# Automatic Question Generation for Literature Review Writing Support

**Conference Paper** · June 2010

DOI: 10.1007/978-3-642-13388-6_9 · Source: DBLP



**3 authors**, including:

Rafael A Calvo
Imperial College London
**317** PUBLICATIONS **5,054** CITATIONS

SEE PROFILE

# Automatic Question Generation for Literature Review Writing Support

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

[1] University of Sydney, Sydney NSW 2006, Australia
[2] University of Sydney, Sydney NSW 2006, Australia
[3] University of Memphis, Memphis TN 38152, USA

**Abstract.** This paper presents a novel Automatic Question Generation (AQG) approach that generates trigger questions as a form of support for students' learning through writing. The approach first automatically extracts citations from students' compositions together with key content elements. Next, the citations are classified using a rule-based approach and questions are generated based on a set of templates and the content elements. A pilot study using the Bystander Turing Test investigated differences in writers' perception between questions generated by our AQG system and humans (Human Tutor, Lecturer, or Generic Question). It is found that the human evaluators have moderate difficulties distinguishing questions generated by the proposed system from those produced by human (F-score=0.43). Moreover, further results show that our system significantly outscores Generic Question on overall quality measures.

**Key words:** Automatic Question Generation, Natural Language Processing, Academic Writing Support

## 1 Introduction

Many studies have shown that most learners have problems recognizing their own knowledge deficits and ask very few questions [1]. Questions are useful to recognize learners's knowledge deficits and improve their learning. When students are asked to prepare a literature review or write an essay, it is often not only to develop disciplinary communication skills but to learn and reason from multiple documents, a skill often called *sourcing* (i.e., citing sources as evidences to support their arguments) and *information integration* (i.e., presenting the evidences in a cohesive and persuasive way).

Simple generic questions are often provided for students to trigger reflection, for example:

- *Have you clearly identified the contributions of the literature reviewed?*
- *Have you identified the research methods used in the literature reviewed?*

Reynolds and Bonk [2] showed that a group of students given generic trigger questions performs better than those students who receive no trigger questions in a writing activity. However, such questions are too general and not likely

to provide strong support in the process of writing on a specific topic. More content-related questions need to be asked and most academics would ask such questions in the process of providing feedback to students.

In the field of Automatic Question Generation (AQG), most of AQG systems [3–5] focus on the text-to-question task, where a set of content-related questions are generated based on a given text. Usually, the answers to the generated questions are contained in the text. For example, Heilman and Smith [4] presented an AQG system to generate factual questions with an 'overgenerating and ranking' strategy based on natural language processing techniques, such as Name Entity Recognizer and Wh-movement Rules, and a statistical ranking component for scoring questions based on features. The target applications of such systems are reading comprehension and vocabulary assessment which may not be suitable for academic writing.

The aim of this study is to scaffold students' reflection on their academic writing with content-related trigger questions which are automatically generated from citations using Natural Language Processing techniques. Table 1 shows examples of generated questions according to the citation category.

Table 1: An Example of Content-Related Trigger Questions produced by AQG system

| Category | Question |
|---|---|
| Opinion | *Why did Cannon challenge this view mentioning that physiological changes were not sufficient to discriminate emotions? (What evidence is provided by Cannon to prove the opinion?) Does any other scholar agree or disagree with Cannon?* |
| Result | *Does Davis objectively show that this classification accuracy gets higher from about 70 % up to 98 % while actors express emotions and computers perform the...? (How accurate and valid are the measurements?) How does it relate to your research question?* |
| System | *In the study of Macdonald, why does workbench tool provide feedback on spelling, style and diction by analyzing English prose and suggesting possible improvements? What are the strength and limitations of the system? Does it relate to your research question?* |

The remainder of the paper is organized as follows: section 2 provides a brief review of the literature focusing on writing support systems and several AQG systems relevant to our approach. Section 3 describes the system design and architecture while section 4 details a pilot study we conducted to assess the quality of the generated questions. Section 5 discusses the results we obtained and gives suggestions on future work.

## 2 Related Area

Research into ways of supporting academic writing includes Sourcer's Apprentice Intelligent Feedback mechanism (SAIF) [6], a computer assisted essay writing tool used to detect plagiarism, uncited quotations, lack of citations, and limited

content integration problems using a rule-based approach and Latent Semantic Analysis (LSA.

