J Braz Comput Soc (2013) 19:59–73
DOI 10.1007/s13173-012-0079-1

ORIGINAL PAPER

A coherence analysis module for SciPo: providing suggestions for scientific abstracts written in Portuguese

Vinícius Mourão Alves de Souza ·
Valéria Delisandra Feltrim

Received: 14 January 2012 / Accepted: 10 May 2012 / Published online: 23 June 2012

Abstract SciPo is a writing tool whose ultimate goal is to assist novice writers in producing scientific writing in Portuguese, focusing primarily on abstracts and introductions from computer science. In this paper, we describe the development of a coherence analysis module for SciPo, which aims to automatically detect semantic coherence aspects of abstracts and provide suggestions for improvement. At the core of this new module are classifiers that identify different semantic relationships among sentences within an abstract and hence indicate semantic aspects that add coherence to the abstract. Such classifiers are based on a set of features extracted automatically from the surface of the text and by Latent Semantic Analysis processing. All classifiers were evaluated intrinsically and their performance was higher than the baseline. We also resorted to actual users to evaluate our coherence analysis module, which has been incorporated into the SciPo system, and results demonstrate its potential to help users write scientific abstracts with a higher level of coherence.

Keywords Scientific writing in Portuguese · Writing tool · Semantic coherence · Automatic analysis of coherence · Latent Semantic Analysis

1 Introduction

Abstracts are said to be a key section in scientific manuscripts (papers, dissertations, theses, etc.). Along with the title, it is used by researchers to promote their work in a given scientific community. As Feltrim et al. [17] point out, a scientific abstract should be carefully tailored so as to be complete (in the sense of providing the necessary information), interesting and informative. It should allow readers to capture the main ideas of the research being described while, at the same time, convince him/her to read the full text.

Scientific texts tend to have a well-defined structure, which can be defined as Introduction–Development–Conclusion [39]. Swales [39] adds that the Development part may unfold in either Materials and Methods and Results or Materials and Methods, Results and Discussion. The same can be said about abstracts, which tend to present a typical structural organization, especially when we consider abstracts from the same knowledge domain. The well-defined nature of the rhetorical structure of abstracts has allowed researchers to propose structure models for abstracts [7,16,21,43]. Although each model has its peculiarities, there is a clear consensus on the typical rhetorical components of such structure as well as on their order of appearance.

Based on models that take into account the rhetorical structure of abstracts, different computational tools have been developed over the past few years to assist authors in writing/revising scientific abstracts. Taking into consideration only systems that focus on the English language, it is worth mentioning AMADEUS (Amiable Article Development for User Support) [3], SciPo-Farmácia [2], Writer's Assistant [34,35], Writing Environment [28], Composer [33,36], Academic Writer [8], Abstract Helper [31,32],

This is a revised and extended version of a previous paper that appeared at STIL 2011, the 8th Brazilian Symposium in Information and Human Language Technology.

V. M. A. de Souza · V. D. Feltrim (✉)
Informatics Department, State University of Maringá,
Av. Colombo, 5.790, 87020-900 Maringá, PR, Brazil
e-mail: valeria.feltrim@gmail.com;valeria.feltrim@din.uem.br

V. M. A. de Souza
e-mail: vsouza@din.uem.br

and Mover [5]. All these systems target non-native novice writers and aim to help users write scientific texts in English. Within the specific context of Portuguese, we cite SciPo (Scientific Portuguese) [15,17], a system designed to help novice writers, specially undergraduate and graduate students from the discipline of computer science, by providing support for the writing of introduction and abstract sections of theses and dissertations. To the best of our knowledge, SciPo is the only system of this nature which targets the Portuguese language specifically.

Among other functionalities, SciPo examines the rhetorical structure of abstracts and introductions submitted for analysis and provides both criticisms and suggestions for improvement. For abstracts, the system relies on the rhetorical structure proposed by Feltrim el al. [16], which comprises six rhetorical components arranged in the following order: Background, Gap, Purpose, Methodology, Result and Conclusion. It provides feedback indicating which parts of the abstract could be improved regarding its structure. However, no attention is paid to semantic aspects of the text, such as coherence, which are essential to the readability and interpretability of the abstract.

Coherence and cohesion are responsible for adding sense to a group of words or sentences. By coherence we refer to what makes a group of words or sentences semantically meaningful. We assume that coherence refers to the establishment of a logical connection between sentences within a text. Thus, it is a principle of interpretability related to a given communicational situation and to the ability of the reader to interpret the meaning of the text. Therefore, it is bounded to the text, even though it does not depend solely on it [23]. On the other hand, meaning can only be established if we use textual elements responsible for connecting words/sentences and hence provide cohesion to the text [12]. Coherence and cohesion are closely related and this is why they are usually treated together. Here, we focus on coherence specifically and treat it as a level of semantic relationship among specific text segments. In line with van Dijk and Kintsch [41], we refer to it as semantic coherence.

We have developed a coherence analysis module (CAM) to identify semantic coherence aspects in abstracts. Here, we examine coherence by focusing on two or more rhetorical components and determining the level of semantic similarity between them. Following Higgins and Burstein [19] and Higgins et al. [20], four types of relationships among rhetorical components are considered and we have termed dimensions. These are: (1) Dimension Title: examines the semantic relationship between the Purpose sentence(s) and the title of the abstract; (2) Dimension Purpose: verifies the semantic relationship between the Purpose sentence(s) and those related to Methodology, Result and Conclusion; (3) Dimension Gap-Background: assesses the semantic relationship between Gap and Background sentences; and (4) Dimension Linearity-break: checks whether there is a break in the logical sense between adjacent sentences. Although aware that there are many aspects of a discourse that contribute to coherence, as pointed out by Foltz et al. [18], our main assumption is that a low level of semantic relationship may be interpreted as an indication of a coherence problem. Thus, this system can be used to complement SciPo's functionalities by providing users with suggestions related to semantic coherence.

To automate the analysis of each dimension, we have developed a number of text classifiers. Such classifiers are based on features that have been extracted automatically from the surface of the text and by Latent Semantic Analysis (LSA) [27] processing. With the exception of Linearity-break, all dimensions rely on the abstract's rhetorical structure, which is automatically detected according to SciPo's rhetorical structure model.

We believe our work brings innovative contributions regarding two aspects: (1) the nature of the corpus in question, since we are dealing with scientific abstracts written in Portuguese, and (2) the kind of application to which we apply automatic coherence analysis. As Burstein et al. [10] point out, there is a small body of work that has investigated the problem of identifying coherence in student writing. What is more, none has focused on scientific writing but instead on essays written by native/non-native English writers with different writing skills. In addition to the composition of the corpus, the context in which our approach applies is also different from most systems presented so far in the literature. To date coherence analysis has been applied mostly within the context of Automatic Essay Scoring [29]. Three scoring systems which consider aspects of coherence when grading essays are worth mentioning: Criterion [9,10,20], Intelligent Essay Assessor [26], and Intellimetric [14]. Unlike these systems, SciPo is a scientific writing support tool, which in other words means that we are not interested in assigning a score to the text. Our aim is instead to detect potential structure and coherence problems and give the writer constructive feedback.

This paper is organized as follows: in Sect. 2, we briefly describe the SciPo system, focusing on its main functionalities implemented so far. In Sect. 3, we detail our corpus and its annotation process as well as our proposed dimensions. In Sect. 4, we focus on the classifiers that comprise the proposed CAM and on the results of its intrinsic evaluation. Section 5 presents the CAM, how it is incorporated into the SciPo system, and the results of its evaluation by actual users. Last but not least, in Sect. 6, we draw some conclusions and offer some suggestions for further investigation.

