Chapter 9
Mining Domain Knowledge for Coherence Assessment of Students Proposal Drafts

Samuel González López and Aurelio López-López

Abstract Often, academic programs require students to write a thesis or research proposal. The review of such texts is a heavy load, especially at initial stages. One feature evaluated by instructors is coherence, i.e. the interrelationship of the various elements of the text. We present a coherence analyzer, which employs latent semantic analysis (LSA) to mine existing corpora to further assess new drafts. We designed the analyzer as part of an Intelligent Tutoring System, considering seven common sections. After mining domain knowledge, experiments were done on graduate and undergraduate corpora to define a grading scale. Another experiment that involved human reviewers was set to validate the process. The technique allowed evaluating the coherence of the different sections, reaching an acceptable result and hinting that the level reached so far is adequate to support online review. An innovative exploration across sections was performed, uncovering a consistent interrelationship, according to methodology authors.

Keywords Coherence · Writing support · Latent semantic analysis · Intelligent tutoring system

Abbreviations

DM	Data mining
ITS	Intelligent tutor system
LSA	Latent semantic analysis
LSI	Latent semantic indexing
NMF	Non-negative matrix factorization
PLSA	Probabilistic latent semantic analysis

S. G. López · A. López-López (✉)
Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro # 1, Santa María Tonantzintla, Puebla 72840, Mexico
e-mail: allopez@inaoep.mx

S. G. López
e-mail: sgonzalez@inaoep.mx

A. Peña-Ayala (ed.), *Educational Data Mining*,
Studies in Computational Intelligence 524, DOI: 10.1007/978-3-319-02738-8_9,

SPM Student progress module
SVD Singular values decomposition

9.1 Introduction

Academic programs or courses in educational institutions often conclude requiring students to elaborate a thesis or research proposal. A customary process followed by students is to write a first draft and then improve it after iterated reviews and recommendations of the instructor.

Some institutions provide guides that support students in structuring the proposal draft. However, this is insufficient in many cases, i.e. students often need help on how to structure and write every aspect of their draft. This demands that the academic advisor or instructor spends extra time on the reviewing process.

Data mining (DM), whose aim is to identify novel, potentially useful and understandable correlations or pattern from data, can adhere to one of two approaches: seek to build models or find patterns.

Educational data mining has similar aim and approaches but working on data obtained from educational settings [1]. An educational setting of interest is college education, where the heavy load of draft review can be ameliorated by the use of information technologies and methods.

One way to achieve this objective is to mine existing corpora of research proposals and theses to build models of different features (e.g. topics, language models, or argumentation) to analyze in new drafts of students. In particular, employing data mining, we can characterize the semantics of the domain of information technologies and computer science to assess coherence in drafts.

This chapter focuses on examining coherence in documents written in Spanish. Coherence is defined as the connection of all parts of a text into a whole [2]: the interrelationship of the various elements of the text. Therefore, coherence within proposal drafts is important because if a document does not have each of the elements related into a whole, or sections are not close to a topic, it would seem incoherent.

In this chapter, we present a global coherence analyzer, which employs Latent Semantic Analysis technique and tool to mine existing corpora of research proposals and theses to further assess proposal drafts of college students in information technologies and computer science. Its main aim is to help students to improve their coherence in drafts during the writing process, especially in the early stages. Furthermore, we intend that this analyzer, implemented in a system, indirectly helps the academic advisor by reducing the time dedicated to the draft review, enabling to focus on content.

We designed the analyzer considering seven common sections in drafts, and is in-tended as part of an intelligent tutor system (ITS), supporting students online. To assess global coherence after mining domain knowledge, experiments were

done on graduate and undergraduate corpora to validate the process. Experiments involved human reviewers to compare the results of the analyzer with those of the reviewers, so we computed agreement measures. From mined domain knowledge, an exploration across sections was performed, as an additional validation procedure.

The results on coherence analysis reported here are parts of a larger project that may help students to evaluate their drafts early, and facilitate the reviewing process of the academic advisor.

The approach contributes to DM with a method to employ the results of latent semantic analysis (LSA) to grade and support students online to improve their writings, and a process for further exploration of mined knowledge.

This chapter is structured as follows. Next Section reviews previous related research. Section 9.3 describes the coherence analyzer. Section 9.4 details the data employed to mine and validate the experiments. Experiments validating the approach are presented in Sect. 9.5. Section 9.6 discusses results and their analysis. Section 9.7 includes an overview of the ITS. Finally, Sect. 9.8 details the conclusions and future work.

9.2 Background

Three themes are central to this research; coherence, data mining employing latent semantic analysis, and previous approaches for the mining of learners essays. We review concepts and related work in the following subsections.

9.2.1 Global Coherence

Coherence is classified based on its scope: global and local. Global coherence means that a document is related to a main topic, i.e. it is not consistent when its elements have no such main topic. Local coherence is defined within small textual units [3]. Recently, [4] reported a study of different factor conducing to cohesion and coherence in texts coming from student discussion forums.

An exploration of how foreign language learners express cohesion and coherence in their writings is reported in [5], employing topical structure analysis. An analysis of several methods for assessing coherence in the context of automated assessment of learners' responses is given in [6]. In [7], the authors define four aspects related to local and global coherence, one of which relates to the topic developed in the essay respect to the required topic by the teacher. Despite focusing on local coherence, [8] highlights specific areas of research for NLP in essay scoring. None of these studies of coherence is on proposal writings and they are predominately to grade essays already written, i.e. not to support directly the writing process.

9.2.2 Latent Semantic Analysis

LSA at first known as latent semantic indexing (LSI) [9], is an automatic indexing and retrieval technique, which was initially designed for improved detection of relevant documents on the basis of search queries. This is a dimensionality reduction technique based on statistical analysis that allows uncovering the implicit (latent) semantics (structure) in a collection of texts. Afterward, Landauer and Dumais developed the LSA technique [10].

They define the LSA as a theory and method for extracting and representing the contextual meaning of words in use, through statistical computation applied to a large corpus.

In [11], they evaluated the textual coherence using LSA technique. This paper shows the coherence prediction by analyzing statement by statement a set of four texts, with a 300-dimensional semantic space, which is constructed based on the first 2,000 characters of each of the 30,473 articles of the Encyclopedia of American Academic Groliers. After separation of the four individual sentences texts, the vector of each text was calculated as the sum of the weights (each term), subsequently being compared with the next vector, so the cosine of these two vectors showed the semantic relationship or coherence.

One of the discussions in this paper is whether the LSA technique is a model of text-level knowledge of an expert or novice. They state that it depends on the training that the LSA system has received in the application domain. This technique focuses on the latent semantic aspect, which is a relevant feature to our work.

Alternative techniques related to LSA are: Probabilistic Latent Semantic Analysis (PLSA) and Non-negative Matrix Factorization (NMF). PLSA [12] has a well-developed statistical foundation, defining a proper generative data model; and uses a generalization of Expectation Maximization algorithm for training, with some gains in performance.

NMF [13] applies a non-negativity constraint when factorizing a term-document matrix, leading to a more intuitive representation of documents as addition of topics. A comparison of four popular text mining methods is reported in [14], including LSA. An alternative way to mine regularities, and in consequence assess coherence, is proposed in [15].

9.2.3 Related Work

Several previous works have focused on evaluating educational aspects using the LSA technique. In the educational field, different kinds of documents are generated, such as documents written by teachers related to learning activities, student essays or textbooks [16]. Our work focuses on proposal drafts of undergraduate students, specifically in the Spanish language.

A different application of mining is presented in [17], where the aim is to reveal the processes involved during collaborative writing. In [18], they take a data mining perspective to do essay scoring, with LSA as one of the methods to consider content.

