REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2020 PUCV, Chile 53(104) 643-663

Assessing Thesis Conclusions by their Connectedness with Goal, Judgment and Speculation

Evaluando Conclusiones de Tesis por su Conectividad con el Objetivo, Juicio y Especulación

Samuel González-López
UNIVERSIDAD TECNOLÓGICA DE NOGALES
MÉXICO
samuelgonzalezlopez@gmail.com

Aurelio López López
INSTITUTO NACIONAL DE ASTROFÍSICA, ÓPTICA Y ELECTRÓNICA
MÉXICO
allopez@inaoep.mx

Recibido: 07-VII-2018 / **Aceptado:** 17-XI-2019
DOI: 10.4067/S0718-09342020000300643

Abstract

Writing a thesis involves complying with certain rules and requirements established by institutional guides of universities. Students, often being too inexperienced to create good written documents, have guidelines to follow when developing their first drafts. This study seeks to help students improve their first writings, based on natural language processing techniques. We focus primarily on the conclusion section of a thesis, a central element when completing a research project. In this paper, a conclusion analyzer that includes three models: goal connectedness, judgment and speculation is presented. Such subsystems try to evaluate the main expected features in conclusions, specifically the connectedness with the general objective, the evidence of value judgments, and the presence of future work as a result of the student's reflection. In the study, we provide initial models, internal exploration of conclusions, and evaluations of our approach. We found across the three features evaluated that graduate level student texts outperformed those of undergraduate level. The behavior provides evidence, that students with more practice in writing a scientific paper or thesis (at the graduate level), have better writing skills.

Key Words: Natural language processing, educational data mining, automated text evaluation, goal connectedness, thesis assessment.

Resumen

Escribir una tesis involucra cumplir con ciertos requerimientos y reglas establecidas en las guías institucionales de las universidades. Los estudiantes tienen pautas cuando desarrollan su primer borrador de tesis, sin embargo es insuficiente para obtener un buen documento. Este estudio busca ayudar a los estudiantes a mejorar sus primeros escritos, basado en técnicas de procesamiento de lenguaje natural. Nos enfocamos principalmente en la sección conclusión de una tesis, un elemento central cuando se finaliza una investigación. En este artículo, presentamos un analizador para las conclusiones que incluye tres modelos: conectividad con el objetivo, juicio y especulación. Estos subsistemas tratan de enfocarse en las principales características esperadas en la conclusión, específicamente la conectividad del objetivo con la conclusión, la evidencia de juicios de valor y la presencia de trabajo futuro como resultado de la reflexión del estudiante. El estudio provee los modelos iniciales, una exploración interna de las conclusiones y una evaluación de nuestro enfoque. Encontramos que los textos de los estudiantes de posgrado obtienen mejores resultados que los de pregrado en las tres características analizadas. Este comportamiento da evidencia de que los estudiantes con mayor práctica en la redacción de documentos científicos o tesis de grado (posgrado) poseen mejores habilidades de redacción.

Palabras Clave: Procesamiento de lenguaje natural, minería de datos en educación, evaluación automática de texto, conectividad del objetivo, evaluación de tesis.

INTRODUCTION

A key requirement for candidate of a degree or professional qualification is the completion of a thesis, a document presenting the prospect's research and main findings on a topic. For inexperienced students, the drafting of document requires usually the guidance of an advisor. These advisors often report that the first draft theses of these students exhibit a variety of deficiencies, ranging from spelling errors to serious content errors. A study by Bitchener and Basturkmen (2006), based on in-depth interviews with supervisors and students (including L2 persons) focused on the perception of the difficulties of students when writing the discussion section of thesis, finding in students uncertainty about the selection of content that will be included in the discussion section. The comment was surprising, considering the time and feedback that students commonly received from supervisors.

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 the following features: an analysis of compliance with the research objectives, a global response to the problem statement, a contrast between results and the theoretical framework, future research work and acceptance or rejection of the established hypothesis (Allen, 1976). A pattern that summarizes what is expected in a conclusion section is provided by the University of New England (UNE Academic Skills Office, 2017). 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 (Figure 1). Such pattern is similar to the conclusion of a scientific article, but more extensive.

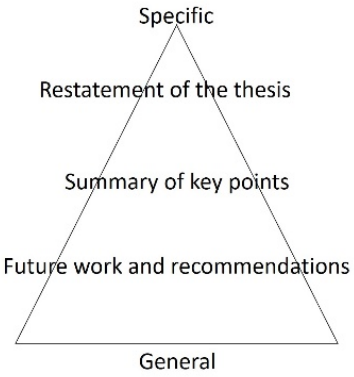

Figure 1. Pattern for conclusions section.

In the pattern of conclusions above, the conclusion starts pointing to the problem solved. In the five-paragraph essay paradigm (Davis & Liss, 2006), the introduction and conclusion share the main topic, this is the theme or subject matter of the essay. The approach is similar to the conclusions section, since the conclusion should be related to the general objective (considering methodological guidelines), in the initial paragraph of the conclusion. In the middle of the triangle, the student must express his/her thoughts and opinions, avoiding a list of results. The Online Writing Lab at Purdue University outlines what to write in a conclusion section, emphasizing that the conclusion must contain well-argued viewpoints and avoid inclusion of additional items that are not contained within the thesis (Purdue Online Writing Lab, 2013). Future work and recommendations included in the conclusion (triangle base) is evidence that the student has gone beyond the solution of the problem and can identify possible important expansions of the work.

