Literador: A Comprehensive Tutoring System for Spanish Writing

Fernando Gutierrez [1], *Christian Soto* [1], *Bernardo Riffo* [1], *María Fernanda Rodríguez* [1], *Ana Vine* [1], *Daniel Mora* [1], *Carolina Calbullanca* [1], *Paola Teppa* [1], *Isabel Cisternas* [2], *Cristian de la Fuente* [1], *Diego Palma* [1], *Antonio Gutierrez* [3]

[1] *Spanish Department, University of Concepcion*

69121 Concepcion, Chile

[2] *Vice-rectory Research and Graduate Studies, University of Bio Bio*

4051381 Concepcion, Chile

[3] *Department of Curriculum, Foundations, and Reading, Georgia Southern University*

Statesboro, Georgia, USA

ABSTRACT

In a professional setting and in adult education, a well-written text needs to convey meaning by presenting ideas in a coherent and cohesive form that facilitates readability, such as the balance in the use of coreferences, the abstraction of the language used, and the lexical diversity of the text. In this work, we proposed Literador, an Intelligent Tutoring System for Spanish writing. By incorporating different Natural Language Processing tools, Literador can analyze and provide feedback on the different aspects of a text, beyond just content. These tools are the Spanish-trained language model BETO, the text complexity analyzer Trunajod, and regular expressions to identify the use of key lexical elements. By combining all of these sources of information, Literador can provide feedback following a strategy that prompts different types of messages depending on the student's level and tries, which also intends to avoid information overflow.

Keywords: Natural Language Processing, Semantic Similarity, Readability

INTRODUCTION

For a text to be considered well-written, ideas need to be presented clearly and coherently, with an adequate vocabulary and syntax that facilitates comprehension. To reach the level of text production that is suited for a professional or an academic setting, a student needs constant practice and meaningful feedback (Bazerman 2013). However, this task cannot be met by simply increasing the number of writing instructors. It is not possible to have an instructor to assist and provide timely feedback to each student. Plus, a human evaluator (i.e., instructor) cannot produce the same level of consistency in the grading, the same rigorousness in the analysis, and the same level of detail in the feedback, over a long period of time. For these reasons is that computer-based platforms for writing are considered a key tool to improve text production (Crossley and McNamara 2016).

In this work, we present Literador, an Intelligent Tutoring System (ITS) for Spanish writing. Through this ITS, we propose the integration of different types of Natural Language Processing tools to provide a better insight into the student's text. The analysis of content is performed by semantic similarity while readability is analyzed through Trunajod. We complement both analysis by including regular expression rules, which can identify key elements. The information generated by these tools is incorporated to produce feedback that is clear and complete.

The rest of this paper is organized as follows. We provide a brief overview of most relevant items in the context of ITS for writing in the Related Works section. In the Method section, we present Literador and the details about the different components that conform it. In the Experiments section we provide details of our initial evaluation and some interesting findings in the current state of development. Finally, we provide some conclusions and indicate future work.

RELATED WORK

Computer-based methods have become key instruments to assist in improving writing skills. Through advances in Natural Language Processing (NLP), many solutions have been proposed based on a wide variety of techniques (He et al. 2009, Gutierrez et al. 2013), in the form of automated grading systems.

The most popular NLP approach for ITS is Latent Semantic Analysis (LSA) (He et al. 2009), which attempts to capture the meaning of a text through Singular Value Decomposition (SVD). However, in recent years, the development of deep neural networks has led to the current state-of-the-art language model BERT (Devlin et al. 2019). This language model is based on the machine learning architecture known as transformers. Because BERT is trained in the tasks of masked words, i.e., predicting missing words in a sentence, and next sentence, i.e., determining if one sentence follows another, it can produce a very accurate model of a language.

Although these machine learning-based methods can produce very accurate results, it is very difficult to determine the specific reason for a score, which limits the possible usefulness of the feedback. To address this limitation, Information Extraction-based (IE) methods have been proposed (Gutierrez et al. 2013). The searching mechanism is in most cases some variation of regular expressions, which is a widely use method to identify sequence of characters in a text based on a specific pattern. Because regular expressions can be combined with NLP annotation techniques, such as part-of-speech recognition, dependency parsing, or Named Entity Recognition, they have been integrated into a large variety of NLP tools (e.g., GATE). Under an IE approach, the text is analyzed following the guidelines of a knowledge representation model, such as a domain ontology (Gutierrez et al. 2013). By incorporating IE into the evaluation method, it is possible to identify and highlight (i.e., feedback) which elements are missing in the text or if there are semantically incorrect statements.

