Eighth IEEE International Conference on Advanced Learning Technologies

Glosser: Enhanced Feedback for Student Writing Tasks

Jorge Villalón[(1)], Paul Kearney[(2)], Rafael A. Calvo[(1)], Peter Reimann[(2)]
University of Sydney, (1)School of Elec. & Inf. Eng., (2)Fac. of Education and Social Work
{villalon, rafa}@ee.usyd.edu.au, {p.kearney, preimann}@edfac.usyd.edu.au

Abstract

We describe Glosser, a system that supports students in writing essays by 1) scaffolding their reflection with trigger questions, and 2) using text mining techniques to provide content clues that can help answer those questions.

A comparison with other computer generated feedback and scorings systems is provided to explain the novelty of the approach. We evaluate the system with Wiki pages produced by postgraduate students as part of their assessment.

1. Introduction

The essay as a student learning activity is common across disciplines and levels in higher education. Indeed, writing is important to almost all knowledge work, and the skills learnt through essay-writing are easily transferable to those used in any knowledge-rich environment. The teaching of writing therefore has become an important part of the curricula at modern universities [1]. In essays, students are expected to show evidence of mastery of specific skills (such as spelling and grammar) as well as of higher level thinking – analysis, argument, and independent thought. To write well, students must involve themselves in reflective thinking about their own work and their own, complex, writing process. However, without the proper support, students have difficulty engaging in high-level reflective thinking [2].

The use of computers to assist the teaching of writing is not new, and tools have been created that support different stages of the writing process. Some software tools generate feedback that is delivered to the author. Other tools that perform Automatic Essay Scoring (AES) have focused on assessment. They are used mainly to overcome time, cost and reliability issues in writing assessment [3]. However this cost-centered design leaves the process of writing out of their scope, focusing only on the product. Most of these systems offer little to no feedback to the student.

According to Daiute [4], the writing process can be divided into three sub-processes: pre-writing, drafting and revising. These sub-processes are not followed in sequential order and occur uniquely for each individual. Providing feedback to students during the process of writing is crucial to the learning process. The system described here is based on the idea that feedback is provided by a reader to a writer, so that the writer can use it for revision. Furthermore, this feedback should support the author's reflective thinking.

In an analysis of technologies that support reflection, Lin introduces the idea of reflective process prompts, and argues that a meaningful prompt to scaffold reflection must make the learner's thinking explicit [5].

Text Mining is an area of artificial intelligence that aims to discover new facts and trends (knowledge) from large collections of text. It combines techniques from areas such as computational linguistics, information retrieval and data mining. It has successfully supported applications such as summarization, question answering and classification, topic detection, among others. These techniques produce valuable information about documents that we argue, can be used as a way to scaffold student reflection.

In this paper we describe "Glosser", a system designed to provide support for the teaching and learning of academic writing in English. The system provides trigger questions, that can be customized to genre-specific goals, and provides feedback content, which we call 'gloss', which is of a non-genre-specific nature, and that students can use in order to more effectively reflect on the questions, and analyze their work and writing process. Ideally, the result would be a) a higher quality outcome and b) an enhanced learning experience. As a basis for our scaffolding we use the MASUS taxonomy [6], created as a pre-test for academic skills in writing. This taxonomy has been

978-0-7695-3167-0/08 $25.00 © 2008 IEEE
DOI 10.1109/ICALT.2008.78

widely used in a number of disciplines on more than 7,000 students.

Section 2 of this paper describes previous work on providing computer generated feedback on student writing. This work provides a theoretical framework and evidence that automatic feedback can improve students' learning. This section also highlights the ways in which our approach is a novel one. Section 3 reviews the text mining techniques used in this project, while Section 4 describes the tool itself and proposes ways in which it can be used to support reflection and reviewing. Section 5 describes an evaluation using wiki pages produced by students enrolled in a postgraduate course. Section 6 concludes.

2. Previous work

Ware and Warschauer [7] review electronic feedback systems for second language writing. They highlight how 'electronic' feedback can be interpreted in several ways, all highly dependent on the approach used to teach writing. They distinguish between writing as the mastery of a compendium of sub skills and writing as a social practice. The former considers electronic feedback as the automated feedback provided by the computer. For the latter, electronic indicates the means by which human feedback is provided. Several existing systems such as wikis are providing support for writing as a social practice, but only AES Systems are supporting the skills of writing. Glosser's feedback, described later, conforms to both interpretations of 'electronic'.

Several studies reported a high agreement between AES systems and human raters [3]. These systems work by creating a model of a good essay that is trained with pre-scored essays on a certain topic, and then assessing new essays by comparing them to the "model" essay. The number of essays required depends on the system, but it's never less than a hundred. Once the model is created it is used with unseen essays to provide a nominal score (a 'category' in technical terms). Several features, both linguistic and statistical, are extracted from the essay, and then used as a way to categorize the essay based on their values. The training phase is used to minimize the errors by adjusting the model in what is called a 'supervised' approach.

