. However in the real world scenario, as a feedback mechanism, lecturers and tutors would generate questions on the basis of other relevant content as well.

Apart from that, the performance depends highly on the extraction of citations, which in turn depends on a Named Entity Tagger. The current Named Entity tagger used was primarily trained on News Text Corpora. Applying it to academic articles led to poor performance. If the tagger were trained on academic articles, author name identification would improve, which would improve the currently low recall for extraction. Also, rules could be added to extract the numbered citations as well to further improve the recall.

All the questions were presented to the students for feedback and evaluation. If the questions were ranked instead and only the best questions were reported, the overall quality would improve. One approach to rank the questions could depend on the location of a citation in the literature review. Certain sections of a literature review contain more important information than certain other sections. For example, the body contains fewer citations than the related work section. The citations in the body of a review would relate more to the proposed approach rather than background information. A weighing scheme could be employed to rank the questions based on the section of the review it has been generated from. Similar weighing scheme could be used for ranking questions in different categories as well. A scoring function to combine these weights could rank the questions as desired.

Moreover, only rule based approach has been used for citation category classification. It affects the scalability of the system.

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

3.1 Proposed Approach

Liu and Calvo further explored usage of machine learning approaches for citation category classification (see Liu et al., 2012b). The same categories of classified citations have been used as before Aim, Opinion, Result, Method, System and Application. Each citation has been represented as 17 generic features and data has been trained on 504 citations from 45 academic papers. Feature selection procedure has been discussed in the following paragraphs.

A cue phrase group is constructed which includes a noun cluster, an adjective words cluster and an adverb cluster. Certain nouns and verbs are often used in more than one category. For example, the words like suggest and observe are used in both Opinion and Result categories. These words are put in a shared cue phrase group. The words in this group are weighted for different categories. Using this approach, 12 cue phrase group features have been defined:

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 and 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 and Method Cue Phrase Group (verb and noun cluster)
11. Own (no cluster)
12. Other Cue Phrase Group (verb, noun cluster and adjective cluster)

Apart from the Cue Phrase features, sentiment features have also been defined. This binary feature checks the existence of positive or negative sentiment words, based on SentiWordNet(see Esuli and Sebastiani, 2006).

Negation features are also used to detect negations in citation sentences. These features are implemented through traditional negation words, restrictive adverbs, negative verbs and adjectives with negative prefixes.

The feature list is also supported by syntactic features like voice and tense of citation sentences, and other features like numeric features and length of citation sentences.

Once a feature set is obtained, supervised training and application is used for citation category classification. After classification, the questions are generated through template matching, in a manner similar to their initial approach.

3.2 Evaluation

In order to evaluate the utility of the questions for literature review writing support, a study was conducted on 33 PhD student writers and 24 supervisors from University of Sydney. Each student submitted a research proposal which was evaluated by another student and respective supervisor to provide feedback. 45 papers were used as the testing data for citation category classification. Maximum of 5 questions each were generated by the two of them and the proposed system. Some generic questions were included for evaluation purposes. Five quality measures and rating on 1 to 5 were used as in the previous study.

The citation extraction rate was found to be 88%. The accuracy of the Named Entity Tagger greatly affects the performance. Higher citation classification accuracy was obtained using this approach as compared to the previous rule based approach. Much higher recall values were observed, 0.71 as compared to a previous 0.37. The improved performance of machine learning approaches in complex syntax of sentences, as compared to Tregex expressions is thought to be the main reason affecting the performance. The citations in the classes aim, result, opinion and application have been identified quite well, except for implicit citations. The system category citations depend on a noun and a verb cluster. Identification is difficult in case of missing verb in shared verb cluster.

Content specific questions significantly outperformed the generic questions and the questions by the peers who may not have sufficient knowledge about the topic of the paper. Though in several dimensions the questions by supervisors were considered to be significantly better than the ones generated, the generated questions were considered equally useful as the ones generated by the supervisor. Also, the questions generated by this AQG system were often thought to be generated by the supervisors and peers. However, the human generated questions were considered to be more concise and correctly written. These results are in accordance with the previous study.

3.3 Limitations and Discussion

The limitations of the previous system all apply to this approach as well, with the exception of rule based classification of citation categories. Further studies were done to explore solutions to these limitations - the main limitation being the dependency of citation for question generation.