Glosser [7], an automated feedback system for students' writing, provides feedback on four aspects of the writing: structure, coherence, topics, and concept visualization. Glosser uses text mining and computational linguistics algorithms that quantify features of the text (supportive content) and a set of trigger questions. The set of trigger questions in Glosser is limited as they must be predefined for each course and they are too general.

AUTOQUEST [3], one of the earliest automatic question generation systems, uses pure syntactic pattern-matching approach to generate content-related questions in order to improve the independent study of any textual material. Recent advances in Natural Language Processing made it possible for more advanced computational question generation models to be proposed: multi-choice question generation [8], factual question generator [9, 4], and medical concept question generator [5]. One of the most relevant works to ours is by Kunichika et. al. [10] who proposed an AQG approach based on both the syntactic and semantic information extracted from the original text. Their approach is based on DCG (Definite Clause Grammar) for grammar and reading comprehension assessment about a story. The extracted syntactic features include subject, predicate verb, modal verb, auxiliary verb, object, voice, tense which were used to transform declarative sentences into interrogative sentences (subject-auxiliary-Inversion). They used three predefined grammatical categories: noun, verb, and preposition to determine the interrogative pronoun for the question. Their empirical results showed that 80% of questions were considered as appropriate for novices to learn English and 93% questions are semantically correct.

## 3 System Design and Architecture

In this section we provide an overview of the system's pipeline architecture shown in Figure 1 and describe each step in detail. The input to the system is a literature review paper and the output is a set of generated questions. The question

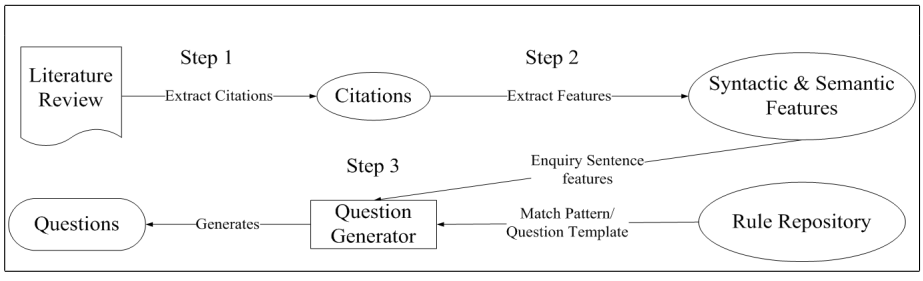


Fig. 1: System Architecture

generation process follows 3 steps shown in Figure 1:

*Step 1. Pre-processing.* The aim of Step 1 is to extract citations from papers. Powley and Dale [11] define 5 types of citation styles: Textual Syntactic, Textual Parenthetical, Prosaic, Pronominal, and Numbered.

A pattern matching technique was used to extract Textual Syntactic and Textual Parenthetical citation style. The regular expression code is shown below.

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

A state of art Named Entity Tagger (NER), LBJ [12], was used to identify citations with Prosaic style and a simple Pronoun Resolver, finding the nearest Name Entity appearing before the pronoun, was used to identify citations with Pronominal style. In the current implementation the Numbered citation style (as in this paper) is not recognized.

*Step 2. Extracting Syntactic and Semantic features.* Syntactic features include subject, predicate verb, auxiliary verb (e.g. be, am, will, have and can) and predicate, voice and tense which are essential to perform subject-auxiliary inversion. We use Tregex on the Phrase Structure Tree derived from the original citation to extract syntactic features. The Stanford Parser is used to parse a sentence into a Phrase Structure Tree. Tregex 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 (NP) or Verb Phrase (VP). The following Tregex expressions are used to extract simple Subject, Predicate Verb, and Predicate from a sentence.

```
Subject: NP > (S > ROOT)   Predicate Verb: /^VB/ > ( VP > ( S >ROOT))
Predicate : VP > (S > ROOT)
```

According to the predicate verb or auxiliary verb we can determine the tense of the sentence and get the verb lemma by using WordNet. We also use the Stanford Parser to derive the Type Dependency relations from a sentence in order to detect the voice of sentences. For example, the nsubjpass dependency between the governor (predicate verb) and dependency (subject) indicates passive voice.

