Analysis of Key Features in Conclusions of Student Reports

Aurelio López-López[1,*], Samuel González-López[2], and Jesús Miguel García - Gorrostieta[1]

[1] National Institute of Astrophysics, Optics and Electronics, Department of Computer Sciences, Santa María Tonantzintla, Puebla, México
{allopez,jesusmiguelgarcia@inaoep.mx

[2] Technological University of Nogales, Department of Information Technologies, México
{samuelgonzalezlopez}@gmail.com

Abstract. This work seeks to help students in improving their first research reports, based on natural language processing techniques. We present a Conclusion model that includes three schemes: Goal Connectedness, Judgment and Speculation. These subsystems try to account for the main expected features in conclusions, specifically the Connectedness with the general objective of the research, the evidence of value Judgments, and the presence of Future work as a result of the student reflection after the inquiry. The article details the schemes, a validation of the approach in an annotated corpus, and a pilot test with undergraduate students. Results of a prior validation indicate that student writings indeed adhere to such features, especially at graduate level. Statistical results of the pilot test showed that undergraduate students in an experimental group achieved improved conclusion content when compared with the control group.

Keywords: natural language processing · automated text evaluation · conclusion formulation · goal connectedness · reports assessment.

1 Introduction

A student report is a document describing the student's research and main findings on a topic. Often such report is further developed into a larger student thesis. Such document requires usually the guidance of an advisor. One study focused on the perceptions of students concerning difficulties when writing the discussion section of reports [1]. The study used in-depth interviews with supervisors and students (including L2) and found that pupils mentioned the uncertainty about what content to include and how discussion sections should be organized. This was surprising, considering the time and feedback that students received from supervisors.

In this paper, we focus on evaluating the conclusion section of student reports and perform a pilot test with undergraduate students. These are part of a larger project that aims to help students to evaluate their early drafts and facilitate the

M. Chang et al. (eds.), *Foundations and Trends in Smart Learning*,
Lecture Notes in Educational Technology, https://doi.org/10.1007/978-981-13-6908-7_5

review process for the academic advisor. Besides, the review time can be reduced improving the quality of feedback provided by the instructor, through allowing the reviewer focusing on the conclusions content [5].

In a conclusion section, a discussion of the results is expected, and students are required to reflect on the whole research work. A good conclusion section should include: an analysis of compliance with the research objectives, a global response to the problem statement, a contrast between the results and the theoretical framework, areas for further research, and an acceptance or rejection of the established hypothesis [2]. A pattern that summarizes what is expected in a conclusion section is provided by the Teaching and Learning Centre at University of New England, Australia (UNE). The pattern goes from the specific to the general, and begins with a reformulation of the problem, followed by key findings, and ending with recommendations and future work. The guide pattern is similar to the conclusion of a scientific article, but more extensive.

In the conclusion pattern, the conclusion starts by pointing to the problem solved. In the five-paragraph essay paradigm [3], the introduction and conclusion share the main topic, namely, the subject matter of the essay. The approach is like the conclusions section, as the conclusion should be related to the general objective (considering methodological guides), in its first paragraph. In the intermediate paragraphs, the student must express his thoughts and opinions, avoiding a list of results. The Online Writing Lab at Purdue University provides an outline for writing conclusion sections, emphasizing that the conclusion must contain well-argued viewpoints and avoid inclusion of additional items that are not contained within the thesis [4]. Future work and recommendations included in the conclusion evidence that the student has gone beyond solving of the immediate problem and can identify possible expansion and implications of the work.

Based on the previous pattern and mentioned desirable features, we aim to use an automatic analysis of conclusions intended to obtain a first diagnostic of frequent problems in student's conclusion writings. For this purpose, we formulate this analysis in terms of three main subcomponents (schemes) that identify the following features of conclusions: *Goal Connectedness, Judgment and Speculation.* Due to the complexity of the task, this work only focuses on the conclusions section, besides of being a key section in a thesis or project.