Given that the coherence analyzer is intended to help students when preparing their text, our work is also related to intelligent writing support systems and tools, such as Glosser [19] that supports students when writing essays by formulating questions and providing content clues to answer them, employing data mining. A more recent work [20] goes a step further by generating questions from student writings citations and content elements.

Despite in essence the coherence analyzer described in the chapter performs student text grading, this is done from text in process of improvement (i.e. prior to submission), not from static given text (post submission) as mainstream essay grading (e.g. [8, 18],). The approach also extends previous applications of LSA with a method to exploit its output to grade and support students online.

9.3 Analyzer Model of Global Coherence

Many text definitions include coherence as a necessary feature. Coherence in proposal drafts of students is important because if it is not present in each of the elements, the central idea loses all meaning. Different approaches have been proposed by researchers, some techniques have focused on the semantic aspect when seeking to achieve overall coherence evaluation, while other studies have worked the syntactic aspect, as a way to attack the local coherence. Both approaches were developed in different ways in [21, 22], but our work focuses in global coherence, as the first step to improve the proposal drafts of undergrad students.

Our model seeks to evaluate the global coherence in seven sections of a proposal draft: problem statement, justification, objective, research questions, hypothesis, methodology, and conclusions. The global coherence refers to the thematic similarity between the section subject to evaluation and the semantic space, mined from an existing corpus in the domain of computer science and information technologies.

For example, if the text under evaluation contains concepts thematically close to biology, their measure of coherence will be poor, since our corpus is of the

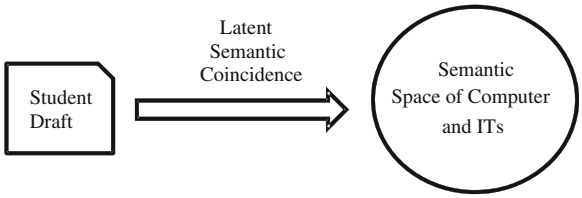

Fig. 9.1 Global coherence

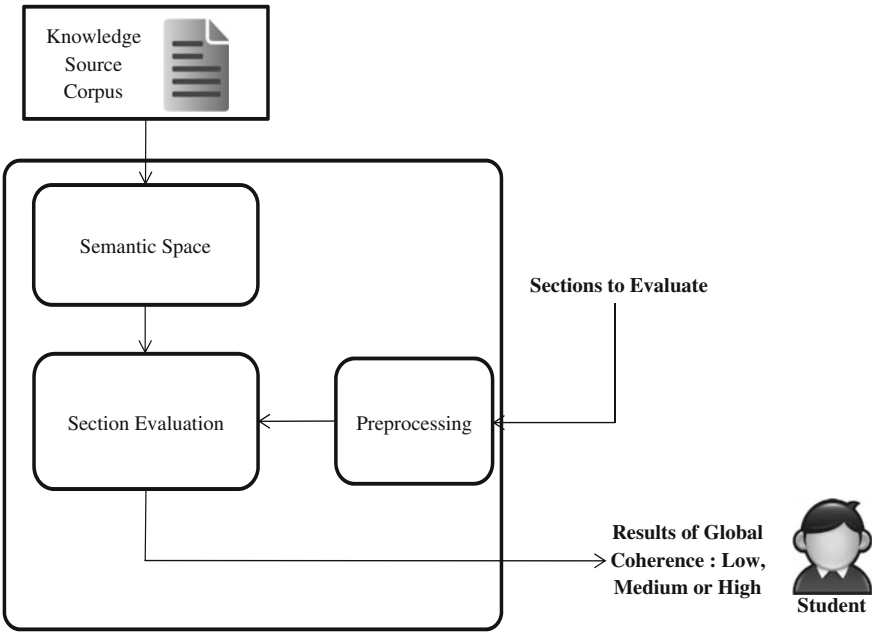

Fig. 9.2 Coherence evaluation model

computer and information technologies domain. Figure 9.1 depicts the concept of global coherence. Under this concept of global coherence, we designed our model as illustrated in Fig. 9.2.

Knowledge Source Corpus. The first step was to gather documents in Spanish, such as student theses and research proposals, previously reviewed and approved. Both kinds of documents were of under-graduate and graduate level. With this corpus, the semantic spaces were extracted for each section, i.e. there were seven corpora to mine. Corpus description is presented in detail in the following section.

Semantic Space. To extract the semantic space, terms of the input elements of a proposal draft were truncated (stemmed). Images, tables and figures were ignored. The goal of the stemming process is to reduce the variations of each word. For example the words "computer" and "computers" (in Spanish computadora, computadoras) are similar, so the process would produce a word stem "comput". We used the Freeling tool for stemming. In this way, many related terms are grouped, reducing the dimensionality of terms. Afterward, each corpus of the sections was processed by removing stop words (empty words), such as articles, prepositions, pronouns, conjunctions, etc. for instance, "of", "the", "by" (in Spanish de, la, por). These stop words were supplied by NLTK-Snowball.

Having the vocabulary of each section, a term-document matrix is built. This matrix was processed to compute weights according to *tf-idf*, where *tf* represents the absolute frequency of appearance of a term in a document, and *idf* is the inverse frequency of the term in the documents of the collection, i.e. the weight of

a term in a document increases if this occurs frequently in such document and decreases if appears in many (most) of the documents.

LSA then reveals the (latent) meaning of words, discarding the words occasionally used in specific contexts and focusing on what is common in all contexts [23]. This is achieved by the core process in LSA, Singular Value Decomposition (SVD). So, after preprocessing the corpus, the algebraic SVD algorithm is used. SVD allows reducing the dimensionality of the original matrix to a more manageable number and also reduces noise or irrelevant information in the matrix. The SVD produces three matrices:

- Orthogonal Matrix U. Obtained by linear processing of number of columns in the original matrix A. This matrix represents terms as vectors in space of words.
- Transpose matrix V^T. Obtained by permuting the rows with columns, providing an orthogonal arrangement of the elements of the row. Through this transposition, documents are represented as vectors in space of words.
- Diagonal matrix Σ. Calculated by linear processing from number of rows, number of columns and the number of dimensions of the original matrix A. The diagonal matrix represents singular values of A. The singular value decomposition of the matrix is illustrated in Fig. 9.3

Once the three matrices are obtained, we can generate the matrix A, but depending on the singular values maintained, it would be a matrix close to matrix A, i.e. an approximation to A with the most relevant information.

Sections to Evaluate. These correspond to the sections that the student wants to evaluate, so they are analyzed one a time (i.e. there is no need to parse sections). Our analyzer allows evaluation of seven sections of a proposal: problem statement, justification, objective, research questions, hypothesis, methodology, and conclusions.

Preprocessing. This part of the model considers the stemming, stop word removal and computation of the *tf-idf* weights on the section to be evaluated. Once these processes have been applied, the text is ready to compare against the corresponding semantic space to measure similarity.

Section Evaluation. The section under evaluation is compared against the corresponding semantic space. For this purpose, the cosine similarity measure is applied to the input vectors obtained from the section and those vectors coding the semantic space. The expression for the computation is:

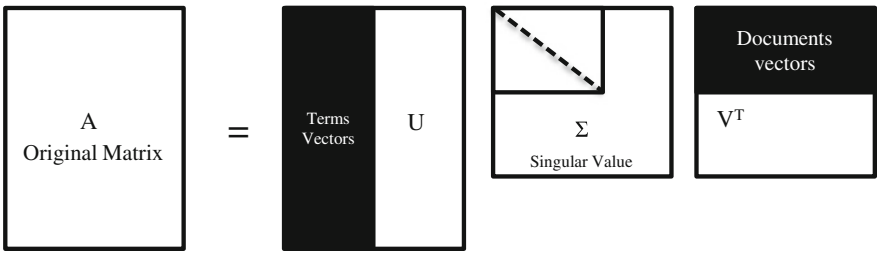

Fig. 9.3 SVD algorithm [24] representation

$$\cos(A, B) = \frac{A * B}{||A|| * ||B||}. \tag{9.1}$$

According to this expression, the similarity is one when the angle between the two vectors is 0°, that is, the vectors are pointing in the same direction and are parallel. This result expresses the highest coherence in the text. We get 0 when the vectors are orthogonal and correspond to no coherence at all.

