Automatic Critiquing of Novices' Scientific Writing Using Argumentative Zoning

Valéria D. Feltrim
University of São Paulo - ICMC/NILC
Av. Trabalhador São Carlense, 400
São Carlos - SP - Brazil
vfeltrim@icmc.usp.br

Simone Teufel
Computer Laboratory
University of Cambridge
JJ Thomson Avenue, Cambridge CB3 0FD, UK
Simone.Teufel@cam.ac.uk

Abstract

Scientific writing can be hard for novice writers, even in their own language. We present a system that applies Argumentative Zoning (AZ) (Teufel & Moens 2002), a method of determining argumentative structure in texts, to the task of advising novice writers on their writing. We address this task by automatically determining the rhetorical/argumentative status and the implicit author stance of a given sentence in the text. Then users can be advised, for instance that a different order of sentences might be more advantageous, or that certain argumentative "moves" (Swales 1990) are missing. In implementing such a system, we had to port the feature detection stage of Argumentative Zoning from English to Portuguese, as our system is designed for Brazilian PhD theses in Computer Science. In this paper we report on the overall system, the porting exercise, a human annotation experiment to verify the reproducibility of our annotation scheme and an intrinsic evaluation of the AZ-part of our system.

Introduction

It is widely acknowledged that academic writing is a complex task even for native speakers, since it involves the complexities of the writing process as well as those specific to the academic genre (Sharples & Pemberton 1992). It can be even harder for novice writers, who are usually not well-acquainted with the requirements of the academic genre. Even when the basic guidelines on scientific writing are explicit and known, it can be very difficult to apply them to a real text.

A number of writing tools have been described in the literature (Sharples, Goodlet, & Clutterbuck 1994; Broady & Shurville 2000; Narita 2000; Aluisio *et al.* 2001) whose goal is to improve the quality of academic texts produced by novice and/or non-native writers. The project SciPo (short for Scientific Portuguese) aims at analysing the rhetoric structure of Portuguese academic texts — in terms of schematic structure, rhetorical strategies and lexical patterns — to derive models for supporting the creation and evaluation of computational writing tools. To make the project feasible, our analysis has focused on specific sections of theses in Computer Science, namely the abstract and introduction, which are the most studied in the literature (Swales 1990; Weissberg & Buker 1990; Liddy 1991; Santos 1996). These particular sections have also been pointed out as the most difficult ones by graduate students at University of São Paulo, in a questionnaire study by the first author. Our reasons for working on this kind of text were threefold: firstly, in the Brazilian University system, theses have to be written in Portuguese, unlike research articles, which are preferably written in English; secondly, there exists a high degree of standardization in Computer Science texts, as in other scientific research areas; and thirdly, SciPo's developers were already familiar with the Computer Science domain.

As we use a corpus-based approach, an analysis of a specific corpus has been carried out by human annotators, based mainly on Swales's (1990) and Weissberg and Buker's (1990) models. Our annotation scheme has the following categories: BACKGROUND, GAP, PURPOSE, METHODOLOGY, RESULTS, CONCLUSION and OUTLINE. This scheme is closer to Swales' original classification than Teufel & Moens' scheme.

The results of this analysis have been used as basis for a computational model using (good and bad) examples and rules and also as a basis for understanding the problems novice writers face when writing in a new genre. We have identified some writing problems that are specific to the academic genre, such as misuse of lexical patterns and verbal tenses, inefficient organization and inappropriate emphasis on some specific components. On the basis of these results, we believe that especially novice writers may benefit from a writing support tool that provides a repository of good and bad examples of structure, writing strategies and lexical patterns. In the next section we introduce the SciPo system.

The SciPo System

Inspired by the Amadeus system (Aluisio *et al.* 2001), SciPo's current main functionalities can be summarized as: (a) a base of authentic thesis abstracts annotated according to our structure model (Feltrim, Aluisio, & Nunes 2003) that can be browsed and searched for all occurrences of a specific rhetorical strategy; (b) support for building a structure that the writer can use as a starting point for the text; (c) cri-

Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

tiquing rules that can be applied to such a structure; and (d) recovery of authentic cases that are similar to the writer's structure. Also, the existing lexical patterns from the case base sentences are highlighted and the writer can easily add such patterns to a previously built structure.

