Automatic Analysis of Semantic Coherence in Academic Abstracts Written in Portuguese

Vinícius Mourão Alves de Souza
State University of Maringá
Maringá, PR, Brazil, 87020-900
vsouza@din.uem.br

Valéria Delisandra Feltrim
State University of Maringá
Maringá, PR, Brazil, 87020-900
valeria.feltrim@din.uem.br

Abstract

SciPo is a system whose ultimate goal is to support novice writers in producing academic texts in Brazilian Portuguese through presentation of critiques and suggestions. Currently, it focuses on the rhetorical structure of texts, being capable of automatically detecting and criticizing the rhetorical structure of Abstract sections. We describe a system that enhances SciPo's functionality by evaluating aspects of semantic coherence in academic abstracts. This system identifies features of sentences based on semantic similarity measures and rhetorical structure. Different machine learning algorithms were trained and evaluated with these features, resulting in three classifiers capable of detecting specific coherence issues on sentences with regard to a rhetorical structure model for abstracts. Results indicate that the system yields higher performance than the baseline for all classifiers.

1 Introduction

This research has been motivated by a need for advanced discourse analysis capabilities for writing tools such as SciPo (short for Scientific Portuguese). SciPo (Feltrim et al., 2006) is a system whose ultimate goal is to support novice writers in producing academic texts in Brazilian Portuguese. Currently, it focuses on Computer Science academic texts and supports the writing of abstracts and introductions. Its functionalities are based on the use of structure models — in terms of schematic structure, rhetorical strategies and lexical patterns — similar to the ones proposed by Swales (1990) and Weissberg and Buker (1990), and authentic examples organized as case bases. Although SciPo provides feedback with regard to the text rhetorical structure in the form of critiques and suggestions, it does not provide considerations about the text semantics, such as aspects related to its coherence, which is a fundamental characteristic for text legibility and interpretability.

We understand coherence as what makes a group of words or sentences semantically meaningful. We assume that coherence refers to the establishment of a logical sense among different sentences of a text. Thus, it is a principle of interpretability related to the communicational situation and to the capability of the reader in calculating the meaning of the text. Therefore, it is bounded to the text, but it does not depend only on the text (van Dijk, 1981).

Aiming at complementing SciPo's functionalities, we have developed classifiers for the automatic detection of specific semantic relations in academic texts in Portuguese, then it can be used by SciPo for providing feedback referring to text coherence. Based on textual features that can be readily read off the text, the classifiers present indications related to semantic aspects that contribute to a high level of coherence.

We believe that our work brings innovative contributions due to the nature of the analyzed corpus, especially by language and rhetorical structure of the texts, and the kind of application to which we intend to apply coherence analysis. As mentioned by Burstein et al. (2010), there is a small body of work that has investigated the problem of identifying coherence in student essays. None of the work cited by Burstein et al. (2010) is focused on academic writing, but on essays written by English writers that may be native/non-native and have different writing skills. This kind of text tends to present more explicit coherence problems than the ones that may occur on a academic writing corpus, as the one used in this work. Academic texts are usually written by students who have domain, at a certain level, on the language (in our case, Por-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1144–1152,
Chiang Mai, Thailand, November 8 – 13, 2011.

tuguese) and on the genre, which can make structure and coherence problems subtle. The more subtle a problem is, more difficult it is to be automatic treated.

Besides the corpus differences, most of systems presented in the literature that realize coherence analysis are in the context of Automatic Essay Scoring (Lapata and Barzilay, 2005), which is also different from our context of work. We cite three scoring systems which considers aspects of coherence when grading essays: *Criterion* (Burstein et al., 2003; Higgins et al., 2004; Burstein et al., 2010), *Intelligent Essay Assessor* (Landauer et al., 2003), and *Intellimetric* (Elliot, 2003). Unlike these systems, SciPo is a writing support system, which means that we are not interested in to ascribe a score to it, but we want the system to be able to detect a possible structure and coherence issues and give some comprehensible feedback to the writer. The three cited systems employ the Latent Semantic Analysis (Landauer et al., 1998) to extract text features related to coherence aspects, and the results reported by them have motivated their use in our work .