2 The SciPo system

SciPo[1] is a web-based system whose primary purpose is to assist novice writers, specially undergraduate and graduate students, in producing scientific writing in Brazilian Portuguese. It focuses mainly on the abstract and introduction sections of dissertations and theses from computer science and was designed to help users structure their texts and make adequate linguistic choices. SciPo allows its users to choose between two working modes:

(i) A top-down process that starts from planning the rhetorical structure and then tackles the writing itself. This mode was inherited from the AMADEUS project [3];
(ii) A bottom-up process in which the system automatically detects and analyses the rhetorical structure of the text submitted.

In fact, these two modes are different starting points for the same cyclical process of refinement given that the rhetorical structure detected and assessed in (ii) can be improved using the resources available in (i).

The system contains four knowledge bases, namely: the Abstract Case Base, Introduction Case Base, Rules and Similarity Measures, and Critiquing Rules. The Abstract Case Base includes 52 examples of schematic structures taken from authentic abstracts as well as the description of the rhetorical components, strategies and lexical patterns for each case. Similarly, the Introduction Case Base contains 48 examples of schematic structures for introductions and the description of the rhetorical components, strategies and lexical patterns for each case. For both case bases, all information was manually annotated according to appropriate rhetorical structure models [4,16]. The user can freely browse these databases and search for occurrences of any given rhetorical structure.

Table 1 shows the rhetorical structure model used for abstracts along with a brief description of the function served by each component. Figure 1 illustrates how an abstract may be annotated according to its rhetorical structure. All lexical patterns have been underlined for emphasis. Given that SciPo's corpus is in Portuguese, for convenience, the example in Fig. 1 has been translated into English.

As for the third knowledge base, Similarity Rules and Measures, they refer to rules established according to similarities among lists (pattern matching) and to nearest neighbor matching [24]. These rules are used to retrieve a given rhetorical structure, as requested by the writer.

Last but not least, the Critiquing Rules are based on prescriptive guidelines for good writing and on structural problems observed in the annotated corpus, as an attempt to anticipate and correct ill-formed structural patterns that the writer might construct. These rules cover two distinct types of problems: content deviations (absence of structural components) and order deviations (occurrence of a given structural component in relation to the overall structure). In short, this base consists of four classes of rules: critical comments on (1) the content and (2) the order, and suggestions for improving (3) the content and (4) the order.

Table 1 SciPo's rhetorical components for abstracts and their functions within the abstract

Component	Function
Background	Presents knowledge already accepted by the scientific community which is used to contextualize the study
Gap	States a research problem/gap within a specific research area, preparing the ground for the purpose of the study
Purpose	States the purpose/aims/goals of the study
Methodology	States the materials and methods used in the study
Result	States the main results of the study
Conclusion	States the study's conclusions/limitations/ contributions

A fifth element of SciPo's architecture is a text classifier which automatically detects the rhetorical structure of an abstract. Named AZPort, it is a Naive Bayesian classifier that implements the Argumentative Zoning approach proposed by Teufel and Moens [40], adapting it to the context of scientific abstracts written in Portuguese. Following the structural components of the rhetorical model proposed by Feltrim et al. [16], AZPort assigns one of the following labels to each input sentence: Background, Gap, Purpose, Methodology, Result, and Conclusion. Further details about AZPort can be found in [15] and [17]. Using AZPort, we are therefore able to incorporate the bottom-up process into SciPo. Figure 2 presents a simplified version of the SciPo's architecture, showing how the bottom-up and top-down processes relate to each other and to the knowledge bases.

As shown in Fig. 2, once the user has decided on a given rhetorical structure, which may have been either automatically detected (bottom-up process) or explicitly constructed (top-down process), he/she receives some feedback from the system. This procedure is repeated as many times as necessary until an acceptable structure has been built. The user can then recover authentic examples from the corpus and use the lexical patterns provided in his/her writing.

3 Corpus and annotation

To determine what kind of semantic relationships would have an impact on coherence in scientific texts written in Portuguese, we have compiled and annotated a corpus of 385 abstracts written in Portuguese by undergraduate students.

[1] The SciPo system as described in the Sect. 2 is available at http://www.nilc.icmc.usp.br/~scipo.

1 Background
"The research article (RA) or paper is one of the most important genres that both scientists and engineers will write."

2 Gap
"When faced with the tasks of reading and writing a complex technical paper, many nonnative scientists and engineers (...) lack an adequate knowledge of commonly used structural patterns at the discourse level."

3 Purpose
"In this paper, we propose a novel computer software tool that can assist these people in the understanding and construction of technical papers (...)."

4 Methodology
"The software uses a supervised learning approach, in which the system first "learns" the characteristic features of text structure in a particular discipline using a small number of training examples."

5 Result
"We can see that the system performs consistently across the different data sets, with an average accuracy of 68%."

6 Conclusion
"The system is tested using research article abstracts and is shown to be fast, accurate, and useful aid in the reading and writing process."

Fig. 1 Example of an annotated abstract according to its rhetorical structure with lexical patterns *underlined* (adaptated from Feltrim et al. [17])

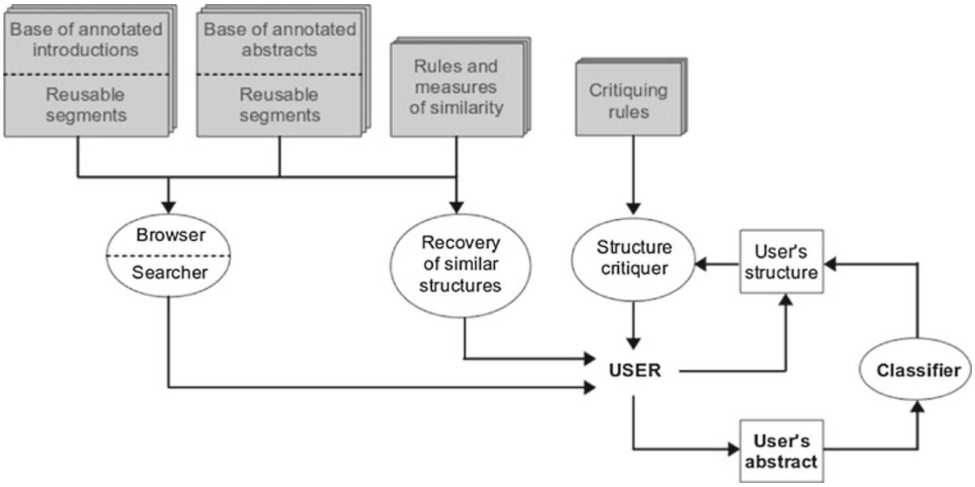

Fig. 2 Simplified version of SciPo's architecture [17]

Since significant differences can be found between European and Brazilian Portuguese in terms of vocabulary and syntactic constructions, we have opted to compile a corpus of abstracts of the latter variety. This is mainly because the SciPo system targets Brazilian students specifically.

All abstracts were extracted from monographs written as one of the requirements for being awarded a BS degree in computer science. These monographs date from 1999 to 2009 and come from different fields within computer science, such as database systems, artificial intelligence, software engineering, computer networks, digital systems, distributed systems, programming languages and image processing. They were collected at three Brazilian universities: the State University of Maringá, where we could collect the abstracts directly from their authors, the State University of Londrina, and the Federal University of Pelotas, where we have collected the abstracts by accessing their digital libraries.[2,3]

Once we had our abstract corpus ready, its annotation was processed in two stages: (1) rhetorical structure annotation, and (2) annotation of coherence-related aspects according to the proposed dimensions. Both annotation processes are described below.

3.1 Rhetorical structure annotation

The first stage of the annotation phase consisted of assigning tags to the abstract title, start and end of each sentence, and their classification according to the six abovementioned rhetorical categories, following the rhetorical components proposed by Feltrim et al.'s structure model [16]. At this point, it is worth mentioning that a rhetorical component may be realized by one or more sentences. For the former case, all the sentences are classified according to the category in question.

