Independent scholar

This short paper introduces LEAF (the Linked Editorial Academic Framework virtual research environment), an enhanced and expanded collaborative editorial platform that supports a variety of digital scholarly projects through a pipeline of integrated tools for collaborative production and publication of scholarly and documentary collections. Funded through the Canada Foundation for Innovation and the Andrew W. Mellon Foundation, LEAF aims to address the challenges that face many who undertake and maintain large-scale collaborative DH projects now: namely, the need to ensure that these projects can remain operational and available to editors and audiences over the long-haul. It is only by sharing physical, software, and human infrastructures across institutions that this can be accomplished. In so doing we can support scalability, interoperability, and preservation while allowing for dynamic, iterative, and collaborative editing, and therefore ensure that our materials, collections, and editions will remain viable and accessible. The LEAF team aims to do this by integrating best practices for text encoding, annotation, and metadata standards. This short paper will report on the development of LEAF and the functionalities that it will provide.

The implementation, and dissemination of LEAF is built upon a collaboration to extend the Canadian Writing Research Collaboratory (CWRC) built by the Universities of Alberta and Guelph (Susan Brown) with Bucknell University (Diane Jakacki), and Newcastle University (James Cummings) as founding partners. This work enhances CWRC's functionality through collaborative software development that will ultimately support multiple instances of the LEAF platform in Canada, the US, and the UK. At Bucknell, this work will inform the Liberal Arts Based Digital Edition Publishing Cooperative and the Bucknell Digital Press, funded by an Andrew W. Mellon Digital Publishing Cooperative Implementation grant that will support an expanding portfolio of peer-reviewed digital editions and edition clusters.

The LEAF platform combines hardware, software, and personnel. LEAF is being built on a solid foundation in terms of its data models, core functionality, and code management, so that it is positioned for extension and long-term sustainability. The platform is based on the Islandora 8 framework, which combines Drupal 8 with a Fedora 5 repository for long-term preservation. The LEAF repository will customize and enhance Islandora to enable digital humanities workflows and publication needs. Enhancements include an innovative web-based editing tool that allows users to employ TEI XML along with Web Annotation and IIIF standards-compatible Linked Open Data annotations that enhance discoverability and interoperability.

The founding LEAF institutions are collaborating to upgrade the existing CWRC environment and produce a fully modular platform that will also be hosted on Bucknell's servers, further tested at Newcastle University, and offered as containerized open-source code freely available for download and installation by other institutions. In particular, LEAF will facilitate the production and publication of dynamic digital scholarly editions and collections, offering multilingual transcription, translation, and image markup. Entirely browser-based, its functionality includes an in-browser XML markup editor, XML rendering tools, built-in text and data visualization tools including the Voyant Tools suite and its Dynamic Table of Contexts Browser. Overall the LEAF platform will provide a sophisticated interface for digital editions in which the XML markup is leveraged for navigation and active reading, and enhanced with Linked Open Data.

Document similarity and topic clues. A historiographical study case

Jolivet, Vincent
vincent.jolivet@chartes.psl.eu
École des chartes, France

Torres, Sergio
sergio.torres@chartes.psl.eu
École des chartes, France

Our University has recently published the longform abstracts of some 3000 institutional theses defended in history since 1849. The corpus is a rich documentary resource for historiographical studies. Unfortunately, there is no standard keyword indexing to browse this large collection and provide the reader with direct access to documents on the same subject. Such a functionality needs specific methods combining keywords, persons, places and in general intergroup patterns whose identification helps determine covered topics and related abstracts across more than a century. For this purpose, the proven clustering methods based on inter-document similarity are very effective, but in practice the interpretation of the similarity scores is difficult: a score describes how similar two documents are, but does not describe why they are similar. We have therefore experimented with methods combining document similarity and keyword extraction, so as to provide the researcher, in addition to a similarity score, with lexical clues facilitating the semantic interpretation of measured similarity.