Results of Global Coherence. Instead of reporting a numerical value as result of the coherence evaluation, the result is expressed in terms of three levels: High, Medium and Low. To achieve this qualitative scale of coherence, a process was applied, setting thresholds to determine each level.

This information was obtained taking as reference the graduate corpus, under the premise that the level of graduate students writing is better than those at undergrad level. Next, we describe the corpus used to extract the domain knowledge, as well as the threshold to define the discrete values for grading.

9.4 Data Description (Corpus)

We gathered a corpus of different elements in proposal documents in Spanish. We distinguished two kinds of student texts: graduate proposal documents, and undergraduate drafts. The first kind of texts includes documents reviewed and approved by faculty, so they are considered as reference or examples for knowledge mining. The second kind of documents is used as test examples. The whole corpus consists of a total of 410 collected samples as detailed in Table 9.1. The corpus domain is computing and information technologies. They were then preprocessed as detailed above.

9.5 Experiments

This experiment focused on evaluating the sections of student's proposal draft from the aspect of global coherence. We selected LSA because it captures the documents' latent semantic, something we want to mine from different sections in a proposal.

Table 9.1 Corpora

Sections	Graduate	Undergraduate
Problem statement	40	14
Justification	40	18
Objectives	60	20
Research questions	40	10
Hypothesis	40	20
Methodology	40	14
Conclusions	40	14
Total	300	110

9.5.1 Experimental Design

An experiment was set to validate the proposed online reviewing process, involving human reviewers to compare the results of the analyzer against their grades, calculating an agreement measure. In particular we computed agreement in terms of Fleiss or Cohen Kappa. From mined domain knowledge, an exploration across sections was performed, exploring their interrelationship.

All the collection was sent for evaluation to two or three instructors serving as reviewers, that have experience in advising students in the preparation of their drafts in the computing and information technologies. The reviewers did not know beforehand the level (graduate or undergraduate) of each sample. Each reviewer was requested to assign a level to each sample, using the scale: High, Medium and Low coherence, where the high level meant that the text has a strong coherence or relationship to the domain of computing and information technologies and the low level meant that the relationship is weak relative to the domain. Two examples of High and Low coherence in the objectives section are given next.

High Coherence. Analyze problems that arise in the system development of software architectures of Enterprise type.

We can observe that the word "systems" and "software" are very close to the domain, including the term "architecture" in the context of the previous terms fit within the domain of computing. Likewise, words with less thematic load such as "development" or "analyze" are often used in the domain.

Low Coherence. Identify the effect of feedback on the learning of the business leader, to allow being more effective.

Notice that even though terms like "learning" or "feedback" may have some proximity to the domain, the words or phrases "business", "leader" or "be more effective" are the central topic and are barely used in the domain of interest.

The assessments provided by our reviewers allowed to exclude those examples in our knowledge mining set considered low by at least two, or those where they did not agree, since they will bias the construction of the semantic spaces. For instance, if an objective was labeled as High by two or three of the human reviewers, this example will be part of our training corpus since, according to the reviewers; such objective is indeed highly coherent with the domain. In case that only one reviewer assigned High grade, this objective will not be part of the training corpus since there is a doubt about its coherence, and can introduce noise into the corresponding semantic space.

On the other hand, the assessments on the test set allow comparing the automatic evaluation of coherence, after extracting the semantic space and defining a grading scale. Once instructors evaluated the whole collection (training and test subsets) detailed in previous section, we then evaluate the level of agreement among them.

These human reviews allow getting the subset of examples subject to mine their knowledge, i.e. those contributing for the construction of the semantic space. Once we computed the semantic spaces, we can set the thresholds that define the scale

after analyzing all samples that human reviewers assigned a high level of coherence (see Table 9.2).

The thresholds for levels High, Medium and Low in our system were established using as a basis the average obtained when evaluating the training corpus (elements labeled with a high level) with a cross-validation. It was a one-fold validation, i.e. the element was removed from the corpus and the semantic space was generated with the remaining examples.

Then, we calculated the standard deviation of the values obtained, and the high level is calculated as the average plus one sigma and low as the average as minus one sigma. Previously, we corroborate the normality of the data, with 95 % of confidence. With the use of one sigma for thresholds, we can ensure that the results will be in a close range to the average obtained with the best documents (labeled as high). In this case if the result is closest to the upper limit, it means that the text is closer to the domain of computing and shows strong evidence of global coherence.

Also having the semantic spaces for the different sections of our mining subset, then one can evaluate automatically the corresponding section in the test subset. Then, we have the elements to evaluate the level of agreement between the grade assigned by the system and by instructors.

9.5.2 Agreement Evaluation

Each section was tagged by human reviewers (two or three reviewers). For each section, Fleiss and Cohen Kappa measures [25] were computed, depending on the case to be presented, i.e. two or three evaluators. In addition, we calculated the Cohen Kappa to evaluate the level of agreement between the analyzer and human reviewers.

We proceed to describe the grades assigned by human reviewers and the level of agreement among them first, and then the result of agreement between the grade assigned by the coherence analyzer and humans. We also provide the values obtained from mined semantic spaces and used to define the thresholds determining the grading levels. These results are presented for each of the section under analysis at a time.

Table 9.2 Training and test corpus

Sections	Training	Test	Tagged as high level
Problem statement	40	14	23
Justification	40	18	20
Objectives	60	20	40
Research questions	40	10	36
Hypothesis	40	20	20
Methodology	40	14	27
Conclusions	40	14	24

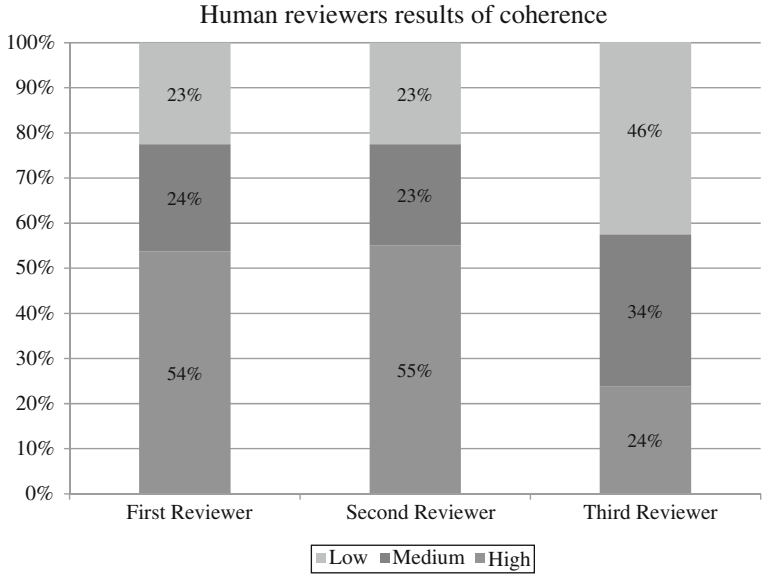

Fig. 9.4 Results of three human reviewers (objective)

Objective. Figure 9.4 shows the percentages of level of coherence assignment by each human reviewer. Note that the first and second human reviewer has similar percentages obtained in each of the levels. The third evaluator presented an inverse behavior to the first two reviewers; we assume that this rater was stricter than the other, when evaluating objectives.

The Fleiss Kappa coefficient of agreement was computed for the three reviewers considering the test corpus. Table 9.3 shows the Fleiss Kappa results for each level, for the objective section.