2 Corpus and Annotation

In order to analyze coherence issues that may occur in academic texts written in Portuguese by undergraduate students, we have collected 385 abstracts of monographs written as part of the requirements for achieving a BS degree in Computer Science. The corpus annotation was processed in two distinctive parts: (i) rhetorical structure annotation and (ii) coherence annotation, as following described.

2.1 Rhetorical Structure Annotation and Analysis

Each abstract has the correspondent work's title attached to it. Also, each sentence was previously delimitated with appropriate beginning/ending tags. Then, we used AZPort (Feltrim et al., 2006) to label each sentence accordingly to its rhetorical status (Teufel and Moens, 2002). AZPort is a Naive Bayesian classifier that renders each input sentence a set of six possible categories, namely Background, Gap, Purpose, Methodology, Result, and Conclusion. These categories correspond to the components that make up the rhetorical structure model proposed by Feltrim et al. (2006) to academic abstracts.

We manually revised the resulting annotated corpus and corrected possible mistakes made by AZPort. Thus, the noise from the automatic annotation of rhetorical structure does not interfere in the coherence annotation. A total of 2,293 sentences were automatically annotated and manually revised. The distribution of categories in the annotated corpus is presented in Table 1.

Categories	Sentence (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

Table 1: Rhetorical categories distribution.

It can be observed in Table 1 that Background is the most frequent category (34.78% of all sentences). The prevalence of category can be explained by the corpus nature. When writing monographs abstracts, writers usually are not limited to a fixed maximum of words, thus they tend to write more sentences contextualizing the work. This is not true for papers abstracts, which tend to be limited in length and, therefore, leading writers to focus on Purpose and Result (Feltrim et al., 2003). In our corpus, Purpose and Result are also frequent categories, accounting for 19.63% and 19.41% of all sentences, respectively. Methodology, Gap and Conclusion categories were less frequent.

2.2 Coherence Annotation and Analysis

Following Higgins et al. (2004), we have tried to identify and annotate semantic relations among specific rhetorical categories, but taking into consideration that we are dealing with abstract sections of academic texts and that we want to use the resulting information as a resource to formulate useful feedback to SciPo users. We came up with an adaptation of the four dimensions proposed by Higgins et al. (2004), resulting in four kinds of relations that we also called dimensions: (i) Dimension Title, (ii) Dimension Purpose, (iii) Dimension Gap-Background, and (iv) Dimension Linearity-Break. Each dimension is described as follows.

2.2.1 Dimension Title

We assume that the title of an academic text should reveal the main topics treated in it. We also assume that the abstract of an academic text should

inform the reader about these topics, even though in a summarized form. The lack of relationship between the abstract sentences and the title may be an evidence of two possible situations: (i) the title is inappropriate for the abstract or (ii) the abstract has coherence problems.

In order to proceed with the corpus annotation, we have assumed that the abstracts titles were always appropriate and then we verified the semantic similarity between each sentence in the abstract and its title. Each sentence was labeled as *high* if it is strongly related to the title. Otherwise, it was labeled as *low*. We have decided to use a binary scale rather than a finer grained one due to the subjective nature of the task. Even with only two possible labels, the agreement between two human annotators measured by the Kappa statistics over a randomly selected subset of 209 sentences of the corpus and was around 0.6 (see Table 4).

Over a total of 2,293 sentences, 1,050 (46.80%) were ranked as been weakly related to the title (*low* sentences) and 1,243 (54,20%) as been strongly related (*high* sentences). The distribution of *high* and *low* sentences among the six possible rhetorical categories is presented in Table 2.

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

Table 2: Dimension Title annotation.