The semantic features include the name of the author and the citation category (one of 'Opinion', 'Result', 'Aim of Study', 'System' or 'Method'), based on a taxonomy of conceptual citation categories proposed by Lehnert et al [13]. For example, Result: a result is claimed in the referenced paper; e.g. *"In [Cohen 87], it is shown that PAs are Turing-equivalent..."*

We use the LBJ NER Tagger to detect authors' names and a rule-based approach to classify the citations. There are many learning materials for academic writing [14] which define three categories of reporting verbs: opinion, aim of study and result. Such reporting verb lists are used in our system to determine the corresponding citation category by matching the predicate verb in a citation with a verb in one of the categories. The matching verb category provides the citation category. If they are no match, a sentiment analysis step is used to detect whether the citation may fall in the Opinion citation category. SENTIWORDNET [15] is used to determine whether the citation contains sentiment words. Tregex expression patterns were developed to detect citations in the System and Method categories. Examples of two Tregex expression patterns are shown below:

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

According to Hyland's citation study [16], there are three main grammatical ways to refer to sources: using reporting verbs, using nouns, and using passive constructions. Sometimes, syntactic structure transformations were needed in order to perform the subject-auxiliary inversion in our final stage. For example, *Wallraff's opinion is that there is a rate of growth...* The citer use the noun:*opinion* to refer to the resource as the citee's opinion. This sentence will be transformed into: *Wallraff states that there is a rate of growth...*

*Step 3. Generation* This is the final step in generating questions with our template-based approach. Once the semantic and syntactic features extracted from a citation match the predefined patterns in our repository of templates the corresponding questions are generated. Table 2 shows the five rules defined in our Rule Repository. Rules 1, 2, or 3, are fired when a citation contains a reporting verb and and fall in one of the following citation categories: Opinion, Result, or Aim of Study, respectively. Rules 4 or 5 are fired when a citation is of type System or Method. We also defined two addition rules, 6 and 7. Rule 6 is fired when a citation does not contain a reporting verb but contains sentiment words. Rule 7 is similar to Rule 6 except the citation does not contain a sentiment word. For example, a citation is extracted in Step 1: *Cannon (Cannon 1927) challenged this view mentioning that, physiological changes were not sufficient to discriminate emotions.* Step 2 identifies the citation category as Opinion by matching the predicate verb (*challenge*) with an entry in our reporting verb database. Step 3 applies Rule 1 to generate a question by matching the pattern that requires the citation contain a reporting verb and of of type Opinion. Table 1 shows the generated questions.

Table 2: The Rule Definition for Patterns and Templates

| Rule | Pattern | Category | Question Template |
|---|---|---|---|
| 1 | Reporting Verb | Opinion | Why +subject_auxiliary_inversion()? What evidence is provided by +subject+ to prove the opinion? Does any other scholar agree or disagree with +subject+ ? |
| 2 | Reporting Verb | 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 | Reporting Verb | 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 | Tregex Rules | Method | In the study of +subject+, why +subject_auxiliary_inversion()? Which dataset does +subject+ use for this experiment? Could the problem have been approached more effectively from another perspective? |
| 5 | Tregex Rules | System | In the study of +subject+, why +subject_auxiliary_inversion()? What are the strength and limitations of the system? Does it relate to your research question? |

## 4 Pilot Study

We explored the ability of our AQG system to generate quality questions by comparing automatically generated questions to those produced by humans. Like

the Bystander Turing Test conducted by Person and Graesser [17], our judges were asked to rate each question along several dimesions of quality. Also, we conducted an evaluation in which judges were asked to ascertain whether the question was generated by a human (lecturer, tutor, generic) or a system. The major difference between the test carried out by Person and Graesser and our evaluation is the application context: we focus on questions for academic writing while they used a snippets of tutorial dialog. Also, our judges were the writers of the source content based on which the questions were generated while their judges did not know the content before the experiment. This is an advantage of our methodology because our judges were experts on the content the questions asked about. Section 4.1 describes the participants and procedure we used in the pilot study. Section 4.2 reports the AQG system performance in terms of the semantic correctness of the generated question as well as the accuracy relative to the citation extraction step. Section 4.3 shows the results along 5 dimensions of quality and of the Bystander Turing test. Section 5 discusses these results.

