Robust and multilingual analysis of historical documents

Antoine Doucet

University of La Rochelle, Avenue Einstein F-17000 La Rochelle, France

Abstract

Many documents can only be accessed through digitization. This is notably the case of historical and handwritten documents, but also that of many digitally-born documents, turned into images for various reasons (e.g., a file conversion or the intermediary use of an analog form in order to manually sign a document, fill out a form, send by post, etc.). Being able to analyze the textual content of such digitized documents requires a phase of conversion from the captured image to a textual representation, key parts of which are optical character recognition (OCR) and layout analysis. The resulting text and structure are often imperfect, to an extent which is notably correlated with the quality of the initial medium (which may be stained, folded, aged, etc.) and with the quality of the image taken from it. In this talk, I will present recent advances in AI and natural language understanding that enable this type of corpus to be analyzed in a way that is robust to digitization. For example, I will show how we were able, in the H2020 NewsEye project to create state-of-the-art results for the cross-lingual recognition and disambiguation of named entities (names of people, places, and organizations) in large corpora of historical newspapers written in 4 languages, written between 1850 and 1950. This type of results paves the way to a large-scale analysis of digitized documents, notably able to cross linguistic borders.

Short Bio

Antoine Doucet is a tenured full Professor in computer science at the L3i laboratory of the University of La Rochelle since 2014. Leader of research group in document analysis, digital contents and images in La Rochelle Université (about 50 people), he also directs the ICT department of the Vietnamese-French University of Science and Technology of Hanoi (USTH). He was until January 2022 the coordinator of the H2020 project NewsEye, focusing on augmenting access to historical newspapers, across domains and languages. He further leads the effort on semantic enrichment for low-resourced languages in the context of the H2020 project Embeddia. His

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'22 Workshop, Stavanger (Norway), 10-April-2022

antoine-doucet@univ-lr.fr (A. Doucet)
https://pageperso.univ-lr.fr/antoine.doucet/ (A. Doucet)
0000-0001-6160-3356 (A. Doucet)

main research interests lie in the fields of information retrieval, natural language processing, (text) data mining and artificial intelligence. The central focus of his work is on the development of methods that scale to very large document collections and that use as few external resources as possible, in order to be applicable to documents of any type written in any language, from news articles to social networks, and from digitized manuscripts to digitally-born documents. Antoine Doucet holds a PhD in computer science from the University in Helsinki (Finland) since 2005, and a French research supervision habilitation (HDR) since 2012.