Based on the previous pattern and desirable features, we aimed for an automatic analysis of conclusions intended to obtain a first diagnostic of frequent problems in student's conclusion writings. With this in mind, we performed this analysis in terms of three main subcomponents (models) that identify the following features of conclusions:

- *Goal Connectedness*: The model seeks to assess whether some of the sentences in the conclusion section have connection with the general objective. This will reveal that the proposed solution to the problem is discussed.
- *Judgment*: Value judgments and reflections expressed by students are key features of a conclusion. With the proposed model in this work, we attempted to assess whether the conclusion has some level of opinion. The idea is to help the

student to undertake a process of examining his/her results so that the conclusion is more than a list of completed activities.

- *Speculation*: Our proposed model identifies the presence of speculative terms in conclusion sentences. Speculative terms represent the reflections of the research already done by the student, we expect that the conclusion shows evidence of future work or possible derivations of it.

We foresee a system with a central Conclusion Analyzer, which integrates the three models described. We take advantage of a corpus to acquire the knowledge of reference, to obtain the optimal features and set score thresholds. After evaluation of a conclusion supplied for analysis, our expected system will send the result to the student, with the goal of showing the diagnosed level reached by its conclusion. The student will be able then to improve his/her conclusion considering the result, before submission to the advisor. For validation, we report the use of the three features in a corpus tagged by two annotators. In addition, we present an analysis of our corpus on the three features selected for this study, revealing a close relationship between Goal Connectedness and Judgment characteristics. This shows evidence that students are indeed connecting their value judgments with the general objective. The results reported here are part of the project named TURET (in Spanish: Tutor Revisor de Tesis) that aims to help students to evaluate their early drafts, and facilitate the review process for the academic advisor. The review time can be reduced and the quality of feedback provided by teacher to student improved, by allowing the reviewer focusing on the conclusions content (Debuse, Lawley & Shibl, 2008). The results of the analysis in this study were obtained from an analyzer system developed by the authors that includes the three features examined in this work. A first version of the conclusion analyzer is already embedded in TURET2.0 hosted in www.tutor.turet.com.mx.

1. Literature review

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 systems use 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 and machine learning algorithms. The system can be used to assign a score or a quality level to a new text directly (Gierl, Latifi, Lai, Boulais & De Champlain, 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 academic advisors and is a complementary tool to the work of a human reviewer. 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. The AWE system in WPal, assesses essay quality using a combination of computational linguistics and statistical modeling. The authors selected different linguistic properties that were used as predictors (Crossley, Varne, Roscoe & McNamara, 2013).

SciPo is a system that analyzes the rhetorical structure of academic texts in Portuguese, in terms of schematic structure, rhetorical, and lexical patterns (Feltrim, Teufel, das Nunes & Aluísio, 2006). Focused on the computer domain, the system seeks to help novice writers to specifically improve the abstract and the thesis introduction. For instance, in the case of the abstract, the authors identify the components through a sequence, that is, the purpose of the study is expected to appear first, then the methodology employed, followed by the results. However, if the abstract of the thesis includes the background, this could confuse the reader. SciPo system suggests the student a reasonable sequence.

A tool developed to support students is e-Rater (Attali & Burstein, 2006). The first version included 60 features in its evaluation process. e-Rater version 2.0 considers a set of intuitive features such as measures of grammar, usage, mechanics, style, organization, development, lexical complexity, and prompt of specific vocabulary usage. The writing analysis tools identify agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical errors. The system was trained on an extensive corpus with the sequences of adjacent words and part of speech tags of the sentences. Another feature of e-Rater is the presence identification of repetitive terms, a property that considerably affects the quality of the text. The system provides feedback to students with information related to the presence or absence of certain elements.

In the work of McNamara, Crossley and McCarthy (2010), the authors aimed to distinguish the differences between essays of undergraduate students that obtained a high score and low score. They used the Coh-Metrix tool and found that essays with a high score showed more complexity of the text and sophisticated language. In addition, and under a holistic approach of quality text, Crossley, Muldner and McNamara (2016) conducted an analysis of four features that together show the presence of the construct 'idea generation' in student essays. Fluency (number of ideas generated in the text), flexibility (new ideas of the author in the text), originality (difference of text ideas to other author's ideas) and elaboration (the level of development of the idea) were the elements analyzed. The corpus was composed by essays written in 25 minutes by first-year undergraduate students, without using external references. Besides, the essay assessment was done by different AWE tools such as WAT (Writing Assessment Tool), or TAACO (Tool for the Automatic Assessment of Cohesion) (Crossley et al., 2016). The results obtained by the authors

indicate that essays with many original ideas (flexible and elaborated) obtained a high evaluation and were significant features for determining the quality of essay.

