

Assessing and Advising on Lexical Richness in an Intelligent Tutoring System

Jesús Miguel García Gorrostieta¹, Samuel González López²,
and Aurelio López-López²

¹ Universidad de la Sierra, Moctezuma, Sonora,
Mexico

² Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla,
Mexico

jmgarcia@unisierra.edu.mx, {sgonzalez, allopez}@inaoep.mx

Abstract. Guiding students on writing is a hard and time consuming chore for advisors, since it requires several iterations before achieving an acceptable level. Normally, when professors advise students close to graduation, most questions are about the structure of the thesis project. Issues such as the correct wording or abuse of certain terms within a title, problem statement, objectives and justification become the main tasks of the instructor. In this paper, we present a web-based intelligent tutoring system (ITS) to provide student advice in structuring research projects. We propose a student model based on a network to follow the progress of each student in the development of the project and personalized feedback on each assessment. This tutor includes a module for assessing the lexical richness, which is done in terms of lexical density, lexical variety, and sophistication. We also establish the methodology for future testing with undergraduate students.

Keywords: E-learning, natural language processing, intelligent tutoring system, lexical richness.

1 Introduction

Guiding and instructing a student on research or thesis writing is a hard and time consuming chore for advisors, since requires several iterations before achieving an acceptable level. There is a need to alleviate the burden of this task, possibly by technologies such as tutoring systems.

An intelligent tutoring system (ITS) is a system that provides personalized instruction or feedback to students without much involvement of instructors. Recent advances in ITS include the use of natural language technologies to analyze student writing and provide feedback as presented in the article by McNamara [1]. Writing Pal (WPal) is an ITS that offers strategy instruction, practice, and feedback for developing writers. There are also intelligent virtual agents able to answer questions for the student related to an academic subject [2]. A dialogue-based ITS called Guru was proposed in [3], which has an animated tutor agent engaging the student in a collaborative conversation that references a hypermedia workspace, displaying and animating images significant to the conversation. Another dialogue-based ITS Auto

Tutor uses dialogues as the main learning activity [4]. All these ITS use Natural Language to interact with the student similarly to the ITS we present in this paper.

Normally when advising students close to graduation, most questions are about the structure of the thesis or research project. Issues such as the correct wording or abuse of certain terms within a title, problem statement, objectives and justification become the main task of the instructor. In this paper, we present a web-based intelligent tutoring system (ITS) to provide student advice in structuring research projects. We propose a student model based on a network to follow the progress of each student in the development of the project and personalized feedback on each assessment. This tutor includes a module for assessing the lexical richness, which is done in terms of lexical density, lexical variety, and sophistication.

There are a variety of methods to evaluate the use of vocabulary (lexicon) in text. One of them is to measure the sophistication of some papers using text word lists. In [5], they used a list of 3000 easy words. For Spanish, some studies use the list provided by the SRA (Spanish Royal Academy) of 1000, 5000 and 15000 most frequent words. Other works have used Yule's K to measure the richness in texts [6], where this kind of measures focus on the word repetitions and this is considered a measure of lexical variety.

We also establish the methodology for future testing with undergraduate students. We cannot ignore the importance of language, especially in writing, when considering the formation of higher education students; one of these stages of formation is related to the generation and application of knowledge through research, which are usually placed in the last semesters of their programs of study.

Each institution adopts various mechanisms that allow students to enter in the field of research, either through business internship, professional practice or in the various forms of professional qualification that presents the possibility of doing a research. However, the process of drafting the research projects is usually not an easy task for students. Therefore, our proposed system intends to assist the work of the instructor and to facilitate and guide students through this process.

The paper is organized as follows. Section 2 describes the model underlying the tutoring systems, while section 3 details the implementation with examples of draft evaluations. We conclude in section 4, discussing further works.

2 The Model

The intelligent tutor in the Domain Module presents material concerning the different elements of the project, such as the problem statement, title, objectives and justification. For each element, a test is applied to validate the reading of materials and practical exercises are applied using the richness Lexical Analyzer to achieve a high level of density, diversity and sophistication in the student text productions. The results of the test and lexical analysis are sent to the Student Progress Module to update the knowledge state of the student in a network. Figure 1 shows the intelligent tutor model.