Our approach is different in that it doesn't attempt to classify the essay into assessment categories. We use the same techniques to extract features from the essays, but we don't use them to train a model; instead, the system uses these features to highlight factors that might affect the quality of the work, and leave it to the learner to reflect on how those features are actually related to the writing issues they have been asked to address. We use the computer as a reader that presents important information extracted from the essay. An advantage of our approach is that we don't need a training set of pre-scored essays, making the system genre-independent and easier to adapt by teachers.

Analyzing the feedback provided by previous systems, a recent review by Dikli [3] found that most of them provide a holistic score for students' essays and sometimes a score for specific features such as organization or sentence fluency, however they provide very poor feedback or none at all.

An early 1996 study by Reynolds and Bonk [8] used generic messages to support the revision activity during writing: the messages appeared as two lines in the bottom of the screen while the author was writing the essay on a word processor. Even though the feedback was theoretically and technologically simplistic, the authors found that the automatic feedback encouraged the students to engage in revision and to make more meaningful changes on their work. Arguably, a generic message alone provides very little information about the learner's knowledge but it might trigger reflection on important issues of their writing. To be of more use to students, it could be supported with evidence extracted from the learner's essay.

Another study by Kakkonen et al. [9] argued that fully automated scoring systems are based on the out-dated educational philosophy of behaviorism. They argued that these systems promote an idea of writing that encourages simplistic second-guessing of the machine, disempowers the student-author, and renders writing tasks "inauthentic". They suggested, instead, the use of Text Mining outputs such as summaries and plagiarism detection to provide more meaningful feedback, but they did not implement the idea.

Recent work by Britt et al. [10] used Text Mining techniques to provide feedback on sourcing and integration for student essays. They detected citations and plagiarized sentences and the program would suggest ways to remedy them. Their approach is similar to ours in that it focuses on the detail of the essay, rather than the whole, but has a narrower focus - citations and plagiarism. We consider the work of Britt and Wiemer-Hastings as work that point in the direction that we have followed in this project.

3. Textual data mining techniques

Text Mining (TM) techniques have been applied to a number of learning and teaching domains: from plagiarism detection, and automatic assessment to question-answering systems [11]. A number of text

mining tools are designed to be used across a wide range of application and others, such as Glosser, are specialized for one particular goal (e.g. automatic feedback). Some TM techniques are based on linguistic approaches and others such as Glosser also use statistics and machine learning.

At the core of our system is the Latent Semantic Analysis technique, created by Deerwester et al. and described in [11]. LSA uses the vector space model representation of a document. A term by text passage matrix is created and then a Singular Value Decomposition (SVD) technique is applied to it, obtaining semantic information about the text. The lesser singular values (eigenvalues) of the decomposition are then discarded (they are considered noise) and the text passages are projected onto the reduced space (called semantic space). Finally distances between terms or text passages can be calculated using the angle between their vectors.

LSA is a powerful technique that has been used primarily for indexing in Latent Semantic Indexing (LSI), however it provides semantic information that can be exploited. In the decomposition, the eigenvalues are sorted by the amount of variance they explain, and each eigenvector corresponds to a topic within a document, starting from the most important topic to the less important. Importance here is basically the extent of the coverage of that topic by the author. Gong and Liu [12] used this idea to extract key sentences from documents. We use the same idea to extract key sentences from a student essay.

Another project by Osinski [13], followed the same approach, he created a semantic space with the results of a search engine. Then he used the topics obtained with the decomposition to obtain key phrases, finally he used the semantic distance to cluster the documents around the topics. They implemented their idea into the Lingo algorithm, which we used in our tool, using sentences rather than web pages. We created a set of topic clusters formed by the sentences that talk about it.

The LSA's semantic distance has been used to measure coherence between paragraphs by Foltz et al. [14] basically it measures the amount of common and semantically related words between the paragraphs, with this rough shifts can be found and presented to the student for analysis.

There are a wide variety of systems used by teachers and students engaged in essay writing. Some are desktop applications (e.g. MS Word) others are web-based. In Glosser, we used existing technologies that could help us both speed the development and then ease the integration of the tool with other software, especially Learning Management Systems. All the basic processing (tokenizing, stemming and removing of stop words) is performed with the Apache's Lucene indexing software, a popular indexing software that is already integrated in the Sakai and Moodle open source LMS. In Glosser, each document is first processed and inserted into an Apache's Lucene index, at insertion time the Porter's stemmer is used and stop words are removed. The document is then split in paragraphs and sentences, and each is inserted into the Lucene index as well. Two semantic spaces are then created for each document: a paragraph based and a sentence based space. On the space creation, term weighting and dimensionality reduction are applied. Finally, several operations are performed on the spaces: key sentences and the last paragraph are extracted, and topic clusters are calculated.

