From Syntactical Analysis to Textual Segmentation

Ana Luisa Leal[1],Paulo Quaresma[1] and Rove Chishman[2]

[1]Departamento de Informática, Universidade de Évora, Portugal
[2]UNISINOS – Universidade do Vale do Rio de Sinos, S. Leopoldo, Brasil
{analu,pq@di.uevora.pt,rove@unisinos.br}

Abstract. In this work a proposal for automatic textual segmentation is described. The proposal uses the output of an automatic syntactic analyzer – Parser *Palavras* – to create textual segmentation. Parse trees are used to infer text segments and a dependency tree of the identified segments. The main contribution of this work is the use of the syntactic structure as source for the automatic segmentation of texts, as well, as the use of inference rules for the textual organization.

1 Introduction

This work presents a proposal to make textual segmentation from syntactic structures. We use the information given by the parser *Palavras* [1] about syntactic structures, and we use a set of rules to obtain the segments and to infer their structure. This structure is called a DTS – dependency tree of segments. We believe that from a DTS, the macrostructure can be constructed. The objective of the author can be recognized through the analysis of the macrostructure offered by these trees.

From the output of *Palavras* it is possible to identify segments, *the concrete manifestation of propositions*, which are *conceptual structures*, compounding the textual structure. Moreover it is possible to obtain the relations that are subjacent to this microstructure. The results show that the proposed approach is able to perform the textual segmentation through the information obtained from the parser *Palavras*. Moreover, the result of the syntactic analysis also establishes the hierarchy of the information.

2 Text, Propositions and Segments – An Unique Relation

2.1 Text – The Conception

Our direct goal in this study is to perform the textual segmentation from the text automatic syntactic analysis, but we may also point out that our major goal is the recognition of the discursive structure in the macropropositions of texts. In this context, we are necessary engaged with the definition of what we

understand as being text. In our work, we define that text is the superficial realization of discourse. So, we recognize discourse as a complex entity that contains propositions in which the author expresses his/her objectives. Text is a concrete realization of the discourse.

2.2 Propositions

The discourse is structured in terms of propositions and, in this sense, it can be recognized as a *big proposition* – macroproposition – of the complete sense, where it is possible to observe units with smaller size that are related in the morphologic, syntactic and semantic levels, to compose the discursive structure. In our work, the propositions are inferred from the output of the parser *Palavras*, which presents as result the text syntactically analyzed and segmented in structural terms. The syntactic information produced by the automatic analyzer is used to identify and classify what can be consider as a segment of the text, which represents propositions of discourse.

2.3 Segments

The segments and subsegments are recognized from the results of the automatic syntactic analysis. The segments represent the propositions of the discourse and through them are established the syntactic and semantic relations that are responsible for the sense of that discourse. We assume that the segments are the smaller units of significance, and these smaller units are what establish the microstructural relations.The propositions are the conceptual forms and the segments are the superficial realization forms of those propositions. However, it is acknowledged that the delimitation of these minimal units of significance represents a serious difficulty to the work that involve the textual segmentation, see Pardo and Nunes [4].

Considering the segmentation process, Carlson and Marcu [2] have also proposed rules that have as base the syntax; they can be applied to texts of different typologies. The rules are normally related to clause/sentences such as: main clause; subordinate clause with discourse cue; complements of attribution verbs; coordinate sentences; temporal clause. The authors call the attention to the relative clauses, appositive and parenthetical, because they must be treated as embedded segments. The difference between our proposal and the proposal of Carlson and Marcu [2] is that with the result of the syntactic analysis produced by parser *Palavras* it is possible to identify the segments of the text and to make a more correct segmentation. After concluding the segmentation, we use rules to relate the identified segments that compound the complete text.

The segments and subsegments play different roles. These segments are identified by Mann and Thompson [3] as nucleus and satellite. There is not a preestablished order to the manifestation of the segments in the text, but there are restrictions in the way the relations are established among them. Considering the segments snd subsegments presented in a text, we stand out the existence of

relations among them. These relations have been presented initially by Mann and Thompson [3], classified as Rhetorical Relations.

3 The Parser *Palavras* – A Proposal to Textual Segmentation

3.1 Segmentation

Our proposal is based in the process of segmentation. This process is structured from the the automatic syntactic analysis and it gives us the segments of the text and as a consequence the propositions of discourse. The parser *Palavras* generates the constitutive blocks of text from which are produced the dependency trees of segments – DTS – and later the rules that determine what is the rhetorical role of the proposition in the discourse.

The segmentation is the stage that precedes the formalization of the dependency trees of segments and of the rules that define and delimit the segments, and it is based in the results of automatic syntactic analysis. From the automatic segmentation, we use rules to identify the segments and its correspondent propositions. The rules represent the possible combinations between the segments and its boundaries, as well as, the relations syntactic-semantics that result of these combinations.

4 Hierarchy, Rules and Relations of Text

The notion of textual hierarchy is important in our work because it is directly related to the identification of the textual macroproposition. In order to recognize and explain the order of the segments it is important to explain how the structural mechanisms of the text are articulated and how they represent the objective of author.

The hierarchy observed in the textual organization gives data to the construction of the dependency trees of segments – DTS – through which we can identify the segments and subsegments. This identification is important to determine which is the main segment and which are accessory/secondary to the thematic chain. In the process of systematization of the rules it is possible to develop a strategy that recognizes only the main segment which composes the subject of the text.

4.1 Rules and Dependency Tree Segments

The rules that we propose to create DTS are built from the result of the segmentation given by the parser *Palavras*. From the syntactic data it is possible to identify the segments of text, and their structure. The tree shows that it is possible to identify the principal segments and the secondary segments, main nodes and the secondary nodes of tree. The identification of the main nodes and the secondaries represent structurally the hierarchy in the organization and disposal

of the segments and subsegments. This structural representation is relevant in our work, because we intend to identify automatically the proposition of the discourse.

4.2 Rhetorical Relations

Rhetorical relations are syntactic-semantic mechanisms that appear in the interior of the propositions represented by the segments. These relations are responsible to the presentation of the information of text. In this work, we do not present in detail all the problems around the rhetorical relations, because we are focusing in the textual segmentation problem from an automatic syntactic analysis view point. However, we believe that the rhetorical relations can also be inferred using the same DTS structure as the main source of information.

The identification of the rhetorical relations between the propositions may help in the process of selecting and excluding the propositions that are not necessary to identify the *big proposition* – macroproposition – of the text, i.e., the structure that synthesizes the objective of the author of a specific text.

5 Final Remarks and Future Work

The research that we propose is extensive and it is related with different knowledge areas, and different cognitive levels. We believe that it is possible to develop a robust system that is capable of articulating these areas and levels. Some of the stages proposed already were concluded and showed satisfactory results. These results give us a good support to the continuity of the study. As final claim, we strongly believe that it is possible to use the text parsing structure to obtain segments and its correspondent propositions, to generate DTS, to infer rhetorical structures, and to obtain the text macrostructure.

References

1. E. Bick. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus University Press, 2000.
2. L. Carlson and D. Marcu. Discourse tagging reference manual. Technical Report 545, ISI Technical Report ISI-TR-545, 2001.
3. W. Mann and S. Thompson. Rhetorical structure theory: toward a functional theory of text organization. Technical report, Technical ReportISI/RS-87-190, 1987.
4. M. Pardo, T.A.S e Nunes. Análise de discurso: Teorias discursivas e aplicações em processamento de línguas naturais. Technical Report 196, Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo, 2003.