It can be observed in Table 2 that Purpose sentences tend to have a strong level of relatedness to the title, since 83.33% of such sentences were ranked as *high*. It is much higher than the average of *high* sentences for other categories, which is 48.79%. Background sentences are the less related to the title, having more than half of the total of sentences (54.95%) ranked as *low*. In fact, these are not surprising results. Background sentences usually appears at the beginning of the abstract with the purpose of establishing the context of the research and, therefore, may not be directly related to the main topics of the research being presented. Instead, it may address questions or state facts of a broader area of study, which will prepare the reader to understand the motivations that led to the presented work. Thus, the detection of a weak relationship between the title and a Background sentence cannot be assumed as a coherence problem. On the other hand, Purpose sentences are expected to address directly the main topics treated by the research and then to be strongly related to the title. This is in accordance with the traditional "general — specific — general" model accepted as standard for scientific texts (Swales, 1990; Weissberg and Buker, 1990), especially introduction and abstract sections. Therefore, the existence of a weak relationship between Purpose sentences and the title probably indicates a coherence issue. With respect to the remaining rhetorical categories (Gap, Methodology, Result, and Conclusion), its relatedness to the title is quite balanced, with an average of 50.5% of *low* sentences and 49.5% of *high* sentences over a total of 1,059 sentences. In our observations, the relatedness of these categories of sentences to the title depends on other aspects than coherence, like the very nature of the research being reported. Thus, we cannot assume that the lack of a strong relationship between a sentence of these categories and the title may indicate a coherence problem.

Taking into account these results, we have concluded that the analysis of this dimension can be used as an indicative of a possible coherence problem in the Purpose rhetorical component of the abstract.

2.2.2 Dimension Purpose

The relationship between a rhetorical component and other components dictates the global coherence of the text (Higgins et al., 2004). Therefore, for an abstract to be easy to follow and understand, the rhetorical components must be related. Taking into consideration the rhetorical structure model used for the annotation of the corpus, it is expected the Purpose component to be related to Methodology, Result and Conclusion components. Thus, we understand that the absence of relationship between each of these components and the Purpose component can be an indication of a coherence problem.

For each abstract in the corpus, we have verified the semantic similarity between the sentences labeled as Purpose and the remaining sentences of the abstract. Each non-Purpose sentence was labeled as *high* if it is strongly related to Purpose; otherwise, it was labeled as *low*. The label *n/a* was assigned to sentences of abstracts which do

not have Purpose sentences. We have measured the agreement between two human annotators by the Kappa statistics over a randomly selected subset of 167 sentences of the corpus and was around 0.8 (see Table 6).

Apart from 573 sentences (426 Purpose sentences and 147 *n/a* sentences distributed among the other five categories), 1,720 sentences were labeled as *high/low* for this dimension. Over this total of sentences, 704 (40.93%) were ranked as been weakly related to the Purpose (*low* sentences) and 1,016 (59,07%) as been strongly related to the Purpose (*high* sentences). The distribution of *high* and *low* sentences among the rhetorical categories is presented in Table 3.

	Sentences	
Categories	High	Low
Background	378	380
Gap	129	079
Methodology	171	082
Result	264	135
Conclusion	074	028
Total	1,016	704

Table 3: Dimension Purpose annotation.

As it can be observed in Table 3, the sentences most related to the Purpose indeed are those labeled as Conclusion, Methodology, and Result. The percentages of *high* sentences for these categories are 72.55%, 67.59%, and 66.17%, respectively. It is worth noticing that the percentage of *high* sentences for Methodology, and Result categories could be even higher, as many sentences of these categories restate the content of the Purpose component by the use of anaphoric expressions, which decreases the level of semantic relationship between the sentences.

Once again, the general nature of Background sentences have placed them as the higher percentage of *low* sentences (50.13%). In fact, Background sentences tend to be closely related to Gap sentences then to Purpose ones, so the low level of relationship between Background and Purpose sentences cannot be assumed as a possible coherence problem.

We have concluded that the analysis of the Dimension Purpose for Methodology, Result, and Conclusion sentences can be used to detect possible coherence problems involving these rhetorical components.