In this presentation we present a pipeline leading with the extraction and formalization of indexing information in order to activate a document-similarity research engine, the evaluation of the scores obtained, as well as the benefits for information retrieval.

Methods

As our corpus is quite large, we preferred unsupervised approaches over supervised. The method is based on a semantic relatedness calculation using vectors, and the pipeline is composed of three steps.

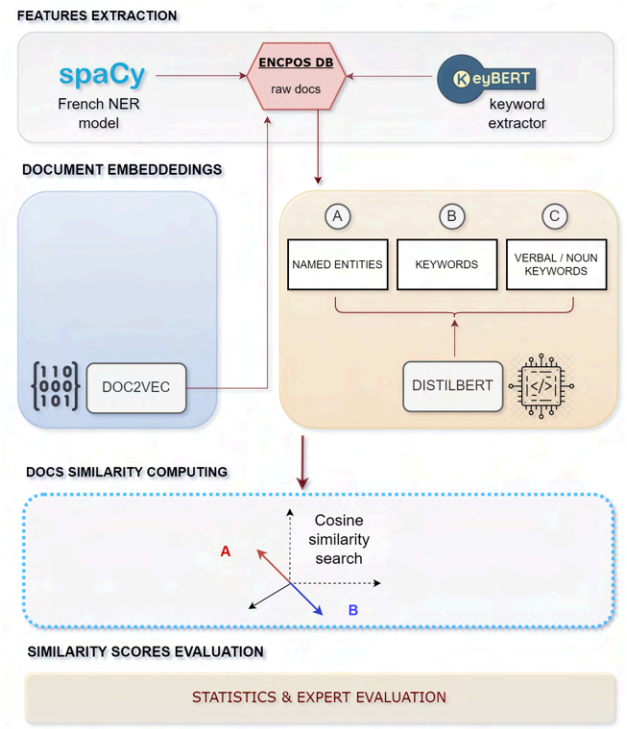

Figure 1:
Three steps for document similarity computing

1. We extract lexical and semantic features: (a) named entities (names, persons and organizations), using a French Spacy (Honnibal, M., Montani, I., 2017: 411) named entity extraction model based on CamemBert (Martin et al., 2019) language-model; and (b) keywords describing each abstract at a section level using KeyBert (Grootendorst, M., 2020), a keyword extractor based on the multilingual DistilBert (Sanh, V. et al., 2019) sentence embeddings library. To do so we apply embedding functions to our texts, mapping raw input data to low-dimensional vector representations. We then calculate the vector distance between the full-text embedding and candidate features embeddings to find the *top-k* candidates (the keywords) that are closest to the full text.

2. Our abstracts are then pre-processed into three versions containing: (a) their named entities, supposing that texts with the same entities are similar at a spatial and chronological level; (b) the keywords extracted during the first step at a paragraph-level and in so doing accounting for inflection variations such as tense and or stylistic elements; and (c) the only verbal and noun keywords, which keep only the phrasal root units to avoid lexical similarities and to summarize the text to its core components. Each one of these representations are later vectorized using Doc2vec (Le, Q., Mikolov, T., 2014: 1188), which generates context-independent embeddings (i.e, it collapses different word-meanings into a single vector), and DistilBert, which leverages Bert (Devlin, J. et al., 2018) to generate context-dependent sentence-level embeddings.

3. Finally, the cosine distance is calculated between the target-text and the database texts expressed as vectors to measure the document similarity score. This is obviously useful, as the score helps to identify related abstracts. Nevertheless, the similarity score doesn't provide all the necessary clues to determine the real performance of the given ranking of documents, and therefore must be evaluated further.