The Student Progress Module (SPM) records the student's progress in the network which is depicted in Figure 2, when the student completes the test, the value of the

test node element is updated and the SPM calculates the student's progress for the parent node using the weights assigned to each question in the test [7].

Similarly as when performing the exercises with the lexical analyzer, the corresponding node in the network is updated and the SPM estimates the student's progress for the parent node using the weights assigned to the lexical density, variety and sophistication in the Lexical Analyzer.

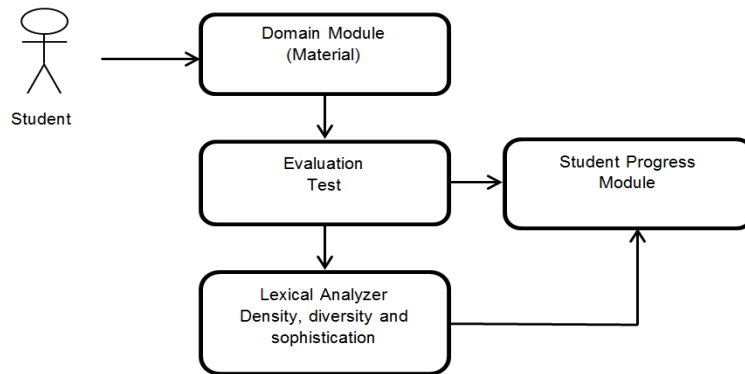


Fig. 1. Model of Intelligent Tutoring System.

Figure 2 illustrates the weights assigned to each node according to the experience of the teacher. For instance, in the Test node of the Statement, a weight of 50% of the parent node problem statement is assigned, which includes 5 questions to verify that the student has read the material. Once the student has correctly answered questions, a 50% of advance in the concept is assigned, as shown in Figure 3. This will enable the student to use the lexical analyzer to perform three exercises which have a combined weight of 50% of the parent node, which is distributed as follows: 20% to lexical density, 20% to lexical diversity, and finally 10% for lexical sophistication.

So, by completing the exercise of lexical density with a high grade, the student would have advanced 70% in the concept, as shown in Figure 6. Also when the student gets a high grade on the exercise of lexical diversity, he will have completed 90% of the problem statement concept, leaving only the exercise of lexical sophistication to complete the 100% of the concept.

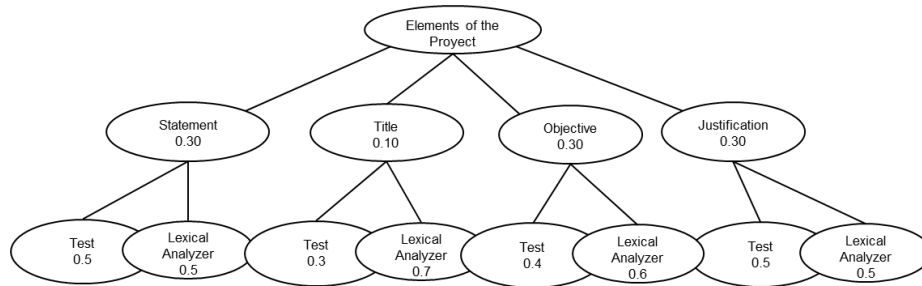


Fig. 2. Network used in the Student Model.

Lexical analysis focuses on the evaluation of three measures: lexical density, lexical variety and sophistication, which together assess lexical richness. The first measure, lexical variety, seeks to measure student ability to write their ideas with a diverse vocabulary. This feature is computed by dividing the unique lexical types (Tlex) by the total of lexical types (Nlex). Tlex refers to the unique terms of content, while Nlex represents total terms of content, both ignoring empty words [8].

The lexical density aims to reflect the proportion of content words in the complete text. This measure is calculated by dividing the unique lexical types or content words (Tlex) by the total words of evaluated text (N), i.e. the number of words before removing stop words.

The third measure is sophistication, which attempts to reveal the knowledge of technical concepts and is the proportion of "sophisticated" words employed. This measure is computed as the percentage of words out of a list of 1000 common words, provided by the SRA. All the measures take values between 0 and 1, where 1 indicates a high lexical value, and values close to zero mean a low value of the lexicon of the evaluated section.