2.2.3 Dimension Gap-Background

As noted earlier, Background sentences tend to be closely related to Gap sentences then to Purpose ones. Thus, it is expected that the Gap component is related with at least one sentence of Background. Therefore, we understand that the absence of relationship between these components can be an indication of a coherence problem.

For each abstract with Gap and Background sentences in the corpus, we have verified the semantic relationship between the sentences of these categories. Each Gap sentence was labeled as *yes* if it is strongly related with some Background sentence; otherwise, it was labeled as *no*.

Apart from 32 sentences belonging to abstracts which do not have Gap/Background sentences, 183 sentences were labeled as *yes/no* for this dimension. Over this total of sentences, 74.86% were ranked as *yes* and 24.14% were ranked as *no*. We have measured the agreement between two human annotators by the Kappa statistics over a randomly selected subset of 46 sentences of the corpus and was around 0.7 (see Table 8).

Taking into consideration the annotation results for this dimension, we have concluded that the analysis of the Dimension Gap-Background can be used to detect possible coherence problems involving the relationship between the rhetorical components Gap and Background.

2.2.4 Dimension Linearity-break

This dimension focuses on detecting linearity breaks between adjacent sentences. Unlike to the other dimensions, Linearity-break is independent of the rhetorical structure of the abstract. A human annotator was instructed to label sentences *yes* when there was a difficulty in establishing a logical connection between the current sentence and its previous and/or its following sentence. Otherwise, the annotator was instructed to label sentences *no*.

Over a total of 2,293 sentences, only 153 were ranked as *yes* (7.14%). This indicates that it is relatively rare to find a sentence which is not related to its adjacencies, as 92.86% of all sentences in our corpus were ranked as *no* with respect to this dimension. In fact, the analysis of this dimension indicates very local coherence issues, which we believe to be more frequent in texts with more serious writing problems than the ones observed in the texts of our corpus.

3 Automatic Analysis of Coherence

As previously stated, the purpose of this work is to develop complementary functionalities for the SciPo system to be capable of identifying semantic coherence related aspects in academic abstracts written in Portuguese. The feedback to be provided by the new functionalities proposed in this work aims at highlighting the presence of potential issues related to semantic coherence in academic abstracts, especially the ones related to Dimension Title, Dimension Purpose, and Dimension Gap-Background.

3.1 Development

For performing the automatic analysis of Dimension Title, Purpose and Gap-Background, we developed classifiers induced by machine learning algorithms and based on features extracted from the text surface and from LSA processing.

The first stage is the annotation of the rhetorical structure of the abstract. In our experiments, we have used abstracts whose automatically assigned rhetorical labels were manually revised. As noted earlier, this is necessary so that the noise from the automatic annotation of rhetorical structure does not interfere in predicting coherence judgments. Nevertheless, in a final version of the semantic coherence analysis module we would use the rhetorical labels assigned by AZPort, and further evaluation of the effect of using these automatically assigned labels is necessary.

The next stage for the semantic coherence analysis concerns the LSA processing. Some pre-processing was required and it proceeds in three steps for all sentences in the corpus: (i) case folding (for data standardization), (ii) stop words removal, and (iii) stemming. These three steps contribute to a better performance of the attributes extracted based on LSA. After data pre-processing and build of a significant semantic space, LSA allows to make comparisons between sentences in order to extract features of the texts. The comparisons took in to account the semantic relation between each pair of sentences based on the LSA model, where the level of similarity is given by the frequency of sentences occurring in similar contexts. For each of the 385 abstracts, we performed all possible comparisons between pairs of sentences within a same abstract, including the abstract title sentences.

