A pesquisa em Geração de língua natural em português na EACH-USP (rascunho)

Ivandré Paraboni

School of Arts, Sciences and Humanities, University of São Paulo (USP / EACH)
Av. Arlindo Bettio, 1000 - São Paulo, Brazil
ivandre@usp.br

Abstract. *Este relatório técnico apresenta uma visão geral de métodos de geração de língua natural aplicados ao português (e em alguns casos ao inglês) em projetos desenvolvidos na EACH-USP, com o objetivo de documentar a produção da área para futura referência, e estabelecer possíveis aplicações na linha de pesquisa de interpretação de língua natural.*

Key words: Geração de língua natural, seleção de conteúdo, expressões de referência

1 Seleção de conteúdo

Um dos primeiros estudos desenvolvidos foi o trabalho em [1] e posteriormente ampliado em [2], em que foi discutida a seleção de conteúdo para geração automática de documentos estruturados em seções, parágrafos etc.

2 Geração de expressões de referência

A área de geração de expressões de referência (GER), ou seleção de conteúdo de descrições definidas, foi a mais desenvolvida no período. Esta linha de investigação partiu de estudos prévios de interpretação de referências pronominais [3–5], reestruturada para tratamento deste fenômeno na tarefa computacional 'oposta', ou seja, de geração.

Uma primeira versão do algoritmo de seleção de conteúdo referencial que leva em conta a redundância lógica destas expressões foi apresentado em [6] e de forma mais completa em [7]. Sua forma final aparece em [8], e uma avaliação com um experimento envolvendo participantes humanos foi apresentada em [9].

No âmbito do grupo de pesquisa da EACH-USP, a continuidade deste estudo aparece em [10, 11] e, com maior destaque, no projeto de construção do córpus Stars/Stars2 para estudo de fenômenos de superespecificação [12–15].

3 Aplicações na interpretação de língua natural

A partir de 2018, o foco da pesquisa fdo grupo deixou de ser a geração de língua natural, e passou a tratar de tarefas de interpretação. O primeiro grande projeto deste tipo abordou o reconhecimento de traços de personalidade a partir

de texto de redes sociais [16–20]. Esta linha de pesquisa foi posteriormente estendida de modo a incluir o problema mais geral de caracterização autoral (de gênero, faixa etária e outras, cf. [21–24]), atribuição autoral [25, 26], detecção de posicionamentos [27, 28] e de discurso de ódio [29].

Atualmente, este último tópico (detecçãod e posicionamentos) é tema de projeto atual, em paralelo a um estudo de detecção de transtornos de saúde mental em redes sociais [30].

4 Considerações

Este relatório apresentou um resumo da pesquisa em GLN realizada na EACH-USP, e de alguns dos seus desdobramentos para a área de interpretação de língua natural atualmente dominante.

Apesar de atualmente menos discutida, entretanto, a linha de pesquisa em GLN não foi completamente abandonada. Em especial, a questão do uso de informação de personalidade nesta tarefa foi apresentado em [31–34], e em outras iniciativas deste tipo, que integram o projeto do córpus b5 para tratamento de modelos computacionais de personalidade humana. Uma visão geral da arquitetura proposta, adaptada de [19], é apresentada na figura 1.

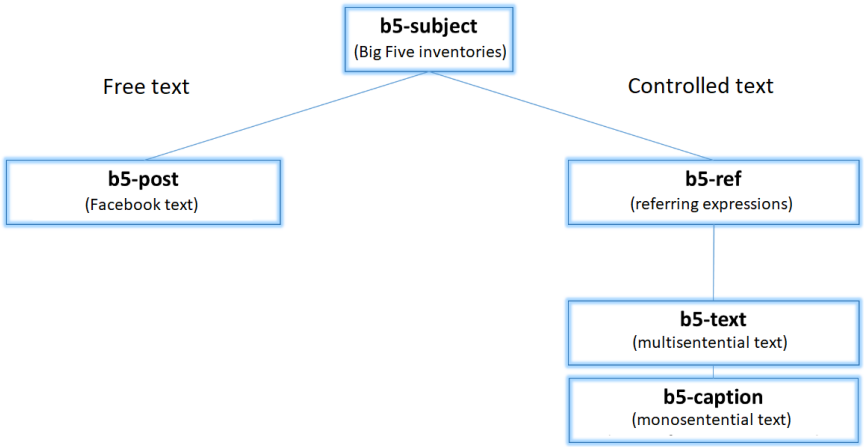

Fig. 1. Estrutura do córpus b5, adaptada de [19].