4 Question Taxonomy and Implications for Automatic Question Generation

4.1 Introduction

The previous studies clearly depict that content specific questions support the students better than generic questions when it comes to writing activities. Using the previous study as the base, Liu and Calvo explored generation of content specific questions without limiting themselves to citations. They proposed a more practical AQG framework to support academic writing in engineering education (see Liu and Calvo, 2011).

Graesser and Person have investigated the taxonomies of different questions asked during the tutoring sessions of college research methods and algebra for 7th graders (see Graesser and Person, 1994). A question is defined as a speech act that is either an inquiry, an interrogative expression, that is, an utterance that would be followed by a question mark in print, or both. According to their study, some frequent questions types asked and examples are the following:

Long answer questions can reflect a student's way of reasoning, misconceptions and amount of knowledge better than short answer questions or yes-no-maybe questions. The questions based on concept, causal relations and procedures are more frequent than those related to verification and judgment. Some simple questions outscore deep questions generated by the supervisors. This shows that conceptual questions are as important as procedural or causal questions. While designing the question templates, all these categories of questions are important to help a student reflect on his knowledge.

Further study was done to find out how are questions generated from text. It was found that 56.8% of the questions were generated from sentence and lexicon levels, 14.4% from discourse levels and 28.8% from the background knowledge level. Most of the questions at concept level are generated at lexicon level.

4.2 Proposed Approach

Using the findings discussed above, an approach for content specific AQG was proposed. The proposed framework consists of the following steps: First of all, sentences are extracted. These sentences are then simplified and parsed to build a term sentence vector space model. From this vector-space model, sentence types and key concepts are identified using sentence classifier and key phrase extractor respectively. Sentence classes could be one of application, result, comparative test and opinion, and a concept could be a research field, algorithm or theory. Once we have the key concepts, the semantic meaning

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

Table 1. Graesser and Pearsons question taxonomy with examples of questions from academic supervisors

of sentence and concept are used to generate questions in different categories through rule based approaches. Thereby, a ranker ranks the questions.

4.3 Discussion

This approach is quite promising, since around 57% of the questions generated by human beings are on sentence and lexical levels. However, it is only a suggested framework, but not a completely developed system. We do not know the actual performance of such a system. It will depend greatly on the performance of sentence extractor, key phrase identifier and sentence classifier. Moreover, the semantic and syntactic correctness of the questions is also unknown. Statistical ranking of questions is only an abstract concept that has been discussed instead of providing any practical implementation suggestions.

5 Using Wikipedia and Conceptual Graph Structures to Generate Questions for Academic Writing Support

5.1 Introduction

Extending the information extraction approach for automatic question generation, Liu and Calvo further proposed incorporating the Wikipedia and conceptual graph structures to generate questions for academic writing support (see Liu et al., 2012a). Along with focusing on finding out the key concepts from many different concepts in an academic paper, they also investigate a way to address system's lack of domain knowledge.

5.2 Proposed Approach

The usage of Wikipedia as a lexical resource has been proposed. This is because it is a free collection of vast and diverse knowledge source available to programmers for use. Due to standard structure of each document, it is relatively easy to create conceptual graph structures using Wikipedia articles.

In order to find the key concepts, a vector space model is built from sentences in the literature review. Lingos algorithm (see Osiński et al., 2004) based on Singular Value Decomposition is then used to find the key phrases. Abbreviated key phrases are expanded for better search results in Wikipedia. Definitions of key concepts are extracted from the Wikipedia articles with the key phrases as titles. The type of the key concept is found using different Tregex rules on the definition sentence. The key concept types fall in one of the concept type classifications from the previous study Research Field, Technology, System, Term and Other. Only the first four have been used for question generation.

From the key phrase classifications, a conceptual graph structure is built with the relations is-a (definition), has-limitation (drawback), has-strength (advantage), apply-to (application) and include-technology (methods used). Sentences and phrases are extracted from Wikipedia pages and classified as the fore-mentioned relations with the key concept using Tregex expression rules. Wikipedia section headings are used to identify the target sentences.

From the triples in the conceptual graph, specific and content-related questions are generated based on questions templates. The question generation approach now is simpler than the previous approach. The current template does not require any natural language transformations like subject-auxiliary-inversion to generate questions. Extracted information need to be directly filled up, hence reducing the no. of errors.