3.2 Attribute Extraction

We extracted a set of 13 features for each sentence in the corpus. We have used the features proposed by Higgins et al. (2004) as a starting point for our owns. All features were automatically extracted and used in the induction of the classifiers. The complete set of features is:

1. Rhetorical category of the target sentence;
2. Rhetorical category of the sentence that precedes the target sentence;
3. Rhetorical category of the sentence that follows the target sentence;
4. Presence of words that may characterize an anaphoric element;
5. Position of the sentence within the abstract, computed based on the beginning of the abstract;
6. Presence of words that may characterize some kind of transition;
7. Length of the target sentence measured in words;
8. Length of the title measured in words;
9. LSA similarity score of the target sentence with its preceding sentence;
10. LSA similarity score of the target sentence with its following sentence;
11. LSA similarity score of the target sentence with the entire abstract title;
12. LSA similarity score of the target sentence with all the sentences of the abstract classified as Purpose; and
13. Maximum LSA similarity score of the target Gap sentence with some Background sentence of the abstract.

Features 1 to 8 are based on the abstract rhetorical structure and other shallow measures. Features 9 to 13 are based on LSA processing. Features 1 to 10 compose our basic pool of features and were used in the induction of all classifiers. Feature 11 was added to the basic pool of features when inducting Dimension Title classifier. For each sentence in an abstract, Dimension Title 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 for the induction of Dimension Purpose classifier. This classifier uses the extracted features to predict, for each sentence

in an abstract, whether it is strongly/weakly related to the Purpose sentences of the target abstract (also *high/low* categories). Feature 13 is extracted only of Gap sentences in abstracts that also have Background sentences. Thus, Dimension Gap-Background classifier uses the basic pool of features plus feature 13 to predict, for each Gap sentence in an abstract, whether it is related with at least one Background sentence (*yes/no* categories).

4 Evaluation of Classification Models

Based on the extracted features, we generated and evaluated classification models for Dimension Title, Purpose and Gap-Background. For each dimension, we trained and tested 15 different machine learning algorithms using the implementations provided by the WEKA (Witten and Frank, 2005), resulting on a total of 45 classifiers. Among the classes of algorithms that we evaluated are decision trees, rule induction, probabilistic models, support vector machines, linear regression, and others. All the classifiers were inducted using 10-fold stratified cross-validation and the set of features. The performance was measured by comparing the system's prediction with one human annotation. We assumed the annotation performed by one of the subjects in the previous annotation experiment as our "gold standard" and used it as training material. The best model for each dimension was used for further experiments and evaluation.

For each dimension, we also report the performance of a simple baseline measure, which always assigns the prevalent category (*high*/*low* or *yes*/*no*) to every sentence.

4.1 Classification Model for Dimension Title

Among the evaluated learning algorithms for Dimension Title, MultiBoostAB implemented based on Webb (2000) presented the best performance. Using C4.5 (Quinlan, 1993) as the base learning algorithm, MultiBoostAB combines boosting and wagging techniques for forming decision committees. The MultiBoostAB classifier achieved F-measure of 0.811 for the *high* category, and 0.782 for the *low* category. We also evaluated the performance of each of our features for this dimension. As expected, feature 11 (LSA similarity score of the target sentence with the entire abstract title) achieved the best performance.

In order to analyze the performance of the classification model with regard to each rhetorical category, we inducted and evaluated six different classifiers, one for each rhetorical category. Each of these classifiers was trained using the abstracts titles and a set of sentences of the target category. Baselines classifiers were also evaluated for each category. The baseline performance for all the Dimension Title classifiers in terms of Precision, Recall, F-measure, accuracy, and Kappa is presented in Table 4. The performance of each Dimension Title classifier also in terms of Precision, Recall, F-measure, Accuracy, and Kappa is presented in Table 5. The Kappa measure shown in Table 4 refers to the agreement between two human annotators. In Table 5, refers to the agreement among each classifier and our "gold-standard".

As shown by the results reported on Table 4 and Table 5, all our MultiBoostAB classifiers outperform the baseline. The best performance, both in terms of F-measure and Kappa, was achieved by the Purpose classifier. The Kappa above 0.8 indicates high agreement between classifier and human annotator.