The reviewers had a Substantial agreement for the Low and High grades, and a Poor agreement in Medium grades. For the results obtained, we conclude that reviewers clearly identified High and Low levels but not those in the middle. The overall level achieved between evaluators was 0.54, this corresponds to a Moderate confidence of agreement for the experiment.

These levels allow automating the evaluation of the coherence analyzer. In particular, for the objective section, we got an average of 0.49 with a standard deviation of 0.17, resulting in the highest threshold of 0.64 and the lowest threshold at 0.28.

Table 9.3 Kappa for test corpus

Kappa	Reviewers (Fleiss)	Analyzer versus reviewers (Cohen)
High	0.6862	0.0000
Medium	−0.0378	0.2609
Low	0.7353	0.4218
Overall	0.5458	0.2237

Once the scale is defined, we evaluated the test samples with the aim to compare the results produced by human evaluators. In this case, Cohen's Kappa is pertinent to compare the level of agreement between human and our coherence analyzer results. Table 9.3 also shows the Cohen's Kappa results for the human versus coherence analyzer, being Fair and Moderate for Medium and Low levels, with a Fair overall agreement. In addition, despite that the High level does not reach an acceptable level yet, low and medium levels of coherence are already detected, giving certain confidence to the instructor of the analyzer can identify objectives with deficiencies.

Problem Statement. For this section, the level of agreement of the three reviewers was very low and only two of them assigned high level grades. Therefore, we decided to consider only two reviewers in the experiment, using for mining their high level grades. The second reviewer did not assign low values as shown in Fig. 9.5, whereas first reviewer assigned the three levels of coherence on the corpus.

As Table 9.4 depicts, there were high values of agreement between reviewers, but only for high and medium grades. For the results obtained, we conclude that reviewers clearly identified High and Medium levels of coherence in this section. The overall level achieved between evaluators was 0.68, this giving Substantial confidence of agreement for the experiment. These levels allow automating the evaluation of the coherence analyzer. For this section, after getting the semantic space, we obtained a low average of 0.127 with and standard deviation of 0.057, leading to setting the thresholds at 0.07 for Low and 0.18 for High.

As observed in the Kappa values between analyzer and reviewers, there is a Perfect level of agreement in High grades but a margin for improvement in the Medium grade since this is Fair as the overall agreement.

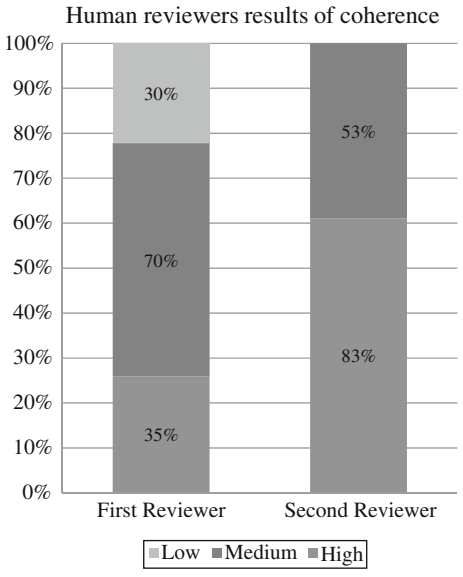

Fig. 9.5 Results of two human reviewers (Problem Statement)

Table 9.4 Kappa for test corpus

Cohen kappa	Reviewers	Coherence analyzer versus reviewers
High	1.000	1.0000
Medium	1.000	0.3300
Low	0.000	0.0000
Overall	0.680	0.4000

Since human reviewers did not agree on tagging problem statements with a low grade in the test set, we cannot expect any agreement with the analyzer. But, to find out if our approach can identify the low grades, we took examples labeled as low in graduate corpus. These examples were not included in the training set, but for exploration purpose, we evaluated the examples with the coherence analyzer and add them to previous results obtained with test set. With these results we computed the Cohen kappa between human reviewers and analyzer.

According to the results, the kappa showed an improvement for low and medium level. High level maintained the level of agreement, the medium and low level of Fair changed to Moderate, with 0.43 and 0.40 respectively. The overall agreement level was 0.49 which represents a Moderate level.

Hypothesis. Figure 9.6 shows the percentages of grades assigned by human reviewers, the first reviewer assigned the High grade more often, while the second reviewer had a more balanced performance. However, this is a normal behavior of human reviewers in the classroom. As in problem statement, we only used two of the human reviewers.

As Table 9.5 details, Kappa results between human reviewers were Acceptable with 0.301, similarly as the Kappa between our analyzer and human reviewers was Acceptable with 0.2558. However, it was lower than in the objective and problem statement sections. For the purpose of automating the evaluation of the coherence analyzer for the hypothesis section, we got an average of 0.636 with a standard deviation of 0.236, resulting in the high threshold of 0.87 and the low threshold at 0.4.

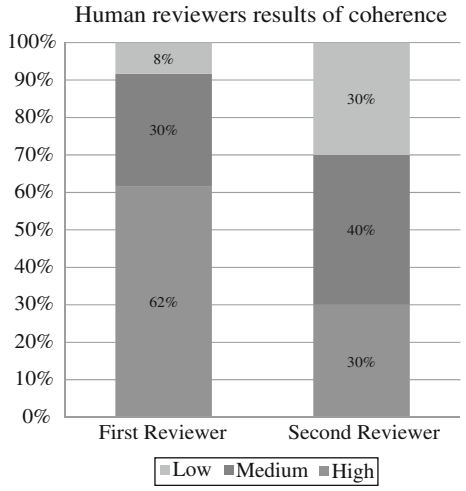

Fig. 9.6 Results of two human reviewers (Hypothesis)

Table 9.5 Kappa for test corpus

Cohen kappa	Reviewers	Coherence analyzer versus reviewers
High	0.3953	0.5294
Medium	0.2528	0.1428
Low	0.0000	0.0000
Overall	0.3010	0.2558

For particular levels, there is a Moderate value in the Kappa scale for High level, among human reviewers and our analyzer. The zero value of agreement among human reviewers affects the outcome of the analyzer to the low level. Although only examples with High level were used to mine, the human reviewers distribution was unbalanced. Low grades in Hypothesis presented a similar complication as the Problem Statement section, where reviewers did not agree tagging examples with low grade. Again, to find out whether our approach can identify low grades, we took the examples labeled as low in graduate corpus.

Then, we evaluated the examples with the coherence analyzer and add them to previous results obtained with test set. When executing Cohen kappa between human reviewers and the analyzer, the values high, medium and low were 0.6363, 0.111 and 0.333, respectively. It was observed that Kappa for High level is "Substantial". The medium level remains at "Slight" level and the Low moved from "Poor" to "Fair". The overall level of agreement was "Fair".

In this case, despite the medium grade did not reach an acceptable level, the low level reach an acceptable agreement. The analyzer can give certain confidence to the instructor that a hypothesis with deficiencies will be identify by our system, and can suggest students to improve their Hypothesis.

Justification. In this section, human reviewers had a more balanced distribution across the three coherence levels. The kappa values achieved were lower compared to the previous sections; even so the level is Acceptable or Fair. Figure 9.7 shows the percentages of levels assigned by the two reviewers. For the justification section, after computing the semantic space, we obtained an average of 0.137 with a standard deviation of 0.066 leading to set the thresholds at 0.07 for Low and 0.2 for High.

An Acceptable level was obtained between the reviewers and the analyzer with 0.39 (Table 9.6). Moreover, high level had a Moderate agreement and the medium level was Acceptable. Observe that the levels of agreement between human reviewers were Fair, despite having a balanced assignment of grades. The reason could be that the high grade was assigned with a similar percentage but not to the same samples.

Unlike the previous two sections, in this section the human reviewers tagged some examples with low grade in the test set, showing a Fair agreement in terms of kappa value.