To annotate the rhetorical structure of each abstract, we have used the aforementioned AZPort classifier. As reported in [17], AZPort was trained and tested by applying 13-fold cross-validation to a set of 52 abstracts from the CorpusDT [16], which comprises 320 sentences. The system's accuracy rate was 74 %. The automatic annotation was also evaluated

[2] Document Archiving and Indexing System of the Computer Science Department, State University of Londrina, available at: http://www2.dc.uel.br/nourau/.

[3] Digital Collections of the Library of Science and Technology, Federal University of Pelotas, available at: http://www.ufpel.tche.br/prg/sisbi/bibct/acervodigital.html.

in relation to that by a human annotator, by calculating the Kappa coefficient K [37] as follows:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is pairwise agreement and $P(E)$ is random agreement. A Kappa value may range from −1 to 1. The former indicates maximal disagreement and the latter suggests perfect agreement. A kappa value of 0 implies that agreement between annotators is no greater than it would be expected by chance, thus following the same distribution as the observed one.

The level of agreement between AZPort and the human annotator was $K = 0.65$. This kappa value is in fact fairly similar to that calculated when the same 52 abstracts were annotated by three human specialists ($K = 0.69$). Even so, AZPort 's categorization was manually revised by one human annotator so as to correct potential errors and hence minimize the chances of any noise from the automatic annotation interfering with the coherence annotation.

A total of 2,293 sentences were automatically annotated and manually revised. Table 2 presents the frequencies of each rhetorical component in the annotated corpus. As can be seen, Purpose sentences are the most frequent, occurring in nearly all abstracts (97.40 %, 375 abstracts). It is followed by Background (68.05 %), Result (55.32 %), Gap (40.51 %), Methodology (37.66 %), and Conclusion (23.11 %).

The distribution of annotated sentences across the six rhetorical components is presented in Table 3. It can be seen that Background is the most frequent category when the number of sentences is considered (35.23 %, 808 sentences), followed by Result (19.67 %), Purpose (18.58 %), Methodology (11.90 %), Gap (9.38 %), and Conclusion (5.24 %). It is also worth noting that while Purpose sentences occur in most abstracts (97.4 %, as shown in Table 2), the number of Purpose sentences (426 sentences) is lower than the number of Background (808 sentences) and Result (451 sentences) sentences. This is explained by the fact that the 426 Purpose sentences are distributed across 375 abstracts, leading to an average of 1.13 Purpose sentences per abstract. By way of contrast, the 808 Background sentences are spread across 262 abstracts, leading to an average of 3.08 sentences per abstract. We believe that this higher number of Background sentences may be explained by the composition of the corpus. When it comes to monograph abstracts, there is no restriction on the maximum number of words and hence authors tend to write more sentences to contextualize their work. The same does not apply to abstracts from scientific papers, which tend to be limited in length, leading writers to focus on Purpose and Result.

Table 2 Frequency of rhetorical components in the corpus

Category	Abstracts (*N*)	Frequency (%)
Background	262	68.05
Gap	156	40.51
Purpose	375	97.40
Methodology	145	37.66
Result	213	55.32
Conclusion	89	23.11

Table 3 Distribution of sentences by rhetorical component

Category	Sentences (*N*)	Distribution (%)
Background	808	35.23
Gap	215	09.38
Purpose	426	18.58
Methodology	273	11.90
Result	451	19.67
Conclusion	120	05.24
Total	2,293	100

3.2 Annotation of coherence-related aspects

In the second stage of the annotation phase, we identified and annotated semantic relationships between specific rhetorical components of scientific abstracts, bearing in mind that the resulting information was intended to be used as a resource to generate useful feedback to SciPo users. For doing so, we have adapted the dimensions proposed by Higgins et al. [20] and proposed four kinds of semantic relationships between rhetorical components which, as mentioned earlier, have been termed dimensions. These are: (1) Dimension Title, (2) Dimension Purpose, (3) Dimension Gap-Back-ground, and (4) Dimension Linearity-break.

To check reproducibility, we conducted annotation experiments with two human annotators who were familiar with the corpus domain and scientific writing. Here again, we used the Kappa coefficient to measure the level of agreement between them.

The four dimensions, originally proposed in [38], are described in the following sections. We also provide statistics on the annotated corpus as well as the Kappa values for the agreement experiments.

3.2.1 Dimension Title

In this dimension, we have examined whether each sentence of the abstract is semantically similar to the title. If the sentence was found to present a high semantic similarity to the title, it was labeled as *high*. Otherwise, it was labeled as *low*.

Table 4 Semantic relationship between sentences from the abstract and the title sentence

Categories	High (*N*)	Low (*N*)
Background	364	444
Gap	104	111
Purpose	355	071
Methodology	139	134
Result	220	231
Conclusion	061	059
Total	1,243	1,050

Table 5 Semantic relationship between sentences from various categories with Purpose sentences

Categories	High (*N*)	Low (*N*)	N/A (*N*)
Background	378	380	050
Gap	129	079	007
Purpose	–	–	426
Methodology	171	082	020
Result	264	135	052
Conclusion	074	028	018
Total	1,016	704	573

Here, we have opted for a binary scale due to the subjective nature of the task. As previously mentioned, we have conducted an agreement experiment with two human annotators. They were initially asked to annotate a subset of ten abstracts. After this training phase, they annotated a subset of 40 abstracts that had been randomly selected from the corpus. This figure represents nearly 10 % of the overall number of abstracts in the corpus and comprises 209 sentences. The resulting Kappa for their level of agreement was approximately 0.6.

Out of a total of 2,293 sentences, 1,243 (54.20 %) were ranked as having *high* semantic similarity with the title and 1,050 (46.80 %) were ranked as *low*. Table 4 presents the distribution of *high* and *low* sentences according their semantic similarity with the title across the six potential rhetorical categories.

As we can observe in Table 4, Purpose sentences tend to present a high level of semantic similarity to the title, since 83.33 % of such sentences were ranked as *high*. This figure is much higher than the average percentage of *high* sentences for all remaining components, which is 48.79 %. In fact, the title should indicate the main topic covered in a scientific text and the same is expected from the purpose of the abstract, even if in a concise form. We understand that lack of semantic relationship between purpose sentence(s) and the title may be interpreted as evidence for two possible situations: (1) the title is inappropriate for the abstract or (2) the abstract may have coherence problems.

On the other hand, Background sentences tend to have a low level of semantic similarity to the title. Over half of the overall number of sentences (54.95 %) was ranked as *low*. We ascribe that to the fact that Background sentences usually appear at the beginning of the abstract so as to place the research within a broader context. Thus, they may not be directly related to the main topic of the research being presented but, rather, address questions or state facts which will prepare the reader to understand the motivations behind the study being presented. We assume that a low level of semantic similarity between the title and Background sentences cannot be viewed as an indication of a coherence problem.

As for the remaining rhetorical categories (Gap, Methodology, Result, and Conclusion), their level of semantic similarity to the title is evenly balanced, with an average percentage of 50.5 % of *low* sentences and 49.5 % of *high* sentences over a total of 1,059 sentences. In this study, we find that the relationship between sentences within these categories and the title depends on aspects other than coherence, such as the very nature of the research being reported. This is mainly why we consider that lack of a strong relationship between Gap, Methodology, Result, and Conclusion sentences and the title cannot be interpreted as an indication of a coherence problem.

3.2.2 Dimension Purpose

For each abstract from the corpus, we have examined the semantic similarity between Purpose sentences and all other sentences of the abstract. If the sentence was found to be closely related to the Purpose component, it was labeled as *high*. Otherwise, it was labeled as *low*. The label N/A was assigned to sentences of abstracts which do not have Purpose sentences or to sentences classified as Purpose themselves. Like in the case of the dimension Title, we have resorted to the Kappa statistics to measure the agreement between two human annotators over a randomly selected subset of 167 sentences. The resulting value was approximately 0.8. The human agreement experiment for this dimension was carried out in the same way as that for the dimension Title.