We propose a system with a central Conclusion Model, integrating the three schemes and for this, we take advantage of a corpus to acquire the reference knowledge, to obtain the best features and set score thresholds. After evaluation of a conclusion supplied for analysis, our system will send the result to the student, with the goal of showing him the diagnosed level reached by the conclusion. The student will be able then to improve his conclusion based on the diagnosis, before submission to the advisor.

We report the use of the three features to assess a corpus tagged by annotators, to validate them, once they have been implemented in a computational tool. In addition, we present the results of a pilot test with undergraduate students of engineering, revealing a correlation between Goal Connectedness and Judgment characteristics. Such outcome provides evidence that students are indeed connecting their value judgments with the general objective.

2 Related Work

Automated Writing Evaluation (AWE) of student texts, also called Automated Essay Scoring (AES), refers to the process of evaluating and scoring written text using a computer system. Such a 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, the researchers have been using artificial intelligence techniques such as natural language processing (NLP) and machine learning. The system can be used to directly assign a score or a quality level to a student text [6]. The use of AWE systems offers students ways to improve their writing in an automated manner, and helps to reduce review time required by academic advisors and is a complementary tool to their work.

Currently, the advances in AWE systems include the use of natural language processing technologies to perform the evaluation of texts and provide feedback to students. In this context, the system Writing Pal (WPal) offers strategy instruction and game-based practice in the writing process for developing writers. WPal assesses essay quality using a combination of computational linguistics and statistical modelling. Different linguistic properties were selected as predictors [7]. Similarly, our work seeks to assess the text features focusing on the conclusion section of a research report, considering three schemes to evaluate it.

In [8], the aim was to distinguish differences between low and high scoring essays of undergraduate students. They used the Coh-Metrix tool and found that essays with a higher score reflected more sophisticated language and text complexity. In addition, using a holistic approach of quality text in [9], the authors conducted an analysis of four features that together evidence the presence of the construct “idea generation” in student essays. Fluency, flexibility, originality, and elaboration were the elements analyzed. The corpus consists of essays written in 25 minutes by first-year undergrad students, without using external references. The essay assessment was done by different AWE tools such as Writing Assessment Tool, and Tool for the Automatic Assessment of Cohesion. The results obtained indicate that essays with many original ideas (flexible and elaborated) obtained a high evaluation and were significant features for determining the quality of essay. In our work, we evaluate elements of a conclusion, as those described in the pattern, with the aim to help students improve their writings. Similarly, as the work described previously, we identified that the conclusions of graduate level obtained high values of connection to the objective, these being more extensive than those of undergrad level.

We found in a collected corpus that conclusions that obtained high values (Goal Connectedness/Judgment/Speculation) after the evaluation corresponded to graduate students, using a corpus of research proposals and theses. These results suggest that graduate students with better writing skills (lexical richness) [10] also achieved satisfactory results in the features examined in conclusions. Hence, the students who successfully completed a master or doctoral degree seem to possess better writing skills than students of college level. In addition, the result of a pilot test supported the conclusion that the experimental group students obtained better results than those in control group, when guided in the conclusions preparation.

3 Methodology and Corpus

The first step of our study was the creation of a subcorpus of the Coltypi 1.0 collection (coltypi.org) which contains student theses, project and research reports. Coltypi includes documents of Graduate level: Master (MA) and Doctoral (PhD) degree; and Undergraduate level: Bachelor (BA) and Advanced College-level Technician (TSU) (a two-year technical study program offered in some countries). The corpus domain is computing and information technologies. Each item of the collected corpus is a document (in Spanish) evaluated previously by a committee.

Table 1. Text Corpus (words in average).

Level	Objective-Conclusion	Words in Conclusion	Words in Objective
Doctoral	26	584	37
Master	126	577	35
Bachelor	101	419	44
TSU	59	353	40

We gathered for each conclusion of the collection the associated general objective. In total, we had 312 conclusions and 312 objectives (see Table 1). Also, we can notice that on average the conclusions of graduate level are longer than those of undergraduate level. However, the objective section tends to be shorter than conclusions section. To validate our model, 30 conclusions were selected with their corresponding objectives, 15 of bachelor and 15 of TSU level. Each conclusion was manually reviewed for the three elements by annotators.