A strategy implemented to raise the agreement results for low grades was using half sigma to define the thresholds. The results improved for low level, but affected the medium level. The kappa values for the High and Low level were 0.33 and zero respectively.

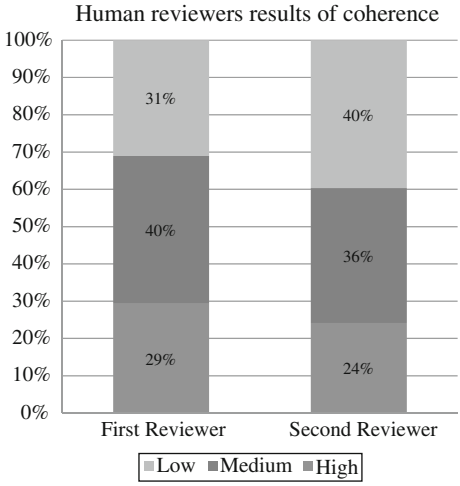

Fig. 9.7 Results of two human reviewers (Justification)

Table 9.6 Kappa for test corpus

Cohen kappa	Reviewers	Coherence analyzer versus reviewers
High	0.2200	0.5800
Medium	0.2075	0.3600
Low	0.2758	0.0000
Overall	0.2283	0.3900

Another attempted alternative to improve results was training a classifier (Naive Bayes), using as input vector the LSA value provided by the semantic space and the grade (class) assigned by the reviewers. As training examples, we used the set of graduate and undergraduate texts, evaluated as low and medium. After training, the classifier had a precision of 0.714 and recall of 0.5 for the low grade. The medium level reached a precision of 0.706 and recall of 0.857. The level of agreement was Acceptable in terms of kappa. These results indicate that the classifier is a promising alternative to predict medium and low grades for this section.

Conclusions. For this section the instructors identified the three grade levels at different rates (Fig. 9.8). The first reviewer was probably more rigorous than the second since assigned 35 % to high level, while the second reviewer duplicated the value, assigning a 65 % to high grade.

As expected from the percentages, the agreement results for this section were not satisfactory. The level of agreement between reviewers was 0.31, corresponding to the Acceptable level.

In this section, we got an average of 0.268 with a sigma of 0.247 allowing to set the thresholds for Low at 0 0.021 and for High level at 0.514. Also there was a 0.1666 level of agreement among human reviewers and the analyzer, this means a Slight level of agreement.

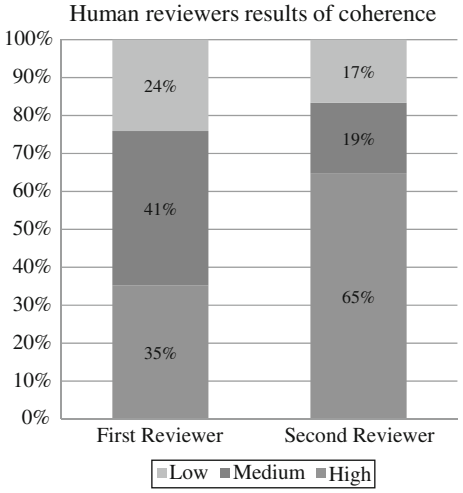

Fig. 9.8 Results of two human reviewers (Conclusions)

High and medium grades were Acceptable according to a kappa of 0.28. The value of agreement was zero for low grade. This was probably due to the low coincidence of examples labeled as high. As observed in results of previous sections, our analyzer results regarding human agreement levels are close, indicating that our analyzer is directly dependent on the level of agreement between humans.

In addition, the kappa level between human reviewers for low level was null, since none of the examples was graded as low (see Table 9.7). But to know whether our approach can identify low grades, we took examples labeled as low in the graduate corpus. The result again was unfavorable, since the values were low, according to previous values.

Subsequently, we decided to try a classifier (Naive Bayes) to improve results. For training, we used examples of the graduate corpus, tagged as medium and low. After training, we obtained the values of precision and recall.

The results were favorable, reaching a precision value of 1 and recall of 0.556 for the medium class, while for the low class reached a precision of 0.556 and recall of 1. Kappa value was of 0.447, higher than using thresholds.

These results indicate that for this section, the use of a classifier for predicting medium and low class seems more promising than using sigma to define the scale. The classifier was trained with medium and low classes, since our analyzer was built with the high class.

Table 9.7 Kappa for test corpus

Cohen kappa	Reviewers	Coherence analyzer versus reviewers
High	0.2857	0.2857
Medium	0.4000	0.2857
Low	0.0000	0.0000
Overall	0.3103	0.1666

Research Questions. Human reviewers had a similar assignment percentage of the low grade level. For medium and high percentages, they were unevenly (Fig. 9.9). This behavior was reflected in the values of Kappa. In the figure, we can notice that the first reviewer assigned a 30 % of medium grades, while the second reviewer assigned 55 %. This led to have an average of 0.432 with a sigma of 0.286, allowing to set the Low Level at 0 0.227 and the High level at 0.638, for this section.

We can observe in Table 9.8 that the Medium grade level had a zero percent agreement, which was expected since the level of agreement was very uneven between reviewers. For high grade, reviewers reached a value of 0.50 and for low grade reviewers obtained a kappa of 0.46, which corresponds to a Moderate agreement.

The agreement results between human reviewers and our analyzer were 0.33 for High and Low grades. This corresponds to an Acceptable level according to the range of kappa. We can notice clearly that the reviewers and our analyzer identified High and Low grades.

Methodology. Figure 9.10 shows that human reviewers had a significant difference between percentages of assignments of coherence level. This distribution was one of the reasons for low levels of agreement. For this section, one can notice quite unbalanced percentages of grades.

The Kappa for High level was 0.19 between reviewers, i.e. Slight agreement. An average = 0.315 with a standard deviation of 0.158 allowed to set the Low grade level at 0.156 and the High grade level at 0.474, for automating the grades of the analyzer for this section. Among human reviewers and our analyzer, a value of 0.12 of agreement was obtained for High grades. Both values are in poor performance based on Kappa. One possible cause is that the undergrad methodologies tend to have fewer steps and a simpler elaboration than graduate level methodologies.

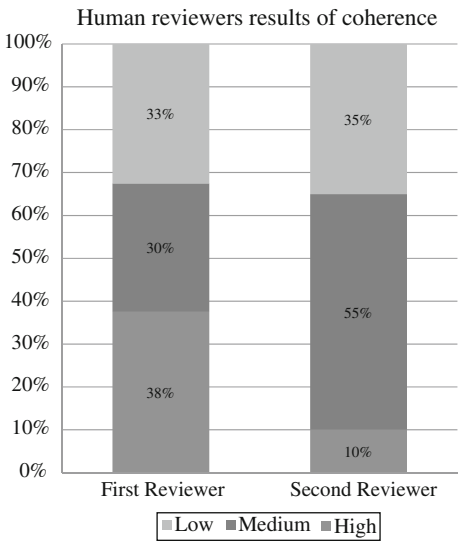

Fig. 9.9 Results of two human reviewers (Research Questions)

Table 9.8 Kappa for test corpus

Cohen kappa	Reviewers	Coherence analyzer versus reviewers
High	0.5000	0.3333
Medium	−0.0230	0.0000
Low	0.4666	0.3333
Overall	0.2727	0.2000

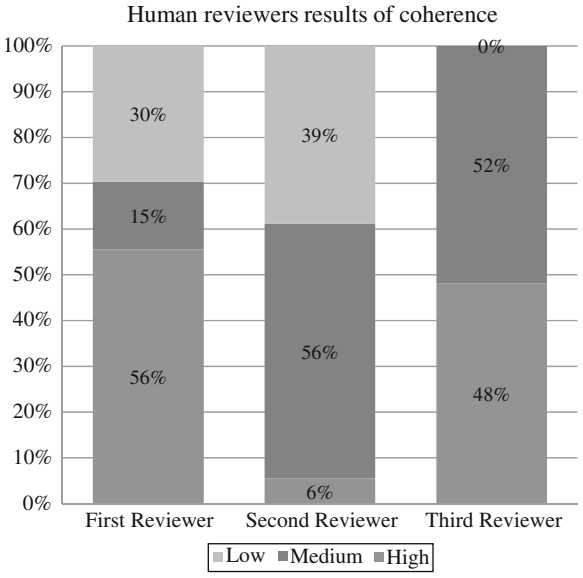

Fig. 9.10 Results of two human reviewers (Methodology)

Finally, Table 9.9 shows that the overall agreement values are lower among reviewers than our analyzer. For Low grade, the agreement amounts to zero. We could not approach this section as a classification task since one of the reviewers did not tag low grades and the rest of the reviewers did not coincide on their grades. One possible cause of this can be the variety in writing in this section that favored a disagreement between human reviewers.