Apart from 573 sentences labeled as N/A (426 Purpose sentences and 147 sentences distributed across all five categories other than Purpose), 1,720 sentences were labeled as *high*/*low* for this dimension. Within these, 1,016 (59.07 %) sentences were ranked as having *high* semantic similarity with the Purpose component and 704 (40.93 %) sentences were ranked as *low*. The distribution of *high* and *low* sentences across all six rhetorical categories is presented in Table 5.

We find that Conclusion, Methodology, and Result sentences tend to present a *high* level of semantic similarity to the Purpose sentences, as shown by their percentage of sentences ranked as such: 72.55, 67.59, and 66.17 %, respectively. It is worth noting that these figures could be even higher since most of these sentences restate the content of the Purpose component. However, for doing so, writers may resort to anaphoric expressions. Since we rely solely on string matching to identify coreferential entities, and entity names introduced in the Purpose component may not always be explicitly reintroduced in Conclusion, Methodology, and Result components, it decreases the level of semantic similarity. Thus, for these specific cases, although we have found a close relationship between sentences from the abovementioned categories and the Purpose, we have labeled them as *low*.

Here again, the general nature of Background sentences can be said to account for the category having the highest percentage of *low* sentences (50.13 %). In fact, Background sentences tend to be closely related to Gap sentences, which in turn are strongly related to the Purpose component. A total of 62.01 % of the analyzed Gap sentences were labeled as *high*. So, we understand that the low level of semantic relationship between Background and Purpose sentences cannot be regarded as a potential coherence problem.

According to Higgins et al. [20], the semantic relationship among the various rhetorical components dictates the global coherence of the text. Thus, an abstract will not be easily readable and entirely understandable if some rhetorical components are not semantically related to each other. Taking into consideration the rhetorical structure model used for the annotation of our corpus, we expect the Purpose component to present a high level of semantic similarity to the Methodology, Result and Conclusion components. Conversely, absence of a close relationship between these components and the Purpose may be an indication of a coherence problem.

3.2.3 Dimension Gap-background

Taking into consideration all Gap and Background sentences from the corpus, we have examined the semantic similarity between these categories within each abstract. Gap sentences were labeled as *yes* if they were found to be closely related to at least one sentence from the Background component. Otherwise, they were labeled as *no*.

With the exception of 32 sentences from abstracts which do not have Gap/Background categories, 183 sentences were considered for this dimension. Within these, 74.86 % (137 sentences) were labeled as *yes* and 24.14 % (46 sentences) were labeled as *no*. Like in the case of all other dimensions, we have used the Kappa statistics to measure the agreement between two human annotators over a randomly selected subset of 46 sentences from the corpus. The result was approximately 0.7. The human agreement experiment for this dimension was carried out in the same way as the aforementioned dimensions.

As previously mentioned, Background sentences tend to be more closely related to Gap than to Purpose sentences. Thus, the Gap component is expected to have a high semantic relationship with at least one Background sentence. In our view, absence of relationship between these components can said to be an indication of a coherence problem.

3.2.4 Dimension Linearity-break

For this dimension, we have examined whether there is a linearity break in the logical sense between adjacent sentences, that is, whether the sentence in question is semantically related to its preceding and subsequent sentences. Unlike all other dimensions, Linearity-break does not dependent on the rhetorical structure of the abstract. A human annotator was instructed to label sentences as *yes* whenever a logical connection between the sentence under analysis and its previous and/or its subsequent sentence was difficult to establish. Otherwise, the annotator was instructed to label sentences as *no*.

Out of 2,293 sentences, 7.14 % (153 sentences) were labeled as *yes* and 92.86 % (2,140 sentences) were labeled as *no*. Within the 153 sentences labeled as *yes*, 26.8 % (41 sentences) are Result sentences, which is the rhetorical category with the highest proportion of *yes* sentences. Gap is the rhetorical category with the lowest number of *yes* sentences, with only 4.57 % (7 sentences) labeled as *yes*.

These results indicate that, within the scope of our study, it is unusual to find sentences which are not related to their surrounding sentences. In addition, we also find that most sentences labeled as *yes* relate with some other part of the text which may not necessarily be their neighboring sentences. This brings extra complexity to the annotation and analysis of this dimension. As a matter of fact, this dimension indicates very local coherence issues which we believe to be frequent in texts with problems more serious than those observed in the texts analyzed here. For these reasons, we have decided to discard this dimension from the automatic CAM, which we describe in the following section.

4 Automatic detection of semantic coherence

As previously stated, the purpose of this study is to develop a complementary module for the SciPo system with a view to identifying aspects related to semantic coherence in scientific abstracts written in Portuguese. This new module is based on three out of the four dimensions presented in the previous section, namely: dimension Title, dimension Purpose, and dimension Gap-Background. For this new functionality to work, the system needs to automatically identify potential coherence problems so that appropriate suggestions can be selected and presented to the writer. Here, the automatic analysis of the aforementioned dimensions is accomplished by means of text classifiers, as we shall see next.

All text classifiers were induced by machine learning algorithms based on features extracted from the surface of the text and by LSA processing [27]. LSA is a well-known statistical method for the extraction and representation of knowledge from corpus. Its basic idea is to create a semantic space in which terms are regarded as similar if they occur in the same context. Similarity between concepts related to two words/sentences can be calculated by the cosine product of vectors that represent the target words/sentences. This is shown in the following equation [30]:

$$\mathrm{sim}(X, Y) = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}}$$

where $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ are vectors with n dimensions and represent the texts to be compared using the *bag of words* model. The similarity value ranges from $[-1, 1]$, where -1 is the lowest possible value of similarity and 1 is the highest. LSA results can be improved by pre-processing the corpus before similarity calculations are performed. In the specific case of our study, the pre-processing phase consisted mainly of case-folding, stemming and stopwords removal.

Since some of our features rely on the rhetorical structure of the abstract, we have used AZPort to automatically label sentences according to their rhetorical category. This automatic annotation was then manually validated so as to correct inadequate labeling and hence avoid noise in the processing of the coherence dimensions. In fact, AZPort is used in the prototype of the semantic CAM and the user can correct its predictions whenever he/she regards it as incorrect.

For inducing the classifiers, we have opted for Platt's [22] Sequential Minimal Optimization (SMO) algorithm. SMO is widely used for training support vector machines (SVM) [42], a machine learning method based on statistical theories in which an optimal hyperplan is created to distinguish categories. The method has been employed in several pattern recognition tasks such as text classification [1], spam detection [13], and coherence analysis [20]. This is why we assume SMO is suitable for the task we have proposed. Details on the training and testing of the SMO algorithm are presented in Sect. 4.2.

4.1 Extracted features

All sentences were automatically analyzed according to a set of 13 features. All features were automatically extracted and used to induce the classifiers. The complete set of features is:

1. Rhetorical category of the sentence under analysis. Possible values are B, G, P, M, R, or C, standing for Background, Gap, Purpose, Methodology, Result, and Conclusion, respectively;
2. Rhetorical category of the sentence that precedes the one under analysis. Possible values are B, G, P, M, R, C, or N/A. The N/A value is assigned when the sentence under analysis is the first one of the abstract;
3. Rhetorical category of the subsequent sentence. Possible values are B, G, P, M, R, C, or N/A. The N/A value is assigned when the sentence under analysis is the last one of the abstract;
4. Presence of words that may characterize an anaphora. Possible values are Yes or No, which are calculated on the basis of a list of Portuguese pronouns that can be used anaphorically, such as "ele/ela (he/she/it)", "deste/desta (of this)", "dele/dela (his/hers)", etc;
5. Position of the sentence within the abstract, estimated in relation to the beginning of the abstract. Possible values are integer numbers starting from 0 (zero);
6. Presence of words that may characterize some kind of transition. Possible values are Yes or No, which are established on the basis of a list of expressions such as "no entanto (however)", "embora (although)", etc;
7. Length of the sentence under analysis, measured in words. Possible values are integer numbers starting from 1 (one);
8. Length of the title, measured in words. Possible values are integer numbers starting from 1 (one);
9. Semantic similarity (LSA) score between the sentence under analysis and its preceding sentence. Possible values are real numbers between -1 and 1. This feature is extracted only when feature number 2 is other than N/A;
10. Semantic similarity (LSA) score between the sentence under analysis and its subsequent sentence. Possible values are real numbers between -1 and 1. This feature is extracted only when feature number 3 is other than N/A;
11. Semantic similarity (LSA) score between the sentence under analysis and the abstract title. Possible values are real numbers between -1 and 1. This feature is extracted only when feature number 1 has the value P;
12. Semantic similarity (LSA) score between the sentence under analysis and the sentence(s) of the abstract classified as Purpose. Possible values are real numbers between -1 and 1. This feature is extracted only when feature number 1 has the following values: R, M, or C;
13. Maximum Semantic similarity (LSA) score between the Gap and the Background sentences of the abstract. Possible values are real numbers between -1 and 1. As some abstracts may not include these categories, this feature is calculated only for abstracts with sentences from both B and G.