4. Triggering reflection with Glosser

The goals of Glosser can be separated into two categories: 1) to trigger reflection on writing quality and 2) to support reflection on the learner's writing process. The former and more traditional goal, which focuses on helping students reflect on, and improve, the document itself, can be scaffolded by using general use rubrics such as the MASUS Criteria, described in [6]. These criteria have been used in a number of educational settings and to support students in a number of disciplines writing text in different genres. The rubric includes 5 quality attributes:

- Use of Source Material
- Structure and Development of Answer
- Control of Writing Style
- Grammatical Correctness
- Qualities of Presentation

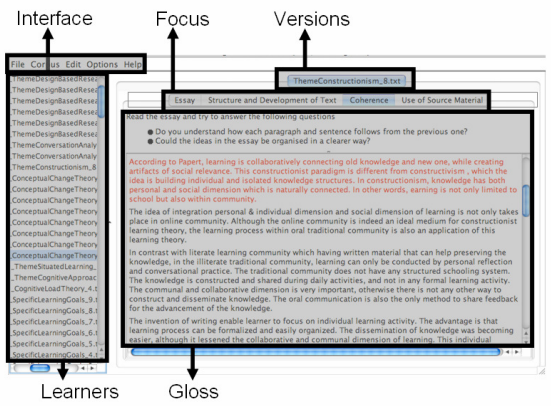

Figure 1: Structure of the Glosser interface

Teachers can adapt the specific criteria in collaboration with a language teaching specialist to their particular discipline. In Glosser we use the first 3

criteria to organize a set of trigger questions that can be customized for a particular learning setting.

The Glosser interface, shown in Figure 1, provides feedback in categories and is structured as a number of tabs that map directly to one of these criteria, and to supportive content that provides different views of the document. Associated with each tab is a number of trigger questions and associated content or gloss. The gloss helps the learner to consider the questions. Its content may be text or images and provides evidence or focus points in the document to assist in answering the questions. Each tab with its associated triggers and gloss, is described in detail below.

4.1. "Essay" tab

This section displays the original text for each version of the document.

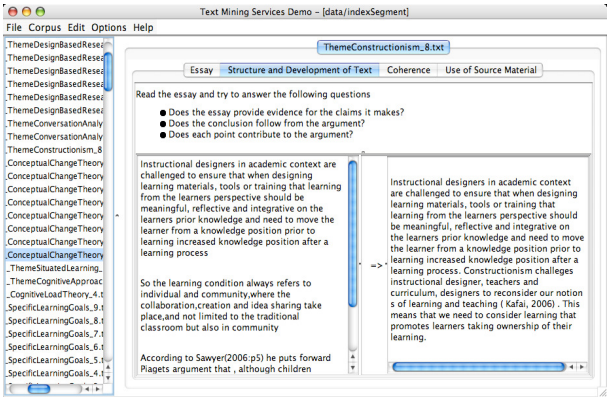

Figure 2: "Structure" section showing trigger questions and supportive content.

4.2. "Structure" tab

The following trigger questions are displayed at the head of this section:

- Does the essay provide evidence for the claims it makes?
- Does the conclusion follow from the argument?
- Does each point contribute to the argument?

Currently, the supportive content for this section (shown in Figure 2) is a number of key sentences on the left hand side, and the concluding remarks (currently the last paragraph before references if they exist) on the right. The goal is to show the student a number of core ideas (as identified by loadings in the LSA space) so the student can evaluate if they provide evidence for their claims.

4.3. "Coherence" tab

The following trigger questions are displayed at the head of this section:

- Do you understand how each paragraph and sentence follows from the previous one?

The supportive content in this section is an identification of pairs of consecutive paragraphs that are too far apart in LSA space, thus indicating possible instances of conceptual incoherence or lack of flow in the text.

The supportive content helps the student to consider a particular point of reflection in a focused and more organized way.

4.4. "Topics" tab

The following trigger questions are displayed at the head of this section:

- Are the ideas used in the essay relevant to the question?
- Are the ideas developed correctly?
- Does this essay simply present the academic references as facts, or does it analyse their importance and critically discuss their usefulness?
- Does this essay simply present ideas or facts, or does it analyze their importance?

In this case, the supportive content is a set of topics that cluster concepts identified by the system as most highly emphasized in the version of the document that is selected.

5. Evaluation

The system was evaluated using a collection of wiki documents written collaboratively by students enrolled in a core subject of the Master of Learning Science and Technology program, at the University of Sydney. Two aspects of the system have to be evaluated and improved: The algorithms accuracy and the impact on student learning.

LSA's modeling accuracy depends only on its own parameters. We followed standard techniques to select them. This evaluation is not discussed here.

