Method to assess students summaries in an intelligent tutor system using coherence and content analysis in a reading comprehension task

Diego Palma[1], Christian Soto[1], Fernanda Rodríguez[1]

[1] Universidad de Concepción

Concepción, 1290 Víctor Lamas Street, Chile

ABSTRACT

In this paper, a discourse-based method that merges syntactic and semantic models for developing an automated system for reading comprehension assessment is proposed. For evaluating semantic content, we use the classical models from the literature: Vector Space Modelling and Latent Semantic Analysis. For evaluating the coherence of a text, we used an entity grid representation of the texts, which extracts syntactic patterns from the texts and relies on the assumption that coherent texts will have similar underlying syntactic patterns. The contribution of this work is twofold: firstly, we develop a new methodology for free-text responses in which we assess student‘s texts by semantic content and coherence. Secondly, we develop an automated system for assessing a student's reading comprehension for Spanish Language using features that can be computed automatically. Experiments show that we can get accuracies of 90% when assessing text content, and of 55% - 60% when assessing text coherence.

Keywords: Natural Language Processing, Educational Tool, Reading

Comprehension System, Coherence Assessment, Word Embeddings, Vector-space Models

INTRODUCTION

In the current literate society, reading comprehension has become a fundamental skill. The development of reading expands the possibilities of progress in school and working life (OECD, 2013). Therefore, understanding the meaning of written words and communicating through text with constructive and critical thinking should be a central goal of schooling around the world (Paris et al., 2009). Unfortunately, there is a wide need to improve reading comprehension skills, according to PISA reading evaluation (San Martín et al., 2012). Specifically, the most recent PISA results show that 28.4% of Chilean students could not achieve basic level of reading comprehension (level 2), while the highest levels were only achieved by 2.3% of the students (level 5 and 6) (OECD, 2016).

Reading comprehension skills can be improved by learning and practicing reading comprehension strategies. These strategies involve the interaction between the students and a tutor (computational or human expert), which can provide feedback (Soto et al., 2019) is a tool developed by the Universidad de Concepción, that helps students practice their reading comprehension skills and is proposed as a tool for improving reading comprehension skills. COMPRENDE has different modules in which a student can practice their reading comprehension skills via meta-cognitive strategies, such as bridging. However, one drawback of COMPRENDE is that it is statically designed, with a set of possible answers pre-defined, and thus, it cannot assess open-ended answers. In this fashion, some other measurements of reading comprehension are not considered.

In this work, we propose an enhancement to COMPRENDE. The proposal consists of an automated summary evaluator that assesses student summaries generated as responses to prompts about a text. The automated assessor incorporates Artificial Intelligence and computational linguistics techniques. By incorporating these techniques, the summary evaluator can assess a text by considering its content and its coherence.

RELATED WORK

Traditionally, automated text assessment methods use shallow features extracted from a text as indicators of its quality. These features often include the frequency of words/sentences, frequency of grammar errors, lexical categories, and readability indices. In addition, some more complex indicators also involved the text’s lexical diversity in terms of the used vocabulary. Text assessment, using the previously mentioned features, is usually seen as a linear regression problem in which each of these shallow features is weighted so as to predict an essay’s score, where the weights are estimated by collecting samples from human-assessed essays. This kind of approach is usually found in commercial assessment systems such as E-rater (Burstein et al., 1998).

On the other hand, semantic-based assessment approaches (Zupanc and Bosnić, 2017) work under the assumption that essays of similar content should receive similar quality scores. In order to estimate similarity between texts, texts are first represented as a word vector using a Vector Space Model (VSM), which is a mathematical representation that considers the most relevant words for each text. Closeness metrics, such as Cosine similarity, are then used to estimate the similarity between word vectors of each text, where one of these texts is a high-quality human assessed essay, extracted from a previously collected set of gold essays. Experiments using this kind of method indicated a (Spearman) correlation r=0.5 when comparing with humans, where human-human correlation achieved r=0.6.

The shallow features approach has the drawback that is not actually assessing intrinsic properties of the text such as cohesion and coherence because assessment depends on extrinsic (e.g. word count) properties that are approximations, and thus, there is no guarantee that the assessment will be accurate in practice. On the other hand, semantic approaches deal with this issue by actually assessing content. Nevertheless, the drawback is that these approaches do not consider sentence ordering or syntactic patterns that could capture some cohesion issues. To overcome the limitation of previously mentioned methods, an approach combining discourse patterns and semantic modeling have been proposed for essay assessment (Palma and Atkinson, 2018b). However, none of the previously mentioned approaches have connected reading comprehension assessment with writing ability in a reading comprehension learning tool.