SciPo contains three knowledge bases, namely the Annotated Case Base, Rules and Similarity Measures, and Critiquing Rules. The Annotated Case Base was built through manual annotation, based on a predefined scheme. This base has 52 instances of schematic structures of authentic abstracts and the same number of introductions, describing the rhetorical components, strategies and lexical patterns of each case. The Rules and Measures of Similarity are based on similarity rules among lists (pattern matching) and on the nearest neighbors matching measure (Kriegsman & Barletta 1993). These rules are used in the case recovery process, when a search is performed according to the user's request of a specific schematic structure. The Critiquing Rules are based on prescriptive guidelines for good writing in the literature and on structural problems observed in the annotated corpus, as an attempt to anticipate and correct problematic structural patterns the writer might construct. The rules cover two distinct problems: content deviations (absence of components) and order deviations (order of occurrence of components and strategies inside the overall structure). Thus, we have four classes of rules: content critiques, order critiques, content suggestions and order suggestions. We use critiques for serious problems as, for instance, absence of purpose. In contrast, we use suggestions for structures that do not have serious problems but can be enriched by adding new components and/or reorganizing the already used ones.

An example of an unfortunate structure is [P M B G P], where the main purpose (first P) is followed by the methodology (M) used to accomplish that purpose. Next, the most natural move would be to present results; however, the writer used a Background component, followed by a Gap (B G), providing more detail of the previously stated purpose and the introduction of yet other purposes. The presence of Background and Gap in the middle of the abstract, separating the main purpose from subsequent detail, confuses the reader, who may lose track of the main purpose of the related research. Also, the sequence Methodology - Background disrupts the cohesion of the text, causing the reader to feel that "something is missing".

Using the resources mentioned above, the writer can build her own structure by choosing components/strategies from a predefined list, get feedback from the critiquing tool until an acceptable structure has been built, recover authentic similar examples and use example lexical patterns (as found in the corpus) in her own writing; cf. the following Purpose sentence (with lexical patterns underlined).

> Este trabalho apresenta o ANIMBS (Animation for MBS), um sistema capaz de visualizar dados gerados por um sistema de simulação de engenharia (SD/FAST) na forma de animações por computador.
>
> [This work presents ANIMBS (Animation for MBS), a system capable of visualizing data generated by an engineering simulation system (SD/FAST) using computer animations.]

SciPo is a powerful system that helps the writer in various ways to organize the structure of his text before the actual writing. However, our aim is to provide a critiquing tool capable of giving feedback on the organization of user's texts after the writing, instead of just aiding its composition. For a tool to supply the user with such information, it has to be able to detect the schematic structure of texts automatically. Such an analysis has been presented, by means of a statistical classifier, by Teufel & Moens (2002). With information about the rhetorical status of each textual part, SciPo could alert the writer to problems like the absence of expected components, unusual ordering of components, and too much emphasis on some components. Also, if the underlying classifier is able to estimate the certainty of its guesses, we believe that a failure to elicit a structure of high certainty could be a clue that the text is unclear or, at least, not written according to the traditional characteristics of the academic genre. In the next section we report on the porting of the Teufel & Moens approach to the Portuguese language.

Argumentative Zoning for Portuguese Texts

For unseen text, the target categories in Argumenative Zoning are estimated by a statistical classifier, based on textual features which can be readily determined in running text. The parameters of the statistical model (a simple Naive Baysian classifier) are learned from human-annotated text. For Portuguese texts we proceed in the same way, adapting the textual features to fit our purposes.

Description of Features

Our first step was to select the set of features to be applied in our experiment. We implemented a set of 7 features, derived from the 16 used by Teufel & Moens (2002): sentence length, sentence location, presence of citations, presence of formulaic expressions, verb tense, verb voice and presence of modal auxiliary.

The Length feature classifies a sentence as short, medium or long length, based on two thresholds (20 and 40 words) that were estimated using the average sentence length present in our corpus.

The Location feature identifies the position occupied by a sentence within a section, in our case, within the abstract. We use four values for this feature: first, medium, 2ndlast and last. We believe that these values characterize common sentence locations for some specific categories of our scheme.

The Citation feature flags the presence or absence of citations in a sentence. As we are not working with full texts, it is not possible to parse the reference list and identify self-citations. Nevertheless, as we are dealing with a thesis corpus, that usually do not contain self-citations, we believe that such distinction will not affect the classification task.

The Formulaic feature identifies the presence of a formulaic expression (see Table 1) in a sentence and the category (within our category scheme) to which an expression belongs. In order to recognize these expressions, we built a set of 377 regular expressions estimated to generate as many as 80,000 strings. The large multiplier is mainly due to the inflectional character of Portuguese. The sources for these regular expressions were phrases mentioned in the literature (translated into Portuguese), and corpus observations, followed by a manual generalization to cover similar constructs.

Category	Formulaic Expression
BACKGROUND	A partir do ano... [*Since the year...*]
GAP	Contudo, é necessário... [*However, it is necessary...*]
PURPOSE	Esta tese apresenta... [*This thesis presents...*]
METHODOLOGY	Nós usamos o modelo... [*We used the model...*]
RESULT	Os resultados mostram... [*The results show...*]
CONCLUSION	Concluímos que... [*We conclude that...*]
OUTLINE	Na seção seguinte... [*In the next section...*]

Table 1: Types and examples of formulaic expressions

Due to the productive inflectional morphology of Portuguese, much of the porting effort went into adapting verb-syntactic features. The Tense, Voice and Modal features report syntactic properties of the first finite verb phrase in indicative or imperative mood.