Looking at the performance of the classifiers for *high* and *low* sentences, it can be observed that most of them perform better for *high* sentences. We ascribe this to the lower level of ambiguity in assigning a sentence as *high*. In fact, our human annotators have found more difficulties in ranking a sentence as being weakly related to the title (*low* sentences) than in ranking it as strongly related (*high* sentences). They claim the existence of a higher level of ambiguity in *low* sentences than in *high* sentences.

As for the superior performance of the Purpose classifier, we attribute that to the strong relationship between the content of Purpose sentences and the title, as previously discussed, and to the fact that Purpose sentences usually are clear and objective, presenting well defined lexical and syntactic markers. In general, it is possible to say that there is less ambiguity in ranking a Purpose sentence as strongly/weakly related to the title than ranking the relationship of a Background sentence to the title.

Both the evaluation results for the classification model and the semantic content of Purpose sentences leads us to employ the Dimension Title automatic evaluation only to sentences rhetorically categorized as Purpose.

	High			Low			Total	
	Precision	Recall	F-measure	Precision	Recall	F-measure	Acc	Kappa (N)
Background (N=808)	0.000	0.000	0.000	0.549	1.000	0.708	0.549	0.750 (87)
Gap (N=215)	0.000	0.000	0.000	0.516	1.000	0.680	0.516	0.577 (46)
Purpose (N=426)	0.833	1.000	0.908	0.000	0.000	0.000	0.833	0.696 (42)
Methodology (N=273)	0.509	1.000	0.674	0.000	0.000	0.000	0.509	0.512 (14)
Result (N=451)	0.000	0.000	0.000	0.512	1.000	0.677	0.512	0.625 (16)
Conclusion (N=120)	0.508	1.000	0.673	0.000	0.000	0.000	0.508	0.500 (4)
All sentences (N=2,293)	0.542	1.000	0.702	0.000	0.000	0.000	0.542	0.610 (209)

Table 4: Baseline performance on Dimension Title.

	High			Low			Total	
	Precision	Recall	F-measure	Precision	Recall	F-measure	Acc	Kappa
Background (N=808)	0.761	0.742	0.751	0.792	0.809	0.800	0.774	0.551
Gap (N=215)	0.748	0.856	0.798	0.844	0.730	0.783	0.790	0.582
Purpose (N=426)	0.977	0.961	0.969	0.818	0.887	0.851	0.948	0.820
Methodology (N=273)	0.703	0.835	0.763	0.787	0.634	0.702	0.736	0.470
Result (N=451)	0.824	0.700	0.757	0.750	0.857	0.800	0.780	0.559
Conclusion (N=120)	0.729	0.836	0.779	0.800	0.678	0.734	0.758	0.515
All sentences (N=2,293)	0.820	0.801	0.811	0.771	0.792	0.782	0.797	0.592

Table 5: MultiBoostAB performance on Dimension Title.

4.2 Classification Model for Dimension Purpose

Among the evaluated learning algorithms for Dimension Purpose, SimpleLogistic, an algorithm of logistic regression implemented based on Sumner et al. (2005), presented the best performance. The SimpleLogistic classifier achieved F-measure of 0.868 for the *high* category, and 0.801 for the *low* category. Once again, the strongest feature was one of the LSA set, feature 12 (LSA similarity score of the target sentence with all the sentences of the abstract classified as Purpose).

In order to analyze the performance of the classification model with regard to each rhetorical category, we inducted and evaluated five different classifiers, one for each rhetorical category except Purpose. Each of these classifiers was trained using Purpose sentences and a set of sentences of the target category. Baselines classifiers were also evaluated for each category. The baseline performance for all the Dimension Purpose classifiers in terms of Precision, Recall, F-measure, Accuracy, and Kappa is presented in Table 6. The performance of each Dimension Purpose classifier also in terms of Precision, Recall, F-measure, Accuracy, and Kappa is presented in Table 7. The Kappa measure shown in Table 6 refers to the agreement between two human annotators. In Table 7, refers to the agreement among each classifier and our "gold-standard".