TAALES2.0 (Tool for the Automatic Assessment of Lexical Sophistication) by Kile, Crossley and Berger, (2018), is an AWE tool that computes numerous indices related to: word frequency (less frequent words are considered more sophisticated or complex), word range (number of documents containing particular elements), n-gram frequency (set of infrequent terms that relate to the quality of the text), n-gram range, n-gram strength of association, contextual distinctiveness (measures the diversity of the context in which a word occurs), semantic network and word neighbors (words that share phonological, phonographic and orthographic similarities). This tool has been applied to L1 and L2 (Second Language) students to predict holistically the lexical proficiency. The authors discarded variables that did not comply with a minimum correlation, in addition to variables that presented multicollinearity. The final model included ten variables, which explain the 58% variance in the lexical proficiency; 51.7% of lexical proficiency was achieved in the previous version of TAALES. The tool does not interact directly with a student, i.e., it was designed so that the researchers use the results that TAALES provides, and then the results will be processed in another step. However, the results of TAALES are promising to measure the proficiency of the text.

Contrasting our research with the previous related systems or approaches, in a similar way as WPal, our work seeks to assess some aspects of the text, but focusing on the conclusion section of a thesis, considering three models to characterize the Goal Connectedness, Judgment and Speculation features. As SciPo, our analyzer sends feedback to the student depending on the progress achieved in each of the dimensions analyzed. Although our model does not generate a sequence, it does consider the level of the opinion of the sentences of the conclusion. Similar to TAALES, our method seeks to identify features that can help to improve student texts. Also, our work differs on the three characteristics evaluated (goal connectedness, judgment, and speculation). Nevertheless, the three features allow helping to satisfy the conclusion requirements.

A machine learning approach has been used for student essays assessment, to find the thesis and conclusion sections in the essay (Burstein & Marcu, 2003). The authors used two annotators to mark the thesis (problem statement) and conclusion section. Among the features used to train the algorithm are the lexical items (words and cue-terms). For instance, the cue term 'in conclusion' is associated with the conclusion section. Other feature considered in training is text position. In contrast, in our corpus the conclusion section is assumed clearly delimited. Our analysis contemplates the content evaluation of the conclusion, considering the recommendations for writing the conclusions. Under a phrase extraction approach, scientific papers like a thesis have been studied, both kinds of document report results obtained after applying a

scientific methodology. In the work of key phrase extraction (You, Fontaine & Barthes, 2013), an analysis is done to identify the most important phrase in a scientific paper. These phrases characterize the document and allow differentiating it from other types of documents, such as a newspaper article. Similarly, we seek to identify features (full sentences or terms) that are proper to a conclusion. These features are relevant considering the pattern of a conclusion described previously.

2. Method and materials

Below, we describe the collection that was used to develop the experiments. In addition, the solution scheme is provided as a conclusion analyzer, which includes three models: Goal Connectedness, Judgment, and Speculation.

2.1. Data description

The corpus for the study contains conclusions of Graduate level: Master (MA) and Doctoral (PhD) degree; and Undergraduate level includes: Bachelor (BA) and Advanced College-level Technician (TSU) degree (two-year technical study program offered in some countries). The corpus was downloaded from the Research Thesis and Proposal Collection - Coltypi2.0 (www.coltypi.org) using the Coltypi query interface in 2017. This collection includes around 968 theses and research proposals in Spanish of graduate and undergraduate levels.

The corpus domain is computing and information technologies. Each item of the collected corpus is a document that was revised at some point by a reviewing committee. Also, we gathered for each of these conclusions the associated general objective, required in the Goal Connectedness model. In total, we have 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 models, 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 (Goal Connectedness, Judgment, and Speculation) by two annotators.

Table 1. Average of words in the objective-conclusion corpus.

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

The annotation process included two annotators, marking the text that reveals the presence of Goal Connectedness and Speculation. To assess the Judgment, a scale of

three levels was established ('Yes, a lot', 'Yes, a little', and 'No opinion'). Each of annotators has experience in the review process of theses. For instance, an undergraduate objective-conclusion tagged by the annotators (see Table 2), where S1 denotes Sentence 1.

Table 2. An undergraduate objective-conclusion pair.

Objective: S1: Design application software in Visual Basic for data acquisition of digital drivers using OPC technology.
Conclusion: S2: This work shows the communication between software and PLC Allen-Bradley Compact Logix, covering processing needs for level control of a boiler. S3: 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. S4: The graphical interface designed is a clear example of the scope that has Visual Basic for design automation technologies and therefore is widely used by international designers. S5: Moreover, it can be seen that the Ethernet communication provides a higher speed compared with the RS-232, using Ethernet we achieve a more reliable monitoring since we satisfied with the information presented on screen, achieving more efficient supervisory control. S6: 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). S7: The interface turns into a tool that efficiently makes the control of a process within any industry to provide the operator updated and organized information, we mention that the basis of this program can be used for control of different variables either temperature, flow or pressure. S8: Thus, we see that the OPC technology offers a variety of tools for client-server connection, showing great amplitude data management.

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

S3: 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.

S4: The graphical interface designed is a clear example of the scope that has Visual Basic for design automation technologies and hence, their wide use by international designers.

