The Coreference Annotation of the CSTNews Corpus

Thiago Alexandre Salgueiro Pardo[1], Jorge Baptista[2], Magali Sanches Duran[1], Maria das Graças Volpe Nunes[1], Fernando Antônio Asevedo Nóbrega[1], Sandra Maria Aluísio[1], Ariani Di Felippo[3], Eloize Rossi Marques Seno[4], Raphael Rocha da Silva[1], Rafael Torres Anchiêta[1], Henrico Bertini Brum[1], Márcio de Souza Dias[5], Rafael de Sousa Oliveira Martins[1], Erick Galani Maziero[1], Jackson Wilke da Cruz Souza[3], and Francielle Alves Vargas[1]

[1] Universidade de São Paulo
[2] Universidade do Algarve
[3] Universidade Federal de São Carlos
[4] Instituto Federal de São Paulo
[5] Universidade Federal de Goiás

taspardo@icmc.usp.br,{jorge.manuel.baptista,magali.duran}@gmail.com, gracan@icmc.usp.br,fernandoasevedo@gmail.com,sandra@icmc.usp.br, {arianidf,eloizeseno,raphaelsilva2500}@gmail.com,rta@usp.br,{henrico.brum,marciosouzadias,martins.rso,egmaziero,jackcruzsouza}@gmail.com, franciellеavargas@hotmail.com

Abstract. We report in this paper the coreference annotation process of the CSTNews corpus as part of a collective task of the IberEval 2017 conference. The annotated corpus is composed of 140 news texts written in Brazilian Portuguese language and counts with several annotation layers, including annotations in the morphosyntax/syntax, semantics, and discourse levels. The annotation, focused on nominal references, was conducted in a semi-automatic way by five teams, achieving satisfactory annotation agreement results.

1 Introduction

Coreference resolution is the task of finding linguistic expressions in a text that refer to the same entity [1]. As an illustration of coreference occurrence, we show below a short text with some coreferent elements in bold. In this short text, the referring expressions "a passenger plane", "the airplane" and "it" refer to the same entity and form a "coreference chain". It is interesting to notice that the coreference resolution task includes pronominal anaphora resolution, which is part of the problem.

> *At least 17 people died after the crash of* ***a passenger plane*** *in the Democratic Republic of Congo. According to an ONU spokeswoman,* ***the airplane*** *was trying to land in the Bukavu airport in the midst of a storm.* ***It*** *failed to reach the runway and fell in a forest 15 kilometers away from the airport.*

5 Final Remarks

All the annotated data is in an XML format, which is a traditional way of marking and making data available. It shall be available in the SUCINTO project website, as it constitutes an additional linguistic annotation layer of the CST-News corpus.

We expect that the produced coreference annotation fosters other research initiatives on discourse processing tasks. For the short term, the new data may help to improve summarization models, specifically those involving coherence and cohesion evaluation, for which the occurrence and distribution of referring expressions are very important features (see, e.g., the entity-based model proposed in [38]).

For future work, concerning the CSTNews corpus, the task of pronominal anaphora resolution remains to be done, as it was not directly tackled in the reported annotation effort.

Acknowledgments

The authors are grateful to FAPESP, CAPES and CNPq for supporting this work.

References

1. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall (2009)
2. Hobbs, J.R.: Resolving pronoun references. Lingua **44**(4) (1978) 311–338
3. Mitkov, R.: Anaphora Resolution. Pearson Education (2002)
4. Haponchyk, I., Moschitti, A.: A practical perspective on latent structured prediction for coreference resolution. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Volume 2. (2015) 143–149
5. Wiseman, S., Rush, A.M., Shieber, S.M.: Learning global features for coreference resolution. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. (2016) 994–1004
6. Grosz, B., Joshi, A., Weisten, S.: Centering: A framework for modeling the local coherence of discourse. Computational Linguistics **21**(2) (1995) 203–225
7. Cristea, D., Ide, N., Romary, L.: Veins theory. an approach to global cohesion and coherence. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. (1998) 281–285
8. Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., Zhang, Y., eds.: Joint Conference on EMNLP and CoNLL – Shared Task, Association for Computational Linguistics (2012)