AN AUTOMATED SUMMARY EVALUATOR FOR READING COMPREHENSION ASSESSMENT

In this work an automated multiple dimensions text assessor is proposed, that considers content and coherence assessment. However, the developed system assesses these two dimensions separately using different criteria. Unlike other state of the art assessors , the contribution of this work is twofold:

1. Contribution One: The approach combines content and coherence assessment for summaries in a reading comprehension tool, being assessed separately.
2. Contribution Two: Summaries' coherence approximations are obtained by using discourse patterns.

DATA-DRIVEN APPROACH TO BUILD AN AUTOMATED SUMMARY EVALUATOR

The architecture of our method for text assessment consists of two phases that are shown on Fig. 1 and Fig. 2. The training phase requires having a corpus of human graded student responses and a corpus of texts related to the domain in which the assessment is performed. Next, texts are processed and a feature extraction phase occurs. In this feature extraction, measurements from texts are gathered. In this work, two types of features are considered: semantic features and coherence features. In the runtime phase, the system is already trained to assess text responses. In this phase, a student submitted a response to the system, the system processes it and then assesses it using the content assessor and the coherence assessor. Then, the system will compute scores for these dimensions and will give the student feedback based on obtained scores.

The problem is treated as a classification problem, in which, the input is the text and the output is the score of the text for content and coherence. To solve this problem, machine learning techniques are used. In particular, for this work, a Random Forest model was trained (Breiman, 2001).

Each text is represented in a Vector Space Model (Peng et al., 2010), which is a mathematical representation of the texts based on term frequency. Then, we apply Latent Semantic Analysis (Landauer, 2003), which is a dimensionality reduction technique that addresses Vector Space Model issues such as detecting synonyms and semantic relationships. To obtain the semantic space, a corpus related to the domain of interest is needed. Finally, students' responses are mapped to this semantic space, by converting the responses into a word vector and projecting it into the obtained semantic space.

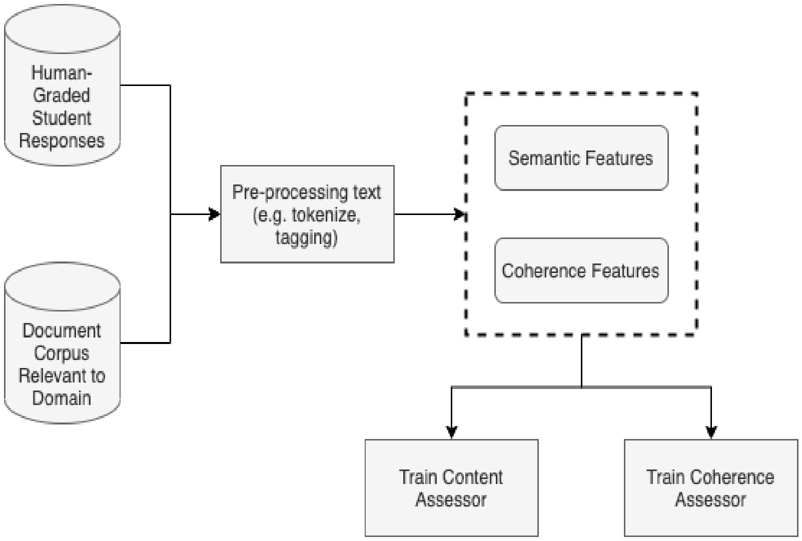
Human-Graded Student Responses
Document Corpus Relevant to Domain
Pre-processing text (e.g. tokenize, tagging)
Semantic Features
Coherence Features
Train Content Assessor
Train Coherence Assessor

Figure 1. Training phase of the automated text assessor.

ASSESSING CONTENT

Each text is represented in a Vector Space Model (Peng et al., 2010), which is a mathematical representation of the texts based on term frequency. Then, we apply Latent Semantic Analysis (Landauer, 2003), which is a dimensionality reduction technique that addresses Vector Space Model issues such as detecting synonym and semantic relationships. To obtain the semantic space, a corpus related to the domain of interest is needed. Finally, students' responses are mapped to this semantic space, by converting the responses into a word vector and projecting it into the obtained semantic space. Representation is shown in equation 1, in which each document consists of a vector of normalized frequencies.

$$\square_{\square} = (\square_{1}, \square_{2}, \dots, \square_{\square}) \, . \tag{1}$$

ASSESSING COHERENCE

To obtain coherence features, each text is represented as an Entity Grid (Barzilay, Lapata, 2008), which is a computational linguistics model based on centering theory (Grosz, et al., 1995). The centering theory is a linguistics discourse model that states that coherence of a text depends on how entities (e.g., noun phrases) in a text are distributed across utterances, and that a text is perceived as less coherent if there are too many changes of focus in the text. This has been proven to have a high degree of association in coherence assessments (Miltsakaki and Kukich, 2004). Entity grids represent text in a grid, in which rows are sentences and columns are entities in the text (e.g. noun phrases, pronouns, etc.). Each cell of the grid is the grammatical role of the entity in a given sentence. The model considers four types of roles: subject (s), object (o), other (x) and non-present (-).