Evaluation

To estimate the relevance of the keywords embeddings method we calculate the similarity scores for all the documents pairs in an all vs all scheme using this method vs the Doc2vec and the Distilbert embeddings methods applied on full texts.

	key-distilbert				Distilbert			
	mean	median	var	stdev	mean	median	var	stdev
doc2vec	0.18	0.15	0.02	0.13	0.16	0.14	0.01	0.12
Distilbert	0.09	0.08	0.01	0.07	-	-	-	-

Figure 2:
Cosine distance statistics in an all vs all (≈ 9M matrix) scheme comparing three embedding methods: our key-distilbert method using the keywords (81 words on average) vs doc2vec and distilbert using the entire document (2013 words on average). var: variance, stdev: standard deviation

The statistics (median, mean, variance, standard deviation) indicate that in general the keyword approach, using 25x less amount of text, generates a similarity score very close (± 0.08 - 0.15) to the ones obtained using the full text (see Figure 2) on both methods, also proposing a time calculation 5x faster. This confirmation opens perspectives for the processing of very large corpora insofar as for close similarity scores.

Our method has another advantage, which was our initial goal: we provide for each pair of abstracts, in addition to a similarity score, a lexicon of the shared features (keywords and/or named entities) that we believe is useful for interpreting the score. For example, for two given abstracts, their high similarity score of 0.83 is enhanced by the lexicon of shared keywords "library", "manuscript", and "abbey"; these shared features seem to be clues for the semantic interpretation of the similarity. Not all cases are so obvious and the question of the relevance of these lexical clues for interpreting thematic similarities between documents is strongly raised.

Additionally, to evaluate the similarity scores as well as the relevance of the features lexicons, we submitted similar documents to experts, asking them to assess their degree of similarity and to rate the relevance of these shared lexicons to describe the link between the documents. This evaluation is ongoing.

Conclusion

Our initial objective was to obtain a reliable measure of the thematic similarity of the abstracts, by providing lexical clues useful for the semantic interpretation of the scores. We are convinced of the effectiveness of the method for the exploration of serial corpora such as cartularies or correspondences. Finally, the data produced are valuable for historiographical study, making it possible to quantify the most and least studied subjects diachronically, in particular through the analysis of the most associated keyword groupings.

Bibliography

Honnibal, M. and Montani, I. (2017). Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Unpublished software application*.

Martin, L. et al. (2019). Camembert: a tasty french language model. arXiv preprint. arXiv:1911.03894.

Grootendorst, M. (2020). keyBERT: Minimal keyword extraction with bert.

Sanh, V. et al. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Le Quoc and Mikolov, T. (2014). Distributed representations of sentences and documents. *International conference on machine learning*. PMLR, 2014. p. 1188-1196.

Devlin, J. et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Information Platform for Linked Data of Regional Historical Materials and its Agent Name Finding Process

Kameda, Akihiro
cm3ak@outlook.com
National Museum of Japanese History, Japan

Goto, Makoto
m-goto@rekihaku.ac.jp
National Museum of Japanese History, Japan

Overview

This presentation describes the construction of a system and the analysis and maintenance of data for the advanced use of the inventory of regional historical resources, especially using interactive annotation of agent names. We are driving a project for the inheritance and preservation of regional historical resources. In order to achieve the objectives of this project, we have developed a data infrastructure for advanced use of the inventories of regional historical resources. In particular, we aimed to create a system in which computers help people to discover information, rather than the conventional system in which people search and browse directly. Specifically, we resolved orthographical variants and integrated values, constructed identifiers and URIs, and described the provenance and components of resources. As a result, we were able to provide a Linked Data for regional historical resources, and we found the design of appropriate information infrastructure and its data generation process. In regional historical resources, there are many people and companies described. Some people can be associated with clans and positioned in family tree diagrams, other people are nameless and their detailed profiles are unknown. Those agent names and relationships among them in the archive of regional historical resources characterize the archive itself. If we only focus on some famous people already known in other documents, it is efficient to bring the dictionary of those names including alternative names and find those names in the archive. However, some names which are quite frequently used in the archive and not so much known in other documents are worth being analyzed and described. So, we extract the candidate names from the archive, list up the information of famous people from external resources