Speculative text marked by annotators in conclusion (ST):

S6: 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).

Judgment level selected by annotators: Yes, a lot

The Kappa agreement between annotators for Goal Connectedness element was 0.923 that corresponds to 'Almost Perfect' (Landis & Koch, 1977). For the Speculation element, the agreement was 0.650 that corresponds to 'Substantial'. Finally, for the Judgment feature, the agreement was as follows: 0.47 ('Moderate'), 0.21 ('Fair'), and 0.44 ('Moderate'), according to the defined scale.

2.2. Conclusion analyzer

Our system has a Conclusion Analyzer, which contains three main models. Goal Connectedness model is responsible for identifying whether a conclusion sentence has a connection with the general objective, as a way of considering the compliance with the research objectives. Judgment model processes each sentence to identify terms with opinion load, evidencing the presence of opinions or value judgments formulated by the students. The final model, Speculation identifies whether the student expressed future work or possible derivations of his/her work. This model uses two lists of speculation terms.

In an analysis of four Learning Analytic Tools implemented in different universities, a 'recommendation' facility is displayed as a key feature, so students can improve their performance, considering the advice provided by the system, using the data associated with the interactions of the students in the use of the tool (Atif, Richards, Bilgin & Marrone, 2013). Our system seeks to help a student with little or partial experience in drafting conclusions, to assess the elements that academic advisors consider in writing such a section. In addition to the Conclusion Analyzer displayed on our model, we also include feedback to the student with recommendations. The suggestions are provided to the learner, depending on the level reached in each of the features evaluated. Each of the recommendations was formulated by our annotators, which are college-level instructors with experience in thesis revision.

2.2.1. Goal connectedness model (GC)

This model of the Conclusion Analyzer seeks to identify whether the conclusion shows some connection with the general objective. We expect that some sentences display this relation. And given that the objective-conclusion pair can be about any subject, we model such relations as looking for the sentence that best covers 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 lemmatized with FreeLing, a library of automatic multilingual processing functions and linguistic text annotation (Padró & Stanilovsky, 2012). For the conclusion section, we extracted its sentences; that is, we got a set of sentences, which were compared individually against the objective (consisting of only one sentence). For computing the

connectedness feature, we do it in terms of coverage, applying the following expression:

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

where S is a list of words of an objective (*So*) and the *i-th* sentence of conclusion (Sc_i), and N is the number of terms in the objective. The value of the sentence with the 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 measure for the previous conclusion given in data description section is:

- ✓ Our connectedness model obtained: S2=0.08, S3=0.25, S4=0.30, S5=0.0, S6=0.25, S7=0.0, S8=0.16
- ✓ The sentence with the highest coverage value obtained by Goal Connectedness model is S4. We found a coincidence between annotators (CT) and our analyzer in S4.

Also, a variant of coverage measure was explored doing a synonyms expansion in the conclusion sentences, to capture other terms used by students. For master level, the gain was of 10%, however for Doctoral, Bachelor and TSU levels the gain was minimal (1%); therefore, we decided to proceed without synonym expansion.

Evaluation in GC model

First, we processed each of the objective-conclusion pair with the Goal Connectedness model and the result was placed on a scale. To build the scale, the graduate level was 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 bachelor level was included (selected at random). Below we show the scale:

Coverage ≥ 0.123 (Average - 1σ)

If the evaluated sentence is above 0.123, the connection between the objective and the evaluated sentence is acceptable, otherwise it is taken as an absence of connection.

Coverage ≥ 0.414 (Average + 1σ)

If the result is above 0.414 the sentence shows a strong connection. We expected that sentences above the minimum acceptable score (0.123) would provide evidence that the student is linking the objective with the conclusion paragraphs properly.

Finally, to validate the scale, we used the corpus tagged by annotators. 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, which correspond to 'Substantial' agreement.

2.2.2. Judgment model (JM)

The goal of this model is to identify whether the conclusion section shows evidence of some opinions. For instance, in the conclusion: *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 about applying conceptual graphs and semantic representations. To consider terms that reflect an opinion or value judgments, we employed the SentiWordNet (Baccianella, Esuli & Sebastiani, 2010). This tool is a lexical resource for English, which aggregates an opinion score to each term (e.g. noun, adjective) depending on the sense. The sense has three numerical scores for objectivity, subjectivity and neutrality. The range of value is between 0 and 1. Each conclusion was translated to English employing Google Translator. 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 3.0. For instance, the term 'possible' presents a 0.37 opinion load, this result is computed regarding the average of all opinion loads (as a noun has two senses and as adjective also has two senses). The synonyms of the terms were not considered. Below, we provide an example, the Opinion load measure in the conclusion given in data section above, produced the following results, where the total displayed is the sum of all terms:

- ✓ **S2**: work(0.048) shows(0.055) communication(0.042) software(0) PLC(0) Allen-Bradley(0) Compact(0) Logix(0) covering(0.064) processing(0.031) needs(0.312) level(0.036) control(0.046) boiler(0). Total = 0.63

Evaluation in JM model

Like in the case of the Goal Connectedness Model, we took the graduate level texts as reference to define a scale. However, in this case we do not apply smoothing, because we have three levels of opinion. For this feature, the conclusions must reach the average level of review (i.e. 'Yes, a little'), this will give evidence that the student is expressing judgments and opinions in the conclusion paragraphs. Next, we show the scale levels:

- ✓ Judgment $\leq$ 7.84 (Average - 1σ), these are conclusions corresponding to the level 'No Opinion'.
- ✓ 7.84 < Judgment < 26.98, these are conclusions corresponding to the level 'Yes, a little'.