While most methods and techniques used for text analysis in an education setting have been centered on evaluating content, some efforts have been made to consider other facts that also affect a text, such as readability. Coherence has been integrated into essay evaluation through the use of Entity Grids (Palma and Atkinson 2018), while Coh-metrix has focused on cohesion and other linguistic cues (Graesser 2004). On the other hand, TERA (Jackson et al. 2016) and Trunajod (Palma et al. 2021) have been proposed to assist teachers in analyzing text. These tools can determine if a text’s complexity (i.e., readability) is adequate for a specific comprehension task or grade level.

METHOD

Literador is an ITS for Spanish writing that consists of modules that incrementally teaches students different techniques to improve writing production. Each module is formed by a set of tasks that are based on a specific theme and topic.

The core task in Literador, and the focus of this work, is extended writing. In this task, a student must produce a short text (150 to 300 words) that must fulfill a specific requirement associated with the topic of the module, such as a technical report or an essay. Internally, extended writing follows Hayes model (Hayes 2012) in the production of the text. Through intermediate steps, the student must define the main ideas, these ideas need to be transcribed into extended text, and finally the student must perform modifications (i.e., edits) to the text to produce a final version.

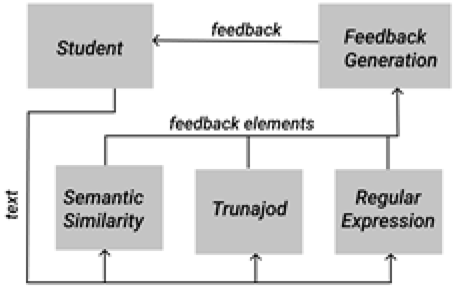

Figure 1. Diagram of components in the extended writing task of Literador.

To provide meaningful feedback for the extended writing task, we propose an analysis system that integrates three NLP tools to analyze the text, plus a feedback generation component (Fig. 1). The first NLP tool, semantic similarity, is used to analyze the content of the text. The second NLP tool, Trunajod, is used to determine the readability of the text. Initially, Trunajod was developed as a tool to assist teachers in identifying text that presented a specific level of difficulty, i.e., identify text that is adequate for targeted grade. To the best of our knowledge, this the first time that Trunajod has been integrated into an ITS to assist students. The third NLP tool is regular expressions, which is used to identify specific elements in the text. Finally, the feedback generation component integrates the information from all three components, and it produces an insightful and succinct feedback message for the student.

In the following sections, we provide more details of these different components that help the extended writing task.

Semantic Similarity

To determine the quality of the text content, Literador uses semantic similarity. This method considers a representation method and a measuring technique.

Although LSA is a common representation method for semantic similarity, in Literador we have opted for the Spanish language model called BETO (Cañete 2020).

It is based on the BERT language model, but trained for Spanish. BETO was trained following considerations similar to BERT-basic over a data set of 3 billion words (+300 million lines of text). This data set consists of texts from Spanish articles from Wikipedia, and OpenSubtitles, among other sources. BETO has shown high performance when used for NLP tasks that require deep language understanding.

To evaluate a text, Literador will use cosine similarity for measuring content. So, Literador will first encode the text with BETO. This will create a numerical representation of the text. Then, it will compare this representation against a predefined set of texts. In this set, there are three types of text: semantically complete text, text with missing information, and text that is associated with the topic, but not the task itself. The student´s text will be labeled based on which of the three types is more similar.

Trunajod

Trunajod is a text complexity analyzer that measures over 50 different types of indicators that are related to readability. These indicators range from surface measurements based on counting of words and average length of sentences, to distance between Named Entities (based on Entity Graphs).

To ease comprehension of the results generated by Trunajod, these indicators have been grouped under five different classes, known as dimensions. The lexical similarity dimension focuses on the level of diversity of the vocabulary presented in the text. This dimension is significantly relevant for Spanish since diverse vocabulary is con-sider a requirement for well-written text. The cohesion dimension measures the integration of ideas across different sentences, which gives a text a sense of whole or unit. The coherency (i.e., connectivity) dimension measure key elements that establish logic connection between the different semantic parts of a text. Concreteness dimension refers to the level of abstraction of the text, i.e., are the words referring to more abstract or concrete objects. Finally, the narrativity dimension indicates if the text is present more as a sequence of actions or more as a description.

There are two aspects that must be taken into consideration when using the values of these dimensions. The first aspect is that a low value in a dimension for one type of text might not be considered low for a different type of text. For example, a well-written technical report might not have the same value of concreteness than a well-written essay on social concepts. The second aspect is that a high value in a dimension might not be adequate for the context. For example, while a text with high lexical similarity might be easier to understand, it might not be adequate for a report or essay.

For the extended writing task, Trunajod provides measurements on each dimension for a text. These measurements will be sent to the feedback component, which must interpret these values based on the feedback strategy. Considering the previously mentioned aspects, it is likely that we need to tune the feedback component regarding readability based on the topics and modules.