### 4.1 Participants and Procedure

A pilot study was conducted on six participants (postgraduate students) from the Faculty of Engineering from whom six literature reviews were collected. The reviews were used the source content for generating the questions. A total of twenty questions (5 each) were generated by the tutor, by a lecturer with expertise in the topic, by our system, and also using generic questions. Each student-author acted as an evaluator in our experiments.

Students were asked to rate the quality of questions generated from his/her literature review paper. Five quality measures inspired by Heilman and Noah [4] were used to evaluate each question: *This question is correctly written (QM1);This question is clear (QM2);This question is appropriate to the context (QM3); This question makes me reflect about what I have written (QM4);This is a useful question (QM5)*. The agreement with each of these statements was marked by the evaluators using a Likert scale were 1 was 'strongly disagree' and 5 'strongly agree'.

### 4.2 System Performance Evaluation and Result

We first assess our system's ability to extract citations from the source content. The dataset contains 1,088 sentences including 221 citations. Table 3 shows that 145 citations have been extracted and the recall is 0.66 in average.

Table 3: Citation Extraction Result

| | Rule 1 | Rule 2 | Rule 3 | Rule 4 | Rule 5 | Rule 6 | Rule 7 | Total |
|---|---|---|---|---|---|---|---|---|
| Number of Citations | 18 | 22 | 12 | 50 | 16 | 29 | 74 | 221 |
| Number of Retrieved Citations | 10 | 12 | 7 | 27 | 10 | 17 | 62 | 145 |
| Recall | 0.56 | 0.55 | 0.58 | 0.54 | 0.63 | 0.59 | 0.84 | 0.66 |

Table 4 illustrates 161 questions generated and the average semantic correctness: 60%. Two human annotators reached substantial agreement as measured by Cohen's kappa coefficient (0.61).

Table 4: Question Generation Result

| | Rule 1 | Rule 2 | Rule 3 | Rule 4 | Rule 5 | Rule 6 | Rule 7 | Total |
|---|---|---|---|---|---|---|---|---|
| Number of Generated Questions | 10 | 9 | 14 | 7 | 6 | 17 | 98 | 161 |
| Number of Correct Questions | 6 | 9 | 9 | 4 | 4 | 10 | 56 | 97 |
| Precision | 0.6 | 1 | 0.64 | 0.57 | 0.67 | 0.59 | 0.57 | 0.60 |

### 4.3 Question Quality Evaluation and Result

Each of the 20 questions, randomly selected, was evaluated by the student-authors. Because we have six authors, 120 questions were evaluated. A one-way ANOVA setting the confidence interval at 95% was conducted to examine whether there are statistical difference in Overall, QM1, QM2, QM3, QM4 or QM5 among questions generated by the lecturer, Tutor, AQG system and Generic. The ANOVA yielded a significant difference in Overall ($F(3,596)=2.63$, $P<0.05$ ), QM3($F(3, 116)=4.085$, $P<0.05$ ), QM4($F(3, 116)=8.65$, $P<0.05$, QM5($F(3, 116)=5.305$, $P<0.05$) and no significant difference in QM1($F(3,116)=2.69$, $P>0.05$) and QM2 ($F(3,116)= 2.335$, $P>0.05$). Follow-up Fishers least significant difference (LSD) tests with 95% confidence interval were performed to determine whether significant differences occurred between the mean scores for each pair of treatments. Figure 2 illustrates the comparisons of mean scores and Table 5 shows that the questions from AQG system significantly outscored Generic Questions in Overall (0.346>LSD=0.283) and QM5 (0.733>LSD=0.633), while questions from the tutor significantly outscored AQG system in QM3 (0.667>LSD=0.593), QM4 (1>LSD=0.648) and Overall (0.6>LSD=0.283). There are no statistically significant differences between questions generated by the lecturer and AQG system. Also, we did not observe any significant differences between the Tutor and AQG system in QM 5 (0.533<LSD=0.633).

The quality of each rule was also evaluated. Fig. 3 shows the average scores. Rule 5 got the highest score (4.3), Rule 4 and Rule 6 took the second place (3.9) and Rule 7 reached the lowest score (3.0). It was also found that Rules 1, 2, 3, 4 and 7 decreased from Quality Measure 4, 5 to Quality Measure 1,2,3 while Rule 5 was stable in along all five quality measures.

