ISSN 1870-4069

A Model for Identifying Steps in Undergraduate Thesis Methodology

Samuel González-López[1], Aurelio López-López[2], Steven Bethard[3], Jesús Miguel García-Gorrostieta[2]

[1] Technological Institute of Nogales, Sonora, Mexico
[2] National Institute of Astrophysics, Optics and Electronics, Tonantzintla, Puebla, Mexico
[3] University of Arizona School of Information, Tucson, Arizona, USA
samuelgonzalezlopez@gmail.com, allopez@inaoep.mx, bethard@email.arizona.edu, jesusmiguelgarcia@inaoep.mx

Abstract. Knowledge generation is an important asset of great economic powers, and knowledge societies are a fundamental part in the development of countries. Mexico is a country that is in the process of development and improvement of its education system, according to the Educational Reform promoted since 2012 by the Federal Government. We identified an area of opportunity at the undergraduate level to help improve the writing of students, specifically in draft theses and research proposals. This work focuses its efforts on analyzing with natural language processing techniques the "Methodology" section, an important element for the development of a thesis, that helps the reader to understand if the techniques and data used are appropriate in an investigation. This paper proposes a Model to identify a series of steps in such a section. In addition, preliminary results of a basic exploration of a collected corpus are presented, pre-processing the text to generate a representation according to Language Models. The corpus contains documents of graduate and undergraduate levels in the computer science and information technologies domain. The preliminary results showed that the information extracted from the corpus serves to adequately differentiate the methodologies of both levels.

Keywords: automated text evaluation, natural language processing, corpus creation, language models, methodology analysis.

1 Introduction

Knowledge generation is an important asset of great economic powers, while knowledge societies are a fundamental part in the development of countries. Mexico is a country that is in the process of development and improvement of its education system, as confirmed by the Educational Reform approved in December 2012 by the Federal Government. This reform establishes that the

State will provide the educational materials and methods, the school organization, and the educational infrastructure for the continuous improvement and the maximum educational achievement of the students.

One of the relevant principles in this reform is that all public and private sectors must collaborate so that education improves and achieves a high quality[4]. Under the dynamics of this principle of improvement and in accordance with the educational reform, an area of opportunity has been identified at the undergraduate level, which includes the support in the writing of documents of students who are finishing their educational program, specifically in documents such as theses and research proposals.

In this paper, we present a component, part of a wider project, which seeks to provide a series of tools to the student. Using these tools, the students can analyze and evaluate their texts, and obtain feedback to improve their writings [5]. In this study, we seek to analyze the element of "Methodology". The methodology has steps and procedures used to develop the research which should provide a step-by-step explanation of the aspects necessary to understand and possibly repeat the research [8]. The methodology should include: the techniques and procedures employed, type of research, population studied, the sample, collection instruments and description of the selection of data, description of the validation instrument, and a description of the statistical analysis process.

In this study, first we analyze one of the elements, specifically the sequence of steps. This feature is implicitly related to the first point since, when showing a procedure, it is presumed that there are a series of steps. However, in this work, the proposed method does not analyze the content. In the following list, we show the elements to be identified to indicate a series of steps. In addition, we identify the actions expressed by the verbs used in the methodology, as well as if there is a logical sequence in the use of verbs. In particular, we focus on:

- **Series of steps:** ordered activities to be carried out in the methodology by applying some technique. They are not necessarily expressed by numbering.
- **Verbs:** words that represent the actions to be performed in the series of steps.
- **Logical sequence:** The student is expected to use verbs with a hierarchical order. Example of expected logical sequence: it would appear first the verb "to explore" (in Spanish "explorar") and then the verb "to implement" (in Spanish "implementar").

The paper is structured as follows. In section 2, we present related work in automatic writing assessment and machine learning. The collection collected and used in the first experiments, and the proposed model to analyze a series of steps in student theses is detailed in section 3. The results of the experimentation are shown in section 4. In section 5, we conclude with some final remarks.

[4] http://www.presidencia.gob.mx/reformaeducativa/#sobre-la-reforma

2 Related Work

The Automated Writing Evaluation (AWE), also called Automated Essay Scoring (AES), refers to the process of evaluating and scoring written text using a computer system. This kind of system builds a scoring model by extracting linguistic features (lexical, syntactic or semantic) on a specific corpus that has been annotated by humans. For this task, researchers have used Artificial Intelligence techniques such as natural language processing and machine learning methods. The system can be used to directly assign the score or the quality level of a student's text [4]. The use of AWE systems offers students a way to improve their writing during the document review process.