Table 9.9 Kappa for test corpus

Kappa	Reviewers (Fleiss)	Analyzer versus reviewers (Cohen)
High	0.1923	0.1250
Medium	0.1900	0.2750
Low	−0.0500	0.0000
Overall	0.1250	0.1764

9.5.3 Across Section Exploration

Given that we mined the semantic spaces for the different sections, we were in the position of performing an analysis among sections. So, our second experiment allowed extracting and identifying a behavior pattern between the different sections evaluated. This exploration was motivated by the relationships that different authors state in research methodology. These relationships are suggested to students when they write their research proposal by their academic advisors.

The relationships found are from the perspective of global coherence, i.e. these are thematic relationships that allow identifying similar concepts. For example, from the corpus of research questions, ten items were taken randomly and were evaluated in the semantic space of the objectives section. The same was done with the remaining sections. It is noteworthy that these inter-sections coincide with what methodology authors suggest. These authors of methodology books suggest that once the objective is defined, this can suggest one or more research questions, which would lead the student to maintain coherence between these elements [26].

The diagram in Fig. 9.11 shows the relationships revealed among the different semantic spaces of sections. The intensity of the gray in the lines represents the strength or degree of relationship, darker color represents higher relation.

Inter-Relations. There is a high relationship among Objective, Research Questions and Hypotheses sections.

- The diagram shows a medium relation between Hypotheses and Research Questions.
- Another aspect shown in the diagram is the low relation between Objectives and Justification elements.
- Also Objective and Conclusions sections show a medium relation.
- Hypothesis, Research Questions and Conclusions showed a medium relation between semantic spaces and elements of their corresponding corpus.

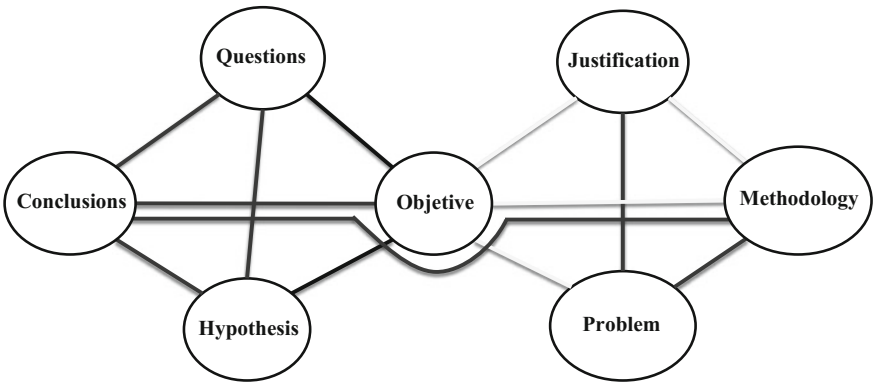

Fig. 9.11 Pattern of inter-relations among sections

These observed behaviors were revealed from corpora. Somehow, the recommendations that instructors provide to their students, lined up when crossing semantic spaces. This knowledge was supposed at the beginning of our experiments, but the detected behaviors reinforce the academic advisors recommendations (from the perspective of global coherence).

9.6 Analysis and Discussion of Results

We observed that the levels of agreement in the Low case is Moderate and Medium level is Fair, the overall level of agreement between humans and the analyzer was Fair. We conclude that the analyzer would have an acceptable support for the student and academic advisor in the process of preparing the proposal draft.

After comparing the statistical results in terms of the Kappa coefficient of agreement, we also performed a qualitative analysis between the results of coherence analyzer and the process of reviewing a proposal draft, i.e. the advisor would expect that the analyzer was a first filter so that when the drafts reach him, at least have a Medium or High Level.

Under this premise, the results of our analyzer match the concept of a strict filtering reviewer, because it provided low and medium values in most test sections. We can observe that if our system does not achieve at this time a higher level of agreement in the High grade level, this is not a problem since the analyzer is being strict to assign the high level.

In the experiment, the analyzer evaluated as Medium the few highest levels assigned by the reviewers. If the analyzer behaves more flexible and allows high level to sections that have to be of a medium or low level, this could cause a burden to the academic advisor, failing to support in review.

Finally we note that between the coherence analyzer and human evaluators, the agreement is Moderate for low levels, bringing confidence that the analyzer is identifying those sections that were classified as Low by reviewers. After assessing coherence, the analyzer as part of a system, can trigger feedback to the student for any of the seven selected sections in the draft. This is further elaborated later on in this chapter.

9.6.1 Across Section Exploration

Given the results depicted in Fig. 9.11, we can suggest that a student should review these three elements together when elaborating a proposal draft: objectives, questions and hypotheses.

The diagram also showed a medium relation between hypotheses and research questions sections, suggesting that the questions should be considered when drafting the hypothesis.

Another detail shown in the diagram is the low relation between Objectives and Justification sections. This relationship can be caused by the varied nature of justifications, since these can be economic, efficiency, capacity, to response to a need for the project, and so on, and could not be related to the stated objectives.

Hence, a student can have some freedom to write independently these two section when writing the proposal. This does not mean that both sections do not have to agree with the Problem Statement.

9.7 System Overview

Despite in essence the coherence analyzer described in the chapter performs student text grading, this is intended for text in process of improvement (i.e. prior to submission). In consequence, our approach final aim is to support students online. So, the coherence analyzer is embedded in an ITS, to enable students to improve their first draft, working on each section at a time, either by typing or pasting for analysis. In addition, students receive feedback so they can improve the document in progress.

9.7.1 Intelligent Tutoring System

The intelligent tutor is illustrated in Fig. 9.12 where the Domain Module includes information (material) concerning the definition of global coherence and what is expected to contain the different sections, regarding the concept of coherence. Also we present material about the structure that should contain a proposal draft.