An in-depth evaluation on student impact requires that students use the tool as part of their learning activity. This will be reported in future work. The current evaluation consisted in validating the meaningfulness of each gloss with respect to the trigger questions. The wiki pages were loaded and processed by Glosser. Over several iterations, the gloss produced for each essay was analyzed by those who had prepared the trigger questions. The qualitative evaluation looked

at how well the gloss scaffolded the analysis of the quality issues highlighted in a particular tab. As a result of this analysis, several parameters for the algorithms were tuned and some of the questions rewritten.

We found that the structure and development of the text functionality highlighted appropriate sentences regarding the main ideas in the essay. In some occasions the sentence did not contain an explicit idea but an elaboration or discussion of it, this is due to the statistic nature of the algorithm that favors long sentences where several concepts are made explicit. The number of sentences to provide is still an open question that we'll study in the near future with real students acting as reviewers.

While evaluating the coherence functionality we found that the technique is prone to simple errors like recognizing a rough shift between a section's title and its first paragraph. However the exercise of analyzing the machine's output helped to focusing the reader attention to potential problems and coherence itself.

The "use of sources" functionality successfully provided a good list of the main topics addressed in the essay and their corresponding sentences. We found that a feature of the Lingo algorithm, assigning the sentences to more than one cluster, provided more meaningful clusters and a simple way to analyze the argumentation.

6. Conclusions

Tools that help students improve their writing are increasingly important as Universities struggle to develop their communication skills. Helping them reflect is one of the most successful strategies. Text Mining and in particular Latent Semantic Analysis can be used to create such tools. By creating new perspectives of a student essay we focus the attention of the reviewer on particular issues of the writing process and therefore support students' reflection.

We implement a system to validate our approach. We use 1) focus points to bring the reviewer's attention to specific issues, 2) trigger questions to help students reflecting about the issues and 3) supporting 'evidence' that we call gloss, which is extracted by the system automatically and provided in the form of text or images to help the students answering the questions and analyze their work focused on each issue. The system was evaluated with a corpus of Wiki pages created by postgraduate students and the results were analyzed by experts in the field showing promising results.

7. Acknowledgements

The authors would like to thank Brian Paltridge and Marie-Louise Stevenson for their contributions to this project. This project has been funded by the Australian Research Council Discovery Grant DP0665064.

8. References

[1] D. R. Russell, *Writing in the academic disciplines, 1870-1990 : a curricular history*. Carbondale: Southern Illinois University Press, 1991.

[2] L. Flower, J. R. Hayes, L. Carey, K. Schriver, and J. Stratman, "Detection, Diagnosis, and the Strategies of Revision," *College Composition and Communication* vol. 37, pp. 16-55, 1986.

[3] S. Dikli, "An overview of Automated Scoring of Essays," *Journal of Technology, Learning and Assessment,* vol. 5, 2006.

[4] C. Daiute, *Writing and Computers*: Addison-Wesley, 1985.

[5] X. Lin, C. Hmelo, C. Kinzer, and T. Secules, "Designing technology to support reflection," *Educational Technology Research and Development,* vol. 47, pp. 43-62, 1999.

[6] H. Bonanno and J. Jones, "Measuring the Academic Skills of University Students - the MASUS procedure, a diagnostic assessment," University of Sydney, Sydney 1997.

[7] P. Ware and M. Warschauer, "Electronic feedback and second language writing," K. H. a. F. Hyland, Ed. Cambridge: Cambridge University Press, 2006.

[8] T. H. Reynolds and C. J. Bonk, "Facilitating college writers' revisions within a generative-evaluative computerized prompting framework," *Computers and Composition,* vol. 13, pp. 93-108, 1996.

[9] T. Kakkonen, N. Myller, and E. Sutinen, "SemiAutomatic Evaluation Features in Computer-Assisted Essay Assessment," in *7th International Conference on Computers and Advanced Technology in Education*, 2004, pp. 456-461.

[10] M. A. Britt, P. Wiemer-Hastings, A. A. Larson, and C. A. Perfetti, "Using Intelligent Feedback to improve Sourcing and Integration in Students' Essays," *International Journal of Artificial Intelligence in Education,* vol. 14, pp. 359-374, 2004.

[11] M. A. Hearst, "The debate on automated essay grading," *Intelligent Systems and Their Applications,* vol. 15, pp. 22-37, 2000.

[12] Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," *24th Conference on Research and Development in Information Retrieval,* pp. 19-25, 2001.

[13] S. Osinski and D. Weiss, "A Concept-Driven Algorithm for Clustering Search Results," *IEEE Intelligent Systems,* vol. 20, pp. 48-54, 2005.

[14] P. Foltz, W. Kintsch, and T. Landauer, "The measurement of textual coherence with Latent Semantic Analysis," *Discourse Porcesses,* vol. 25, pp. 285-307, 1998.