1. TTILE

Semantic approaches for facilitating interoperability between data driven sources: A case study

2. A SHORT DESCRIPTION

2.1 Introduction

Harmonization and standardization of metadata is increasingly prioritized in Europe. In order to achieve the harmonization of data, a semantic layer is missing to enable better harmonization between this metadata, and to enrich these datasets with additional information. Sometimes, with metadata, fields are missing to express information beyond an application domain. Therefore, there is a need to add a new semantic layer, such as ontologies.

New Artificial Intelligence technologies, such as semantic technology, help us to address the issue of harmonization between data and additional information from datasets. The use of semantic technologies has gained importance in the creation and use of data exchange standards, as well as in problem-solving arising from heterogeneity and poor interoperability. The concept of ontology and knowledge graphs have attracted increasing attention in information science for their ability to achieve a shared knowledge representation. These technologies are key to delivering services and data in a standardized way. As a result, the use of a knowledge graph is proposed. They do not only introduce a shareable and reusable knowledge representation but can also add new knowledge about the domain.

This paper is organized as follows. Section 2 presents related work on knowledge graph construction and language models. Section 3 describes the purpose and motivation of our work through a use case. Semantic approaches for automatic enrichment of knowledge graphs are outlined in Section 4. Finally, Section 5 presents conclusions and highlights future lines of work.

2.2 Related work

2.2.1 Knowledge Graph

The knowledge graph has gained a lot of attention in recent years. In 2012, Google applied one to improve search engines capabilities and enhance the user's search quality and search experience. Then, the knowledge graph has been used in many fields of application[1]. The knowledge graph is graph-based knowledge representation and organization method, which uses a set of subject-predicate-object triplets to represent the various entities and their relationships in a domain. Most previous works tried to construct the knowledge graph manually and others automatically. However, manually requires enormous domain expert time and effort. From content unstructured is complex to construct a knowledge graph automatically. In recent years, thanks to the rapid progress of big data and natural language processing (NLP) technologies, automatically mining knowledge become a promising research challenge. In recent years, papers include named entity recognition (NER), entity normalization, relation extraction/ranking and graph embedding. In addition, research[2] explore the creation of knowledge graph from textual data by applying deep learning.

2.2.2 Language Models

Natural Language Processing (NLP) concerns the automatic generation and understanding of human languages. Language Models are the core foundations of NLP. They determine word

[1] Zhao et al., 2018. Architecture of Knowledge Graph Construction Techniques. Volume 118 No. 19 2018, 1869-1883.

[2] Wang et al, 2020. Language Models are Open Knowledge Graphs.

probability by analyzing text data. Language Models have evolved from simple frequency counts, such as bag-of-words, n-grams and term frequency-inverse document frequency (TF-IDF), to more advanced representations that use neural networks to learn the latent structure of language. In recent years, the transformer architecture had made significant breakthroughs in several natural language understanding tasks with the revolution of pre-trained language models such a BERT[3], GPT-3[4], RoBERTa[5], T5[6], and XLNet[7]. in which the context of the information is deemed.

2.3 Purpose and Motivation

In this section, we describe a case study that illustrates the motivation of our article, which is in the context of a European project called EuHubs4Data[8].

2.3.1 Case Study: EuHubs4Data

The importance and impact that data have on the European economy, industry and society is nowadays unquestionable. Data-driven innovation is a key driver of growth and jobs to boost European competitiveness. In this context, to bring it closer to the industry, the role of competence center and digital innovation hubs (DIH) offering support around big data services, products and application are crucial. With the aim of linking and establishing collaboration among existing initiatives in the domain of big data, the need arises to establish a data catalog and data management services that will federate and ease the exploitation of Big Data resources.

At present, we rely on a collection of 160 relevant data-driven sources and datasets brought by EUHubs4Data members to the project (especially from twelve DIHs) and leverage existing open data platforms and repositories at EU, country, and regional levels, including high-value datasets identified by the European Commission[9].