5.3 Evaluation

Analysis and evaluation was done on 20 engineering students and two domain experts supervising them. 82 questions were generated by the two supervisors and 154 questions by the AQG system using the literature reviews written by the students. Only 96 of the system generated questions were used for evaluation purposes. 5 Generic questions were also used from each paper. The following 6 quality measures were used for evaluation:

1. This question is correctly written.
2. This question is clear and not ambiguous
3. This question helped me develop a deeper understanding of important concepts or topics related to my project.
4. This question helped me think or learn about topics or concepts that I did not consider before.
5. This question prompted me to reconsider The structure of my literature review.
6. This question helped me think about how I can revise and improve my literature review.

289 unique key phrases were identified from the 20 literature reviews. Apart from the category Other, there were more Technology (39) and Research Field (20) key phrases than Term (12) and System (14). During the key phrase classification step, the category "Term" was found with a greater difficulty as compared to other categories. It may be due to more no. of implicit definitions in this category. Another problem encountered during classification was assignment of multiple categories to a sentence. Random choice was made to proceed with single category classification.

As expected, participants could easily identify the generic questions. However, following a similar trend as previous studies, distinguishing between human and computer generated questions was considered difficult. 44% of the computer generated questions were wrongly identified as supervisor generated, while 41% of the supervisor questions were wrongly identified as being from the computer system.

The questions generated by the proposed AQG system were considered to be pedagogically as useful as supervisor-generated questions, and more pedagogically useful than generic questions. This shows that the system was successful in identifying important key concepts in the literature reviews, and providing meaningful questions about those key concepts. Also, the computer-generated questions were rated as relatively less clear (QM1 and QM2). Part of the reason might be that some of the phrase lists extracted were not informative enough.

To understand the differences in evaluation in different groups of people, participants were divided into two groups the first and second year students occupying the first group and the third and fourth year students composing the second group. It was found that the first group thought that the computer generated questions were more useful for learning and understanding new concepts. This shows that Wikipedia knowledge base is more useful to students who are new to a research area.

The usefulness of the five relation types was also evaluated. It was found that the five relation types were considered to be useful for learning important concepts. The Has-Strength and Has-Limitation questions were more useful because these question types addressed critical analysis issues on literature review writing.

5.4 Limitations and Discussion

One of the major limitations of this concept-based automatic question generator is its domain dependency. Only a limited no. of concepts are used to generate questions. These concepts are quite frequent in fields related to Science, but infrequent in other academic fields. This makes it less adaptable across different domains. More concept types could be explored and integrated to improve the adaptability.

Moreover, the usage of Wikipedia section headings to identify the target sentences for conceptual graph structures is not very optimal. The section headings which do not have similar terms as key phrases are skipped. More efficient approach could be used here, for example using some natural language analysis to find paraphrasing in section headings. A good tool to detect synonymy or paraphrase could lead to better conceptual graph constructions. Also, instead of rule based search, an approximate search could be integrated to create a more exhaustive concept-graph structure.

6 Conclusions, Analysis and Discussion

The four research ideas surveyed here span different approaches for automatic question generation for literature review writing support.

The first approach discusses automatic question generation through citation extraction. One of the major limitations of the system is the usage of only citations to generate content specific questions. In the real world scenario, lecturers and tutors often generate questions that are not related to citations. Moreover, only rule based approaches have been used for citation classification. This makes the system less scalable and accurate. Also, the no. of participants in the study were limited. Despite these limitations, the study showed that the content specific questions were considered to be more useful than generic questions. Also, despite some semantic and syntactic incorrectness of the questions, the participants found it difficult to distinguish between the questions generated by the proposed system vs. the ones generated by humans. The questions have not been ranked. If a ranking scheme was employed, for example, weighing questions based on location of the corresponding citation in the paper, a better subset of questions could be obtained.

The second approach makes one small change to the previous approach and incorporates machine learning approaches for citation category classification. With otherwise same limitations as earlier, similar results were obtained. However, the accuracy of citation category classification improved over the previous result, and the questions generated were comparatively well-formed. In order to improve the performance over the previous step significantly, one major change is required: Using information apart from only those presented by citations.