- ✓ Judgment ≥ 26.98, these are conclusions that correspond to the level 'Yes, a lot'.

Regarding the previous example, we computed the sum of S2+S3+S4+S5+S6+S7+S8 (0.63+2.39+1.05+2.43+1.18+2.24+1.63=11.55). This result corresponds then to 'Yes, a little' and is close to the value assigned by annotators (i.e. 'Yes, a lot').

After obtaining the scale, 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. In order to test agreement with the Kappa coefficient, we used the value determined by each annotator in each level of the scale. For instance, if for the first objective-conclusion both annotators agree that the conclusion does not present any type of opinion, then that would represent a match for the Kappa coefficient. Therefore, both annotators were assigned with a coincidence for the calculation of Kappa.

2.2.3. Speculation model (SM)

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 (Kilicoglu & Bergler, 2008), that include modal auxiliaries, epistemic verbs, adjectives, adverbs, and nouns (see Table 3).

Table 3. Speculative words.

Feature type	Speculative words
Modal auxiliaries	may, might, could, would, should
Judgment verbs	suggest, indicate, speculate, believe, assume
Evidential verbs	appear, seem
Deductive verbs	infer, deduce
Adjectives	likely, probable, possible
Adverbs	probably, possibly, perhaps, generally
Nouns	possibility, suggestion

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. The dataset contains annotations at the token level for negative and speculative keywords (Vincze, Szarvas, Farkas, Móra & Csirik, 2008). The Corpus was tagged by two independent linguists following guidelines.

To obtain this list, we extracted from the XML file all terms tagged as speculation, such as suggesting and could:

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

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, getting a list of 227 speculative terms.

Evaluation in SM model

We processed each of the conclusions counting the speculative terms in each sentence. A scale for this feature was not defined, only the coincidence between the text marked by the annotator and the sentence with the maximum number of speculative terms. For instance, in the conclusion given in data section, the annotators marked the future work (ST). Our analyzer identified 'recommendation' as a speculative term on S6. In this case, we found a match between the analyzer and the annotators. The annotator marked all sentences (list of terms), while our analyzer identified in the sentence the speculative term(s), indicating that the selected sentence expresses the future work or derivations. After analyzing the annotated pairs using the criterion just described, 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 correspond to 'Almost Perfect' agreement.

3. Internal exploration of conclusions

During the analysis of the conclusions with the Conclusion Analyzer, we performed an exploration of the three selected features, as a way of validating the conclusion pattern (Figure 1). The exploration goal was to identify which part of the conclusion presents the maximum measure of Goal Connectedness, Judgment and Speculation. The initial hypothesis is that the Goal Connectedness and Judgment features tend to appear in the initial paragraphs of the conclusion. Also, we expected that Speculation tend to appear at the end of the conclusion, and not necessarily when the student states their value judgments. These behaviors are derived from the features suggested by the conclusion pattern (Figure 1). We identified the number of the sentence from the test corpus with the maximum value of each of the measures, as long as the maximum was above the thresholds for each feature.

Below we show a chart with the 30 conclusions (annotated corpus) assessed with the Conclusion Analyzer. The x axis represents the number of sentences identified in the conclusions (from one up to nine sentences). The y axis indicates the size (in proportion) of the analyzed conclusions, i.e., if the point is closer to one, this means that the point appears close to the start of the conclusions; otherwise it was found close to the end. For instance, for the first sentence, 16 points are shown within a circle, that is in 16 times the connectedness/judgment/speculation obtained the

maximum value nearly at the beginning of the conclusion. Figure 2 depicts that the points corresponding to Goal Connectedness (dots) appear close to points of Judgment (cross), and the points tend to be located in the initial sentences of conclusions. From the fourth sentence, the Speculation feature (triangles) begins to predominate. The Pearson correlation coefficient between Goal Connectedness and Judgment was of 0.65. The correlation value between Goal Connectedness and Speculation was 0.17. Between Judgment and Speculation, the correlation was 0.28. The correlation results show that Goal Connectedness and Judgment are found close in the conclusion section. Besides, the two elements appear more often in initial sentences. Low correlation of Speculation with GC and Judgment elements appears as green triangles to the right of the x-axis in Figure 2. So, the results confirm our hypothesis.

