SEMI-AUTOMATIC EVALUATION FEATURES IN COMPUTER-ASSISTED ESSAY ASSESSMENT

Tuomo Kakkonen, Niko Myller, Erkki Sutinen
Department of Computer Science
University of Joensuu
Finland

Abstract

The role of assessment and evaluation has recently changed in a way which will have fundamental consequences in applying information and communication technologies (ICTs) to analyzing learning outcomes. Rather than helping the teacher only to get her students' final grades from an entirely automated assessment system, the idea is also to support the student to reflect her learning process as early as possible and point out the strong and weak aspects of it. This paradigmatic shift from a teacher-centered assessment towards learner-centric process evaluation offers interesting challenges to educational technologists, especially in the area of essay evaluation. We present a system that evaluates the content of essays based on Latent Semantic Analysis (LSA). The system applies LSA to compare the conceptual similarity between the essays and selected text passages from the course material covering the essay assignment-specific subject matter. In addition to grade, we use LSA to provide writer with more detailed feedback about the covered points in the essay as an introduction to the semi-automatic essay grading.

Key Words

Semi-automatic essay assessment, automated feedback, Computer-assisted learning and instruction

1. Introduction

Computer-assisted assessment refers to the use of computers in assessing students' learning outcomes. To reduce the costs of essay grading, methods to automate the assessment process have been developed. The need for computer-assisted assessment of learning outcomes is two-fold. Teachers need to automate the assessment and evaluation process especially in mass courses. On the other hand, a student wants to get feedback and assess his or her own learning process before an examination. Evaluation is a broad concept which covers both formal and informal feedback, carried out either explicitly or implicitly.

While automating the essay grading process is not a novel idea, it has been utilized only occasionally. Despite the impressive results, automated scoring systems are not widely accepted and used by educators [1]. This might reflect the mixed feelings and unconscious attitudes among teachers. For most of them, essay grading seems an island of their professional pride, worth while defending. At the same time, many teachers struggle with their teaching loads and other duties. In addition, university teachers should devote more time to research, which means decreasing the time used for routine-like tasks which also essay grading might become.

Given teachers' increasing work loads, the reasons for the successful teacher resistance against automatic essay grading must be searched for also in the pedagogical thinking. Although fully automated scoring can be feasible in some occasions, we hypothesized that a system supporting only such a simple form of evaluation is based on the outdated idea of behaviorism. There is a need for utilizing computers not just for the grading, but for giving feedback and supporting the learner. An assessment system reacts to a poor essay by a negative feedback and makes the student to fight, or learn, for a better grade. We concluded that a system with semi-automatic grading features would get a more welcome response from teachers, if it were inherently a part of a constructive learning process.

Instead of grading a submitted essay in a black box, a semi-automatic essay evaluation environment would help a learner while he is authoring an essay by working together with him. It parses the language, compares it to available learning materials, analyzes the style, grammar, vocabulary, structure, and argumentation of the essay, identifies its key sentences, and detects potential plagiarism. The student is all the time aware of the evaluation process and can intervene to it. The semi-automatic approach means also that the system works as cognitive tool that helps the student to progress as an essay author (see Figure 1).

In this paper, we describe an essay assessment system based on a commonly known information retrieval technique, Latent Semantic Analysis (LSA). The grade is computed by using both human-graded essays and assignment-representative text from a textbook. We start

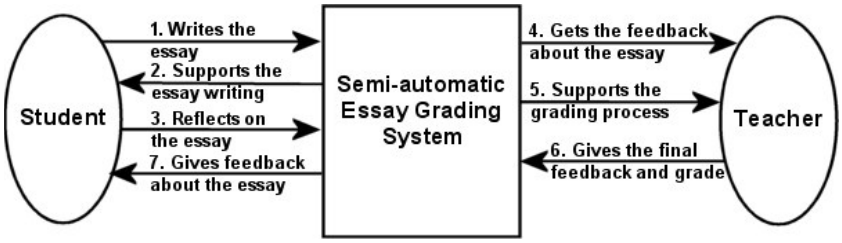

Figure 1: The structure of the semi-automatic system.

by giving a short overview to the approaches developed to computerize the process of grading and giving feedback. In Section 3 we describe our LSA-based grading method and results of our experiments with the system. Section 4 introduces the methods supporting automated feedback and other semi-automatic assessment features and our experiment with LSA-based feedback method. Finally, we conclude the findings and outline some future opportunities.

