81. Creating a Knowledge Base of Research Methods from Archaeology Publications

Vayianos Pertsas (Athens University Of Economics and Business)*; Nikolaos Kapralos (Athens University Of Economics and Business); Ioanna Ntountoudi (Athens University Of Economics and Business)

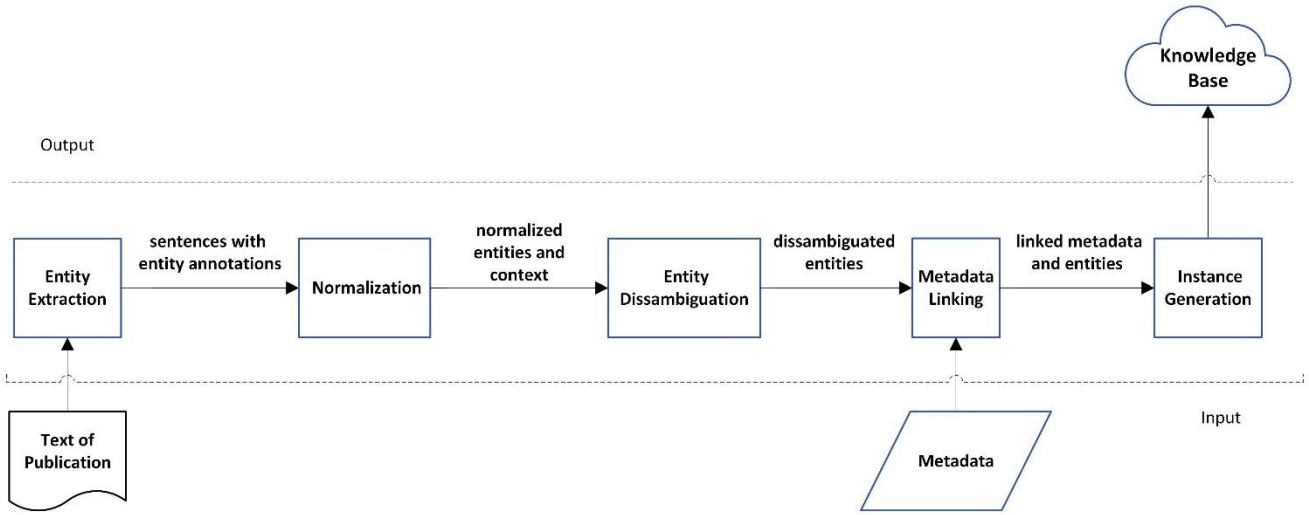
Access to knowledge captured in research publications constitutes a major information need of scholars across disciplines, a phenomenon dramatically exaggerated in the past decades due to the explosive growth rates of publications across scholarly domains. This situation calls for new strategic reading methods that transform the essence of knowledge encoded in textual form into structured formats like Knowledge Graphs (KG), thus changing the way researchers engage with literature [1]. This type of encoded information offered through Knowledge Bases (KB) can alleviate the task of maintaining a bird's-eye-view of research works across disciplines, something particularly useful for interdisciplinary fields like Digital Humanities, while keeping up to date with state-of-the-art methodologies in a specific domain such as Archaeology.

In this paper we present a digital workflow for creating such a KB of research methods in Archaeology, i.e. entities of variable length that appear in text with a proper name (e.g. "radiocarbon dating", "stable isotope analysis", etc.) and denote how research is conducted. Specifically, our workflow takes as input unstructured text from research articles' abstract and main body and through a series of steps that employ Deep Learning and LLM Prompting techniques in combination with inference rules: 1) identities and extracts textual spans denoting research methods; 2) normalizes them; 3) disambiguates them through Wikipedia and Wikidata; 4) links them with other entities extracted from publication metadata and 5) transforms the encoded knowledge into instances of a KB implemented in NEO4J . The following figure shows the modular architecture of the presented workflow. The entire process of KG creation is ontology driven, meaning that the definitions of the entity and relationship types that comprise the KB schema are provided from Scholarly Ontology (SO), a CIDOC-CRM compatible conceptual framework specifically designed for modeling and documenting scholarly work [2].