In order to make the content specific questions independent of citations in the text, a new framework was proposed. As a basis of this framework, it was found that concept, causal and procedural questions are more frequent than judgmental and verification questions. Based on these question types, the levels in text (lexical, discourse, background knowledge) were identified from which most of the questions are generated by tutors and supervisors. A framework was proposed to extract key concepts from sentences to generate the frequent questions using lexical level, which spans 56.8% of all questions. Later, the framework proposed to rank the questions before using them as feedback questions. However, this is only a framework and not a developed system. Hence, it does not address any actual limitations in the execution of the recommended steps. Overall the framework is quite promising if the performance of such a system is acceptable within limitations.

The final approach that has been discussed is based on the framework that has been suggested in the previous step. Wikipedia is used as a lexical resource to identify relations between key concepts and other concepts in the text. This approach is quite promising, because it explores the key concepts in a literature review and a set of related sentences to generate questions. The conducted study showed that the questions generated were considered to be pedagogically as useful as the questions by supervisors. These questions were found to be much more useful than the generic questions. Also, the first and second year student who do not have sufficient knowledge about the concepts discussed found the questions to be more useful than the older students who knew more about the topic. This observation is quite interesting and makes one wonder about the applicability about the Wikipedia knowledge sources across different target groups. There are certain limitations associated with this system as well. One of the major limitations of this approach is the procedure of generating conceptual graph structure. The system tries to find the relevant sentences by evaluating the presence of key concept terms in the Wikipedia section headings. This disregards all the phrases that do not contain the cue phrases in their section headings. This approach could be improved to a more intensive search to find relevant sentences. Moreover, algorithms for category classification for sentences that match multiple categories should be improved for better accuracy instead of random selection. Also, the concept categories should be expanded to include the categorization for literature reviews across multiple fields.

The research ideas discussed are by no means exhaustive or perfect. They provide only limited support in literature review writing activities. For efficient support, a three step guidance mechanism could be employed:

1. Support prior to writing a literature review - insight on structure of review, content and mechanisms to employ to write an effective review.

2. Feedback about the content immediately after writing a review.

3. Feedback about updates in review after the previous feedback. An insight about the utility of previous feedback can be obtained.

All the approaches discussed are pipeline approaches. Poor performance by any single component in the pipeline can reduce the performance of the entire system. Special care needs to be taken to train corresponding classifiers in similar domains as the test data set. However, tremendous amount of research ongoing in the field of computational linguistics promises better performance of individual system components. It creates a huge potential for such applications and promises a lot of improvement in the upcoming years. Better accuracy is expected to be attained to be able to substitute human-centered feedback mechanism.

References

Davey, B. and McBride, S. (1986). Effects of question-generation training on reading comprehension. *Journal of Educational Psychology*, 78(4):256.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422.

Gavelek, J. R. and Raphael, T. E. (1996). Changing talk about text: New roles for teachers and students. *Language Arts*, pages 182–192.

Graesser, A. C. and Person, N. K. (1994). Question asking during tutoring. *American educational research journal*, 31(1):104–137.

Levy, R. and Andrew, G. (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. Citeseer.

Liu, M. and Calvo, R. A. (2011). Question taxonomy and implications for automatic question generation. In *Artificial Intelligence in Education*, pages 504–506. Springer.

Liu, M., Calvo, R. A., Aditomo, A., and Pizzato, L. A. (2012a). Using wikipedia and conceptual graph structures to generate questions for academic writing support. *Learning Technologies, IEEE Transactions on*, 5(3):251–263.

Liu, M., Calvo, R. A., and Rus, V. (2010). Automatic question generation for literature review writing support. In *Intelligent Tutoring Systems*, pages 45–54. Springer.

Liu, M., Calvo, R. A., and Rus, V. (2012b). G-asks: An intelligent automatic question generation system for academic writing support. *Dialogue & Discourse*, 3(2):101–124.

Osiński, S., Stefanowski, J., and Weiss, D. (2004). Lingo: Search results clustering algorithm based on singular value decomposition. In *Intelligent information processing and web mining*, pages 359–368. Springer.

Powley, B. and Dale, R. (2007). Evidence-based information extraction for high accuracy citation and author name identification. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 618–632. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

Reynolds, T. H. and 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.

Singer, H. and Donlan, D. (1982). Active comprehension: Problem-solving schema with question generation for comprehension of complex short stories. *Reading Research Quarterly*, pages 166–186.

Steward, B. (2004). Writing a literature review. *The British Journal of Occupational Therapy*, 67(11):495–500.