2. Background

It has been shown in several studies that computers can be successfully used for scoring free-text responses. In the previous decades, several approaches have been applied for automated essay grading. The best known are Project Essay Grade (PEG) [2], e-rater [3] and approaches based on LSA, e.g. [4, 5]. All these methods have had their primary focus on numerical assessment. They take human ratings of essay quality as the basis for creating an assignment-specific scoring model. Concerns has been voiced that machine scoring of essays will guide students to write non-creative, "flat" essays [6]. To overcome such criticism, the research has been directed towards more direct and transparent measures of essay quality. The earliest systems, such as PEG, measured the surface features, such as the length of the essay and the number of commas [6]. More recent approaches, such as e-rater [3], text categorization approach by Larkey [7] and the systems based on LSA, e.g. [8, 9], have focused more on the content of the essays.

The development in the field of natural language processing and information retrieval opens many possibilities for automated essay assessment. One of the main goals on the recent studies has been the shift from the holistic grading towards analytical assessment. The aim is not only to give a single grade representing the quality of an essay, but to analyze the different aspects of the text and provide the writer more detailed feedback and instructions. Shermis et al. [10] have extended PEG with methods they call trait rating, to grade essays with regards to five different aspects: content, organization, style mechanics and creativity. Critique [11], which has its foundations in e-rater, is analysis software that detects errors in grammar and writing style and aims to locating discourse elements from the essay. LSA-based methods [9, 12, 13] have proven successful in analyzing the content of essays and enabling the feedback to be given to the writer. This technique has also been applied to intelligent tutoring [14]. The current approaches supporting analytical assessment and enabling more detailed feedback, those that can also be used for semi-automatic essay assessment will be discussed in more detail in Chapter 4.

3. The Grading System

We have developed an automated grading system based on LSA [15]. LSA is a method originally developed for information retrieval providing means for determining the similarity of the meaning of words and text passages. The power of LSA lies in the fact that it is able to extract the meaning of words and text passages starting from word co-occurrence data, without need of human intervention, for example construction of logical rules. Compared with other methods used for essay assessment, LSA has the following advantages: it focuses on the conceptual content of the essay, not the surface features or content based simply on keyword frequencies and it allows assignment-specific scoring model to be calibrated with relatively low amount of pre-scored essays. Thirdly, in addition to scoring based on the comparison to the human-scored essays, LSA can be used for comparing the essays to domain-representative text.

We will shortly describe the technical details of the method. A more detailed description of the LSA may be found in [16] and [17]. LSA represents words and passages in a "semantic space". The text is presented as a matrix, in which rows stand for unique words and columns for *contexts* where the words occur. A context can be for example a sentence or a paragraph. The preprocessing methods in LSA include typical information retrieval techniques stemming, term weighting and use of stopword list. The essence of LSA is the dimension reduction based on singular value decomposition. Singular value decomposition is a form of factor analysis, which reduces the dimensionality of the original word-by-context matrix and thereby increases the dependency between contexts and words [17].

In our approach, we use part of a relevant learning material, like textbook, to train the system with assignment-specific knowledge. The motivation for using the textbook as source for creating the semantic space comes from the assumption that the student's knowledge is usually acquired by reading the course content and thus

student's knowledge can be measured as the degree of semantic similarity between the essay and the parts of the textbook covering the assignment-specific knowledge [11, 18]. The essays whose content matches more closely to the content of the course should be given a higher grade. Figure 2 illustrates the idea of the grading process of our system.

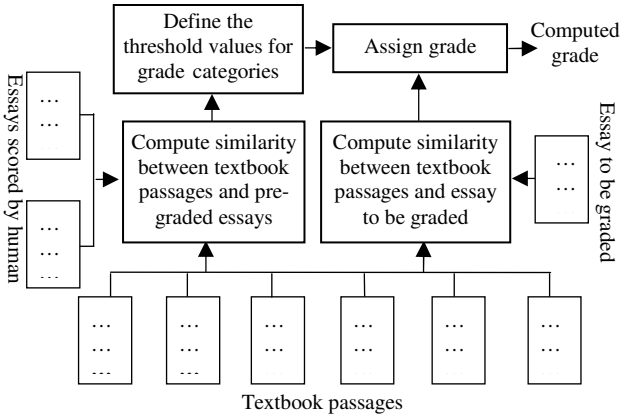

Figure 2. The grading process.

