Mining of Conclusions of Student Texts for Automatic Assessment

Samuel González-López, Aurelio López-López

Instituto Nacional de Astrofísica, Óptica y Electrónica, Coord. Cs. Computacionales. 72840 México
{sgonzalez, allopez}@inaoep.mx

Abstract

Writing a thesis involves complying with certain requirements and rules established by institutional guides. So, students have guidelines to follow when developing their first draft. Generally this draft presents deficiencies, which has to be polished with the help of the academic advisor to reach an acceptable document. However, this task is repeated every time a student prepares his thesis, becoming extra time spent by the student and the advisor. Our work aims to help the student improve the writing, based on natural language processing techniques. For the current study, we focus primarily on the conclusions section of a thesis, a central element when concluding the research. In this paper we present a Mining Component that includes three models: Coverage, Opinion and Speculation. Our system seeks to assess a conclusion taking into account the general objective, the evidence of value judgments, and the presence of future work as a result of reflection of the student. We provide initial models and their evaluations.

Introduction

The conclusion of a career by students often involves the elaboration of a thesis or research proposal text. These documents must comply with the drafting characteristics established by institutional guidelines and books of methodology. Moreover, they have to satisfy the appropriate structural features of each section of a thesis. However, the experience of instructors is that these theses exhibit a variety of errors, ranging from misspellings to content faults.

A study about the perception of the difficulties of students when writing the discussion section of thesis (Bitchener & Basturkmen 2006), that applied depth-interviews to supervisors and students, commented about the uncertainty of students about the content that should include the discussion section and how it has to be organized. This information was surprising to supervisors, considering the time and feedback that students received. In the conclusion section, a discussion of the results is expected, and that the students ponder about the whole research work.

In particular, a good conclusion has to include the following features: an analysis of compliance with the research objectives, a global response to the problem statement, a contrast between results and theoretical framework, future research work and acceptance or rejection of the established hypothesis (Allen 1976).

With the aim of diagnosing some common problems when writing conclusions, we developed a system with three main subcomponents (models) that identify the following features of conclusions:

- ✓ *Coverage*: The model seeks to assess whether any of the sentences of the conclusion section have some connection with the general objective.
- ✓ *Opinion*: Value judgments and reflections elaborated by students are key features of a conclusion. With the proposed model in this work, we seek to assess whether the conclusion has an acceptable level of opinion.
- ✓ *Speculation*: Our proposed model identifies speculative terms in conclusion sentences. As a result of the reflections of the research done, we expect that the conclusion shows evidence of future work or possible derivations.

The system has a central Mining Component, which integrates the three models described above. We take advantage of a corpus to acquire the knowledge of reference, to obtain the best features. Then, we use these features to assess the corpus tagged by annotators, as a way to validate the Mining Component. The reported results are part of a larger project aimed to help students to evaluate early their drafts, facilitating so the review process of the advisor.

Background

Automated Writing Evaluation (AWE), also called Automated Essay Scoring (AES), of student texts refers to the process of evaluating and scoring written text using a computer system. This system builds a scoring model by extracting linguistic features on a specific corpus that has

been tagged by humans. For this task, the researchers have been using artificial intelligence techniques such as natural language processing and machine learning. The system can be used to assign directly a score or a quality level to a new student text (Gierl et al. 2014).

The use of AWE systems offers students ways to improve their writing during the review process of documents. The AWE system helps to reduce the review time dedicated by advisors, and is a complementary tool for human reviewer. Currently, the advances in the 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 (Crossley et al. 2013). This AWE system assesses essay quality using a combination of computational linguistics and statistical modeling. Different linguistic properties were selected and used as predictors. Similarly, our work seeks to assess the quality of the text, but focusing on the conclusion section of a thesis.

In the work of (McNamara, Crossley, & McCarthy 2010), the authors looked for distinguishing the differences between essays that obtained a high score and low quality of undergraduate students. They used the Coh-Metrix tool and found that essays with a high quality score showed more complexity of the text and sophisticated language.

System Overview

Our system has a Mining Component, which contains three main models. Coverage model is responsible for identifying whether a conclusion sentence has a connection with the general objective, in terms of the main concepts. This as a way to take into account the recommendations of authors of research methodologies books. Opinion model processes each sentence to identify terms with an opinion load, evidencing the presence of opinions or value judgments formulated by the students. The idea is to help the student to undertake a process of analyzing results and that the conclusion is not just a list of achieved activities. The final Speculation model identifies whether the student expressed future work, or possible derivations of his work.

After evaluation of a conclusion supplied for analysis, our system reports the result to the student with the aim of showing the diagnosed level reached. The student may improve his/her conclusion regarding the result.

Data Description

The corpus contains conclusions of graduate (Master and Doctoral degrees) and undergraduate level (Bachelor and Advanced College-level Technician - a two year technical study program offered in some countries - (TSU) degrees). The domain is computing and information technologies. Each item of the collected corpus is a document (graduate proposal and theses in Spanish) that was evaluated at some point by a reviewing committee. Also, we gathered for each of these conclusions the associated general objective. In total, we have 312 conclusions and objectives (Table 1).