Features 1–8 rely on the abstract's rhetorical structure and other shallow measures. Features 9–13 are based on LSA processing. Features 1–10 make up our basic pool of features.

Feature 11 was added to the basic pool of features when inducing the classifier for dimension Title. For each Purpose sentence in an abstract, this classifier uses the extracted features to predict whether it is strongly/weakly related to the title (*high/low* categories).

Similarly, feature 12 was added to the basic pool of features when inducing the classifiers for dimension Purpose. Since the dimension Purpose applies to sentences from three different rhetorical categories—Result, Methodology and Conclusion—different experiments were carried out to induce each classifier. For each Result, Methodology, and Conclusion sentence in an abstract, these classifiers use the extracted features to predict whether it is strongly/weakly related to the Purpose sentence(s) (also *high/low* categories).

As mentioned above, feature 13 is extracted only for abstracts that include both Gap and Background sentences. Thus, in the induction of the classifier for dimension Gap-Background, we have used the basic pool of features plus feature 13. For each Gap sentence in an abstract, the classifier predicts whether it is associated with at least one Background sentence (*yes/no* categories).

We have conducted feature selection experiments for all aforementioned classifiers. The results as well as the intrinsic evaluation of each classifier are described below.

4.2 Feature selection and intrinsic evaluation of the classifiers

Using the annotation presented in Sect. 3 and the set of features extracted from the corpus, we have generated and evaluated five classifiers: one for dimension Title, three for dimension Purpose [(namely, Purpose (M); Purpose (R), and Purpose (C)], and, finally, one classifier for dimension Gap-Background.

The feature selection experiments were carried out in the Weka learning environment [44], and so were the training and testing of all classifiers. For the feature selection experiments, we have adopted the *Wrapper* method in conjunction with the *Best-First* search. The SMO algorithm with the *PolyKernel* kernel was used to select features as well as induce classifiers. All classifiers were induced using tenfold stratified cross-validation with parameter *Filtertype* of the SMO algorithm set to the value "*Standardize training data*", to normalize the numerical attributes so that their average is zero and the variance interval is unitary.

Table 6 presents the set of features with the best performance in the feature selection experiments by classifier. It can be noticed that features extracted by LSA processing appear in all best performance feature sets. Table 6 also shows the Kappa resulting values between the human annotation and the classifiers as well as their corresponding accuracy values.

Table 6 Feature selection results, Kappa agreement between human annotation and classifiers and accuracy values by classifier

	Attributes	Kappa	Acc. (%)
Dim. Title	1, 3, 4, 5, 6, 7, 11	0.871	96.48
Dim. Purpose (M)	1, 2, 6, 9, 12	0.683	86.17
Dim. Purpose (R)	1, 10, 11, 12	0.763	89.47
Dim. Purpose (C)	1, 12	0.748	90.19
Dim. Gap-Background	1, 6, 13	0.679	88.52

From the figures presented in Table 6, we can conclude that all classifiers showed satisfactory levels of Kappa, taking into consideration the subjective nature of the task. The Dimension Title classifier had the best performance, with $K = 0.871$. The lowest value ($K = 0.679$) was recorded for the dimension Gap-Background classifier. It is nevertheless regarded as a good level of agreement. We also find that all classifiers achieved high accuracy rates, with values between 86.17 and 96.48 %. However, raw accuracy is a measure which does not take into account the number of successes and errors across the predicted classes and hence a more detailed analysis of the classifiers performance is required. Table 7 shows the performance of classifiers in terms of Precision, Recall, F-Measure and Macro-F.

For comparison purposes, we also present the results of a simple baseline by classifier. It is calculated separately for each classifier by assigning the majority class as output (Table 8). In both Tables 7 and 8, *high* and *low* classes refer to the classifiers of dimensions Title, Purpose (M), Purpose (R), and Purpose (C), while classes *yes* and *no* refer to the classifier of dimension Gap-Background.

As can be seen in Tables 7 and 8, all classifiers outperform their corresponding baselines. This is particularly evident for the classifier for dimension Title, which also showed the best results for the measures presented in Table 6. We also find that the F-measure value for the classes *high/yes* is consistently higher than the values for the classes *low/no*. Although there is an imbalance in the corpus that may favor the classes *high/yes*, we believe the behavior of the classifiers can be explained by the lower level of ambiguity in the annotation of *high/yes* sentences in comparison with *low/no* sentences. In fact, for all dimensions, our human annotators have found it more difficult to rank sentences as being weakly related to others than to rank them as having a high relationship. They argued that *low/no* sentences seem to show a higher level of ambiguity than *high/yes* sentences.

In some specific cases, such as the Purpose (M) and Purpose (R) classifiers, this ambiguity can be justified by other factors. In the first case, the content of Methodology sentences introduces new terms concerning names of techniques and methods. Such terms tend to be proper names and lead the semantic similarity between these sentences and the

Table 7 Performance in terms of Precision, Recall, F-measure and Macro-F by classifier

	High/yes			Low/no			Macro-F
	Precision	Recall	F-measure	Precision	Recall	F-measure	
Dimension Title	0.975	0.983	0.979	0.912	0.873	0.892	0.936
Dimension Purpose (M)	0.895	0.901	0.898	0.790	0.780	0.785	0.842
Dimension Purpose (R)	0.914	0.928	0.921	0.855	0.830	0.842	0.882
Dimension Purpose (C)	0.921	0.946	0.933	0.846	0.786	0.815	0.874
Dimension Gap-Background	0.903	0.949	0.925	0.821	0.696	0.753	0.839

Table 8 Baseline performance in terms of Precision, Recall, F-measure and Macro-F

	High/yes			Low/no			Macro-F
	Precision	Recall	F-measure	Precision	Recall	F-measure	
Dimension Title	0.833	1.000	0.908	0.000	0.000	0.000	0.454
Dimension Purpose (M)	0.675	1.000	0.795	0.000	0.000	0.000	0.398
Dimension Purpose (R)	0.661	1.000	0.840	0.000	0.000	0.000	0.420
Dimension Purpose (C)	0.725	1.000	0.840	0.000	0.000	0.000	0.420
Dimension Gap-Background	0.748	1.000	0.855	0.000	0.000	0.000	0.428

Purpose, which is calculated by feature 12, to be classified as *low*. This contradicts the assessment of the human annotator whose analysis goes beyond the text surface. Similarly, the content of Result sentences usually introduces names of metrics for evaluating the results. These terms cause the semantic similarity between Result and Purpose sentences to be classified as *low* and here again contradict the human annotator, who has a deeper understanding of the sentences. Even so, the overall performance of the classifiers was reasonably good, with macro-F values between 0.839 and 0.936. Macro-F takes into account the F-measure values for both *high/yes* and *low/no* classes.

