Building a conversational query system to navigate in a multi-modal corpus

Application to the exploration of a corpus in digital humanities

a PhD subject proposal

Supervisors: Jean Lieber and Emmanuel Nauer
LORIA UMR 7503 Université de Lorraine, CNRS, Inria
Team K (`k.loria.fr`)
`jean.lieber@loria.fr`, `emmanuel.nauer@loria.fr`

This call for applications to a PhD studentship is subject to funding approval. Please contact us as soon as possible with a motivated letter and a resume if you are interested.

Scientific context

Given a corpus of documents of significant size, its exploration can be assisted with a computer system that can find the relevant subset of documents, using exact or approximate search. The development of such a query engine is frequently based on the following assumptions:

(A1) The users of the system are familiar with the query language and of the vocabulary used in the application, which means either that they have enough skill with complex querying and a knowledge of this vocabulary, or that this query language is simplified to make it easier to access (e.g. consists of forms to be filled);

(A2) The users have at the start of the search a precise and intangible goal for this research;

(A3) The access to the document is based on a sole modality, e.g. according only to annotations, or according only to search in plain text.

The goal of this subject proposal is to build and study a querying system that overcomes the limitations involved by these assumptions.

Applicative context

The Henri Poincaré correspondence is a corpus of digital humanities that contains about 2000 letters sent or received by this famous mathematician and is maintained by the AHP-PReST laboratory.[1] These letters have been digitized, retranscribed in plain text, and semantically annotated in RDFS. The annotations of a letter describe general information about it (sender, receiver, writing date, etc.) as well as about its content (topics, links to other letters, etc.).

This corpus is accessible through a website[2] and a SPARQL query interface is used to interrogate it [1]. The difficulty of using SPARQL for users that are not computer scientists and the need to know the vocabulary used in the annotations led to several specific user interface development [2]. Moreover, classical querying is also based on the assumption that the user has a clear and formal idea of the query at the start of the process, whereas a user may have at the start only a general or vague idea that has to be made precise. This has led, in the last years, to the implementation of a conversational query system which organizes hierarchically the letters, using formal concept analysis [3] on letter properties, on the basis of the CreChainDo system [4]. The users can interact with this system to specify their needs on the basis of the clustering of properties, and thus

[1] `https://poincare.univ-lorraine.fr`

[2] `henripoincare.fr`

iteratively search the corpus to discover links between different elements in the letters. Therefore, this approach contributes to address the limitations involved by assumptions (A1) and (A2).

However, this system currently exploits only metadata embedded in semantic annotations and thus fail to address the issue linked with assumption (A3). The main goal of historians is to analyze the content of the letters to group them into thematic volumes. Five volumes have already been published (e.g. letters to mathematicians, letters to geodesists). This requires exploring the full text of the letters in parallel of acquiring knowledge about concepts (mathematical functions, institutions, people, etc.) relevant to a particular study.

Objectives of the PhD

The objective of the PhD is to design a conversational search system using several modalities in a complementary way. In particular, such a system has to be implemented and deployed to help historians in building analysis of the correspondence of Henri Poincaré on a given topic (e.g. Henri Poincaré and research in algebra or the family life of Henri Poincaré). The three modalities here are RDF annotations, plain texts, and also mathematical expressions that have been retranscribed in LaTeXformat.

Acquiring knowledge from full texts. A first work addresses knowledge engineering and document retrieval. Indeed, when historians have to explore letters on a given topic they have to search relevant documents addressing the topic and organize and structure their knowledge according to this topic. The use case is typically the following. The user will first search for documents with a general query looking for documents containing the word “mathematics” but such a general query will not return all documents about mathematics because the word “mathematics” does not appear in some letters dealing with mathematics. When the conversational system classifies the letters containing “mathematics”, related terms (e.g. differential equation, Navier-Stockes equations) will be presented to the user, and access to the full texts enables the user to also identify related terms. The objective is here to enrich the conversational tool to give historians functionalities for organizing the terminology around the searched topic, creating concepts, and structuring these concepts into ontologies (e.g. “Navier-Stockes equations” are “differential equations”) or through semantic relationships (“Navier-Stockes equations” describe “Newtonian fluid motion”). These new concepts and their organization can then be exploited by the conversational system to reorganize documents. For example, a concept about “mathematics for physics” is supposed to appear for letters about “Navier-Stockes equation”.

Similarity retrieval. Exploiting knowledge, built by the historian or coming from external resoources (the K team is currentlty working on a RDFS search engine, so collecting all possible knowledge RDF graphs on the web) allows to build a similarity retrieval. This similarity retrieval function can be implemented using the relation between concepts, taking into account some weightings depending on the type of relation.

Handling mathematical expressions. The letters from Henri Poincaré correspondence contain mathematical functions, in the transcription, encoded in LaTeX. The exploitation of these functions allows to identify the (not textual) presence of some concepts in the texts. For example, `\nabla` (∇) is a differential operator, which has to be linked to “differential equation”. Some mathematical expressions (e.g. equations) can be classified in an ontology of equations, but this recognition process supposes a flexibility in the recognition as two equations expressing the same knowledge can be expressed in various ways (using different unknowns, different symbols, ordering the terms differently etc.). For example, such a system should recognize that $x^2 + 3x - 5 = 0$ and $t^t + 3t = 5$ are two expressions for the same equation. For this purpose, integrating a term rewriting tool in the search engine could be a way to handle such situations. Another possibility would be to explore the application of principles and techniques of natural language processing to the mathematical language.

Expected results

Even if the thesis application domain is the correspondence of Henri Poincaré, the idea is to build a generic conversational system with specific functionalities to jointly structure knowledge to documents on a given topic using multiple modalities of the documents (annotations, plain text, mathematical expressions, etc.) and relations between these modalities.

References

[1] O. Bruneau, S. Garlatti, M. Guedj, S. Laubé, and J. Lieber. SemanticHPST: Applying Semantic Web Principles and Technologies to the History and Philosophy of Science and Technology. In Fabien Gandon, Antoine Zimmermann, Catherine Faron-Zucker, John Breslin, Serena Villata, and Christophe Guéret, editors, *The Semantic Web: ESWC 2015 Satellite Events* , volume 9341 of *Lecture Notes in Computer Science*, pages 416–427, Portoroz, Slovenia, May 2015. Springer International Publishing.

[2] Olivier Bruneau, Nicolas Lasolle, Jean Lieber, Emmanuel Nauer, Siyana Pavlova, and Laurent Rollet. Applying and Developing Semantic Web Technologies for Exploiting a Corpus in History of Science: the Case Study of the Henri Poincaré Correspondence. *Semantic Web – Interoperability, Usability, Applicability*, 12(2):359–378, 2021.

[3] Bernhard Ganter, Gerd Stumme, and Rudolf Wille, editors. *Formal Concept Analysis, Foundations and Applications*, volume 3626 of *Lecture Notes in Computer Science*. Springer, 2005.

[4] E. Nauer and Y. Toussaint. CreChainDo: an iterative and interactive Web information retrieval system based on lattices. *International Journal of General Systems*, 38, 2009.