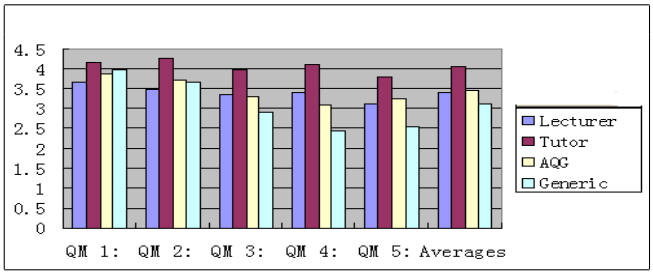


Fig. 2: Comparisons of normalized mean scores

Each evaluator was asked to ascertain who wrote this question: Lecturer, Tutor, System or other. In order to clearly evaluate the participants' classification ability between a Human and a System, we did not take the Generic Question into consideration. Therefore, only 15 questions evaluated by a participant were considered. We use the balanced F-score to evaluate the classification result and

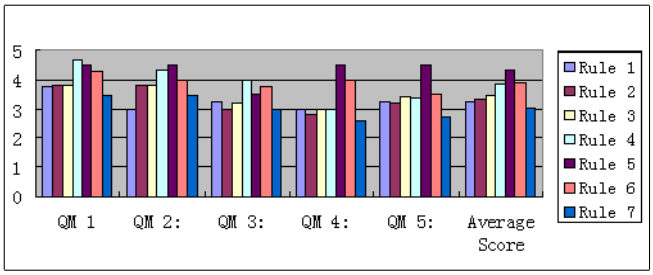


Fig. 3: Comparisons of scores for each rule

Table 5: Fisher's least significant difference (LSD) tests with 95% confidence interval

| | Lecturer v. AQG | Tutor v. AQG | AQG v.GQ |
|---|---|---|---|
| QM3(LSD=0.593) | 0.067 | 0.667 | 0.367 |
| QM4(LSD=0.648) | 0.333 | 1 | 0.633 |
| QM5(LSD=0.633) | 0.133 | 0.533 | 0.733 |
| Overall(LSD=0.283) | 0.047 | 0.6 | 0.346 |

F-score is defined as follows: F-score=2*Precision*Recall/(Precision+Recall). Table 6 shows the participants' average performance on the classification, which found that they achieved F-score of 0.43 on AQG system, F-score of 0.24 on Lecturer and F-score of 0.18 on Tutor category.

Table 6: Confusion Matrix (Average)

| Response / Real | Tutor | Lecturer | AQG System |
|---|---|---|---|
| Tutor | 0.7 | 2.7 | 1.6 |
| Lecturer | 0.8 | 1.2 | 3 |
| AQG | 1.4 | 1.0 | 2.6 |

## 5 Discussion

This paper presents a novel Automatic Question Generation approach to support literature review writing and also describes a pilot study evaluating the system performance along several dimensions—the Citation Extraction Ability and Semantic Correctness of the generated questions and Question Quality—and comparing it with humans and generic questions.

The study has a few limitations including a relatively small number of subjects (6) and questions (120). Furthermore, it only evaluates a set of very specific types of questions that refer to only one aspect (citations) out of the many involved in literature review. In a real teaching scenario, the human assessors (tutor and lecturer) would prepare questions on other issues besides the citations. For the future, we plan on having pedagogical experts involved to help with the question formulation as well as with the evaluation. Despite these shortcomings, we believe that the dataset is large enough and the evaluation meaningful because we use real academic writings, i.e. student-written literature review papers, as

our dataset and the evaluators have higher education background and are very familiar with the source content, as being the authors of the review papers used, and thus being in a better position than others to judge quality of questions.

Within these limitations, this pilot study suggests that the AQG system can produce questions that are as helpful to promote students' reflection on their academic writing as those by human tutors. The most significant finding from this pilot study was that writers found it moderately difficult to distinguish between questions generated by humans and automatically generated questions. This claim is supported by the fact that students perceive approximately as much value in automatically generated questions as in those written by the lecturer.