To sum up, we conclude that the positive results obtained in the evaluation of all classifiers allow us to use them in the CAM proposed in this study, specifically to automatically detect semantic coherence aspects evaluated by dimensions Title, Purpose and Gap-Background. In the next section, we explain how the proposed dimensions are used to generate suggestions for improving coherence in scientific abstracts written in Portuguese.

5 The CAM

As reported in Sect. 2, SciPo assists novice writers in producing scientific abstracts in Portuguese by offering criticisms and/or suggestions regarding aspects of their rhetorical structure. Here, we intend to extend SciPo's functionalities so that it can also provide feedback on semantic coherence. Given the proposed coherence dimensions (Sect. 3), the developed classifiers and their good performance (Sect. 4), we have prototyped a CAN (henceforth, CAM) to be incorporated into the SciPo system. CAM identifies potential issues related to three out of the four coherence dimensions discussed earlier and selects the appropriate feedback from a collection of predefined coherence suggestions.

Figure 3 presents the new SciPo system architecture, including CAM, which is highlighted by the dashed rectangle. CAM comprises a *base of coherence suggestions*, a set of *coherence classifiers* (as previously mentioned, one for dimension Title, three for dimension Purpose, and one for dimension Gap-Background), and a *coherence advisor*, which selects the appropriate suggestion(s) based on potential coherence problems.

The coherence analysis process starts by resorting to the AZPort classifier to automatically detect the abstract's rhetorical structure. If any structural problem is detected, the user gets criticisms and/or suggestions from SciPo so that he/she can correct it. Otherwise, the detected rhetorical structure as well as text itself are passed on to CAM for LSA processing and feature extraction. Based on the extracted feature values, the five classifiers analyze each sentence of the abstract and their results are sent to the coherence advisor. In case of a potential coherence problem, the appropriate suggestion(s) will be selected by the advisor and presented to the user. The refinement cycle continues until either the system has no suggestions to offer or the user has decided to stop the process.

It is important to stress that the user can freely reject the coherence suggestions made by the system. In fact, semantic aspects are controversial by nature so we cannot rule out the possibility that user and system may disagree on the coherence problems identified. With this issue in mind, we have

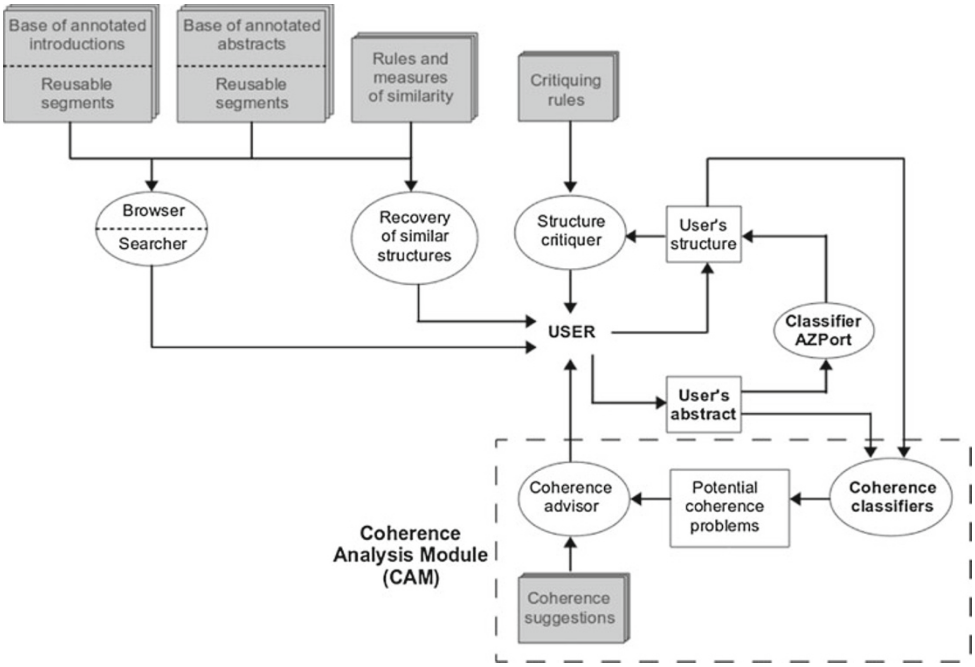

Fig. 3 SciPo's architecture including CAM

decided to leave the user free to accept or reject the suggestions offered by CAM. This freedom of choice given to the user had already been implemented in the SciPo system, and we have decided to maintain it in CAM.

5.1 Coherence suggestions

Coherence suggestions are presented to the user whenever one or more coherence classifiers return a *low/no* value. According to the classifier in question, the coherence advisor then selects the appropriate suggestions out of the set presented in Table 9. In this table, we show the five coherence suggestions elaborated according to the dimensions discussed earlier as well as a brief explanation that is presented to the user along with the suggestion. For convenience, all suggestions and their explanations in Table 9 have been translated into English, although in SciPo they are presented in Portuguese.

5.2 Evaluation by users

To evaluate CAM in its context of use, i.e., as part of the SciPo system, we have conducted an experiment with actual users to check how effective the coherence suggestions are in scientific abstract writing. The experiment was carried out with eight MSc students in computer science from the State University of Maringá. All students have written or were in the process of writing their master's dissertation in the year 2010 and hence had already finished writing its abstract or had at least a draft of it.

All users were asked to use CAM in the refinement of their abstract/draft. Before doing so, they were presented to the main purposes of CAM, along with a brief explanation of the components that make up the rhetorical structure of a scientific abstract. It is worth mentioning that users were not familiar with the concepts of rhetorical components and structure. After using CAM, all users were asked to complete a questionnaire reporting, among other points, their impressions on CAM, including easiness to use, relevance of the presented suggestions, alterations in the coherence level of his/her abstract after using CAM. In Fig. 4, we present a summary of the users' answers to the questionnaire.

As explained earlier, CAM uses the AZPort classifier to detect the abstract rhetorical structure. In this experiment, users were asked to correct AZPort output whenever they felt appropriate. Out of a total of 63 sentences classified by AZPort, 26 (41.3 %) were corrected.

During the experiment, six users were presented with one coherence suggestion for their abstracts. For three of them it referred to low relationship between Title and Purpose. For two of them the suggestion was for the low relationship between Purpose and Methodology. In one case, the suggestion concerned the low relationship between Purpose and Result. Two users were presented with two suggestions for their abstracts. For one of them, one suggestion referred to the low relationship between Title and Purpose, and another to the low relationship between Purpose and Methodology. For the other user, the suggestions were for the low relationship between Purpose and Methodology, and between Purpose and Result.

Table 9 Coherence suggestions and their explanations as presented by CAM

Dimension	Suggestion and explanation text
Title	The Title and Purpose sentences should be more closely related Purpose sentences are expected to indicate the main objective of the research. Similarly, the Title of abstract should "summarize" the main objective, using in limited number of words (one or two lines of text) so that it can be immediately understood by the reader. Check whether the title of your abstract describes succinctly the main objective cited in Purpose sentences or whether the Purpose sentences refer to the content presented in Title. Lack of relationship between the Title and the Purpose sentences could result in a coherence problem in your abstract. Consider rewriting either the Title or the Purpose sentences of your abstract
Purpose (M)	Methodology and Purpose can be more closely related Methodology sentences indicate materials and methods which have been used or served as the basis of the research. Thus, methods are usually described or at least indicated in the abstract. Highly coherent abstracts are those in which the description or indication of the methods used are related to the main objective of the study being presented, thus justifying their use. Check if the content of Methodology and Purpose sentences are related. It may be necessary to rewrite the Methodology sentences
Purpose (R)	Result and Purpose can be more closely related Result sentences describe any artifact developed by the author or indicate the results of experiments and evaluations. In highly coherent abstracts, result description are expected to be closely related to the main objective of the research so that the reader understands the relevance and contributions of the study in relation to the objectives presented in Purpose sentences. Check if the content of Result and Purpose sentences are related. It may be necessary to rewrite the Result sentences
Purpose (C)	Conclusion and Purpose can be more closely related Conclusion sentences "close" the text and are intended to offer recommendations, contributions and to highlight the value of presented research. To do it successfully, Conclusion sentences are expected to make reference to the content of Purpose sentences so that the reader associates them with the main objective of study. Check if the content of Result and Purpose sentences are related. It may be necessary to rewrite the Conclusion sentences
Gap-background	Gap and Background should be related Gap sentences indicate some research questions that are worth investigating. Coherent abstracts are therefore expected to include at least one sentence that contextualizes these research questions before presenting them. Check if the content of Gap and Background sentences are related. It may be necessary to rewrite these sentences

