A Two-Step Summarizer of Brazilian Portuguese Texts

Thiago I. Carbonel[1], Eloize M. Seno[1], Thiago A.S. Pardo[1],

Jorge César Coelho[2], Sandra Collovini[2], Lucia Helena M. Rino[1], Renata Vieira[2]

[1] Núcleo Interinstitucional de Lingüística Computacional (NILC/São Carlos)
DC/UFSCar – CP 676, 13565-905 São Carlos, SP, Brazil
http://www.nilc.icmc.usp.br
taspardo@icmc.usp.br,eloize@mail.fpte.br
{thiagocarbonel,lucia}@dc.ufscar.br

[2] UNISINOS - Av. Unisinos, 950, 93022-000 – São Leopoldo, RS - Brazil
{cesar.coelho,sandrinha.collovini}@gmail.com,
renata@exatas.unisinos.br

Abstract. We describe a two-step text summarizer of texts written in Brazilian Portuguese that comprises a discourse analyzer and a summary structurer. The system tackles the problem of dangling anaphors in the summaries based upon the Rhetorical Structure and Veins Theories. The system potentialities deserve discussion due to its insertion in ongoing work on AS focused on co-referential chains.

1 Introduction

Automatic text summarization is the task of automatically producing shorter versions of source texts [3]. As such, summaries must convey the central idea of the corresponding texts. However, they usually present coherence and cohesion problems, amongst others. Dealing with them requires deep processing on text meaning and representation. Our system handles text structures in a knowledge-rich approach to tackle a cohesion problem that may damage coherence of automatic summaries: breaks in coreferential chains (CRCs). A CRC conveys inter-related text units through referential means. Only anaphors are considered: one of the units conveys the anaphor and another one, previously occurring, its referent. In this context, a CRC break occurs in a summary when the anaphoric unit is included in a summary but its antecedent is not. The summarizer and a preliminary assessment are described, respectively, in Sections 2 and 3.

2 The system architecture

Our summarizer, *RHeSuma-2*, plugs DiZer [9], which produces RST trees [4] of plain texts (Figure 1), onto RHeSumaRST [14], which summarizes the trees yielding the *target sum-*

maries in BP aiming at avoiding CRC breaks. Both modules are independent; RHeSumaRST is language-independent, but DiZer is so far genre-dependent.

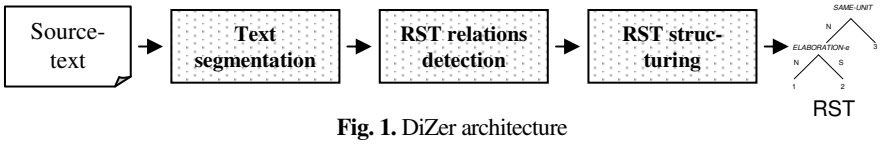

Fig. 1. DiZer architecture

An RST tree has text segments as leaves and RST relations between segments as internal nodes. According to Marcu [5, 7], leaves convey *elementary discourse units* (EDUs). DiZer linguistic patterns (c.a. 750 patterns) are derived from cue-phrases, as suggested in [6]. Text segmentation simply delimits single clauses as EDUs, with no parsing involved. Eventually, DiZer yields ambiguous RST trees of a source text. Conversely, it identifies no relation if none of the patterns apply, issuing an ELABORATION relation instead.

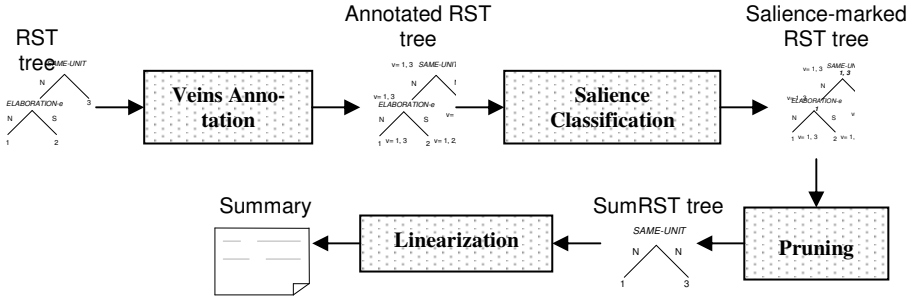

Fig. 2. RHeSumaRST architecture

RHeSumaRST pruning (Figure 2) uses heuristics [14] to identify superfluous information. It relies on RST to address coherence and informativity and on the Veins Theory, or VT [2] to tackle CRC breaking. Both theories are complementary: RST focuses on the salience of information [5]; VT focuses on the *domain of evocative accessibility* of each EDU, or its "vein". After vein-annotating a source RST tree and calculating the salience of each node, once an EDU is included in the summary, RHeSumaRST assures the inclusion of its full vein. This strategy is completely blind, concerning anaphora resolution. So far, linearization has been extractive: EDUs are simply juxtaposed.