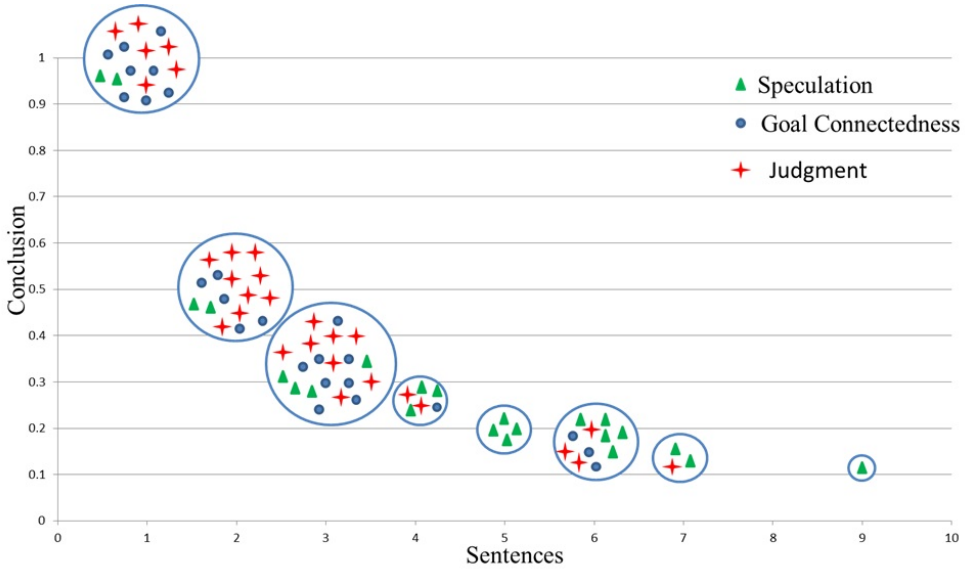

Figure 2. Explored features.

In addition, we performed an exploration of the whole corpus identifying the position of the features Goal Connectedness, Judgment and Speculation, for the different scholar levels. According to the Conclusions pattern, the Connectedness is located at the beginning, the Judgment at the center and future work (Speculation) at the end of the conclusion. Below we present the percentages found for Goal Connectedness-Judgment and Judgment-Speculation (see Figure 3). The percentage (Found) represents the number of conclusions where comparisons were done, otherwise included as (Not found).

Figure 3 depicted the proximity between the position of each identified feature and the conclusions pattern. Moreover, we note that the graduate level has a higher percentage than undergraduate level, i.e., students of doctorate and master level wrote

the conclusion section adhering to a structure like the triangle pattern in Figure 1. This structure tends to relax in undergraduate levels (BA and TSU).

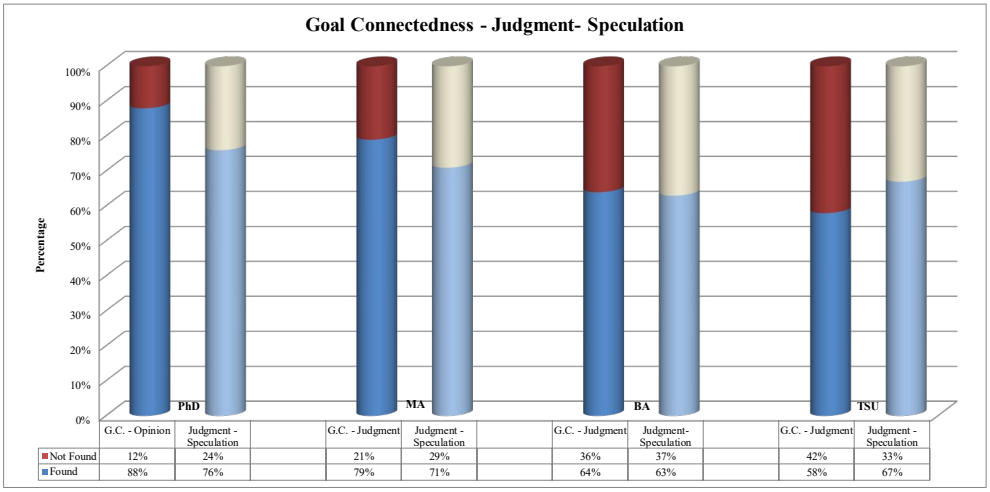

Figure 3. Explored features in whole corpus**.**

3.1. Overall Corpus Analysis

Furthermore, we conducted an analysis of the whole corpus using the models described above. The objective was to identify the levels of Goal Connectedness, Judgment and Speculation in graduate and undergraduate levels (see Table 4). The value of Goal Connectedness is the average of the maximum values of each sentence-conclusion pair of the corpus. The judgment 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.

Table 4. Corpus analysis between levels.

Level	Goal Connectedness	Judgment	Speculation
Graduate	0.30	20.50	3
Undergraduate	0.20	14.50	2

These results show that graduate students have better performance in connecting the conclusion with the objective and do so expressing more detail about their judgments and opinions. Besides, the significance and Power tests were applied for each feature between graduate and undergraduate level (Two-Sample T-Test. α = 0.05). A statistical test result with P-value less than 0.05 suggests the null hypothesis should be rejected. However, a P-value greater than 0.05 indicates that graduate and undergraduate groups do not show a statistically difference. For Power test, a value above 0.80 expresses the probability of rejecting the null hypothesis correctly. The

value of Power test is directly related with the Type II error (false negative). Below we show the hypothesis:

$$H_0: \mu_{Graduate} = \mu_{Undergraduate}$$

$$H_1: \mu_{Graduate} \neq \mu_{Undergraduate}$$

For the three features (Goal Connectedness, Judgment and Speculation) the P-value was 0.001 for each test. Hence, the Null Hypothesis was rejected with this result. Also, we computed the Power test with the goal of verifying the significance of previous result. The values of Power test for the features were 0.99, 0.96 and 0.96 respectively, showing reliability to the rejection of the null hypothesis. Therefore, the statistical analysis of the three features confirms the existence of a significative difference between both groups.

