Contribution ID: **42** Type: **Poster**

Ontology-Based Literature Classification and Knowledge Extraction in the domain of perovskite-based solar cell manufacturing

Tuesday 10 October 2023 14:30 (15 minutes)

The important increase in efficiency of perovskite-based solar cells (PSCs) in the last decade is a result of scientific work, which produced a huge quantity of literature and data-sets (between 2014 and 2022 almost 30,000 reports were published). The aim of this work is to elaborate an ontology which can primarly be used to classify literature paragraphs according to the subject discussed within. Furthermore, the annotation of data-sets using terms of the envisaged ontology can help making them more understandable, interoperable, and reusable.

The approach is to firstly develop an ontology with a team of domain experts and then elaborate knowledge extraction concepts from literature based on it.

From a technical point of view an ontology is a representational artifact, comprising a taxonomy as proper part, whose representations are intended to designate some combinations of universals, defined classes, and certain relations between them. A taxonomy on its role is a hierarchy definition between terms denoting types (or classes) linked by subtype relations. The knowledge extraction process consists of the following steps: **(1)** acquiring relevant text resources, **(2)** processing the text into individual terms (also known as tokenization), **(3)** document segmentation and paragraph classification, **(4)** recognizing tokens as classes of information, **(5)** entity relation extraction, and **(6)** named entity linking. The ontology being developed will facilitate some of the tasks in the knowledge extraction process (mainly entity recognition and disambiguation (4) and predicate determination (to build *(subject, predicate, object)*-triplets) by using link prediction during realization of (5)) by strengthening the context to an ontological commitment and by making the correspondence between *sign*, *concept*, and *thing* (or *symbol*, *thought or reference* and *referent*) more robust, especially when named entities and their corresponding classes in the ontology should be precisely recognised.

On the other hand both metadata and content of several scientific articles are represented as RDF triplets. Metadata of newly published scholarly papers are almost structured and based on unique identifiers (such as ORCID or ROR), which make the task of retrieving them much easier. But the knowledge acquisition from textual content reserve some problems and pitfalls during its transformation to RDF triplets. Using an ontology for this purpose helps overcoming ambiguity issues by applying scales of similarity between graphs defined through ontological facts, and ones extracted from texts by means of token-neighbourhood criteria.

The ontology being developed and the content, augmented with its metadata, are then represented in one property graph. In property graphs it is allowed a set of property–value pairs, called attributes and a label to be associated with nodes and edges, offering additional flexibility when modelling data.

The result is a knowledge graph (KG), a graph of data whose nodes represent entities of interest and whose edges represent potentially different relations between them. Both factual and ontological knowledge are existing side-by-side in this KG. In this data-space, classification using link prediction (CULP) transforms recognition problems to those of link prediction (where we try to find the link between an unlabeled node and the proper class node for it).

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

Ontology, Knowledge Extraction, Knowledge Graph, Perovskite

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary author: Mr KOUBAA, Mohamed Anis (Institute for Automation and applied Informatics)

Co-authors: SCHMIDT, Andreas (IT4EDM/IAI); STUCKY, Karl-Uwe (KIT); SUESS, Wolfgang (KIT)

Presenter: Mr KOUBAA, Mohamed Anis (Institute for Automation and applied Informatics)

Session Classification: Poster session

Track Classification: Poster session