Each entity grid j, for a document di corresponds to a feature set Φ(xij)={p1(xij),p2 (xij),…,pm (xij)}, where m is the total number of possible transitions and pk(xij) is the probability of transition k occurring in the text. Transitions can be of arbitrary length; however, best results are obtained considering transitions of length 2 (Barzilay, Lapata, 2008), and thus, discourse patterns considered are of the form {ss}, {so}, etc.

The key rationale to use these features is that the distribution of entities in coherent summaries should show some regularities on its entity grid representation, and thus, texts that exhibit these regularities should be graded better than texts that do not exhibit these regularities. For example, a coherent text should have a high density of transitions {--}.

The model will be embedded into the COMPRENDE tool, for reading comprehension activities. The activity works as follows, the student should read a text and answer an open-ended question regarding that text, which is basically a summary of the source text. The student answer is a free text response, thus, the system will do some checks such as a minimum string length and number of sentences, so the system can capture relevant aspects of the student's summary. Finally, the system assesses the text based on two dimensions, content and coherence and gives feedback to the student regarding his/her response. The students may use this feedback to improve their answers and during this process, the hypothesis is that students' reading comprehension skills will improve.

VALIDATION

To validate the tool, experiments were conducted in order to validate two important factors:

1. Does the tool have the required accuracy in assessing content and coherence?
2. Do the coherence and content assessment have an impact in reading comprehension learning?

The first question is relevant because the automated tool results need to be nearly as accurate as a human assessor. The second question is relevant, because the reason for automating the summary evaluation is because it is hypothesized that the learning comprehension tool will have an impact in increasing the reading comprehension level of students.

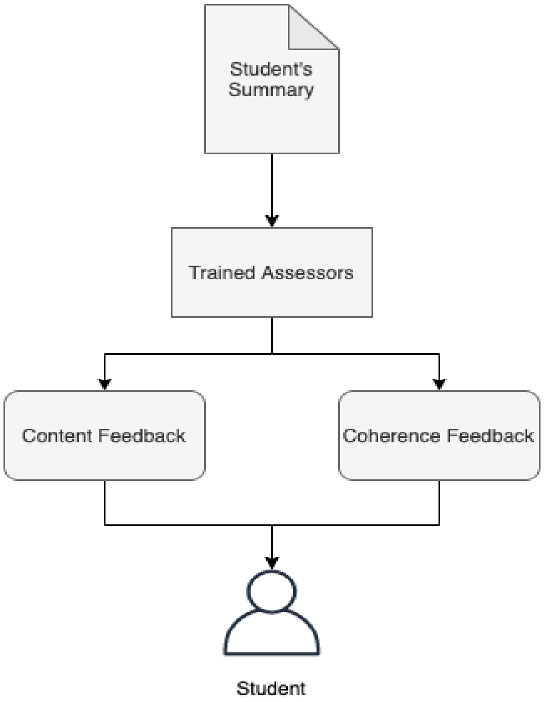

Figure 2. Execution phase of the assessor.

AUTOMATIC SUMMARY EVALUATION ACCURACY

Two different domains were considered and several students' responses were gathered for the experiments. To generate the semantic spaces, a corpus of 1000 documents were collected for each domain. Furthermore, each student response was graded by human experts using a defined criterion to assess content and coherence of responses.

Table 1: Collected data to validate the developed model.

Text Domain	# Student Responses	Size of Corpus
Water	300	1000
Adolescent Games	150	1000

To generate the automated assessors several machine learning models were examined; however, best results were obtained using the Random Forests model. To assess the performance of our model, we used cross-validation (k = 10). To assess performance of the developed model, we considered accuracy, which is computed by the ratio between correctly assessed responses (same score that a human expert would have assigned) and total number of responses. Table 2 summarizes results. The system obtained over 0.8 accuracy on assessing content and close to 0.7 accuracy on assessing coherence. Lower content responses were obtained in adolescent games, and this is explained because fewer texts were available for this domain. We theorize that gathering more data from students could improve the content model accuracy on this domain. Regarding coherence model accuracy, in both domains it obtained similar results. A possible explanation for this is that how humans perceive coherence cannot be captured solely by relying on the entity grid model because the model is limited only to entity distributions across the text. To improve this accuracy, the model can be enhanced by considering other linguistics resources that trigger coherence to some degree, such as discourse markers.