After doing statistical tests between the graduate and undergraduate level, we performed a detailed analysis. The purpose was to identify at specific study levels, the behavior of each of the three characteristics evaluated. In Table 5, we show the values obtained in each assessed characteristic.

Table 5. Corpus analyzed by study level.

Degree	Goal Connectedness	Judgment	Speculation
PhD	0.29	20.54	3
MA	0.31	20.65	3
BA	0.20	15.65	2
TSU	0.21	13.18	2

Notice that doctoral and master levels are very close but clearly above BA and TSU. However, to confirm whether the results are significant, we performed a further analysis. The objective was to calculate the P-value to know if the null hypothesis is rejected or not.

Also, we calculate the Power test of each result. 18 hypothesis tests (Two-Sample T-Test. $\alpha = 0.05$) and 18 Power tests were carried out. For instance, the hypothesis stated for TSU and PhD degrees in the Connectedness feature is:

$$H_0: \mu_{TSU} = \mu_{PhD}$$

$$H_1: \mu_{TSU} \neq \mu_{PhD}$$

Table 6 shows the results of each of the tests performed, the rows labeled with P-value correspond to the values that identify whether groups are identical or different, while the rows labeled with Power test show the values of the probability of correctly rejecting the null hypothesis, i.e. that the groups are different.

Table 6. Statistical results between Graduate and Undergraduate study levels.

Statistical	Goal Connectedness			Judgment			Speculation		
	BA	**MA**	**PhD**	**BA**	**MA**	**PhD**	**BA**	**MA**	**PhD**
P-value - TSU	0.25	0.001*	0.061	0.161	0.001*	0.039*	0.005*	0.002*	0.001*
P-value - BA		0.001*	0.002*		0.009*	0.509+		0.210+	0.013*
P-value - MA			0.572			0.341			0.080*
Power Test - TSU	0.64	0.980*	0.081	0.53	0.98*	0.890*	0.810*	0.990*	0.980*
Power Test - BA		0.990*	0.980*		0.96*	0.210+		0.430+	0.970*
Power Test - MA			0.180			0.503			0.830*

* Significance and Power test: High

\+ Significance and Power test: Low

In the Goal Connectedness feature, we found evidence that the MA degree differs strongly from BA and TSU with a P-value of 0.001 and power test values of 0.98 and 0.99, respectively, confirming the null hypothesis rejection. Between PhD and BA degree, we find similar behavior. This indicates that the difference is significant; that is, the conclusions written by graduate students have a better connection of the objective with its conclusions than TSU and BA, as expressed by the Connectedness feature. Moreover, we note that BA-TSU and PhD-MA degrees are quite close, but the Power test values indicate that the number of elements of each sample is insufficient to establish the equality of means.

In Judgment feature, we identified a similar behavior between the degrees of MA and BA-TSU, with high values of significance, showing that students of MA degree include more sentences of reflections that BA-TSU students. A finding was a medium level of significance 0.504 between PhD and BA degree, showing proximity between groups, i.e., the null hypothesis is not rejected. However, revising the Power test we found a low value (0.21). This result suggests increasing the size of the collection of the doctoral degree. Finally, in Speculation column, the PhD degree differs from BA and TSU, with a high value of significance. We might assume that doctoral student includes in the conclusion section the future work or derivations of their work in greater proportion than students of BA and TSU degree. This result is expected in doctoral theses, since they are primarily focused on research. However, further work is a feature that should be included in any thesis, regardless of study level. Also, we found a case where MA and BA degree obtained a significance value of 0.21; therefore the null hypothesis is not rejected, but the Power test value is low (0.43). We can notice that most of the results of this detailed analysis demonstrate that the PhD and MA levels show better results in the three characteristics evaluated than BA and TSU levels. These results suggest that graduate level students with better writing skills (González-López & López López, 2015) also achieved good results in the features examined in conclusions. Hence, the students who completed successfully a master or doctoral degree seemed to provide them with better writing skills than students of college level.

CONCLUSIONS

In this paper, we have presented a model to evaluate the connectedness with the general objective, the evidence of value judgments, and the presence of future work. The model considers specific features proposed by methodology books and institutional guides for writing the conclusion section of academic texts. We also take advantage of the knowledge contained in our theses corpus, which was previously reviewed by different academic advisors, and extract features from it using different models. In the three evaluated features we found that texts by graduate students outperformed those by undergraduate ones. This behavior provides evidence that students with more practice writing a scientific paper or thesis (graduate level), possess better skills. Furthermore, our models can help to improve writing thesis papers of undergraduate students or inexperienced learners, mainly in the features of Goal Connectedness and Speculation, since the achieved Kappa levels were substantial or better.