AWE systems help to reduce the review time dedicated by academic instructors, i.e., they are complementary tools to the reviewer's work. Currently, advances in AWE systems include the use of Natural Language Processing technologies to perform the evaluation of student texts and provide feedback to students. Under this context, the Writing Pal (WPal) system offers instructions and practice based on game theory. The WPal system evaluates the quality of the essay using a combination of linguistic computation and statistical models. The authors of this system selected different linguistic properties that were used as predictors [2]. In a similar way, our work evaluates the quality of the text but focusing on the methodology section.

The Machine Learning approach has been used to assess student essays, with the aim of finding the main topic and conclusion in essays [1]. The authors used two annotators to identify the main topic and the conclusions section. Among the features used to train the algorithm are words and phrases. For example, the phrase "in conclusion" is associated with the conclusions section. Another feature considered is the position of the text. In contrast, in our corpus, we assume that the methodology is clearly delimited by a subtitle. Our proposed method includes sending recommendations to students to improve the methodological approach of their project. In previous work, we have presented a system called TURET (Tutor for the Writing of Thesis -in Spanish "Tutor para la Redacción de Tesis"), which analyzes the lexical richness of seven sections of a thesis, applying natural language processing techniques [7].

In the phrase extraction approach, scientific articles similar to a thesis have been studied. In the work of [9], a sentence extraction method was developed to identify the most relevant phrases, which define the document and differentiate it from other types of documents. In a similar way, our work identifies elements that represent the methodology and capture it through a language model.

3 Method

A collection was created using the ColTyPi site, which stores theses and research projects from the area of information technology. In Table 5, we present the corpus extracted from ColTyPi [6].

The graduate level is composed of Doctoral and Master theses. The Undergraduate level is composed of Bachelor and Advanced College-level Technician

Table 1. Collection of methodology texts.

Sets	Graduate Level	Undergraduate Level
Train	90	56
Test	18	14
Total	108	70

(TSU) theses. The theses and research proposals of the collection have been reviewed at some point by a review committee. ColTyPi includes texts in Spanish in the area of Information Technologies. However, the proposed method does not limit to thesis of this area, but it does have a qualitative cut. In Table 2, we show an example of an item in the collection labeled by an annotator (person with experience in reviewing thesis).

Table 2. Annotated (translated) example of Methodology Collection [Graduate Level].

Element	Methodology	Logical Sequence
1	To develop the proposed work, a set of steps was followed to ensure each of the objectives presented. The following are the needs surpassed for the development of the research: *1.* ***To compile*** *the Bibliographic and detailed analysis of existing disambiguation approaches.* *2.* ***To characterize*** *the language families and their relationship with the Spanish language.*	Yes

Table 2 shows three features that a methodology must include. The "series of steps" in italics, the "verbs" used (in bold), and "YES" in the logical sequence column. Observe that the methodology contains the three elements, however, sometimes the methodology does not show a series of steps and even without a numbering. The collection used in this study has been tagged by two annotators with experience in reviewing theses, reaching F-scores of 0.9 and 0.89 in Set of Steps and Verbs respectively, and Kappa of 0.46 (Moderate) in Logical Sequence.

Figure 1 shows the main components that are proposed to identify a series of steps in the Methodology section. The central part of the Analyzer is the language model, which captures the main characteristics of the different methodology sections in the collection. Below, we detail the components used in the proposed method.

Transitional Devices (TD): During the corpus analysis, the use of specific terms have been observed in the methodologies. These devices work as a connection bridge between sentences. The online writing lab at Purdue University [5] identifies the following categories: add, compare, test, show an exception, show

[5] https://owl.english.purdue.edu/owl/resource/574/02/

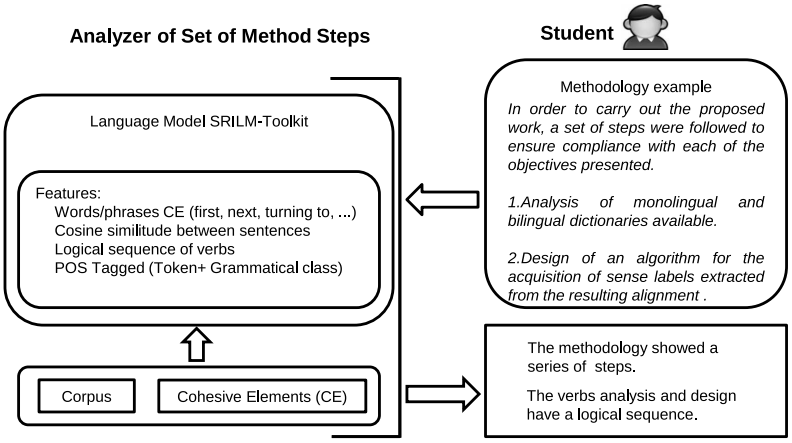

Fig. 1. Analyzer Model "Series of steps in Methodology".

time, repeat, emphasize, show sequence, give an example, summarize. We focus on the "Sequence" category, presented in Table 3.