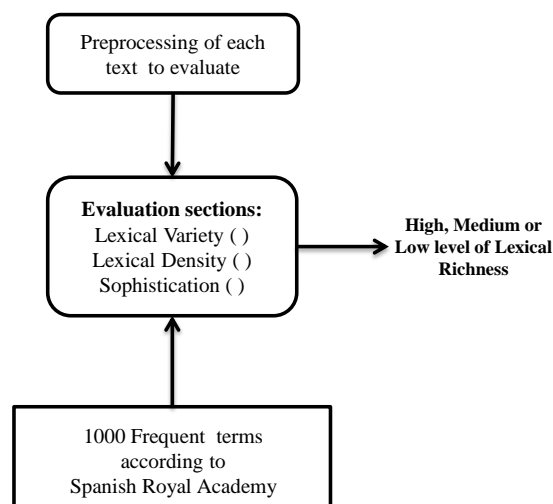


Fig. 3. Model of Lexical Analyzer.

The preprocessing of the text was filtering and removing empty words from a list provided by the module of NLTK-Snowball. Stop words include prepositions, conjunctions, articles, and pronouns. After this step, only content words remained, which allowed the calculation of the three measures. Finally, the results produced by the Lexical Analyzer are sent to the Student Progress Module, so the intelligent tutor manages the results achieved by the student.

A scale ranging in High, Medium and Low in lexical richness has been established based on our previous work [9], where we analyzed research proposals and theses of graduate and undergraduate students.

3 The Intelligent Tutoring System

The intelligent tutoring system is developed in PHP for easy access via web and the network structure is stored in a MySQL database, the lexical analyzer is developed in Python because of the easy access to processing tools of natural language. The analyzer uses the open source tool FreeLing¹ for stemming words and then analyzes the density, diversity and sophistication in the text.

Figure 4 shows the graphical interface of the tutoring system in which we observe the button to the main menu to access the elements of the project (in Spanish *Elementos del proyecto*) inside we find links to access the problem statement, title, objectives and justification. For each element, there are three sections: material, test and practical evaluation. In this figure, we can also notice the progress section (in Spanish *Avance*) in the left side, reporting the progress in the concept. As we can see, to enter the practical evaluation, the student must first successfully complete the test receiving a 50% advance in the concept and 15% in the complete project.

Fig. 4. Lexical Analyzer for Density (in Spanish).

The section of practical evaluation is also depicted in figure 4, where the student writes his problem statement to be analyzed for lexical density. First, the analyzer performs a tokenization of words, then a classification based on the 1000 most common words of Spanish are done to identify stop words and the rest as content words. Density analysis measures the balance between content words and stop words,

¹ This software is available at <http://nlp.lsi.upc.edu/>.

if the text has too many stop words it will have a very low density, if the text has just a few stop words compared to content words it will have a high density.

As we can see in Figure 5 the feedback of the lexical density analysis and the level assigned to the problem statement proposed by the student is "Low Density" (in Spanish *Densidad Baja*) due to the large number of stop words relative to content words, the system sends a message to the student with a feedback according to the level assigned.

The message displayed is "we suggest reviewing of the text, there are few content words, try to reduce the terms outlined in red" (in Spanish *Se sugiere revisar el texto, ya que existen pocas palabras de contenido, procura reducir los términos subrayados en rojo*) in the paragraph, we observe stop words underlined to indicate to the student that it is necessary try to reduce them, and a progress bar is presented to indicate the progress of his writing graphically, in this case a 50.98% of advance.

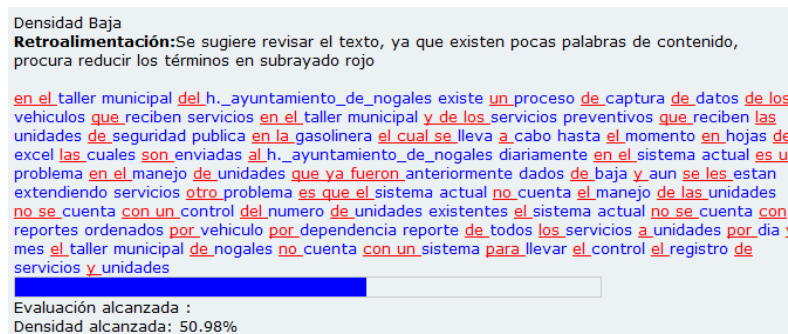


Fig. 5. Detailed feedback of Lexical Analyzer for Density (in Spanish).