3 Summarizing Brazilian Portuguese Texts: A Case Study

A blackbox evaluation (i.e., only comparing inputs and outputs of automatic systems) of RHeSuma-2 and two other systems was carried out using a test corpus[1] comprising 47 news texts. DiZer was not retrained because we aimed at prospecting difficulties and potentialities of the full system as it is and re-customizing it would be time-consuming and should take advantage of the former goal. The study focused on the systems potentialities to con-

[1] Selected from the Rhetalho Corpus [11].

vey no dangling anaphors in the summaries. Certainly RHeSumaRST depends on the quality of its input structures, thus on DiZer performance. However, the posed problem has no influence on DiZer. The use of the summarizers Topline and Salience replicated a former experiment setting [13]: Salience is actually Marcu's summarizer [5]; Topline is a baseline that prunes every satellite of an RST tree, leaving only central information [4]; compression rate is 70%. The test corpus was first parsed by PALAVRAS [1][2], then hand-annotated for anaphoricity using MMAX [8] and coreference encoding guidelines [13]. This step aimed at distinguishing new and old DDs, so that the latter could be linked to their antecedents, as defined in [2]. An *ideal corpus* of coreference annotated source texts was thus produced upon the agreement of three annotators, amounting to 940 DDs that embedded c.a. 300 anaphoric ones (c.a. 32% of all).

Overall, the summary corpus embedded 232 CRCs and only 37 CRC breaks were produced by the three systems together (16% CRC breaking rate). Considering each system independently, CRC breaks should be resolved for more coherent summaries: Topline, which has no focus on coherence, presented the greatest rate of breaks. Concentration of breaks in a summary was also relevant, since the less CRC breaks, the more coherent the summary. Both RHeSuma-2 and Salience had just one break per summary, whilst Topline had 16 breaks in 14 summaries (14% rate). Still it is questionable if RHeSuma-2 is actually of some use, for the effort of its software engineering and potential of scalibility, which is smaller than that of Salience, which is much simpler than RHeSuma-2. However, when considering relative CRC breaks, there was only a 9% loss of coherence in RHeSuma-2, against 13% in the Salience results.

4 Final remarks

Clearly, RHeSuma-2 did not perform well, if we consider the overall figures. 28% of its summaries introduced CRC breaks and only 17 summaries (36%) were considered cohesive and coherent. Its performance may improve if DiZer is retrained under the intended source texts conditions. Actually both DiZer segmentation and discourse analyzer have been under reformulation by first parsing the source text dor discourse analysis. This is very likely to improve recognizing the EDUs boundaries and determining proper cue phrases, as well as determining more adequate RST relations and increase correctness on the nuclearity of inter-related spans, which implies directly the quality in determining veins.

Acknowledgments

The work reported has been funded by CNPq (ProCaCoSA Project – Proc. Nro. 50703020044).

[2] http://visl.sdu.dk/visl/pt/parsing/automatic [Fev2006].

References

1. Bick, E.: The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Arhus University, Arhus (1997)
2. Coelho, J.C.; Muller, V.; Collovini, S; Vieira, R.; Rino, L.H.M. Resolving Portuguese Nominal Anaphora. In Vieira, R.; Quaresma, P.; Nunes, M.d.G.V.; Mamede, N.J.; Oliveira,C.; Dias, M.C. (eds.), Proc. of the 7th International Workshop PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006, Lecture Notes in Artificial Intelligence, 3960 (2006) 160-169
3. Cristea, D.; Ide, N.; Romary, L.: Veins Theory: A Model of Global Discourse Cohesion and Coherence. In: Proc. of the Coling/ACL (1998) 281-285
4. Mani, I.: Automatic Summarization. John Benjamin's Publishing Company (2001)
5. Mann, W.C.; Thompson, S.A.: Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RS-87-190 (1987)
6. Marcu, D.: The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. PhD Thesis, Department of Computer Science, University of Toronto (1997)
7. Marcu, D.: Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury (eds.), Advances in Automatic Text Summarization. The MIT Press (1999) 123-136
8. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. The MIT Press. Cambridge, Massachusetts (2000)
9. Müller, C.; Strube, M.: MMAX: A tool for the annotation of multi-modal corpora. In: Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems , Seattle, Wash., USA, August 5 (2001) 45-50
10. Pardo, T.A.S; Nunes, M.G.V.; Rino, L.H.M. DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. In A.L.C. Bazzan, S. Labidi (eds.) Advances in Artificial Intelligence. *Lecture Notes in Computer Science*, 3171 (2004) 224-234
11. Pardo, T.A.S.; Seno, E.R.M. Rhetalho: um corpus de referência anotado retoricamente. In Anais do V Encontro de Corpora. São Carlos-SP, Brasil. 25 a 26 de Novembro (2005)
12. Pereira, F.C.N. and Warren, D.H.D.: Definite Clause Grammars for Language Analysis – A Survey of the Formalism and Comparison with Augmented Transition Networks. Artificial Intelligence, N. 13 (1980) 231-278
13. Poesio, M.: Coreference.mate dialogue annotation-deliverable d2.1. Technical report, http://www.ims.uni.stuttgart.de/projekte/mate/mdag (2000)
14. Seno, E.R.M.; Rino, L.H.M. Summarizing RST trees focusing on referential chains: A case study. In the *Proc. of the III Workshop em Tecnologia da Informação e da Linguagem Humana* – TIL'2005. São Leopoldo – RS. Julho (2005a)
15. Seno, E.R.M., Rino, L.H.M.: Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In: *Proc. of Workshop on Crossing Barriers in Text Summarization Research – RANLP'05* (2005b) 70-75