Table 3. Transitional Devices in Methodology.

Category	Transitional Devices
Sequence	First, second, third, next, then, following, at this point, now, before, later at this moment, subsequently, finally, consequently, previously simultaneously, therefore.

For example, the cosine similarity between the two weighted term vectors of sentences in (1) is 18.26%. In a series of steps, it is expected to find this type of result, thus giving a series of transitions of similarity between the steps.

(1) 1. To compile the Bibliographic and detailed analysis of existing disambiguation approaches.
2. To characterize the language families and their relationship with the Spanish language.

Logical Sequence of Verbs: This feature captures the order of the verbs used in the series of steps. We used the Bloom Taxonomy , consisting of three hierarchical models (cognitive, affective and psychomotor). We employed the cognitive model. In the work of [3], they proposed an adapted taxonomy for the computer tools implementation.

Part of Speech (PoS) Tagging (Term+Grammatical Category): The grammatical position of each term within sentences captures some patterns in the use of the terms, specifically in Methodology. For example, the use of a verb

(possibly appearing at the beginning of sentence) that indicates the action to be performed in the series of steps.

Language Model "Series of Steps in Methodology": The Statistical Language Model extracts models that estimate the probability of word sequences. In our approach, each training element of the language model contains the features: transition devices, cosine similarity, logical sequence of verbs and the grammatical categories (PoS tags). The objective of the model is to capture information that identifies a series of steps. To build the model, we employed the SRILM [6] tool. With the trained model, we expect to evaluate new methodology formulations of students. It is worth mentioning that each feature listed provides a value, which is part of the vector of features that the model takes for training.

4 Results

In the first stage of development of the proposed model to identify a series of steps, the corpus validation was carried out. This validation was with the objective of identifying if the corpus of graduate and undergraduate levels differed. Otherwise, the corpus would not be useful to construct the language model with the features described in Section 3. In this first experiment, we used the corpus detailed above and the SRILM tool for the model construction.

The first step was the construction of three language models, taking into consideration only one of the characteristics of PoS Tagging. The first model was built using the graduate level training group consisting of 90 elements, the second language model was built with 56 elements of the undergraduate level, and the third group was built with 70 conclusions. The third group was created to test the behavior of models 1 and 2 with different texts, out of the corpus of methodologies.

Each methodology element was lemmatized, to obtain the root of the word using the FreeLing tool. The purpose of lemmatizing is to obtain more coverage when training the language model and minimize the burden of processing. We employed sequences of 2, 4 and 8 terms (n-grams) for the training with a smoothing value of 0.01 (to avoid the effect of null probabilities). The bigrams model (two terms) achieves the best performance, having the lowest values of perplexity. A low value of perplexity indicates that the language model is capturing the analyzed sequences.

The model comparison was developed by evaluating the test sets in the models, with the hypothesis that each model would better identify the test set corresponding to its level. Figure 2 shows the results achieved when evaluating the test groups in each model.

In Figure 2, we observe that the graduate model and the undergraduate model have a similar performance (25 and 28 of perplexity, respectively). Between these two models, the graduate model has a lower average perplexity in tests performed with graduate texts, while the undergraduate bigram model has

[6] www.speech.sri.com

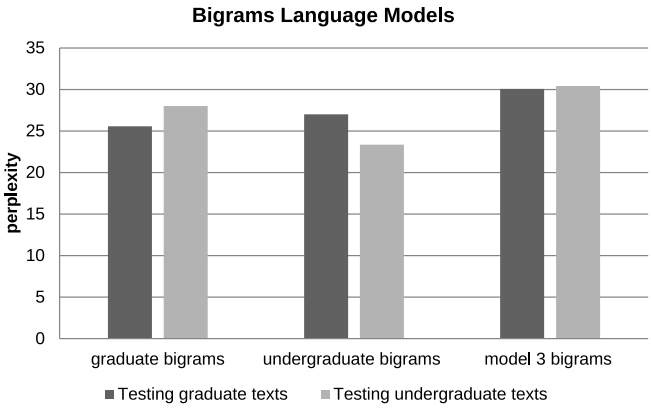

Fig. 2. Comparison chart between bigram models.

a lower perplexity in tests with undergraduate texts. This result is expected since each model obtained low perplexity with the sets of the same level, that is, the models are recognizing their respective level. The third model shows the highest average perplexity (30 and 31 of perplexity) with both graduate and undergraduate texts, which indicates that the models can be differentiated between the texts of each level of different texts.