Level	*Objective-conclusions*
Doctoral	26
Master	126
Bachelor	101
TSU	59

Table 1. Corpus

Of the corpus just described, 30 conclusions were selected for validation with their corresponding objectives, 15 of bachelor and 15 of TSU level. Each conclusion was tagged by two annotators. The tagging process included marking the text that reveals the presence of Coverage (gray text) and Speculation (underline text). To assess the Opinion, a scale of three levels was established ("Yes, a lot", "Yes, a little", and "No opinion"). Each of our annotators had experience in the review process of theses. For instance, sentences of an undergrad objective-conclusion pair tagged by the annotators are:

Objective:

S1: *Develop a system of monitoring control and power of light in common areas through a programmable logic controller (PLC).*

Annotated Conclusions:

S2: *It was possible to establish the communication between the software (LabVIEW) and hardware (PLC), to minimize energy used in labs, cubicles and common areas presented.*

S3: *So the power control system based on PLC presented meets the objectives as well as minimizing energy use, is user friendly and may be expanded to multiple cubicles , labs and common areas.*

Opinion level: *Yes, a little*

The Kappa agreement between annotators for Coverage element was 0.92 that corresponds to *Almost perfect*. For Speculation element was 0.65 that corresponds to *Substantial*. For the Opinion scale, the agreement was: 0.47 (*Moderate*), 0.21 (*Fair*), and 0.44 (*Moderate*).

Coverage Model

This model seeks to identify whether the conclusion shows connection with the general objective. We expect that some sentences display this relation. In the first step, we remove empty words from documents of graduate and undergraduate level, in conclusion section and general objective. Empty words, also called stop words, include prepositions, conjunctions, articles, and pronouns. Also

each term was stemmed with the FreeLing tool. For the conclusion section, we used a group of sentences, while in objectives we used the full text, that is we considered an objective as one sentence. For computing coverage, we applied the following expression:

$$\text{Coverage}(C) = \frac{\#(So \cap Sc_i)}{N}$$

where S is a list of words of an objective (So) or a sentence *i* of conclusion (Sc_i), and *N* is the number of terms in the objective. The value of the sentence with highest coverage is kept. The result is in a range from 0 to 1, where a value close to 0 means that sentence is far from the objective.

For example, the Coverage measures for the conclusions sentences given in previous section were 0.25 for **S2** and 0.50 for **S3**, taking S3 as the conclusion discussing the objective.

Evaluation:

For evaluation, we used the corpus tagged by annotators. We processed Coverage of each of the objective-conclusion pair and the result was placed in a scale. To build the scale, the graduate level was used as a reference of Coverage, that is after processing each objective-conclusion pair, the average of all results was computed. However, to smooth out the scale, a group of 50 elements of bachelor level was included (selected at random). Below we show the scale:

Coverage >= 0.12 (Average - 1σ). This indicates that the connection between the objective and the evaluated sentence is acceptable, otherwise is taken as an absence.

Coverage >= 0.41(Average + 1σ). This corresponds to a strong connection. We expect that sentences exceed the minimum acceptable (0.12), giving evidence that the student is properly linking the objective with the conclusion paragraphs.

Finally, after evaluation of the tagged corpus (30 objective-conclusions), we computed the Fleiss Kappa between our analyzer and the annotators, obtaining a result of 0.799, corresponding to *Substantial* agreement.

Opinion Model

The goal of this model is to identify whether the conclusion section shows evidence of opinions. For example:

It was demonstrated that the use of conceptual graphs and general semantic representations in text mining is feasible, especially beneficial for improving the descriptive level results.

We can observe that terms as *feasible* and *beneficial* imply an opinion.

To take into account terms that reflect an opinion or value judgments, we employed SentiWordNet, a lexical resource for English, which associates an opinion score to each term depending of the sense (e.g. noun, adjective), with three numerical values for objectivity, subjectivity and neutrality (each between 0 and 1). Each conclusion was translated to English employing Google Translator (Aiken et al. 2009), and then, empty words were removed and the value for each sentence was computed, searching each term in SentiWordNet 3.0. For instance, the Opinion load measures (non null) in the conclusion given above:

S2: Possible(0.37) make(0.13) communication(0.04) minimize energy(0.21) use(0.07) common(0.29)
Total = 1.11

The term *possible* presents a 0.37 opinion load, this result is computed regarding the average of all opinion loads (as a noun has 2 senses and an adjective has 2 senses). The total displayed is the sum of all terms. We expect that in conclusion (S2+S3) an acceptable load was expressed.

Evaluation:

Similar to the Coverage Model, we took as reference the graduate level texts to define a scale. However, in this case we did not smooth, since we have three levels of opinion. For this element, the conclusion has to reach the average level of review (i.e. "Yes, a little"), this will give evidence that the student is expressing judgments and opinions. Below we show the scale:

- ✓ Opinion <= 7.84 (Average - 1σ), these are conclusions corresponding to the level "No Opinion".
- ✓ 7.84 < Opinion < 26.98, these are conclusions presenting the level "Yes, a little".
- ✓ Opinion > =26.98, these are conclusions that correspond to the level "Yes, a lot".