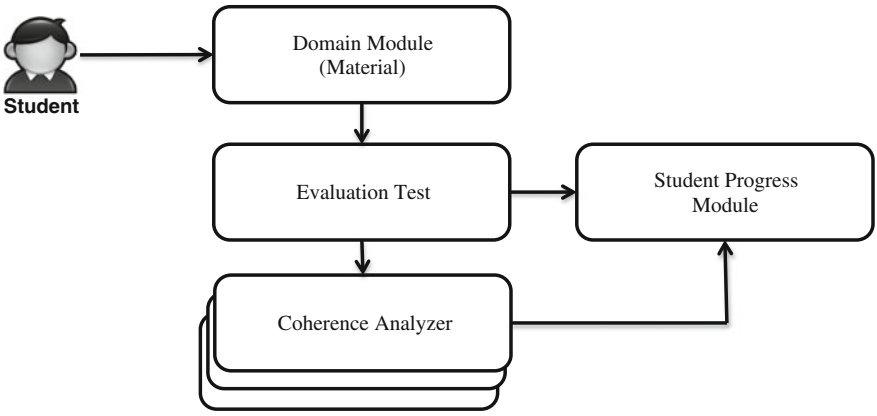

Fig. 9.12 Model of intelligent tutoring system

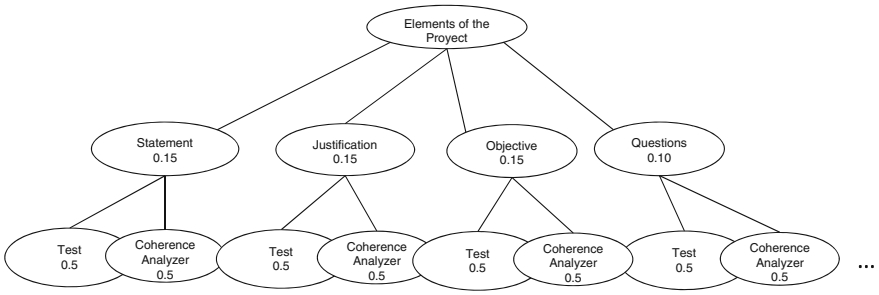

Fig. 9.13 Network used in student progress model of tutor

A test is applied to validate the reading of materials and then practical exercises are suggested and applied using the Coherence Analyzer to achieve a high level of coherence in the student text productions. The results of the test and coherence analysis are sent to the student progress module (SPM) to update the knowledge state of the student, represented in a network. The SPM records the student's progress in the network representation, which is partially depicted in Fig. 9.13 (only four of the seven sections are illustrated to avoid clutter the diagram). When the student completes the test, the value of the test node element is updated and the SPM calculates the student's progress for the parent node, considering the weights assigned to each question in the test.

Similarly, when doing the exercises with the Coherence Analyzer, the corresponding node in the network is updated and the SPM estimates the student's progress for the parent node using the weights assigned. Figure 9.13 illustrates the weights assigned to each node according to the experience of instructor.

For instance, in the Test node of the Problem Statement, a weight of 50 % of the parent node (Statement) is assigned, which includes five questions to verify that the student has read the pertinent information.

Fig. 9.14 Coherence analyzer (in Spanish)

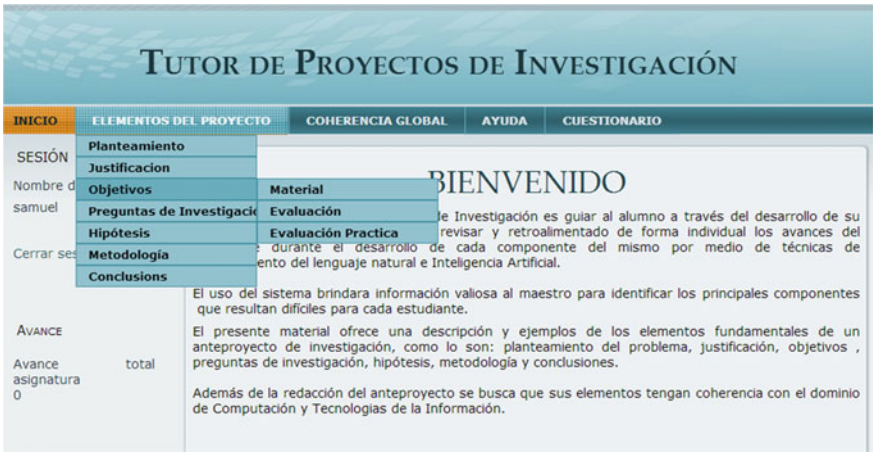

Fig. 9.15 Main menu of ITS (in Spanish)

9.7.2 Web Interface

The ITS is developed in PHP for easy access via web and the network structure is stored in a MySQL database, the coherence analyzer is developed in Python given the easy access to processing tools for natural language.

Figure 9.14 shows the graphical interface (in Spanish) of the tutoring system in which we can observe the login section to the left.

Figure 9.15 depicts the menu on the top to access the elements (sections) of the writing project (in Spanish Elementos del proyecto). For each element, there are three sections: material, test, and practical evaluation.

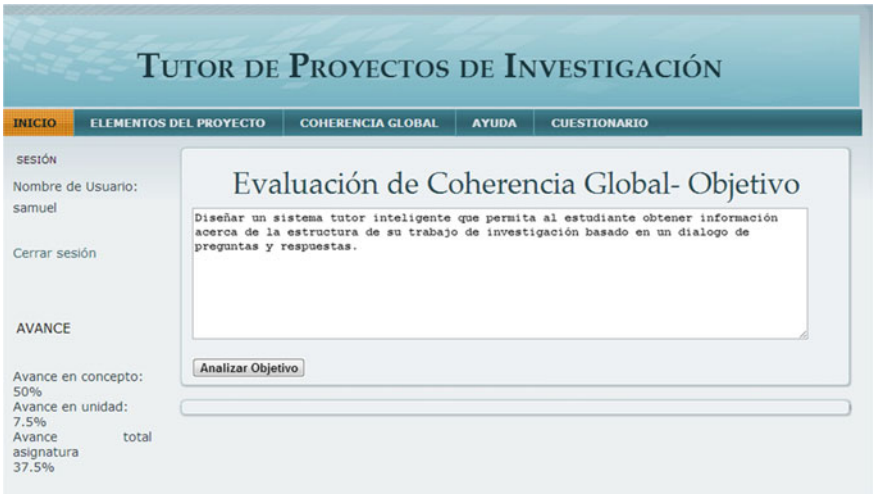

Fig. 9.16 Coherence analyzer (in Spanish)

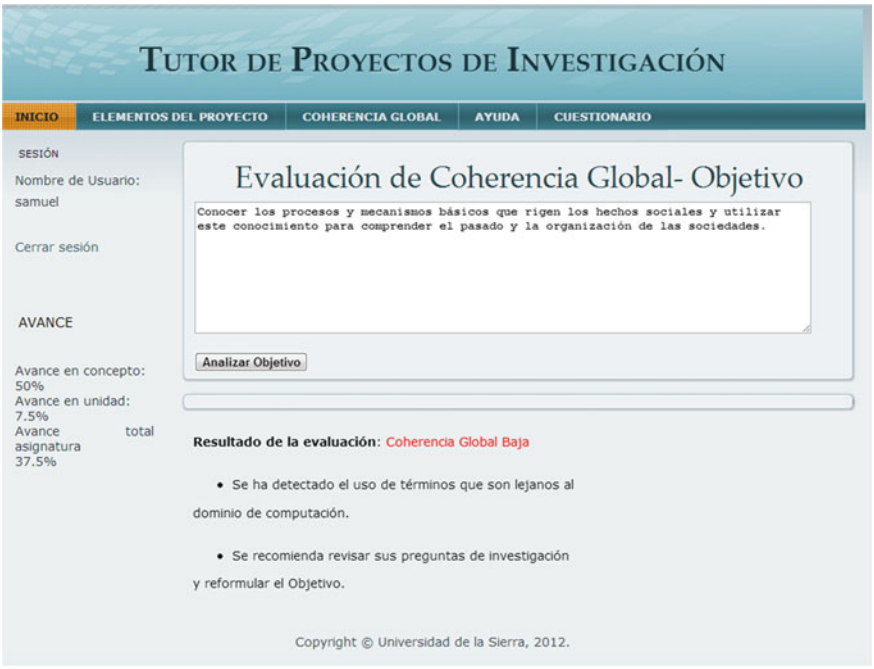

Fig. 9.17 Results of coherence analyzer (in Spanish)

Figure 9.16 below shows the coherence analyzer and the report of student progress (in Spanish *Avance*) in percentages in the bottom left part of the screen. This screen snapshot also illustrates an objective text ready for coherence analysis.