Tense may assume 14 values, namely NOVERB for verbless sentences, IMP for imperatives or some identifier in the format SimpleTense-(not)perfect-(not)continuous, where SimpleTense refers to the tense of the finite component in the verb phrase, and (not)perfect/(not)continuous flag the presence of perfect/continuous auxiliary "*ter*|*haver*/*estar*". As verb inflection in Portuguese has a wide range of simple tenses – many of which are rather rare in general and even absent in our corpus – we collapsed some of them. As a result, SimpleTense may assume one single value of past/future, to the detriment of the three/two morphological past/future tenses. In addition, SimpleTense neutralizes mood distinction. The Voice feature may assume NOVERB, PASSIVE or ACTIVE. Passive voice is understood here in a broader sense, collapsing some Portuguese verb forms and constructs that are usually used to omit an agent, namely (i) regular passive voice (analogous to English, by means of auxiliary "*ser*" plus past participle), (ii) synthetic passive voice (by means of passivizating particle "*se*") and (iii) a special form of indeterminate subject (also by means of particle "*se*"). The Modal feature may assume NOVERB or flag the presence of a modal auxiliary.

Corpus Annotation Experiment

In order to verify the reproducibility of our annotation scheme and whether our annotated corpus could be used as a valid training material, we performed an experiment with human annotators. Based on our annotation scheme and using specific annotation guidelines similar to the original AZ guidelines, we trained 3 human annotators. The annotators were already knowledgeable of the corpus domain and familiar with scientific writing, so the training focused on the definitions of category (as stated in the written guidelines). Our corpus presents a high number of sentences with "overlapping argumentative roles", which often leads to doubt about the correct category to be assigned. Therefore, the full understanding of the guidelines is very important since they state strategies to deal with conflicts between categories. We used 6 abstracts in the training phase, which was performed in 3 rounds, each round consisting of explanation, annotation, and discussion. We found that the training phase was crucial to calibrate the annotators' knowledge about the annotation task. After training, the annotators were asked to annotate 46 abstracts sentence by sentence, assigning exactly one category per sentence. We use the Kappa coefficient K (Siegel & Castellan 1988) to measure reproducibility among k annotators on N items. In our experiment, items are sentences. The use of the Kappa measure is appropriate in this kind of task since it discards random agreement.

The results show that our scheme is reproducible (K=.69, N=320, k=3). Considering the subjectivity of this task, these results are well acceptable. In a similar experiment, (Teufel, Carletta, & Moens 1999) measured the reproducibility of their scheme as slightly higher (K=.71, N=4261, k=3). One reason why our agreement rate is lower than theirs might be that our scheme refines their OWN category into more specific categories METHODOLOGY, RESULTS and CONCLUSION, which increases the complexity of the task. Collapsing these three categories increases our agreement significantly (K=.82, N=320, k=3). From this we conclude that trained humans can distinguish our set of categories and that we have data which is reliable enough to be used as training material.

Automatic Annotation Results

Our training material is a collection of 52 abstracts from theses in Computer Science (366 sentences; 10,936 words). The abstracts are automatically segmented into sentences using xml tags; Citations in running text are also marked with a xml tag. The text is POS-tagged according to the NILC tagset (Aires *et al.* 2000). The target categories for our experiment were provided by one of the subjects of the annotation experiment described above. As a baseline, we considered a random choice of categories weighted by their distribution in the corpus.

We used the simple Naive Bayesian classifier from the Weka System (Witten & Frank 2000) for our experiments. We measured the system performance comparing the system's prediction with a human ("gold standard") annotation. For our corpus, the similarity between the two was K=.58, when compiled with a 13-fold cross-validation, and K=.56 when using 66% of the data for training and remainder for test. This is encouragingly high amount of agreement (compared to Teufel & Moens' figure of K=.45). Such a good result is in part due to the fact that we are dealing with abstracts (instead of full papers) and that all of them have the same knowledge domain (Computer Science). This result is also much better than the baseline (K=0 and percentage accuracy of 20%).