First, we create a comparison material, or semantic space, based on the relevant textbook by constructing a word-by-context matrix representing the selected textbook passages. Our experiments consisted of essays and textbook written in the Finnish language. To compare the similarity of an essay to the textbook passages covering the assignment-specific knowledge, first a query vector of the same form as each of the vectors in the word-by-context matrix is constructed. We compare the query vector representing an essay to each text passage of the textbook to calculate the similarity score by using the standard LSA similarity measure, the cosine of the angle between the document vectors to measure the similarity of the meaning between the documents. The similarity score for the essay is calculated as the sum of the similarity between it and each of the relevant textbook passages.

By calculating the similarity scores between each pre-scored essay and the textbook chapters, we can define cut-off points for each grade category based on the scores given by the human assessor. Now, with grade categories defined, we can determine a grade for each essay to be graded by calculating the similarity score between the essay and the textbook passages. The similarity score is defined with exactly the same method as with the pre-scored essays. After calculating the similarity score for the essay, we simply examine which grade category the essay belongs to according to the defined threshold values.

We have tested our system with a set of 143 essays from an undergraduate course in education, which where graded by one assessor on a scale form zero to six. The assignment was especially suitable for our method, because in the assignment there was a reference to a certain textbook chapter. The grading model was trained by using a relevant chapter from the course textbook. For the experiments, essays were randomly divided into two separate test sets. In the first experiment, 70 essays were used for defining the threshold values for grade categories and 73 essays were graded. In the second experiment, grading model was based on 86 essays and 57 essays were graded. Table 1 shows results of some of our experiments.

Table 1. The results of the grading experiments.

Material for creating the scoring model	**Num. of graded essays**	**Results**		
		Exact	***Exact or adjacent***	***Correlation***
26 paragraphs and 70 essays	73	39.7	83.6	0.78
144 sentences and 70 essays	73	35.6	84.9	0.80
26 paragraphs and 86 essays	57	36.8	77.2	0.81
144 sentences and 86 essays	57	38.6	78.9	0.82

In Table 1, the first column shows the structure of the semantic space and the number of pre-scored essays used. The results of the experiment are shown for the dimensionality that produced the most accurate scores. The proportion of cases where the same score was assigned by the system and human grader and the proportion of the essays where the grade given by the system was at most one point away are shown in columns labeled "exact" and "exact or adjacent". The last column of Table 1 shows the Spearman rank correlation between the scores given by human and the system.

The results in Table 1 show the Spearman rank correlations varying from 0.78 to 0.82 between the grades given by the system and the human grader in the optimal dimension when entropy based weighting model was used. The Spearman rank correlation around 0.80 is comparable to the results achieved by the other automated assessment systems based on LSA and to those generally achieved by two human judges. For example Landauer et al. [4], Lemaire & Dessus [9] and Folz et al. [12] have reported inter-rater correlations ranging from 0.64 to 0.84 and correlations 0.59…0.89 between the LSA-based system and human graders, LSA-based usually performing at least as well as the human graders.

4. Analytical Assessment and Semi-automatic Features

As discussed before, automatic essay grading can create problems; especially the absence of feedback does not provoke learning. A similar problem can also appear when a teacher grades the essays because the feedback can be difficult to formulate and the reasons for the given grade are not that clear [1]. The *semi-automatic essay grading* can aid in three ways: it can assist the teacher to grade the essays, it can support the student during the essay writing process and it can make the grading process more visible in a sense that some criteria for grading and feedback about the essay are available for the student. In

this section, we present some features that a system for semi-automatic essay grading should support and also describe our experiments with a LSA-based feedback method. Burstein et al. [11] have addressed some of these issues in their system, Criterion^{SM}. In table 2, we summarize the similarities and differences of essay assessment/grading done by humans or with computer-based, fully and semi-automatic, methods.

Table 2. Comparison of manual, semi-automatic and automatic essay assessment.