Regular Expressions

As mentioned, regular expressions are used in Literador to identify key words and phrases in the students' writing. They are used to complement what is done by semantic similarity and Trunajod, enriching the feedback.

Regular expressions in Literador are used as part of two tasks. The first task refers to identifying key content elements. While semantic similarity can indicate how complete is the content of a text, it might not be clear about what part of the information in the text is missing. For example, if a student text presents only one of two important ideas in the text, it will be labeled as incomplete by semantic similarity. How-ever, it is not possible to infer from the label if a key idea is missing or something else. In this case, a set of regular expression rules can determine if both ideas are present in the text, which leads to more fine-grained feedback.

The second task is identifying elements that affect the readability of the text. While Trunajod might use some regular expressions to estimate some of its metrics, it provides a numerical value regarding this information. To complement this information, we have specific regular expressions to identify elements, such as a connector that serves to bridge two sections of a text. If a connector is missing, the feedback can indicate to the student that the text's referential coherence is low, and that might be caused by missing connectors.

Feedback Strategy

The previous methods can produce a significant amount of information that, if presented directly to the student as feedback, can be overwhelming and not informative. In Literador, we have defined a feedback strategy that intends to present information in a concise and clear fashion.

The feedback strategy considers three key factors in order to incorporate information from the NLP tools in the feedback message. The first factor is the thematic goal of the task. In most cases, this goal can be achieved by presenting the required content in the text. If the content in the text is incomplete, it will be informed to the student. This factor can be considered as fundamental, and it can affect the student's progression if not met.

The second factor refers to the readability of the text. If the text meets all the requirements regarding content, readability becomes the focus of the feedback message. In the case where the feedback is centered on missing content elements, the readability feedback is presented as complementary information. In this last case, it will only focus on the dimensions of coherence and cohesion because they tend to have a higher impact on the clarity of the text.

The third and final factor is the number of attempts submitted by the student to answer the task. In the first attempt, the feedback messages will intend to elicit the correction through cues. However, after a certain number of attempts, the feedback will explicitly indicate the issues present in the text. The objective with this approach is to allow the student to identify the issues in the text with minimal guidance. In the case that the student cannot identify the issue, Literador gives the information the student needs to keep progressing through the modules. Currently, Literador provides three attempts before providing the explicit feedback; we are evaluating alternatives in the case of modules that have more complex themes.

EVALUATION AND RESULTS

For this work, we will focus the evaluation of our proposed approach on the use of Trunajod as a method to assess text production. The reason for this approach to evaluation is that Trunajod was not envisioned as a tool to help text production. It was designed and developed to help teachers determine the complexity of a given well-written text, and if this text is adequate for a specific age (i.e., grade). This is not the case for the other NLP tools used by Literador.

Although BETO has not been used in the context of an ITS of Spanish writing, less accurate language models combined with semantic similarity have already shown their impact in tutoring and evaluation systems. Similar argument can be made for regular expressions, where rule-based Information Extraction has successfully been incorporated into evaluation systems. For this reason, we have focused the evaluation in this work to determine if Trunajod can help provide guidance for text production, and to determine if some tuning of Trunajod is needed

4.1 Evaluation Setting

From the five dimensions measured by Trunajod, we will focus on lexical similarity, cohesion, and concreteness. We have selected these dimensions, because the two lasting dimensions are being addressed through other factors in Literador. Coherence is covered by the language model since BERT and BETO are trained in the next sentence task. On the other hand, narrativity will be limited (i.e., guided) by the type content that is being prepared for Literador.

The evaluation is based on two small sets of text. The first data set consists of 37 short reports (150 words) from college students about in-field work. The second data set consists of 99 short essays (300 words) from high-school students.

Both sets were labeled by a group of six experts. They independently labeled (i.e., evaluated) each text with low, medium, or high, on all three dimensions. We aggregated the evaluation of the experts following majority voting (i.e., mode) to determine the final value for each dimension of a text.

In the case of Trunajod, we discretized the measures of each dimension following a bins approach (i.e., three bins). This transformation made the values obtained from Trunajod comparable to the labels assigned by the group of experts.

Results and Discussion

From Table 1, we can see the accuracy of determining the value of a text under each of the three dimensions evaluated. Trunajod could measure concreteness most accurately. For the other dimensions, the difference between the experts and Trunajod seems related to the distribution of labels. The experts seem to have given more texts the label low while most Trunajod´s labels are centered in medium.