Further analysis of our results shows that the classifier performs well on all categories apart from the category OUTLINE, cf. the confusion matrix in Table 2. This should not surprise us, since we are dealing with an abstract corpus which has very few OUTLINE sentences (only 6 sentences in the entire corpus). Regarding the other categories, the best performance of the classifier is for PURPOSE sentences (*F-measure*=.825), followed by gap sen-

Machine								
		BACKGROUND	GAP	PURPOSE	METHODOLOGY	RESULTS	CONCLUSION	OUTLINE
Human	BACKGROUND	48	7	0	2	19	0	1
	GAP	7	24	0	0	5	0	0
	PURPOSE	3	0	52	0	9	1	0
	METHODOLOGY	1	0	0	27	17	0	0
	RESULTS	6	1	9	4	93	3	1
	CONCLUSION	0	0	0	0	14	6	0
	OUTLINE	0	0	0	0	5	1	0

Table 2: Confusion matrix: human vs. Weka Naive Bayesian classifier

tences (*F-measure*=.706). We attribute the high performance on these categories to the presence of strong discourse markers on these kind of sentences (modelled by the Formulaic feature).

If we look at the contribution of single features, we see that Formulaic feature is the strongest one. This is in accordance with Teufel and Moens' observations, who counted Agent and Action features amongst the strongest features overall. In our case, the Formulaic feature comprises those two features.

The results for automatic classification are reasonably in agreement with our previous experimental results for human classification. We also observed that the confusion classes of the automatc classification are similar to the confusion classes of our human annotators. The performance of the classifier is lower than human, but promising. Therefore we have reason to believe that the classifier can be used as part of a critiquing tool. The success of the system will be independently determined by user experiments.

Conclusion

We have reported on the porting of Argumentative Zoning from English to Portuguese. The features which were mostly affected by this porting were the Verbal Tense, Modality and Voice features, and the Formulaic Expressions feature. We report here the results of two experiments: 1. agreement results for human annotation, and 2. an intrinsic evaluation of automatic annotation, which are similar but numerically slightly better than Teufel and Moens' original results for English. This is an encouraging result, particularly as the porting was performed in a matter of weeks.

The framework in which we use Argumentative Zoning is that of an automatic Critiquing Tool for Scientific Writing in Portuguese (SciPo). Being able to automatically determine the rhetorical status of a sentence puts us in a position to implement a fully automatic critiquer, in addition to the currently implemented guided writing assistance.

Acknowledgements

The authors would like to thank the financial support from the Brazilian funding agencies CAPES and FAPESP as well as the annotators involved in the human annotation task for their work and suggestions. Special thanks to Jorge Marques Pelizzoni, for his invaluable help.

References

Aires, R. V. X.; Aluisio, S. M.; Kuhn, D. C. S.; Andreeta, M. L. B.; and Oliveira Jr., O. N. 2000. Combining multiple classifiers to improve part of speech tagging: A case study for brazilian portuguese. In *Proceedings of the SBIA'2000*.

Aluisio, S. M.; Barcelos, I.; Sampaio, J.; and Oliveira Jr., O. N. 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*, 257–260.

Broady, E., and 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*. London: CILT. 131–151.

Feltrim, V.; Aluisio, S. M.; and Nunes, M. d. G. V. 2003. Analysis of the rhetorical structure of computer science abstracts in portuguese. In Archer, D.; Rayson, P.; Wilson, A.; and McEnery, T., eds., *Proceedings of Corpus Linguistics 2003, UCREL Technical Papers, Vol. 16, Part 1, Special Issue*, 212–218.

Kriegsman, M., and Barletta, R. 1993. Building a case-based help desk application. *IEEE Expert* 18–26.

Liddy, E. D. 1991. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing and Management* 27(1):55–81.

Narita, M. 2000. Corpus-based english language assistant to japanese software engineers. In *Proceedings of MT-2000 Machine Translation and Multilingual Applications in the New Millennium*, 24–1–24–8.

Santos, M. B. d. 1996. The textual organisation of research paper abstracts. *Text* 16(4):481–499.

Sharples, M., and Pemberton, L. 1992. Representing writing: external representations and the writing process. In Holt, P., and Williams, N., eds., *Computers and Writing: State of the Art*. Oxford: Intellect. 319–336.

Sharples, M.; Goodlet, J.; and Clutterbuck, A. 1994. A comparison of algorithms for hypertext notes network linearization. *International Journal of Human-Computer Studies* 4(40):727–752.

Siegel, S., and Castellan, N. J. J. 1988. *Nonparametric Statistics for the Behavioral Sciences*. Berkeley, CA: McGraw-Hill, 2nd edition.

Swales, J. 1990. *Genre Analysis: English in Academic and Research Settings. Chapter 7: Research articles in English*. Cambridge, UK: Cambridge University Press. 110–176.

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

Teufel, S.; Carletta, J.; and Moens, M. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99)*, 110–117.

Weissberg, R., and Buker, S. 1990. *Writing up Research: Experimental Research Report Writing for Students of English*. Prentice Hall.

Witten, I., and Frank, E. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.