Essay grading	Manual	Semi-automatic	Automatic
Pedagogical background	Instructionist/ Behaviorist	Constructiovist/ Instructionist	Behaviorist
Implementation	A teacher	A system and possibly pre-graded essays or other material and a teacher	System and pre-graded essays or other material
Support for learning	Some support	High support	Very little support
On-line/ off-line	Usually off-line	Online	Usually online
Feedback	Grade and optionally some verbal comments	Grade, verbal and graphical comments	Only grade
Personalization	Very high	High	Very low

Table 2 shows the comparison between different methods of essay grading. The pedagogical background of the different approaches is ambiguous but the most probable background or backgrounds are presented. The manual grading is *instructionist* in the sense that it is teacher-centered and the evaluation is based on the criteria made by the teacher. But if only a grade for the essay is given, as it is usual, the result is close to *behaviorism*. The automatic essay grading supports only *behaviorist* learning where the feedback is immediate and the learner is only given a grade as the prize for a correct essay or punishment for an incorrect essay. The semi-automatic essay grading is based on the teacher's evaluation of the paper and thus promotes *instructionism*. However, it also supports *constructivism* where the student constructs the essay and the tools to analyze the essay can scaffold the writing process and makes the process of grading more visible. Furthermore, the student has a possibility to reflect on her own writing process, and thus a possibility to influence to the grading but also learn at the same time.

The implementation of manual grading is simple. All that is need is a teacher that masters the essay area. However, this demands much labor from the grader. For automatic and semi-automatic grading a system, that takes time to be developed, is needed. These approaches normally require a set of pre-graded essays. Moreover, semi-automatic essay grading is still based on the grading of the teacher but hopefully can ease the work of the grader.

The manual essay grading can support learning if the verbal comments are given and the feedback also addresses other than the surface structure of the essay. Automatic essay grading promotes learning very little because the grade of the essay does not tell much about the areas of the subject that should be further studied. It only indicates if the essay was good or bad. Semi-automatic essay grading supports learning on several levels. During the essay writing the system can scaffold the writing process. The possibility for reflection and self-evaluation about the essay supports also the learning as viewed by the constructivism. Finally, the feedback given by the system and the grader can indicate same areas for improvement.

Manual essay grading is usually done off-line and the essays to be assessed are hand written. To benefit from computerized essay assessment system, essays should be written to electronic form by the students. As addressed before, although manual assessment offers teacher virtually unlimited possibilities for giving feedback, the real-life limitations usually restrict the amount of informal feedback to a few short lines. In fully automated grading, the feedback is given only in numeric form, while the semi-automatic approach aims to also generating additional comments and suggestions for the essay writer and even offering him/her possibility to participate in the assessment process. With both computer-based methods the feedback can be given to writer immediately after submitting the essay or even during the writing process. In manual assessment, there typically is a delay from several days to few weeks before the student gets any feedback.

The greatest advantage of the human-made assessment compared to the current computer-based methods is its high level of personalization. If the teacher knows the strengths and weaknesses of the student, she can direct the feedback towards the needs of the essay writer, giving supporting comments and advice. In fully automated grading, personal aspect is completely lacking. In our point of view, the personalization of feedback and the possibility of the student to take part in the assessment process should be in the focal point on the semi-automated assessment environment.

4.1 Ideas and Implementation

Next we collect the different approaches for semi-automatic essay grading and evaluate their meaning in the different stages of the essay writing and grading.

- There exists several ways to give *feedback about the essays*. Most commonly the feedback is related to the coverage of different topics of the essay. This can be done for example by using LSA [12]. Other possibility for feedback is to comment the lexical and syntactical structure of the essay. The feedback is especially important for the writing process but can be also valuable for the essay grader.

- *Summary of the essay* can be also done with different methods. Burstein and Marcu [19] used Rhetorical Structure Theory (RST) and discourse parsing to identify the most important parts of the text and to form a summary using them. Miller [20] has also applied LSA for summarization with similar results. The summary can help the grader to find the relevant points covered by the essay and this can be also given as feedback for the writer.

- The essay summarization is also related to the *highlighting of the most relevant and irrelevant sentences* from the essay. The simple statistical method, LSA or the discourse structures, such as Rhetorical structure theory (RST), can be used to find the important and unimportant sentences [21]. The identification of sentences' relevance can be useful for supporting both the essay writing as well as the essay grading.