The issue with cohesion seems to be different. Trunajod had the lowest accuracy for this dimension in the 300-word data set. In particular, the labels from this case did not have the difference in distribution observed in the other dimensions. When observing the distribution of labels for cohesion among experts, there were significant differences that do not match the distribution for the labels in the other dimensions. It is possible that for evaluating this dimension some context regarding topic or purpose might be required.

Table 1: Results from evaluations on the dimensions of lexical similarity, cohesion, and concreteness for the small data set (150 words) and large data set (300 words).

Label	**Accuracy**
Lexical similarity (150 words)	48.6%
Cohesion (150 words)	56.7%
Concreteness (150 words)	70.2%
Lexical similarity (300 words)	60.6%
Cohesion (300 words)	39.3%
Concreteness (300 words)	57.5%

The results reflect several interesting phenomena that have to do with the characteristics of the evaluation of the subjects in relation to Trunajod. In the first place, it must be said that the evaluation of the experts is more consistent with the dimension of *concreteness* since it refers to the identification of specific words, therefore this may be more evident in the text, since they are the easiest words to identify. However, we infer a natural tendency for experts to lose consistency and control of the evaluative process as we analyze more abstract and complex phenomena. This occurs with *lexical similarity* and even more so with *cohesion*, where the difference between expert evaluations is significant. Complementarily, in terms of specifics, it is easier to evaluate this phenomenon taking into account a limited length of the written product (better 150 words than 300). Something similar could happen when *cohesion* is analyzed. By considering short writing, evaluators can analyze the specific relationships between one sentence and another. However, as the sentences in the text increase, the evaluator loses awareness of how closely sentence 1 is related to sentence 3 or 5. In the case of *lexical similarity*, something interesting occurs; evaluators can be more precise in longer text because the phenomenon of word similarity, although it could be relatively easy to identify, only manifests this dimension (or has the possibility of manifesting itself) in texts that are not short. This is due to the fact that in Spanish a series of successive sentences is required to repeat the referent but with different words, otherwise we would fall into badly evaluated redundancies from the point of view of the quality of the writing.

CONCLUSION

In this work, we present the ITS Literador for Spanish writing. Through Literador, we propose the integration of NLP tools BETO for analyzing the quality of the content, Trunajod to measure readability, and regular expressions to determine the use of key words and phrases. The information produced by these tools is processed following a feedback strategy that is focused on guiding the student, rather than providing the final answer. Our initial evaluations of incorporating Trunajod indicate that, although some tuning is still required, it is possible to obtain reasonable results, adequately categorizing text regarding its readability. We already know that certain dimensions of writing can be better evaluated taking into consideration short texts and, on other occasions, more elaborate texts. Our challenge is to propose tasks that take ad-vantage of the findings identified in this study, providing feedback that helps to differently improve the way in which the writer operates with the various dimensions of the text that he is writing. This tuning of Trunajod will be one of the main objectives in our future work.

ACKNOWLEDGMENTS

This work is funded by the Chilean Agencia Nacional de Investigación y Desarrollo (ANID) through the grant FONDEF ID20I10290.

REFERENCES

Bazerman, C. (2013). Understanding the lifelong journey of writing development, Journal for the Study of Education and Development, Volume 36 No. 4, pp. 421-441.

Crossley, S.A., McNamara, D. (2016). Adaptive Educational Technologies for Literacy Instruction, Routledge, New York.

He, Y., Hui, S.C., Quan, T. T (2009). Automatic summary assessment for intelligent tutoring systems, Computers and Education, Volume 53 No. 3, pp 890–899.

Gutierrez, F., Dou, D., Martini, A., Fickas, S., Zong, H. (2013) “Hybrid Ontology-Based Information Extraction for Automated Text Grading,” proceedings of the 12th International Conference on Machine Learning and Applications, pp. 359-364.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805.

Palma, D., Atkinson, J. (2018). Coherence-Based Automatic Essay Assessment, IEEE Intelligent Systems, Volume 33 No. 5, pp. 26-36.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments & Computers, Volume 36, pp. 193-202.

Jackson, G. T., Allen, L. K., & McNamara, D. S. (2016). COMMON CORE TERA, Adaptive Educational Technologies for Literacy Instruction, Volume 49.

Palma, D., Soto, C., Veliz, M., Karelovic, B., Riffo, B. (2021). TRUNAJOD: A text complexity library to enhance natural language processing, Journal of Open Source Software, Volume 6 No. 60.

Hayes, JR. (2012). Modeling and Remodeling Writing. Written Communication, Volume 29 No.3, pp. 369-388.

Cañete, J., Chaperon, G., Fuentes, R. Ho, J.H., Kang, H., Pérez, J. (2020) “Spanish Pre-Trained BERT Model and Evaluation Data,” proceedings of the Practical ML for Developing Countries: learning under limited/low resource scenarios.