In the Entity Extraction step, we perform token classification using a pretrained transformer model (RoBERTa-base) for vector representation of text, in tandem with a transition-based parser for the NER task [3]. Both components are fine-tuned (transformer) and trained (NER) in a manually annotated dataset comprised of 10,000 sentences sampled from research articles from Humanities disciplines that were retrieved from JSTOR repository (years 2000-2021) and contained in total 5,450 textual spans labeled as Method. For the Normalization step we created a prompt template that uses as input the sentence context and the extracted label and prompts the LLM for a normalized label (i.e. expansion of acronyms, decomposition of composite phrases, etc.). For the Entity Disambiguation step, we employ the GENRE system that takes as input the normalized entity and context and yields the corresponding Wikipedia URL in the output. Then, from the related Wikidata page, we extract the method's description and alternate names. Evaluation of entity extraction and disambiguation modules was conducted using a dataset of 1,000 sentences sampled exclusively from 751 Archaeology papers that were manually annotated and disambiguated for the task at hand. The idea was to create a broad sample in order to test the ML models into as many writing styles as possible. Evaluation scores reached 82% and 81% (F1 measure), for the entity extraction and entity disambiguation respectively. In the Metadata Linking step we query ORCID through the provided API in order to link the authors of the

paper with their corresponding author ID when available, while mapping the rest of the information extracted from publication metadata of the article into the corresponding SO entities and relationships, namely providing instances for the SO classes: Person (authorID -retrieved from ORCID), Organization (author affiliations – retrieved from ORCID), Article (articleID from the metadata), Topic (author keywords) and Aggregation (Journal / Conference Proceedings where the article was published). Finally, the KG Creation step handles the creation of an instance of KB in NEO4J, as well as the indexing of various methods' properties utilizing NEO4J's indexing technology for faster retrieval. Specifically, we examined more than 50 different cypher queries that can be utilized using this KB schema, by measuring retrieval times as well as the lexical characteristics and the frequency of appearance of each search term inside them. Evaluation showed that the optimum indexing strategy -in order to save resources while maintaining efficiency and fast retrieval speeds- is to create a text index for "name" property and a full-text index for "description" property of Methods nodes respectively. In the output, each created instance is inserted / merged into the KB after a series of duplicate checks based on its unique identifiers.

Through the presented digital workflow, unstructured text from research articles can be transformed into a structured form and knowledge regarding research methods can be extracted, disambiguated and encoded into a Knowledge Base capable of answering semantically complex queries such as "Given a specific Method, show similar ones - based on its description- that have been mentioned in articles concerning Anthropology, Archaeology and Paleontology" or "Show the top N methods in Archaeology and Paleontology that are referenced in articles by scholars affiliated to a specific Organisation (e.g. Oxford)". Future work includes the expansion of the presented digital workflow with modules -specifically finetuned and tested in Archaeology papers- that extract more entities and relationships from SO (i.e. research activities, denoting the actual research events -such as an archaeological excavation- or steps thereof, research goals, denoting the activities' objectives and research findings, denoting their results). In addition, we plan to leverage the NEO4J integration with LLM frameworks such as Langchain in order to employ advanced retrieval techniques such as graph-based Retrieval Augmented Generation.



References:

[1] Renear, A. H., and Palmer, C. L. 2009. "Strategic Reading, Ontologies, and the Future of Scientific Publishing." Science 325 (5942): 828–832. https://doi.org/10.1126/science.1157784.

[2] Pertsas, V. and Constantopoulos, P. 2016. “Scholarly Ontology: modelling scholarly practices” International Journal on Digital Libraries 18 (3): 173–190. https://doi.org/10.1007/s00799-016-0169-3.

[3] Pertsas, V. and Constantopoulos, P. 2023. “Ontology-Driven Extraction of Contextualized Information from Research Publications.” In Proceedings 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. 108–118. https://doi.org/10.5220/0012254100003598."