In Figure 3, which shows perplexity for each undergraduate text in the test set, we can notice that the "bigrams 3 model" (in green), obtains high values which indicates that is not capturing the sequences as well as the "bigrams model of undergraduate", which achieves lower values of perplexity. Therefore, bigrams model of undergraduate responds better to its corresponding set, that is, of undergraduate level.

5 Conclusion

In this paper, we presented a method to evaluate the methodology section of students writings. The results of these first experiments give us a guideline for the development of further methods. We observed that although the distances between the perplexity results of the models were small, they were however consistent in the test set. Therefore, the differences found between graduate and undergraduate groups with the n-gram language model gives evidence that in a future stage of experimentation, the proposed model will be feasible.

During the qualitative analysis of the methodology statements of both levels, we observed that some of them did not present a list of numbered steps, rather, a writing in separate paragraphs. Our model expects to identify this type of methodology statements since transitional devices are expected to be found.

Finally, in future work, we expect to develop the proposed method using the annotated corpus to perform agreement tests between the annotators and

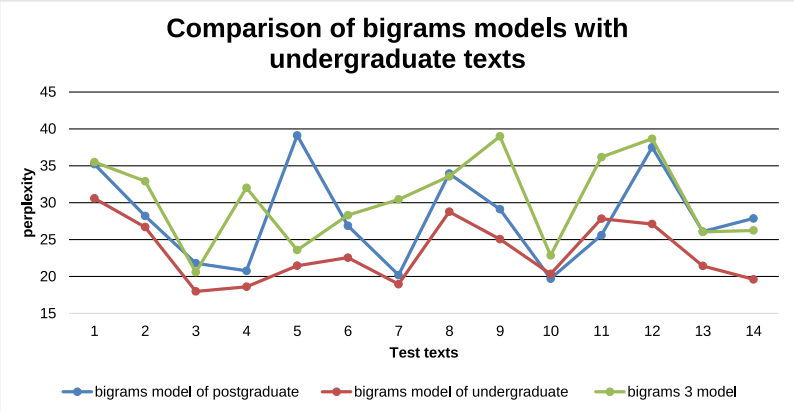

Fig. 3. Comparison between models with the 14 undergraduate texts in the test set.

the analyzer. We also plan to integrate the proposed method into a TURET2.0 system to analyze and evaluate student drafts online.

References

1. Burstein, J., Marcu, D.: A machine learning approach for identification thesis and conclusion statements in student essays. Computers and the Humanities 37(4), 455–467 (2003)
2. Crossley, S.A., Varner, L.K., Roscoe, R.D., McNamara, D.S.: Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In: International Conference on Artificial Intelligence in Education. pp. 269–278. Springer (2013)
3. Fernández-Sánchez, P., Salaverría, A., Perez, E.M., Valdes, V.: Taxonomía de los niveles del aprendizaje de la ingeniería y su implementación mediante herramientas informáticas pp. 522–527 (2012)
4. Gierl, M.J., Latifi, S., Lai, H., Boulais, A.P., De Champlain, A.: Automated essay scoring and the future of educational assessment in medical education. Medical education 48(10), 950–962 (2014)
5. González López, S., López-López, A.: Supporting the review of student proposal drafts in information technologies. In: Proceedings of the 13th annual conference on Information technology education. pp. 215–220. ACM (2012)
6. González-López, S., López-López, A.: Colección de tesis y propuesta de investigación en tics: un recurso para su análisis y estudio. In: XIII Congreso Nacional de Investigación Educativa. pp. 1–15 (2015)
7. González-López, S., López-López, A., García-Gorrostieta, J.M., Espinoza, I.R.: Turet2. 0: Thesis writing tutor aimed on lexical richness in students' texts. Intelligent Learning Environments 129, 9–17 (2016)
8. Muños Razo, C.: Como elaborar y asesorar una investigación de tesis. Libertad: Pearson México (2011)
9. You, W., Fontaine, D., Barthès, J.P.: An automatic keyphrase extraction system for scientific documents. Knowledge and information systems 34(3), 691–724 (2013)