The annotation process included two annotators, marking the text that reveals the presence of Goal Connectedness, Judgment and Speculation. Each of our annotators had experience in theses review. Next, we show some sentences of undergraduate objective-conclusion tagged by the annotators.

Goal Connectedness (GC) text marked by annotators in a conclusion section:

As we noted earlier, each driver manufacturer has a different method of accessing the internal information, therefore for this reason, the software designed *should be adapted to the driver manufacturer, considering slight changes in the routing of the items (variables) located within the controller memory.*

Speculative text marked by annotators in conclusion (SsP):

Furthermore, as recommendation observe that the GUI can be modified at any time with the right software, with the use of the OPC library (open technology).

For Judgment Model the annotators only write: *Yes or Not presence of Judgment*

The annotator task is complex since each academic reviewer has his own criteria for tagging, adding a certain level of subjectivity to the task. The Kappa agreement between annotators for Goal Connectedness was 0.923 which corresponded to "almost perfect" [11]. For Speculation was 0.650 which corresponded to "substantial". Finally, for Judgment, the agreement was 0.72 (also "substantial").

4 Model Overview

The second step was the construction and model evaluation for the conclusion section. Our Model has a Conclusion Analyzer, which contains three main schemes (see Figure 1) and seeks to help students with little or partial experience in drafting conclusions, to assess the elements that academic advisors deem important. In addition to the Conclusion Analyzer displayed on our model, we also include student feedback and recommendations. The suggestions are provided to the student, depending on the level reached in each of the features evaluated. Each of the recommendations was formulated by our annotators, which are higher education instructors with experience in research report and thesis review.

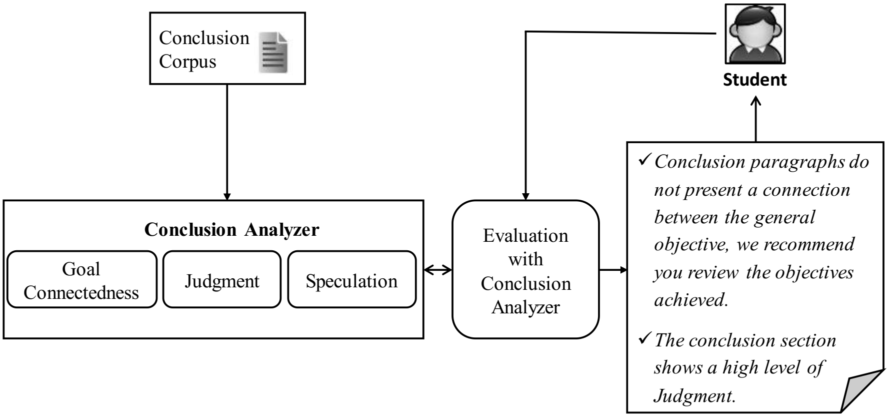
Conclusion Corpus
Student
Conclusion Analyzer
Goal Connectedness
Judgment
Speculation
Evaluation with Conclusion Analyzer
✓ Conclusion paragraphs do not present a connection between the general objective, we recommend you review the objectives achieved.
✓ The conclusion section shows a high level of Judgment.