Nesta arquitetura, observa-se que o córpus consiste de textos livres (provenientes de redes sociais) e controlado (obtido por meio de experimentos com partiicpantes humanos a partir de estímulos conrolados), neste caso divididos em conjuntos de expressões referenciais, textos descritivos e legendas de figuras.

References

1. Paraboni, I., van Deemter, K.: Issues for the generation of document deixis. In: Procs. of workshop on Deixis, Demonstration and Deictic Belief in Multimedia Contexts, in association with the 11th European Summers School in Logic, Language and Information (esslli99). (1999) 44–48
2. Paraboni, I., van Deemter, K.: Towards the generation of document-deictic references. In: Information sharing: reference and presupposition in language generation and interpretation. CSLI Publications (2002) 329–352
3. Paraboni, I.: Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa. Master's thesis, PUCRS, Porto Alegre (1997)
4. Paraboni, I., de Lima, V.L.S.: Possessive pronominal anaphor resolution in Portuguese written texts. In: Proceedings of the 17th international conference on Computational linguistics-Volume 2, Association for Computational Linguistics (1998) 1010–1014
5. Cuevas, R.R.M., Paraboni, I.: A machine learning approach to Portuguese pronoun resolution. In: IBERAMIA-2008, Lecture Notes in Artificial Intelligence 5290, Lisboa, Portugal, Springer-Verlag (2008) 262–271
6. Paraboni, I.: An algorithm for generating document-deictic references. In: Procs. of workshop Coherence in Generated Multimedia, associated with First Int. Conf. on Natural Language Generation (INLG-2000), Mitzpe Ramon. (2000) 27–31
7. Paraboni, I., van Deemter, K.: Generating easy references: the case of document deixis. In: INLG-2002, New York. (2002) 113–119
8. Paraboni, I.: Generating references in hierarchical domains: the case of Document Deixis. PhD thesis, University of Brighton (2003)
9. Paraboni, I., Masthoff, J., van Deemter, K.: Overspecified reference in hierarchical domains: measuring the benefits for readers. In: Proceedings of the fourth international natural language generation conference (INLG-2006), Sydney, Australia, Association for Computational Linguistics (2006) 55–62
10. Pereira, D.B., Paraboni, I.: Statistical surface realisation of Portuguese referring expressions. In: Gotal-2008, Lecture Notes in Artificial Intelligence 5221, Gothenburg, Sweden, Springer-Verlag (2008) 383–392
11. de Lucena, D.J., Paraboni, I., Pereira, D.B.: From semantic properties to surface text: The generation of domain object descriptions. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial **14**(45) (2010) 48–58
12. Paraboni, I., Yamasaki, A.K., da Silva, A.S.R., Teixeira, C.V.M.: Generating underspecified descriptions of landmark objects. In: Text, Speech and Dialogue (TSD-2014), Lecture Notes in Artificial Intelligence 8655, Brno, Springer (2014) 76–83
13. Teixeira, C.V.M., Paraboni, I., da Silva, A.S.R., Yamasaki, A.K.: Generating relational descriptions involving mutual disambiguation. In: Computational Linguistics and Intelligent Text Processing (CICLing-2014), Lecture Notes in Computer Science 8403, Kathmandu, Nepal, Springer (2014) 492–502
14. Paraboni, I., Galindo, M., Iacovelli, D.: Stars2: a corpus of object descriptions in a visual domain. Language Resources and Evaluation **51**(2) (2017) 439–462
15. dos Santos Silva, D., Paraboni, I.: Generating spatial referring expressions in interactive 3D worlds. Spatial Cognition & Computation **15**(03) (2015) 186–225
16. Silva, B.B.C., Paraboni, I.: Learning personality traits from Facebook text. IEEE Latin America Transactions **16**(4) (2018) 1256–1262