- *The structure, coherence and cohesion of the text* are important measures for the assessment of the writing style [22]. The structure of the text can be identification from the different types of sentences that can be identified from the discourse structures of the text [21, 23] or with a trained LSA. This can be especially useful during the authoring process to support the correct formation of the discourse structures. To identify bad coherence and cohesion of the text, [24, 25] have used the centering theory. Moreover, Foltz, Kitsch and Landauer [26] used LSA to determine the coherence and cohesion between two adjacent sentences or paragraphs.

- The teacher should be allowed to *search for terms that should coexist nearby or in same sentence with each other*. This could be done with a technique called proximal nodes [27]. In this way, the teacher could look for terms that should be in connection with each other.

- Students should also have a possibility for *self-evaluation of the essay*. They should be able to analyze their essay and that should be also taken in consideration in the grading. This can support the learning process of the student as she can reflect on the writing process.

- *The detection of plagiarism* is also very important in the case of the essay grading. There are two separate issues that should be addressed: copying from the other similar essays and coping from the reference materials. LSA can be used to inform when two essays from same topic are too much a like or otherwise suspicious [5]. A sliding dictionary can be used to inform when the vocabulary changes drastically that can be an indications of the coping from the reference material. The techniques for detection of coherence can be used to find abrupt changes of the topic that can be a sign of the coping from reference materials [24, 25].

As a first step towards semi-automation, we have augmented our existing essay grading system with a feedback module. As described in section 3, our grading system is based on the comparison of essays and course materials. It is quite straightforward to extend such a system, so that it can provide information about the topics that writer has covered or missed in his/her writing. Others e.g. [9, 12] have introduced similar type of LSA-based feedback methods.

For providing the feedback for topics that the essay writer had covered, we added to our grading method shown in Figure 2, the possibility to divide the textbook sections used as the comparison material, to different subtopics. As each text passage in the comparison material is marked with appropriate subtopic marker, the system can define, based on average of the similarity scores in the subtopic, how well the writer has covered the topic. According to our preliminary results such a method works well in practice, enabling the system to pinpoint writer the topics that she/he needs more training in.

By the method just described, the system is able to measure which topics the writer has covered, but not to detect off-topic information or to locate the positions in the text, where the information is found in. As mentioned earlier RST [21] is one option to be used for locating the most important sentences. LSA has also been used in sentence-level detection of the most important structures in the essays [12].

5. Conclusions and Future Work

We have presented a system for automatically grading essays written in the Finnish language. The results indicate that LSA can be successfully applied to Finnish and that grading based on the course content, such as the textbook can yield good results. A final grade alone, although useful for assessment purposes, is not very helpful for the students, thus methods enabling more analytical assessment and giving feedback are needed.

We have also presented preliminary results on augmenting our grading system with feedback module, which helps the students so that they can identify the parts of the course content where they need more practice.

Our long-term goal is to create a semi-automatic essay assessment and evaluation environment which allows both a student and his/her peers and teacher to process and analyze an essay and identify its strong points and shortcomings. This goal could be realized as an examination aquarium or automate where a student could take any exam at any time. In such a system, a student could for example set an agent to work for him in the evaluation process. Thus, the agent would carry out in a

real time the tasks that are particularly relevant to the student, from his learning point of view. It is crucial that the evaluation process is not hidden from the user, but she can intervene to it at any time.

References

[1] M. Hearst et al., The Debate on Automated Essay Grading, *IEEE Intelligent Systems, Trends & Controversies feature*, 15(5), 2000, 22-37.

[2] E.B. Page & N. S. Petersen, The computer moves into essay grading, *Phi Delta Kappan*, 76(7), 1995, 561-565.

[3] D.E. Powers, J. Burstein, M. Chodorow, M.E. Fowles & K. Kukich, *Comparing the validity of automated and human essay scoring (GRE No. 98-08a, ETS RR-00-10)* (Princeton, NJ: Educational Testing Service, 2000).

[4] T.K. Landauer, D. Laham, B. Rehder & M.E. Schreiner, How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans, *Proc. of the 19th annual meeting of the Cognitive Science Society*, Mawhwah, NJ, 1997.

[5] P.W. Foltz, D. Laham & T.K. Landauer, The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2), 1999.