Fig. 1. Model for Conclusion Assessment

Goal Connectedness Scheme (GC): This scheme seeks to identify whether the conclusion shows some connection with the general objective. We expect that some sentences display this relation. So, we target such relations looking for the sentence that best cover the objective. In the first step, we remove function words in input documents, i.e., in conclusion section and general objective. Function words, also called stop words, include prepositions, conjunctions, articles, and pronouns. Also, each term was stemmed with FreeLing (nlp.lsi.upc.edu/freeling), a library of automatic multilingual processing functions, that provides analysis and linguistic text tagging. For the conclusion section, we used a group of sentences, while in objectives we used the full text, i.e. we consider an objective as one sentence. For computing the Connectedness feature, we do it in terms of coverage, applying the expression in Table 2. To evaluate the GC, we processed each of the objective-conclusion pairs with the Goal Connectedness scheme and the result was placed in a scale. To build the scale, the graduate texts were used as a reference, i.e., we processed each objective-conclusion pair, and after that, the average of all results was computed. However, to smooth out the scale, a group of 50 elements of undergrad level was included (selected at random). Finally, to validate the scale, we used the corpus tagged by annotators. After evaluation of the annotated corpus (30 objective-conclusions), we computed the Fleiss Kappa between our analyzer and annotators, obtaining a 0.799 value, corresponding to a "substantial" agreement.

Judgment Scheme (JS): The goal of this scheme is to identify whether the conclusion section shows evidence of opinions. To consider terms that reflect an opinion or value judgments, we turned to SentiWordNet 3.0 since there is no such extensive resource for Spanish. The tool is a lexical resource for English, which aggregates an opinion score to each term (e.g. noun, adjective) depending of the sense. The sense has three numerical scores for objectivity, subjectivity and neutrality. The range of values is between 0 and 1. Each conclusion was translated to English employing Google Translator (A study of four services using Spanish to English translation showed that Google was superior [12]). After translation, empty words were removed and the value for each sentence was computed. To obtain the measure of each sentence, we search each term in SentiWordNet. To evaluate the JS, we took again the graduate level texts as reference to define a scale. However, in this case, we do not smooth, as we have three levels of opinion. For this feature, the conclusions must reach the average level of review, this will give evidence that the student is expressing judgments and opinions in the conclusion paragraphs. We computed the Fleiss Kappa between the results of our analyzer and annotators (30 objective-conclusions pairs), reaching 0.65, a "substantial" level.

Speculation Scheme (SpS): The model aims to identify evidence of sentences that describe future work or derivations of the research. For this purpose, we resort to two lists of speculative terms. The first list includes lexical features provided by [13]. The second list was obtained from the "Bioscope corpus", consisting of three parts, namely medical free texts (radiology reports), biological full papers, and biological scientific abstracts. Both lists are independent of our corpus. The dataset contains annotations at the token level for negative and speculative keywords [14]. The corpus was tagged by two independent linguists following guidelines. After extraction of speculative terms, we combined the two lists, with the goal of gathering a more exhaustive list. Each term of the merged list was translated, producing a list of 227 speculative terms. To evaluate the Speculative feature, we processed each of the conclusions counting the speculative terms in each sentence. Only the coincidence level between the text marked by the annotator and the sentence with maximum number of Speculation terms was described. After analyzing the annotated pairs using the criterion described, we computed the Fleiss Kappa between the results of our analyzer and the annotators (30 pairs), obtaining a result of 0.887, i.e. “almost perfect” agreement.

Table 2. Text Corpus.

Model	Parameters	Expression
Goal Connectedness	Absence of connection <0.12 0.12 < Acceptable < 0.41 Strong connection > 0.41	$C = \frac{\#(So \cap SC_i)}{N}$ C = Coverage *So* = List of words in objective SC_i = Sentence i of conclusion *N* = Number of terms in the objective

Judgment	No Judgment < 7.84
	Yes, presence of Judgment
	> 7.84 < 26.98
		$T = \sum Wi \left(\frac{On + Op}{N}\right)$
		T = Score*
		On = Negative Score; Op = Positive Score
		N = Number of occurrences (noun, pronoun)
		Wi = each word of sentence

5 Conclusion Analysis in Practice

After the corpus exploration and evaluation of methods to assess conclusions, an online system was developed with the goal of validating the models and identifying if the tool could help students to improve their writings. The computational tool TURET2.0 (In Spanish: Tutor Revisor de Tesis) is hosted at tutor.turet.com.mx. Any student can register and use the system. In addition, TURET2.0 has a section that explains its use and provides support material for the student. The support material gives the student an explanation of the elements evaluated by the system.