Table 2: Performance results of the developed models.

Text Domain	Content Model Accuracy	Coherence Model Accuracy
Water	0.9	0.65
Adolescent Games	0.75	0.63

CONCLUSIONS

In this work, we propose a novel method that assesses summaries based on multiple dimensions in a reading comprehension tool. The approach combines semantic model-based features and centering techniques in order to predict the summary's quality using a random forest model.

Experiments showed that our proposed method achieves results with high accuracy when compared to human assessment, even though the predicted measurements are as complex as text coherence. This result may be due to the similarity with the task performed by humans such as a human reading the response and assessing based on content and based on intrinsic properties that are perceived as the coherence of a text.

These results show promising possibilities in this type of techniques to generate reading and writing evaluations and providing feedback to students. Normally, this type of technique allowed the learner to help only with tasks of filling in sentences and short answers. However, now we can take a step forward, allowing the evaluation of more elaborative comprehension processes, capturing more complex aspects of the student´s answer, such as the global coherence of comprehension. The merit of this tool is obtaining a high level of accuracy compared with human`s evaluation, using both content and coherence measures. This finding is not just important to future educational tools that train literacy abilities but it also helps to generate a computational modelling about mental processing involved in complex tasks such as summarizing.

Future work should aim to enhance the system to more domains, and also propose new predictors to help the system learn the scoring model. Examples of these enhancements are using features such as discourse markers and readability indices, among others. Also, the machine learning strategy can be enhanced using more complex models such as neural networks.

ACKNOWLEDGMENTS

Thanks to the FONDEF ID20I10290 project: "Literator: An intelligent tutor that enhances reading and writing skills" funded by ANID, Chile.

REFERENCES

Barzilay, R., and Lapata, M. (2008). Modeling local coherence: An entity-based approach. Computational Linguistics, 34(1), p.1–34.

Breiman, L. (2001). Random Forests. Machine Learning, [online] 45(1), pp.5–32. Available at: https://link.springer.com/article/10.1023/A:1010933404324.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D. and Wolff, S. (1998). COMPUTER ANALYSIS OF ESSAY CONTENT FOR AUTOMATED SCORE PREDICTION: A PROTOTYPE AUTOMATED SCORING SYSTEM FOR GMAT ANALYTICAL WRITING ASSESSMENT ESSAYS. ETS Research Report Series, 1998(1), pp.i–67.

Grosz, S. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. Computational Linguistics, 21(2), p.203–225.

Landauer, T. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. Automated essay scoring: A cross-disciplinary perspective.

Miltsakaki, E. And Kukich, K. (2004). Evaluation of text coherence for electronic essay scoring systems. Natural Language Engineering, 10(1), pp.25–55.

Palma, D. and Atkinson, J. (2018). Coherence-Based Automatic Essay Assessment. IEEE Intelligent Systems, 33(5), pp.26–36.

Peng, X., Ke, D., Chen, Z. and Xu, B. (2010). Automated Chinese Essay Scoring using Vector Space Models. 2010 4th International Universal Communication Symposium.

San Martin, E. (2012). ¿Cuán relevante es el aporte de diversos usos de TIC para explicar el rendimiento lector en PISA? Modelando el aporte neto TIC en Chile, Uruguay, España, Portugal y Suecia.

OECD, E. (2014). PISA 2012 results: What students know and can do–Student performance in mathematics, reading and science. Paris: OECD.

Palma, D. and Atkinson, J. (2018b). Coherence-Based Automatic Essay Assessment. IEEE Intelligent Systems, 33(5), pp.26–36.

Peng, X., Ke, D., Chen, Z. and Xu, B. (2010). Automated Chinese Essay Scoring using Vector Space Models. 2010 4th International Universal Communication Symposium.

PISA 2015 Results (Volume I). (2016). PISA. OECD.

San Martín, E. (2012).¿Cuán relevante es el aporte de diversos usos de TIC para explicar el rendimiento lector en PISA? Modelando el aporte neto TIC en Chile, Uruguay, España, Portugal y Suecia.

Soto, C., Gutierrez de Blume, A.P., Rodríguez, M.F., Asún, R., Figueroa, M. and Serrano, M. (2019). Impact of Bridging Strategy and Feeling of Knowing Judgments on Reading Comprehension Using COMPRENDE: an Educational Technology. TechTrends, 63(5), pp.570–582.

Zupanc, K. and Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. Knowledge-Based Systems, 120, pp.118–132.