After correcting the paragraph, the analyzer indicates a high level (in Spanish *Densidad Alta*) and activates the access link to the analysis of lexical diversity (in Spanish *Análisis de diversidad léxica*), as shown in Figure 6, the feedback indicates that the statement problem is in balance between stop and content words, with a 66.67% of lexical density.

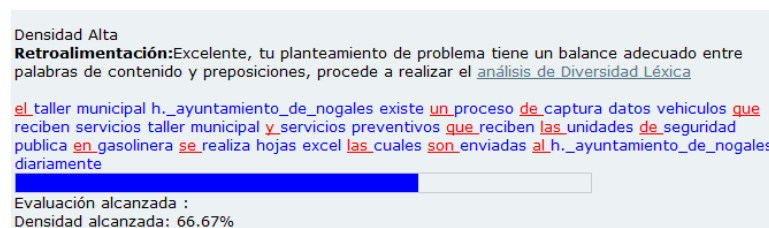


Fig. 6. Detailed feedback from Lexical Analyzer, for High Density in text (in Spanish).

Figure 7 shows the lexical analyzer for diversity which are content words that are repeated several times such as "services" (in Spanish *Servicios*) and "units" (in Spanish *unidades*). This case has a medium level of diversity with a feedback to the student "There are still repetitive words of content, modify your text, avoid using the

reach the highest level to improve his writing skill. Also we will adapt the interface of the ITS to have an improved use on mobile devices.

We also plan to extend the ITS assessing additional aspects in drafts such as coherence and specific language usage for particular sections of proposals.

Acknowledgments. The second author was supported by Conacyt scholarship 1124002, while the third author was partially supported by SNI.

References

1. McNamara, D. S., Raine, R., Roscoe, R., Crossley, S., Jackson, G. T., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P. M., & Graesser, A. C.: The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In: Applied natural language processing and content analysis: Identification, investigation, and resolution. Hershey, PA: IGI Global. pp. 298-311 (2012)
2. Rospide, C.G. & Puente, C.: Virtual Agent Oriented to e-learning Processes. In: Proceedings of 2012 International Conference on Artificial Intelligence. Las Vegas, Nevada (2012)
3. Olney, A.; D'Mello, S. K.; Person, N. K.; Cade, W. L.; Hays, P.; Williams, C.; Lehman, B. & Graesser, A. C.: Guru: A Computer Tutor That Models Expert Human Tutors. In: Stefano A. Cerri; William J. Clancey; Giorgos Papadourakis & Kitty Panourgia (eds.) 'ITS', Springer, pp. 256-261 (2012)
4. Graesser, A.C., D'Mello, S.K., Craig, S.D., Witherspoon, A., Sullins, J., McDaniel, B., and Gholson, B.: The Relationship between Affective States and Dialog Patterns during Interactions with Autotutor, *J. Interactive Learning Research*, vol. 19(2), pp. 293-312 (2008)
5. Schwarm, S. and Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05), pp. 523-530, (2005)
6. Miranda, A., and Calle, J.: Yule's Characteristic K Revisited. *Language Resources and Evaluation*, 39(4), pp. 287-294 (2005).
7. Sucar, L.E., and Noguez, J.: Student Modeling. In: O. Pourret, P. Naim, B. Marcot (eds.), *Bayesian belief networks: a practical guide to applications*, Wiley, pp.173-186 (2008)
8. Roberto, J., Martí, M., and Salamó, M.: Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. *Procesamiento de Lenguaje Natural*, 48, pp. 97-104 (2012)
9. González, L. S. and López-López, A.: Supporting the review of student proposal drafts in information technologies. In: Proceedings of the 13th annual conference on Information technology education (SIGITE '12). ACM, New York, pp. 215-220 (2012)