In the system interface, the student submits the objective and conclusion of his report. Subsequently, the system sends the results of the analysis back to the student indicating if the score reached is acceptable. The student can repeat the analysis and each attempt is recorded. In case of no evidence of Judgment, the system provides the following text “Opinion is very important in a conclusion, to achieve an acceptable level of judgment, improve the conclusion by incorporating sentences that contain your value judgments”. In case of Goal Connectedness was strong, the system sends the message “The connection value is strong between your objective and your conclusion. Congratulations, you have achieved an excellent score”. The system was created with Django, Python, and libraries for text analysis.

5.1 Pilot Test

We designed and performed a pilot test to assess the impact/benefit of using an online application focused on Goal Connectedness, Judgment and Speculation in a conclusion section of a research report. The experiment involved undergraduate engineering students. Also, we considered two randomly selected groups, one experimental, and other for control, each with 15 students. The two groups received instructions on how to write a conclusion section. Students were informed of each key feature, using the triangle pattern of conclusion section. The control group had a traditional monitor, that is, an academic advisor reviewing their documents, while the experimental group had access to the intelligent tutor 24 hours a day. All documents produced by both groups were evaluated with TURET2.0 to compare the results among them. The foremost hypothesis to be validated in this pilot test was: "The use of an online application, allow students in the experimental group generate documents with better parameters, in terms of the features.

One can notice that the experimental group produced higher values on each feature than control group. These results provide evidence that students of experimental group reach twice the values of measures. It was also observed in the experimental group that on average, the number of attempts of TURET2.0 use was 8. However, when we observed the standard deviation, in the control group we found that it was lower than the experimental group. This could indicate that the control group is more uniform in performance. It is possible that in the experimental group some students using a technological tool (TURET) allow them to achieve superior results, while other students have an average performance on the test. Also, we performed a statistical analysis to validate the results. We applied a hypothesis test for two independent samples with different standard deviation.

The confidence level was 95%. We carried out the hypothesis test for each measure. For the three features, the null hypothesis was rejected with P-values of 0.046 (Goal Connectedness), 0.020 (Judgment), and 0.024 for Speculation feature. These statistical results indicate that the null hypothesis is rejected for the three characteristics. The TURET2.0 system allowed students to achieve higher measures than the students in the control group.

In addition, a correlation analysis was performed among the three characteristics in the two groups. In Table 3, we can observe a correlation of the experimental group which is quite close to the correlation identified in the annotated corpus. The characteristics of Goal Connectedness-Judgment show a positive correlation with significance in the annotated corpus and in the experimental group, i.e. a value of 0.609. The result of Goal Connectedness-Speculation shows that there is no correlation, as is the case of the annotated corpus. We can assert that the students wrote conclusions with a closeness to the pattern of conclusions, since the correlation numbers were close to those of the annotated corpus.

Table 3. Experimental and Control Group Correlations

Features	Experimental Group		Control Group	
	Correlation	P-value	Correlation	P-value
G. Connectedness-Judgment	0.609	0.016	0.36	0.187
Judgment-Speculation	0.535	0.04	0.042	0.881
G. Connectedness-Speculation	0.223	0.424	0.339	0.216

For the students of the control group no correlations were found, which indicates that control students should continue working with the writing of their conclusions, to reach acceptable values. We also applied a satisfaction survey based on Technology Acceptance Model [15] to assess the opinion of the experimental group on using the online analyzer, in the aspects of usefulness, ease of use, adaptability and intention to use the system. Students answers were based on a five-point Likert scale ranging from 1 ("Strongly disagree") to 5 ("Strongly agree").

The average results were: 4.46 for system usefulness, 4.33 in system ease of use, 4.25 in system adaptability and 4.11 for intention to use the system. That is, the four aspects of the survey were above 4 points ("Agree"), so we can conclude that the analyzer was found useful, easy to use, adapted to their level and students have the intention to use it. However, in student comments, it was found that some of them felt the registration was complex, primarily because of its registry process.