[6] W. Wresch, The Imminence of Grading Essays by Computer – 25 years later, *Computers and composition*, 10(2), 1993, 45-58.

[7] L. Larkey, Automatic Essay Grading Using Text Categorization Techniques, *Proc. of 21st Annual International Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998.

[8] T.K. Landauer, D. Laham, B. Rehder & M.E. Schreiner, How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans, *Proc. of the 19th annual meeting of the Cognitive Science Society*, Mawhwah, NJ, 1997.

[9] B. Lemaire & P. Dessus, A System to Assess the Semantic Content of Student Essays, *Journal of Educational Computing Research*, (24)3, 2001, 305-320.

[10] M.D. Shermis, C.M. Koch, E.B. Page, T.Z. Keith & S. Harrington, Trait Ratings for Automated Essay Grading, *Educational and Psychological Measurement*, 62(1), 2002, 5-18.

[11] J. Burstein, M. Chodorow & C. Leacock, Criterion^{SM}: Online essay evaluation: An application for automated evaluation of student essays, *Proc. of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco, Mexico, 2003.

[12] P.W. Foltz, S. Gilliam & S. Kendall, Supporting Content-based Feedback in Online Writing Evaluation with LSA, *Interactive Learning Environments,* 8(2), 2000, 111-129.

[13] P. Wiemer-Hastings & A.C. Graesser, Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions, *Interactive Learning Environments*, 2000, 149–169.

[14] P. Wiemer-Hastings, K. Wiemer-Hastings & A.C Graesser, Approximate natural language understanding for an intelligent tutor, *Proc. of the 12th International Florida Artificial Intelligence Research Symposium*, Menlo Park, CA, USA, 1999.

[15] T. Kakkonen & E. Sutinen, Automatic Assessment of the Content of Essays Based on Course Materials. *To appear in the Proc. of International Conference on Information Technology: Research and Education 2004 (ITRE 2004)*, London, UK, June 2004.

[16] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer & R. Harshman, Indexing By Latent Semantic Analysis, *Journal of the American Society for Information Science*, 41(6), 1990, 391-407.

[17] T.K. Landauer, P.W Foltz & D. Laham, An Introduction to Latent Semantic Analysis, *Discourse Processes*, 25(2&3), 1998, 259-284.

[18] P.W. Foltz, M.A. Britt & C.A. Perfetti, Reasoning from multiple texts: An automatic analysis of readers' situation models, *Proc. of the 18th Annual Cognitive Science Conference*, USA, 1996.

[19] J. Burstein and D. Marcu, Towards Using Text Summarization for Essay-Based Feedback, *Le 7e Conference Annuelle sur Le Traitement Automatique des Langues Naturelles TALN'2000*, Lausanne, Switzerland, October, 2000.

[20] T. Miller, *Generating coherent extracts of single documents using latent semantic analysis* (University of Toronto, Master's Thesis, Graduate Department of Computer Science, 2003).

[21] D. Marcu, Discourse trees are good indicators of importance in text. In I. Mani & M. Maybury (eds.), *Advances in Automatic Text Summarization*, The MIT Press, USA, 1999.

[22] M.D. Shermis, H.R. Mzumara, J. Olson & S. Harrington, On-Line Grading of Student Essays: PEG goes on the World Wide Web, *Assessment & Evaluation in Higher Education*, 26(3), 2001, 247-259.

[23] J. Burstein, D. Marcu & K. Knight, Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing* 18(1), 2003, 32-39.

[24] E. Miltsakaki & K. Kukich, Automated evaluation of coherence in student essays, *Proc. of the Workshop on Language Resources and Tools in Educational Applications, 2nd International Conference on Language Resources and Evaluation*, Athens, 2000, 7-14.

[25] E. Miltsakaki & K. Kukich, The role of centering theory's rough-shift in the teaching and evaluation of writing skills, *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, 2000, 408-415.

[26] P.W. Foltz, W. Kintsch & T.K. Landauer, Analysis of Text Coherence Using Latent Semantic Analysis, *Discourse Processes*, 25(2-3), 1998, 285-307.

[27] G. Navarro & R. Baeza-Yates, Proximal nodes: A model to query document databases by content and

structure, *ACM Transactions on Office and Information Systems*, 15(4), 1997, 401-435.