17. dos Santos, V.G., Paraboni, I., Silva, B.B.C.: Big five personality recognition from multiple text genres. In: Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415, Prague, Springer-Verlag (2017) 29–37
18. Silva, B.B.C., Paraboni, I.: Personality recognition from Facebook text. In: 13th International Conference on the Computational Processing of Portuguese (PROPOR-2018) LNCS vol. 11122, Canela, Springer-Verlag (2018) 107–114
19. Ramos, R.M.S., Neto, G.B.S., Silva, B.B.C., Monteiro, D.S., Paraboni, I., Dias, R.F.S.: Building a corpus for personality-dependent natural language understanding and generation. In: 11th International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan, ELRA (2018) 1138–1145
20. dos Santos, W.R., Ramos, R.M.S., Paraboni, I.: Computational personality recognition from facebook text: psycholinguistic features, words and facets. New Review of Hypermedia and Multimedia **25**(4) (2019) 268–287
21. Hsieh, F.C., Dias, R.F.S., Paraboni, I.: Author profiling from facebook corpora. In: 11th International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan, ELRA (2018) 2566–2570
22. Pavan, M.C., dos Santos, V.G., Lan, A.G.J., ao Trevisan Martins, J., dos Santos, W.R., Deutsch, C., da Costa, P.B., Hsieh, F.C., Paraboni, I.: Morality classification in natural language text. IEEE transactions on Affective Computing (2020)
23. Flores, A.M., Pavan, M.C., Paraboni, I.: User profiling and satisfaction inference in public information access services. Journal of Intelligent Information Systems (2021) –
24. Delmondes Neto, J.P., Paraboni, I.: Multi-source BERT stack ensemble for cross-domain author profiling. Expert Systems (2021) –
25. Custódio, J.E., Paraboni, I.: EACH-USP ensemble cross-domain authorship attribution. In: Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF-2018) vol.2125, Avignon, France (2018)
26. Custódio, J.E., Paraboni, I.: Stacked authorship attribution of digital texts. Expert Systems with Applications **176** (2021) 114866
27. dos Santos, W.R., Paraboni, I.: Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text. In: Recents Advances in Natural Language Processing (RANLP-2019), Varna, Bulgaria, INCOMA Ltd. (2019) 1069–1075
28. Pavan, M.C., dos Santos, W.R., Paraboni, I.: Twitter Moral Stance Classification using Long Short-Term Memory Networks. In: 9th Brazilian Conference on Intelligent Systems (BRACIS). LNAI 12319, Springer (2020) 636–647
29. da Silva, S.C., Ferreira, T.C., Ramos, R.M.S., Paraboni, I.: Data driven and psycholinguistics motivated approaches to hate speech detection. Computación y Systemas **24**(3) (2020) 1179–1188
30. dos Santos, W.R., Funabashi, A.M.M., Paraboni, I.: Searching Brazilian Twitter for signs of mental health issues. In: 12th International Conference on Language Resources and Evaluation (LREC-2020), Marseille, ELRA (2020) 6113–6119
31. Paraboni, I., Monteiro, D.S., Lan, A.G.J.: Personality-dependent referring expression generation. In: Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415, Prague, Springer-Verlag (2017) 20–28
32. Ramos, R.M.S., Monteiro, D.S., Paraboni, I.: Personality-dependent content selection in natural language generation systems. Journal of the Brazilian Computer Society **26**(2) (2020)
33. Neto, G.B.S., Paraboni, I.: Reescrita sentencial baseada em traços de personalidade. Linguamática **12**(1) (2020) 49–61
34. da Silva Rocha, D., Paraboni, I.: Building referring expression corpora with and without feedback. Language Resources and Evaluation **54**(4) (2020) 875–891