As we had expected, the AQG system outscored Generic Questions because the content-related questions were more helpful than the generic questions. Surprisingly, we found that our AQG system slightly outperformed the Lecturer, which may be explained by some factors. First, students may intend to give higher scores to a Lecturer. Second, it took a lot of effort for a lecturer to create 30 questions in total for six literature review papers across different topics. This might affect the lecturer's performance on creating pertinent questions. Finally, the length of template-questions, longer on average than questions generated by the lecturer, may affect the evaluation.

There are two main reasons for generating incorrect semantic questions(40% inaccuracy):1 The NER component and 2: Citation Category Classifier. Because the LBJ NER tagger was primarily trained on News Text Corpora it might affect its the performance on academic articles. Our current Citation Category Classifier is based on a rule-based approach which is simple but not scalable. As we can see from Figure 3 and Table 4, we may need to add extra patterns to Rule 5 to generate more questions while also improving the question templates in Rules 1, 2, 3 and 4 in order to achieve higher scores on Quality Measures 4 and 5. In addition, more citation categories might be explored which could improve the performance for Rule 6 and Rule 7.

Future work will focus on ranking the generated questions, combining a Machine Learning approach with a rule-based approach to improve the citation category classification, training the LBJ NER tagger on a large collection of academic papers, and upgrading the taxonomy of citation category in our system. It is also planned to integrate the AQG system into our peer review system which will be used for students in Research Method course next semester.

## 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 a University of Sydney TIES grant and Australian Research Council Discovery Project DP0986873.

## References


1. Graesser, A.C., Person, N.K.: Question asking during tutoring. American Educational Research Journal **31** (1994) 104–137
2. Reynolds, T., Bonk, C.: Computerized prompting partners and keystroke recording devices: Two macro driven writing tools. Educational Technology Research and Development **44(3)** (1996) 83–97
3. Wolfe, J.H.: Automatic question generation from text - an aid to independent study. SIGCUE Outlook **10**(SI) (1976) 104–112
4. Heilman, M., Noah, S.A.: Question generation via overgenerating transformations and ranking, The 14th Annual Conference on Artificial Intelligence in Education Workshop on Question Generation (2009)
5. Wang, W.M., Hao, T.Y., Liu, W.Y.: Automatic question generation for learning evaluation in medicine. in LNCS **4823** (2008) 242–251
6. Britt, M.A., Wiemer-Hastings, P., Larson, A.A., Perfetti, C.A.: Using intelligent feedback to improve sourcing and integration in students' essays. Int. J. Artif. Intell. Ed. **14**(3,4) (2004) 359–374
7. Villalon, J., Kearney, P., Calvo, R.A., Reimann, P.: Glosser: Enhanced feedback for student writing tasks. In: Proc. Eighth IEEE International Conference on Advanced Learning Technologies ICALT '08. (July 1–5, 2008) 454–458
8. Mitkov, R., Ha, L.A.: Computer-aided generation of multiple-choice tests. In: Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing, Morristown, NJ, USA, Association for Computational Linguistics (2003) 17–22
9. Rus, V., Cai, Z.Q., Graesser, A.C.: Experiments on generating questions about facts. In: CICLing '07: Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing, Berlin, Heidelberg, Springer-Verlag (2007) 444–455
10. Kunichika, H., Katayama, T., Hirashima, T., Takeuchi, A.: Automated question generation methods for intelligent english learning systems and its evaluation, Proc. of ICCE01 (2001) 1117–1124.
11. Powley, B., Dale, R.: Evidence-based information extraction for high-accuracy citation extraction and author name recognition. In: Proceedings of the 8th RIAO International Conference, Pittsburgh, PA, (2007)
12. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: CoNLL 2009. (2009)
13. Lehnert, W., Cardie, C., Riloff, E.: Analyzing research papers using citation sentences, Proceedings of the Twelfth Annual Conference of the Cognitive Science Society (1990) 511–518
14. University of Glasgow: Writing an assignment:reporting verbs. Technical report, University of Glasgow (2009)
15. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06. (2006) 417–422
16. Hyland, K.: Academic attribution: citation and the construction of disciplinary knowledge. Applied Linguistics **20** (1994) 341–367
17. Person, N.K., Graesser, A.C.: Human or computer? autotutor in a bystander turing test. In: ITS '02: Proceedings of the 6th International Conference on Intelligent Tutoring Systems, London, UK, Springer-Verlag (2002) 821–830