An NLP-paradigm for Retrieving Historical Documents from the Romanian Language

Daniela-Maria Cristea, Dana Lupsa, Gabriel-Bogdan Trofin

Computer Science Department, Babes-Bolyai University, Cluj-Napoca, RO-400084

danielacristea@cs.ubbcluj.ro, box2_2001@yahoo.com

This paper introduces the need in historical document analysis to use natural language processing techniques to aid historians and scientists by organizing metadata, enabling them to search for appropriate considerations among all processed documents. We consider its use for academic goals and for the benefit of the community through various elements such as science usefulness, accessibility to late medieval documents to research events or private history of 14-16 century. Our approach provides high accuracy detection of the language used in a text and should work with at least one sentence length texts. The information is assessed based on the recognition of known words in different languages and the greatest probability of choosing the one.

References

[1] D. Cristea, C. Forascu, Linguistic Resources and Technologies for Romanian Language, *Computer Science J. of Moldova*, **14,** 1 (40), (2006) pp.34–73.

[2] S. Ruder, An overview of gradient descent optimization algorithms, *J. of arXiv:1609.04747v2*, (2016).

[3] Ruder, Sebastian and Bingel, Joachim and Augenstein, Isabelle and Søgaard, Anders, Sluice networks: Learning what to share between loosely related tasks, *J. of arXiv preprint arXiv:1705.08142*, (2017).

[4] V. Mirzaee Abar, *An Ontological Approach to Representing Historical Knowledge*, Vancouver, BC: University of British Columbia, 2004.

[5] R. Grishman, *Computational Linguistics: An Introduction*, Cambridge University Press, 1988.

[6] R. Mitkov, *The Oxford Handbook of Computational Linguistics*, Oxford University Press, 2006.

[7] S. Bird, E. Loper, E. Klein, *Natural Language Processing with Python*, OReilly Media Inc., 2009.

[8] Karttunen, M. Arnold, Zwicky, *Introduction Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, Cambridge University Press., 1985.

[9] A. Pesah, A. Wehenkel, G. Louppe, *Recurrent Machines for Likelihood-Free Inference*, O'Reilly Media Inc., 1985