No suggestions were offered for the relationship between Gap and Background nor for the relationship between Conclusion and Purpose. In addition to the fact that the analyzed abstracts did not present problems in relation to these components, we should also bear in mind that most abstracts did not include Gap and Conclusion sentences. Although we had previously explained about the importance of all rhetorical components and users were given the chance to refine his/her abstract's rhetorical structure using SciPo, some opted for not including Gap and Conclusion in their abstracts. We believe that this can be partially explained by the fact that some users had submitted a finished version of abstract to the system, rather than a draft, and were reluctant to modify it.

All users have accepted the suggestions presented by CAM. Six have accepted all the suggestions provided and two have accepted them partially. In the latter case, the system continued to present suggestions even after the rewriting and adaptation of the abstract and the refinement process was ended by the user.

As regards the relevance of the coherence suggestions presented by CAM, five users considered them as *relevant*, two users found them *very relevant*, and one user regarded them as *irrelevant* (see Fig. 4a).

When comparing the initial abstract to its revised version, after the refinement process guided by the coherence suggestions, four users reported that the coherence level of the final abstract *increased substantially*, while three users considered that the coherence level of the final abstract *increased reasonably*. One user found that the adjustments in the abstract did *not alter* its level of coherence (see Fig. 4b).

Three users had *some doubts* about how to adjust/rewrite the abstract in accordance with the suggestions presented by CAM. One user had *many doubts*. On the other hand, four

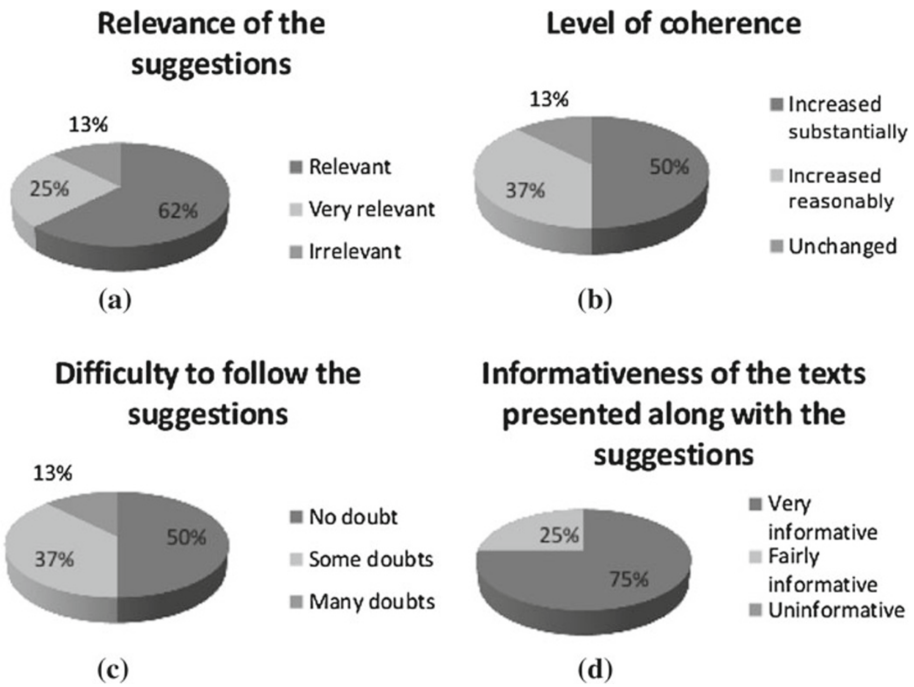

Fig. 4 Summary of the users' answers to the questionnaire applied after the experiment with CAM

users reported that they had *no doubts* about how to rewrite and adequate the abstract according to the coherence suggestions (see Fig. 4c). We believe that the reported difficulties in adjusting the abstracts are inherent to the writing process, since most users (six of them) ranked the information and suggestions provided as *very informative*; two users ranked them as *fairly informative*, and even those with difficulties in rewriting the abstract did not considered the suggestions as *uninformative* (see Fig. 4d).

6 Conclusions

This study aimed to implement and evaluate a method for automatically detecting potential problems regarding semantic coherence in scientific abstracts written in Portuguese. The method is based on the evaluation of three coherence dimensions. More specifically, we have developed a complementary module to the SciPo system, namely CAM, which identifies potential coherence problems and generate appropriate suggestions for semantic aspects of the abstract section.

CAM is the prototype resulting from the induction and evaluation of several machine learning models (classifiers) applied to the proposed coherence dimensions, plus a base of suggestions. We initially have proposed four dimensions: (1) Title, (2) Purpose, (3) Gap-background, and (4) Linearity-break. However, due to the reduced number of examples with a Linearity-break in the annotation process, we could not induce a classifier for this dimension. By way of contrast, for the other three dimensions, the classifiers presented good results which allowed us to use them to identify potential coherence problems and, as a result, implement them as part of CAM. Another important point to stress here is that CAM performs real time text classification with no runtime efficiency issues. Even though CAM classifiers depend on the AZPort system, when it comes to extracting features and classifying texts, CAM classifiers remain reasonably fast. The slight delay in the process is hardly felt by the user.

Given the difficulties found in the induction of a classifier for Dimension Linearity-break, in future studies, we intend to explore the entity-based model proposed by Barzilay and Lapata [6] for extracting new features that may be helpful for such dimension. In addition, during the training and testing of the classifiers, we have performed a feature selection phase using the SMO algorithm with a view to removing redundant features, thus improving classifiers performance. However, we believe that further study on the SMO parameters could optimize their values and consequently provide even better classification results. In a near future, we also consider experimenting with algorithms from other approaches, such as Bayesian models and decision trees.

Another point to be considered in the classifiers' performance is the imbalance of classes. In future investigations, it is our intention to explore techniques for artificially balancing classes, such as Undersampling, which eliminates instances of the majority class [25], and Oversampling, which replicates instances of the minority class [11]. The impact of these techniques is worth analyzing since the manual annotation of a larger corpus has a high cost.

Another issue to be addressed in future studies is the adequacy of the dimensions to other sections of a scientific work, such as introduction and conclusion. In addition to differences related to rhetorical structure, these sections are usually longer and present more variations than abstracts in terms of both structure and content.

Finally, we conclude that the classifiers evaluations both intrinsic and as part of CAM demonstrate the potential of the proposed dimensions to support the writing of scientific abstracts. The experiment with actual users, although preliminary, has shown that CAM can provide relevant suggestions and offer potentially useful guidance for writing abstracts with a high level of coherence.

Acknowledgments We would like to thank the Brazilian Funding Body CNPq (Process # 135669/2009-0) for partially funding this research. We are also grateful to Danilo Machado Jr. for his invaluable work in the corpus annotation.