The results reported on Table 6 and Table 7 show that all our SimpleLogistic classifiers outperform the baseline. The best performance, both in terms of F-measure and Kappa, was achieved by the Gap classifier. The Kappa for this classifier is 0.754, which indicates a good level of agreement between classifier and human annotator. Apart from Background classifier, all four classifiers performed well. As discussed earlier, it is not surprising that the Background classifier present a weaker performance, as the semantic content of Background sentences usually are general, and, therefore, semantically distant from the Purpose.

Taking into account the F-measure values only for *high* sentences, the best performance was achieved by the Conclusion classifier. In most cases, Conclusion sentences that are strongly related to Purpose, reintroduce the topics stated in the Purpose, even if in a broader context. Again, it is accordance with "general—specific—general" model for scientific texts.

It can also be observed on Table 7 that the Methodology classifier presents the second worse performance on this dimension (it outperforms only the Background classifier), despite the strong relationship between the Methodology and Purpose components. We ascribe this to the characteristics of Methodology sentences, which usually introduce new nouns to the abstract, such as names of techniques, metrics, and other. These newly introduced nouns cause a low LSA score between Methodology and Purpose sentences, contradicting the human annotator whose analysis considers more than just the text surface.

	High			Low			Total	
	Precision	Recall	F-measure	Precision	Recall	F-measure	Acc	Kappa (N)
Background (N=758)	0.000	0.000	0.000	0.501	1.000	0.667	0.501	0.644 (87)
Gap (N=208)	0.620	1.000	0.765	0.000	0.000	0.000	0.620	0.804 (46)
Methodology (N=253)	0.675	1.000	0.805	0.000	0.000	0.000	0.675	0.811 (14)
Result (N=399)	0.661	1.000	0.795	0.000	0.000	0.000	0.661	0.818 (16)
Conclusion (N=102)	0.725	1.000	0.840	0.000	0.000	0.000	0.725	1.000 (4)
All sentences (N=1,720)	0.592	1.000	0.742	0.000	0.000	0.000	0.592	0.815 (167)

Table 6: Baseline performance on Dimension Purpose.

	High			Low			Total	
	Precision	Recall	F-measure	Precision	Recall	F-measure	Acc	Kappa
Background (N=758)	0.786	0.804	0.795	0.801	0.782	0.791	0.792	0.586
Gap (N=208)	0.901	0.915	0.908	0.857	0.835	0.846	0.884	0.754
Methodology (N=253)	0.879	0.895	0.887	0.772	0.744	0.758	0.845	0.645
Result (N=399)	0.889	0.909	0.899	0.814	0.778	0.795	0.864	0.694
Conclusion (N=102)	0.897	0.946	0.921	0.833	0.714	0.769	0.882	0.691
All sentences (N=1,720)	0.852	0.885	0.868	0.824	0.778	0.801	0.841	0.669

Table 7: SimpleLogistc performance on Dimension Purpose.

Both the evaluation results for the classification model and the results from the manual annotation process leads us to employ the Dimension Purpose automatic evaluation to sentences categorized Methodology, Result, and Conclusion.

4.3 Classification Model for Dimension Gap-Background

Considering the evaluated learning algorithms for Dimension Gap-Background, DecisionTable implemented based on Kohavi (1995) presented the best performance. The classifier achieved F-measure of 0.935 for the *yes* category, and 0.795 for the *no* category. We evaluated the performance of each of our features and feature 13 (Maximum LSA similarity score of the target sentence with some Background sentence of the abstract) achieved the best performance. The baseline performance and the DecisionTable classifier in terms of Precision, Recall, F-measure, Accuracy, and Kappa is shown in Table 8.

As shown the Table 8, our classifier outperforms the baseline. Furthermore, the Kappa measured between the classifier and our "gold-standard" was 0.731, which indicates high agreement between the classifier and the human annotator.