6 Conclusions

We have presented a Model that uses NLP techniques to evaluate the conclusion section, that was designed to consider specific features of writing. Our model could help improve the writing of research report by undergrad students or inexperienced learners, regarding Goal Connectedness and Speculation, since the achieved Kappa levels were substantial or better.

The pilot test with engineering students in the systems area allowed us to bring the developed schemes to a real environment. We can identify, as a result of the pilot test, that the students of the experimental group showed interest in using the tool and improving their writing. Such interest was observed in the average number of the times that the students used TURET2.0. However, it could have also been due to the competition generated amongst the students of the experimental group when using the system, as results can be improved when using the tool.

One of the constructs that were best evaluated in the satisfaction survey was the usefulness that motivates us to continue with this project. The intention to use construct was the lowest, so strategies to increase this metric were sought, for example, the incorporation of serious games strategies [16]. The results of the correlation analysis between the two groups (control and experimental) validated to some extent the similarity with the pattern of conclusions detailed in the introduction. One finding was that the Goal Connectedness and Judgment measures showed a positive correlation with significance, such as that found in the annotated corpus, where the documents were theses or research projects reviewed previously by a qualified committee.

Furthermore, we are also planning to include metrics to assess whether a conclusion contains a certain level of originality and elaboration [17]. The working hypothesis is that the conclusions of graduate level contain more original ideas.

References

[1] Bitchener, J., Basturkmen, H.: Perceptions of the diculties of postgraduate l2 thesis students writing the discussion section. Journal of English for Academic Purposes 5(1), 4-18 (2006)

[2] Allen, G.R.: The graduate students' guide to theses and dissertations: A practical manual for writing and research. (1973)

[3] Davis, J., Liss, R.: Efective academic writing 3. Oxford: Oxford University Press (2006)

[4] Lab, P.O.W.: Introductions, body paragraphs, and conclusions for an argument paper. Website, https://owl.english.purdue.edu/owl/resource/724/04/, consulted January 30, 2016

[5] Debuse, J.C., Lawley, M., Shibl, R.: Educators' perceptions of automated feedback systems. Australasian Journal of Educational Technology 24(4) (2008)

[6] Gierl, M.J., Lati, 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)

[7] 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: Intl Conf. on Artificial Intelligence in Education. pp. 269-278. Springer (2013)

[8] McNamara, D.S., Crossley, S.A., McCarthy, P.M.: Linguistic features of writing quality. Written communication 27(1), 57-86 (2010)

[9] Crossley, S.A., Muldner, K., McNamara, D.S.: Idea generation in student writing: Computational assessments and links to successful writing. Written Communication 33(3), 328-354 (2016)

[10] González-López, S., López-López, A.: Lexical analysis of student research drafts in computing. Comput. Appl. Eng. Educ. 23(4), 638–644 (2015)

[11] Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometric 32(1), 159-174 (1977)

[12] Aiken, M., Ghosh, K.,Wee, J., Vanjani, M.: An evaluation of the accuracy of online translation systems. Communications of the IIMA 9(4), 67-84 (2009)

[13] Kilicoglu, H., Bergler, S.: Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. BMC Bioinformatics 9(11), S10 (2008)

[14] Vincze, V., Szarvas, G., Farkas, R., Mora, G., Csirik, J.: The bioscope corpus:biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 9(11), (2008)

[15] Tobing, V., Hamzah, M., Sura, S., Amin, H.: Assessing the acceptability of adaptive e-learning system. In: 5th Intl Conf. on eLearning for Knowledge-Based Society. vol. 16, No. 3 (2008)

[16] Long, Y., Aleven, V.: Gamification of joint student/system control over problem selection in a linear equation tutor. In: Intl Conf. on Intelligent Tutoring Systems. pp. 378-387. Springer (2014)

[17] Crossley, S.A., Muldner, K., McNamara, D.S.: Idea generation in student writing: Computational assessments and links to successful writing. Written Communication 33(3), 328-354 (2016)