9. Cardoso, P.C.F., Maziero, E.G., Castro Jorge, M.L.R., Seno, E.M.R., Di Felippo, A., Rino, L.H.M., Nunes, M.G.V., Pardo, T.A.S.: CSTNews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: Proceedings of the 3rd RST Brazilian Meeting. (2011) 88–105
10. Vieira, R., Gonçalves, P.N., Souza, J.G.C.: Processamento computacional de anáfora e correferência. Revista de Estudos da Linguagem **16**(1) (2008) 263–284
11. Collovini, S., Carbonel, T.I., Fuchs, J.T., Coelho, J.C., Rino, L.H.M., Vieira, R.: Summ-it: Um corpus anotado com informações discursivas visando a sumarização automática. In: Proceedings of V Workshop em Tecnologia da Informação e da Linguagem Humana. (2007) 1605–1614
12. Fonseca, E., Vieira, R., Vanin, A.: Improving coreference resolution with semantic knowledge. In: Proceedings of the International Conference on Computational Processing of the Portuguese Language. (2016a) 213–224
13. Fonseca, E., Vieira, R., Vanin, A.: Corp: Coreference resolution for portuguese. In: Proceedings of the International Conference on Computational Processing of the Portuguese Language - Demonstration Session. (2016b) 9–11
14. Fonseca, E., Sesti, V., Atonitsch, A., Vanin, A., Vieira, R.: CORP: Uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências. LinguaMÁTICA **9**(1) (2017) 3–18
15. Radev, D.R.: A common theory of information fusion from multiple text sources, step one: Cross-document structure. In: Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue. (2000) 74–83
16. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: A theory of text organization. Technical Report Technical Report ISI/RS-87-190 (1987)
17. Hearst, M.: Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics **23**(1) (1997) 33–64
18. Owczarzak, K., Dang, H.T.: Who wrote what where: Analyzing the content of human and automatic summaries. In: Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages. (2011) 25–32
19. Baptista, J., Hagège, C., Mamede, N.: Identificação, classificação e normalização de expressões temporais do português: A experiência do segundo HAREM e o futuro. In Mota, C., Santos, D., eds.: Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. (2008) 33–54
20. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
21. Bick, E.: The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)
22. Silveira, S.B., Branco, A.: Enhancing multi-document summaries with sentence simplification. In: Proceedings of the 14th International Conference on Artificial Intelligence. (2012) 742–748
23. Ribaldo, R., Cardoso, P.C.F., Pardo, T.A.S.: Exploring the subtopic-based relationship map strategy for multi-document summarization. Journal of Theoretical and Applied Computing **23**(1) (2016) 183–211
24. Cardoso, P.C.F., Pardo, T.A.S.: Multi-document summarization using semantic discourse models. Procesamiento del Lenguaje Natural **56** (2016) 57–64
25. Nóbrega, F.A.A., Pardo, T.A.S.: Update summarization for portuguese. In: Proceedings of the 6th Brazilian Conference on Intelligent Systems (To appear). (2017)
26. Dias, M.S., Pardo, T.A.S.: A discursive grid approach to model local coherence in multi-document summaries. In: Proceedings of the 16th Annual SIGdial Meeting on Discourse and Dialogue. (2015) 60–67

27. Camargo, R.T., Di Felippo, A., Pardo, T.A.S.: On strategies of human multi-document summarization. In: Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology. (2015) 141–150
28. Maziero, E.G., Hirst, G., Pardo, T.A.S.: Semi-supervised never-ending learning in rhetorical relation identification. In: Proceedings of the Recent Advances in Natural Language Processing. (2015) 436–442
29. Braud, C., Coavoux, M., Søgaard, A.: Cross-lingual RST discourse parsing. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Volume 1. (2017) 292–304
30. Ponti, E.M., Korhonen, A.: Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. In: Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics. (2017) 25–30
31. Di Felippo, A., Nenkova, A.: Phrase generalization: a corpus study in multi-document abstracts and original news alignments. In: Proceedings of the 10th Linguistic Annotation Workshop. (2016) 151–159
32. Cardoso, P.C.F., Pardo, T.A.S., Taboada, M.: On the contribution of discourse to topic segmentation. In: Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue. (2013) 92–96
33. Maziero, E.G., Castro Jorge, M.L.R., Pardo, T.A.S.: Revisiting cross-document structure theory for multi-document discourse parsing. Information Processing & Management **50**(2) (2014) 297–314
34. Souza, J.W.C., Di Felippo, A.: O corpus cstnews e sua complementaridade temporal. In: PROPOR Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish. (2014) 105–109
35. Oliveira, H.G., Gomes, P.: Eco and onto.pt: a flexible approach for creating a portuguese wordnet automatically. Language Resources and Evaluation **48**(2) (2014) 373–393
36. Sarmento, L., Pinto, A.S., Cabral, L.: REPENTINO - a wide-scope gazetteer for entity recognition in portuguese. In: Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language. (2006) 31–40
37. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics **22**(2) (1996) 249–254
38. Barzilay, R., Lapata, M.: Modeling local coherence: An entity-based approach. Computational Linguistics **34**(1) (2008) 1–34