Regarding the previous example, the sum of **S2** and **S3** (1.11+1.34=2.45) fits with *No opinion* level. This result is close to the "value" assigned by annotators (i.e. *Yes, a little)*.

After evaluation, we computed the Fleiss Kappa between the results of our analyzer and annotators (30 objective-conclusions pairs). We obtained a *Fair* agreement for *Yes, a lot* (0.30), and for *Yes, a little* (0.21). For *No opinion* level (0.46), a *Moderate* agreement was obtained.

Speculation Model

The model identifies evidence of sentences that describe future work or derivations of the research. For this purpose, we merged two lists of speculative terms. The first list includes lexical features provided by (Kilicoglu and

Bergler 2008), that include modal auxiliaries, epistemic verbs, adjectives, adverbs and nouns. The second list, the "Bioscope corpus", consists of three parts, namely medical free texts (radiology reports), biological full papers and biological scientific abstracts. The dataset contains annotations at the token level for negative and speculative keywords (Vincze et al. 2008), tagged by two independent linguists following guidelines. To obtain this list, we extracted from the XML file, the terms tagged as speculation type (e.g. the terms *suggesting* and *could*):

<cue type="speculation" ref="X1.6.2">suggesting</cue>
<cue type="speculation" ref="X1.7.1">could</cue>

After extraction of speculation terms, we combined the two lists, with the goal of gathering a more complete list. Terms that appear in both lists were weighted by 2 and those terms that only appear in a list were given the value of 1. Weighted terms indicate higher speculation in the sentences. Each term of the merged list were translated to Spanish, producing a list of 227 speculative terms.

Evaluation:

To compute the speculation measure, we counted only the number of speculative terms in each sentence of the conclusion (i.e. a scale was not stated), only the coincidence between the text marked by the annotator and the sentence with maximum value of speculation terms.

For instance (conclusion of data section):

S2: The analyzer did not find speculative terms, neither the annotators.

S3: The annotator marked the future work, also our analyzer identified "may" as a speculative term.

Finally, we computed the Fleiss Kappa measure between the results of our analyzer and the annotators (30 objective-conclusions), obtaining a result of 0.887 which corresponds to *Almost Perfect* agreement.

Corpus Mined

We conducted an analysis of the whole corpus using the models described above. The goal was to identify the levels of Coverage, Opinion and Speculation in the graduate and undergraduate levels. The Coverage value is the average of the maximum values of each conclusion of the corpus. The Opinion value is the average of the sum of each conclusion. In Speculation for graduate level, the sentence with the highest speculation (average) was around three terms while the undergraduate level had around two terms.

Level	Coverage	Opinion	Speculation
Graduate	0.3	20.5	3
Undergraduate	0.2	14.5	2

Table 2. Corpus mined

We can notice that the graduate level has better values than undergraduate level (see Table 2). Besides, a significance test was performed for each measure between graduate and undergraduate level (Two-Sample T-Test. $\alpha = 0.05$). For the three features, the p-value was 0.001. These results show that graduate students connect better the conclusion with the objective and express more detail about their judgments, opinions and possible derivations.

Conclusion

In this paper, we have presented a system that uses natural language processing techniques to mine specific features of writing for the conclusion section emphasized by authors of methodology or institutional guides. We found in the three features evaluated that graduate level students texts outperformed those of undergraduate level. This behavior provides evidence that students with more practice writing (graduate level), possess better skills.

We plan to increase the number of examples of the corpus to improve the level of agreement between our system and that of the annotators, specifically for opinion. Moreover, we are planning to conduct a pilot test with students of TSU level, with the aim to verify if our system indeed helps students improving their writing.

References

Bitchener, J. and Basturkmen, H. 2006. Perceptions of the difficulties of postgraduate L2 thesis students writing the discussion section. *Journal of English for Academic Purposes* 5: 4-18.

Allen, G. 1976. *The Graduate Students' Guide to Theses and Dissertations: A Practical Manual for Writing and Research*, USA.: Jossey-Bass Inc Pub.

Gierl, M., Latifi, S., Lai, H., Boulais, A., and De Champlain, A. 2014. Automated essay scoring and the future of educational assessment in medical education. *Medical Education* 48:950-962.

Crossley, S., Varne, L., Roscoe, R., and McNamara, D. 2013. Using Automated Indices of Cohesion to Evaluate an Intelligent Tutoring System and an Automated Writing Evaluation System. In Proceedings 16th International Conference AIED, 269-278. Memphis, TN.: Springer Berlin Heidelberg.

McNamara, D., Crossley, S., and McCarthy, P. 2010. Linguistic Features of Writing Quality. *Written Communication* 27: 57-86.

Aiken, M., Ghosh, K., Wee, J., and Vanjani, M. 2009. An Evaluation of the Accuracy of Online Translation Systems. *Communications of the IIMA* 9:67-84.

Kilicoglu, H., and Bergler, S. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics* 9.

Vincze, V., Szarvas, G., Farkas, R., Móra, G., and Csirik, J. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics 9*.