References

1. Aizawa A (2001) Linguistic techniques to improve the performance of automatic text categorization. In: Proceedings of the 6th natural language processing pacific rim symposium (NLPRS-01), Tokyo, Japan, pp 307–314
2. Aluísio S, Schuster E, Feltrim VD, Pessoa A, Oliveira O (2005) Evaluating scientific abstracts with a genre-specific rubric. In: Proceeding of the 12th international conference on artificial intelligence in education (AIED 2005) IOS Press. The Netherlands, Amsterdam, pp 738–740
3. Aluísio SM, Barcelos I, Sampaio J, Oliveira ON Jr (2001) How to learn the many unwritten “rules of the game” of the academic discourse: a hybrid approach based on critiques and cases. In: Proceedings of the IEEE international conference on advanced learning technologies (ICALT 2001), pp 257–260
4. Aluísio SM, Oliveira ON Jr (1996) A detailed schematic structure of research paper introductions: an application in support-writing tools. Procesamiento del Lenguaje Nat 19:141–147
5. Anthony L, Lashkia GV (2003) Mover: a machine learning tool to assist in the reading and writing of technical papers. In: IEEE transactions on professional communication 46(3):185–193
6. Barzilay R, Lapata M (2008) Modeling local coherence: an entity-based approach. Comput Linguist 34(1):1–34
7. Booth WC, Colomb CG, Williams JM (2000) A arte da pesquisa. Martins Fontes, Brazil
8. Broady E, Shurville S (2000) Developing academic writer: designing a writing environment for novice academic writers. In: Broady E (ed) Second language writing in a computer environment. CiLT/AFLS, London, pp 131–152
9. Burstein J, Chodorow M, Leacock C (2003) Criterionsm online essay evaluation: an application for automated evaluation of student essays. In: Proceedings of the 15th annual conference on innovative applications of artificial intelligence, pp 3–10
10. Burstein J, Tetreault J, Andreyev S (2010) Using entity-based features to model coherence in student essays. In: Proceedings of the human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics . Association for Computational Linguistics, Los Angeles, pp 681–684
11. Chawla NV, Bowyer KW, Hall LD, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
12. de Beaugrande RA, Dressler WU (1981) Introduction to text linguistics. Longman Publisher Group, New York
13. Drucker H, Wu D, Vapnik VN (1999) Support vector machines for spam categorization. IEEE Trans Neural Netw 10(5):1048–1054
14. Elliot S (2003) Intellimetric: from here to validity. In: Shermis MD, Burstein JC (eds) Automatic essay scoring: a cross-disciplinary perspective. Lawrence Erlbaum Associates, Hillsdale, pp 71–86
15. Feltrim VD (2004) Uma abordagem baseada em córpus e em sistemas de crítica para a construção de ambientes web de auxílio à escrita acadêmica em português. Ph.D. thesis, Instituto de Computação e Matemática Computacional - Universidade de São Paulo, São Carlos, SP
16. Feltrim VD, Aluísio SM, Nunes MGV (2003) Analysis of the rhetorical structure of computer science abstracts in portugese. In: Archer D, Rayson P, Wilson A, McEnery T (eds) Proceedings of Corpus Linguistics 2003, UCREL Technical Papers, vol 16, part 1, special issue, pp 212–218
17. Feltrim VD, Teufel S, Nunes MGV, Aluísio SM (2006) Argumentative zoning applied to criquing novices scientific abstracts. In: Shanahan JG, Qu Y, Wiebe J (eds) Computing attitude and affect in text: theory and applications. The Netherlands, Dordrecht, pp 233–246
18. Foltz PW, Kintsch W, Landauer TK (1998) The measurement of textual coherence with latent semantic analysis. Discourse Processes 25:285–307
19. Higgins D, Burstein J (2007) Sentence similarity measures for essay coherence. In: Proceedings of the 7th International workshop on computational semantics (IWCS-7), Tilburg, The Netherlands, pp 1–12
20. Higgins D, Burstein J, Marcu D, Gentile C (2004) Evaluating multiple aspects of coherence in student essays. In: Human language technologies: the 2004 annual conference of the north american chapter of the association for computational linguistics. Association for Computational Linguistics. Boston, pp 185–192
21. Huckin T, Olsen LA (1991) Technical writing and professional communication for non-native speakers. McGraw-Hill Humanities/Social Sciences/Languages, New York
22. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK (2001) Improvements to Platt’s SMO algorithm for SVM classifier design. Neural Comput 13(3):637–649
23. Koch IV, Travaglia LC (2003) A coerência textual. Editora Contexto, São Paulo
24. Kriegsman M, Barletta R (1993) Building a case-based help desk application. IEEE Expert Intell Syst Appl 8(6):18–26
25. Kubat M, Matwin S (1997) Addressing the curse of imbalanced training sets: One-sided selection. In: Machine learning: proceedings of the 14th international conference (ICML’97). Morgan Kaufmann, Nashville, pp 179–186
26. Landauer T, Laham D, Foltz P (2003) Automated scoring and annotation of essays with the intelligent essay assessor. In: Shermis MD, Burstein JC (eds) Automated essay scoring: a cross-disciplinary perspective. Lawrence Erlbaum Associates, New Jersey, USA, pp 87–112
27. Landauer TK, Foltz PW, Laham D (1998) Introduction to latent semantic analysis. Discourse Process 25:259–284
28. Lansman M, Smith JB, Weber I (1993) Using the “writing environment” to study writer’s strategies. Comput Compos 10(2):71–92
29. Lapata M, Barzilay R (2005) Automatic evaluation of text coherence: models and representations. In: Proceedings of the international joint conferences on artificial intelligence, pp 1085–1090
30. Ljungstrand P, Johansson H (1998) Intranet indexing using semantic document clustering. Department of Informatics - Göteborg University, Göteborg, Sweden, Master’s thesis
31. Narita M (2000) Constructing a tagged EJ parallel corpus for assisting japanese software engineers in writing english abstracts. In: Proceedings of the 2nd international conference on language resources and evaluation (LREC 2000), pp 1187–1191

32. Narita M (2000) Corpus-based english language assistant to japanese software engineers. Proceedings of MT-2000 machine translation and multilingual applications in the new millennium pp 24-1–24-8
33. Pemberton L, Shurville S, Hartley T (1996) Motivating the design of a computer assisted environment for writers in a second language. In: Diaz de Ilarraza Sanchez A, Fernandez de Castro I (eds) Computer aided learning and instruction in science and engineering. Lecture Notes in Computer Science, vol 1108. Springer, Berlin, pp 141–148
34. Sharples M, Goodlet J, Clutterbuck A (1994) A comparison of algorithms for hypertext notes network linearization. Int J Hum Comput Stud 40(4):727–752
35. Sharples M, Pemberton L (1992) Representing writing: external representations and the writing process. In: Holt P, Williams N (eds) Computers and writing: state of the art. Intellect, Oxford, pp 319–336
36. Shurville S, Hartley AF, Pemberton L (1997) A development methodology for composer: a computer support tool for academic writing in a second language. In: Knorr D, Jakobs E-M (eds) Text production in electronic environments. Lang, Frankfurt, pp 171–182
37. Siegel S, Castellan NJ Jr (1988) Nonparametric statistics for the behavioral sciences, 2nd edn. McGraw-Hill, Berkeley
38. Souza VMA, Feltrim VD (2011) An analysis of textual coherence in academic abstracts written in portuguese. In: Proceedings of 6th corpus linguistics conference: CL 2011, Birmingham, UK
39. Swales JM (1990) Genre analysis: english in academic and research settings. Cambridge University Press, Cambridge, Cambridge applied linguistics series
40. Teufel S, Moens M (2002) Summarising scientific articles - experiments with relevance and rhetorical status. Comput Linguist 28(4):409–446
41. van Dijk TA, Kintsch W (1983) Strategies of discourse comprehension. Academic Press, New York
42. Vapnik VN (2000) The nature of statistical learning theory. Springer Verlag, New York
43. Weissberg R, Buker S (1990) Writing up research: experimental research report Writing for students of English. Prentice Hall Regents, Englewood Cliffs
44. Witten IH, Frank E (2005) Data mining: practical machine learning tools and technique. Morgan Kaufmann, San Francisco