The report generated by the coherence analyzer can be seen in Fig. 9.17. In this case the level of coherence found in the text is Low. In consequence, the tutor makes suggestions to the student, who has to rewrite the objective text. Once the student reaches a High level in coherence, the progress on the left side is updated, and he can move to work the next section of the draft.

9.8 Conclusions

The mining technique allowed evaluating the global coherence of seven sections in proposal drafts, reaching an acceptable result of the percentage of agreement respect to human reviewers. It was crucial to have a gold standard to compare our results.

The exploration across sections performed after mining domain knowledge, uncovered a consistent interrelationship among them, according to methodology authors. This was a newly developed technique for additional exploration and validation of mined knowledge.

We will continue increasing the size of the corpus, so that the analyzer has a wider coverage, since the computing and information technologies domain is quiet extensive and constantly growing. We also need additional good examples for certain sections (e.g. conclusions or justification) to mine and improve their assessment.

In these initial experiments, the evaluation of coherence analysis was important to identify the student level, but could be improved by evaluating additional aspects in texts such as lexical richness [27] or local coherence. This will help students to improve their writing, and academic adviser would have more time to review the contents of the proposal documents.

We expect that this computational tool generates in students a motivation to develop their proposal drafts and this analyzer will contribute to the advance in their writings. We currently have a web interface for the student to evaluate the draft in coherence analysis. Bringing our model to a different domain does not seem too challenging, neither moving it to a different language, assuming similar language processing resources and corpus are available.

The approach discussed in this chapter contributes to the following topics: (a) web mining of educational sources; (b) mining of assessment produced by the learner educational system interactions; (c) DM applied to the personalization of educational content and services; and (d) information repositories oriented to the educational field.

As far as we are aware of related work, this coherence analysis is the first to mine existing resources by proposal sections and specific for computer science and information technologies. Besides coherence, we also plan to mine language models to guide in the formulation of specific sections in proposal texts. Also we are in the process of developing a method to identify answers to methodological questions within the elements and objective justification of a proposal draft. In addition, it has the potential to be extended to other engineering domains (e.g. electrical, electronics, control, mechanical, etc.).

We foresee an experiment that includes a pilot test with a control and experimental group of students.

Acknowledgments We thank the reviewers: Rene Castro M., Claudia I. Esquivel L., J. Miguel García G., Ramón Cárdenas G., Israel Chávez G., Orlando Madrid M., and Raúl Beltran Q. This research was supported by CONACYT, México, through the scholarship 1124002 for the first author. The second author was partially supported by SNI, México.

References

1. Luan, J.: Data mining and its applications in higher education. New Dir. Inst. Res. **2002**(113), 17–36 (2002)
2. Vilarnovo, A.: Coherencia textual: ¿Coherencia Interna o Coherencia Externa? Estudios de Lingüística **6**, 229–240 (1990)

3. Louwerse, M.M.: A concise model of cohesion in text and coherence in comprehension. Revista Signos **37**(56), 41–58 (2004)
4. Skogs, J.: Subject line preferences and other factors contributing to coherence and interaction in student discussion forums. Comput. Educ. **60**(1), 172–183 (2013)
5. Medve, V.B., Takac, V.P.: The influence of cohesion and coherence on text quality: a cross-linguistic study of foreign language learners written production. In: Piechurska-Kuciel, E., Szymańska-Czaplak, E. (eds.) Language in cognition and affect. Second language learning and teaching, pp. 111–131. Springer, Heidelberg (2013)
6. Yannakoudakis, H., Briscoe, T.: Modeling coherence in ESOL learner texts. In: 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pp. 33–43. Association for Computational Linguistics, Stroudsburg (2012)
7. Higgins, D., Burstein, J., Marcu, D., Gentile, C.: Evaluating multiple aspects of coherence in student essays. In: Human language technology conference/North American chapter of the Association for Computational Linguistics, pp. 185–192. Association for Computational Linguistics, Boston (2004)
8. Miltsakaki, E., Kukich, K.: Evaluation of text coherence for electronic essay scoring systems. Nat. Lang. Eng. **10**(1), 25–55 (2004)
9. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**, 391–407 (1990)
10. Landauer, T., Dumais, S.: A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. **104**, 211–240 (1997)
11. Foltz, P., Kintsch, W., Launder, T.: Textual coherence using latent semantic analysis. Discourse Process. **25**, 285–307 (1998)
12. Hofmann, T.: Probabilistic latent semantic indexing. In: 22nd international conference on research and development in information retrieval, pp. 50–57. ACM, NY (1999)
13. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
14. Lee, S., Baker, J., Song, J., Wetherbe, J.C.: An empirical comparison of four text mining methods. In: 43rd Hawaii international conference on system sciences, pp. 1–10. IEEE Computer Society, Washington (2010)
15. Zhang, M., Yang, H., Ji, D., Teng, C., Wu, H.: Discourse coherence: lexical chain, complex network and semantic field. In: Ji, D., Xiao, G. (eds.) Chinese Lexical Semantics. LNCS, vol. 7717, pp. 756–765. Springer, Heidelberg (2013)
16. Dessus, P.: An overview of LSA-based systems for supporting learning and teaching. In: Dimitrova, V., Mizoguchi, R., du Boulay, B., Graesser A. (eds.) Conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modeling, pp. 157–164. IOS Press, Amsterdam (2009)
17. Southavilay, V., Yacef, K., Calvo, R.A.: Process mining to support students' collaborative writing. In: Baker, R.S.J.D., Merceron, A., Pavlik Jr., P.I. (eds.) 3rd International Conference on Educational Data Mining, pp. 257–266. International Educational Data Mining Society, Pittsburgh (2010)
18. Jiang, H., Huang, G., Liu, J.: The research on CET automated essay scoring based on data mining. In: Zhou, M., Tan H. (eds.) Advances in Computer Science and Education Applications. Communications in Computer and Information Science, vol. 202, pp. 100–105. Springer, Heidelberg (2011)
19. Villalón, J., Kearney, P., Calvo, R.A., Reimann, P.: Glosser: enhanced feedback for student writing tasks. In: International conference on advanced learning technologies, pp. 454–458. IEEE Computer Society, Washington (2008)
20. Liu, M., Calvo, R.A., Rus, V.: Automatic question generation for literature review writing support. In: Aleven, V., Kay, J., Mostow, J. (eds.) Intelligent Tutoring Systems. LNCS, vol. 6094, pp. 45–54. Springer, Heidelberg (2010)

21. Higgins, D., Burstein, J.: Sentence similarity measures for essay coherence. In: 7th international workshop on computational semantics, pp. 77–88. Tilburg University, Tilburg (2007)
22. Vasile, R., Nobal, N.: Automated detection of local coherence in short argumentative essays based on centering theory. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing. LNCS, vol. 7181, pp. 450–461. Springer, Heidelberg (2012)
23. Kintsch, W.: On the notions of theme and topic in psychological process models of text comprehension. In: Louwerse, M., Van Peer, W. (eds.) Thematics, Interdisciplinary Studies, pp. 157–170. John Benjamins Publishing, Amsterdam (2002)
24. Berry, M.W., Dumais, S.T., O'Brien, G.W.: Using linear algebra for intelligent information retrieval. Soc. Ind. Appl. Math. **4**, 573–595 (1995)
25. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
26. Hernández, R.: Metodología de la Investigación. Mc Graw Hill, México (2006)
27. García-Gorrostieta, J.M., González-López, S., López-López, A., Carrillo, M.: An intelligent tutoring system to evaluate and advise on lexical richness in students writings. In: Hernández-Leo, D., Ley T., Klamma, R., Harrer, A. (eds.) EC-TEL 2013. LNCS, vol. 8095, pp. 548–551. Springer, Heidelberg (2013)