Looking at the performance of the classifier, it can be observed that most of them perform better for *yes* sentences. We ascribe this to the presence of anaphoric references in Gap sentences, which decrease the level of semantic relationship. Furthermore, we have a smaller number of sentences ranked as *no* (24.14%).

Evaluation results for the classification model and the results from the manual annotation process encourage us to employ the automatic evaluation of Dimension Gap-Background to sentences rhetorically categorized as Gap in abstracts that have both Background and Gap sentences.

5 Conclusions and Future Work

This work mainly proposes to present four coherence-related dimensions that can be incorporated to the SciPo system. We believe such a proposal to be novel in the context of academic writing, especially in Portuguese.

We also presented how the three dimensions can be automated by using classification models. Dimension Title, Purpose and Gap-Background models present good results and should be incorporated to SciPo as new functionalities. On the other hand, taking into consideration the annotation process, we observed difficulties to label the sentences with regard to the Dimension Linearity-break. Therefore, due to the annotation ambiguity and the low number of examples found, we do not present the classification model for Linearity-break in this work. We believe that such a dimension can be applied to future works in a corpus with can provide more examples of linearity break as, for instance, texts generated by automatic summarizers. In addition, an alternative to be considered in analyzing Dimension Linearity-break is the use of the Entity-grid model proposed by Barzilay and Lapata (2008), which treats local coherence aspects.

	Yes			No			Total	
	Precision	Recall	F-measure	Precision	Recall	F-measure	Acc	Kappa (N)
Baseline (N=183)	0.748	1.000	0.855	0.000	0.000	0.000	0.748	0.725 (46)
DecisionTable (N=183)	0.922	0.949	0.935	0.833	0.761	0.795	0.906	0.731 (183)

Table 8: Baseline performance versus DecisionTable classifier on Dimension Gap-Background.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2003. Criterion online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 681–684. Association for Computational Linguistics.

Scott Elliot. 2003. Intellimetric: From here to validity. In M.D. Shermis and Jill Burstein, editors, *Automatic Essay Scoring: A Cross-Disciplinary Perspective*, pages 71–86, Mahwah, NJ. Lawrence Erlbaum Associates.

Valéria D. Feltrim, Sandra Maria Aluísio, and Maria das Graças Volpe Nunes. 2003. Analysis of the rhetorical structure of computer science abstracts in portugese. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of Corpus Linguistics 2003*, volume 16, part 1, special issue of *UCREL Technical Papers*, pages 212–218.

Valéria D. Feltrim, Simone Teufel, Maria das Graças Volpe Nunes, and Sandra Maria Aluísio. 2006. Argumentative zoning applied to criquing novices scientific abstracts. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*, pages 233–246, Dordrecht, The Neherlands. Springer.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 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.

R. Kohavi. 1995. The power of decision tables. *Machine Learning: ECML-95*, pages 174–189.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. 2003. Automated essay scoring and annotation of essays with the intelligent essay assessor. In M.D. Shermis and Jill Burstein, editors, *Automated Essay Scoring: A Cross Disciplinary Perspective*, pages 87–112, Mahwah, NJ. Lawrence Erlbaum Associates.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *In the Intl. Joint Conferences on Artificial Intelligence*, pages 1085–1090.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Marc Sumner, Eibe Frank, and Mark A. Hall. 2005. Speeding up logistic model tree induction. In Alípio Jorge, Luís Torgo, Pavel Brazdil, Rui Camacho, and João Gama, editors, *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005)*, volume 3721 of *Lecture Notes in Computer Science*, pages 675–683. Springer.

John Swales. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge, UK.

Simone Teufel and Marc Moens. 2002. Summarising scientific articles — experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–446.

Teun A. van Dijk. 1981. *Studies in the Pragmatics of Discourse*. Mouton, The Hague/Berlin.

Geoffrey I. Webb. 2000. Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, 40:159–196.

Robert Weissberg and Suzanne Buker. 1990. *Writing up Research: Experimental Research Report Writing for Students of English*. Prentice Hall.

Ian Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, 2nd edition.