Even though the Goal Connectedness model reached a substantial level agreement with annotators, for further work, we consider analyzing the relation between the objective and the conclusion sections under a topic approach, possibly by applying an LSA technique. Other possibility to assess the relations is as Quan, Liu, Lu, Ni and Wenyin (2009) where each topic represents a concept, and the concept represents a set of words with an associated probability. For instance, if we are contrasting the objective 'Develop software to identify twitter opinions reaching a high agreement' against the sentence in conclusion 'a low agreement was reached by the software', in our current exploration of connectedness model, the words high and low are different, but both sentences are undoubtedly related by the topic. For the Judgment feature, as future work we consider to identify whether the orientation is positive or negative, with the goal of carrying out a content-focused analysis, then analyze whether the conclusions show a positive result when compared with the general objective. We intend to include a syntactic parsing before using SentiWordNet and thus we could find the specific weight of each term, according to its syntactic role. Also, we plan to increase the number of examples of the corpus to improve the level of agreement between our system and annotators, specifically for judgment.

Furthermore, we are also planning to include metrics to assess whether the conclusion contains a certain level of originality and elaboration, like the work of Crossley et al., (2016). The working hypothesis is that the conclusions of graduate level contain more original ideas than undergraduate level. For speculation, as future work, we plan to extend the analysis to consider speculative phrases. The results of this exploratory study set the stage for moving to classrooms. We are planning to conduct a pilot test with students, with the aim of verifying if our system indeed helps them to improve their writing in conclusions. This information will help to guide our

project to focus on improving the Conclusion writing to have an impact on students, and consequently in instructors.

REFERENCES

Allen, G. (1976). *The graduate students' guide to theses and dissertations: A practical manual for writing and research*. San Francisco CA, USA: Jossey-Bass Inc Pub.

Atif, A., Richards. D., Bilgin, A. & Marrone, M. (2013). *Learning analytics in higher education: A summary of tools and approaches.* Proceedings 30th ascilite conference, Sydney, Australia.

Attali, Y. & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment, 4*(3) [on line]. Retrieved from: https://ejournals.bc.edu/index.php/jtla/article/view/1650

Baccianella, S., Esuli, A. & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *LREC*. European Language Resources Association.

Bitchener, J. & Basturkmen, H. (2006). Perceptions of the difficulties of postgraduate L2 thesis students writing the discussion section. *Journal of English for Academic Purposes*, *5*(1), 4-18.

Burstein, J. & Marcu, D. (2003). A Machine learning approach for identification thesis and conclusion statements in student essays. *Computers and the Humanities, 37*(4), 455-467.

Crossley, S., Varne, L., Roscoe, R. & McNamara, D. (2013). *Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system.* Proceedings 16th International Conference AIED, Memphis, TN, USA.

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

Davis, J. & Liss, R. (2006). *Effective academic writing 3: The essay*. USA: Oxford University Press.

Debuse, J. C. W., Lawley, M. & Shibl, R. (2008). Educators' perceptions of automated feedback systems. *Australasian Journal of Educational Technology, 24*(4), 374-386.

Feltrim, V. D., Teufel, S., das Nunes, M. G. V. & Aluísio, S. M. (2006). Argumentative zoning applied to critiquing novices' scientific abstracts. In J. Wiebe & Y. Qu (Eds.), *Computing Attitude and Affect in Text: Theory and Applications* (pp. 233-246). Dordrecht: Springer.

González-López, S. & López López, A., (2015). Lexical analysis of student research drafts in computing. *Computer Applications in Engineering Education, 23*(4), 638-644.

Gierl, M., Latifi, S., Lai, H., Boulais, A.P. & De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, *48*(10), 950-962.

Kilicoglu, H. & Bergler, S. (2008). Recognizing speculative language in biomedical research articles: A linguistically motivated perspective. *BMC Bioinformatics,* 9(Suppl 11), S10.

Kile, K., Crossley, S. & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030-1046.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometric, 33*(1), 159-174.

McNamara, D., Crossley, S. & McCarthy, P. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.

Padró, L. & Stanilovsky, E. (2012). *FreeLing 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference*. Turkey: Istanbul.

Purdue Online Writing Lab (2013). *Introductions, body paragraphs, and conclusions for an argument paper* [on line]. Retrieved from: https://owl.english.purdue.edu/owl/owlprint/659/

Quan, X., Liu, G., Lu, Z., Ni, X. & Wenyin, L. (2009). Short text similarity based on probabilistic topics. *Knowledge and Information Systems, 25*(3), 473-491.

UNE Academic Skills Office (2017). *Writing pattern for conclusion paragraphs* [on line]. Retrieved from: https://aso-resources.une.edu.au/academic-writing-course/paragraphs/conclusion-paragraphs/

Vincze, V., Szarvas, G., Farkas, R., Móra, G. & Csirik, J. (2008). The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics, 9*(Suppl 11), S9.

You, W., Fontaine, D. & Barthes, J. P. (2013). An automatic keyphrase extraction system for scientific documents. *Knowledge and Information Systems, 34*(3), 691-724.