The collected data sources come from different regions, application domains, formats, types and have different metadata vocabulary standards, so it is necessary to devise a strategy or methodology for harmonizing data sources and datasets dynamically. Specifically for metadata, which is the most powerful enabler for facilitating interoperability between data driven sources. And unfortunately, there is a lack of knowledge about metadata vocabularies, clear and unique recommendations, and practices.

Any resource needs to be described with relevant metadata to facilitate its identification, selection, and reuse. Metadata is data about data, describing characteristics of groups of data and its suppliers. The concept of ontology has attracted increasing attention in information science because of its ability to achieve a representation of shared knowledge. Ontologies are

---

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[6] Roberts, A., Raffel, C., Lee, K., Matena, M., Shazeer, N., Liu, P. J., ... & Zhou, Y. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

[7] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

[8] EuHubs4Data project (https://euhubs4data.eu/).

[9] EU Commission, Report on high-value datasets from EU institutions, SC17DI06692, 2014.

of great importance in the creation and use of data exchange standards, as well as in problem-solving arising from heterogeneity and poor interoperability.

The use of ontologies allows for the definition of the structure of knowledge graphs, which link data to data and entities to entities in a meaningful way for people and interpretable by machines and systems. Knowledge graphs unify, integrate and make all content and resources easily accessible and searchable. The construction of a knowledge graph is proposed in order to achieve integration between data, metadata, and knowledge. This integration will contribute to better interoperability between data and metadata, as well as contribute to a better service offering.

2.4 Knowledge graph enrichment

How to automatically enrich knowledge graphs without human supervision is a major challenge. This section describes the IDS model as a foundation for the construction of the knowledge graph. Then, we present the outlines of a new semantic approach to data harmonization.

2.4.1 IDS Model

Different standards and vocabularies currently available for describing metadata (e.g., Dublin Core, VoID, OMV, DCAT, MOD) were analyzing with the aim of building a knowledge graph to achieve integration between data and metadata. After that, the IDS Model was selected as the core of the knowledge graph to be built and to model the harmonization of data in the context of the EUHubs4Data project. A brief description and the criteria used for its adoption are detailed in the following.

The International Data Spaces (IDS) Information model[10] uses relevant concepts from the International Data Spaces[11], extending concepts from external ontologies (DCAT, SKOS, FOAF, Owl-Time, PROV…) and allows the definition of metadata on resources, which in our case, are the datasets.

The IDS model contains many concepts, but we are going to focus on those involved in modelling the dataset as *DataResource*. This means that we will not only focus on the classes *Dataset*, *Digital Content* and *Described* and relationships, but also on concepts and relationships related to the data provider, the representation of the dataset (json, RDF, csv...), the dates of dataset, the language, and the data source. Used in other project tasks, as a current active asset, the IDS model was chosen as the core of the knowledge graph.

2.4.2 Metadata approach

Most of the data driven sources and datasets include their description, the region they cover, the URL (except for private sources), the URL of the RDF file (metadata), the format, the license and information related to data protection and privacy.

Figure 1 – Example of dataset metadata

A question that needs to be addressed is how to enrich the knowledge graph with metadata. Our metadata approach starts by searching all the datatype properties of the knowledge graph

[10] https://github.com/International-Data-Spaces-Association/InformationModel
[11] https://internationaldataspaces.org/

(in this case the IDS model) that are related to the *DataResource* class and its superclasses and that are most similar to a metadata property.

Then, the metadata needs to be mapped to the *DataResource* class. If the input metadata matches any property of the IDS model, the knowledge graph would be enriched with the provided property description information.

If the input metadata does not directly map to any property of the IDS model, semantic similarity between words is used to compute semantic similarity scores for concepts, words, and entities using knowledge-based semantic similarity metrics.

If the semantic similarity metric is above a threshold, the knowledge graph is enriched with the information provided by the metadata. However, if it is lower a threshold, at this moment, before discarding it, validating the associated comment or other rules that allow further enriching the knowledge graph is analyzed.

Furthermore, if the language does not appear among the input metadata provided for a dataset, it can be detected and the knowledge graph enriched with relationships.

2.4.3 Textual approach

Another approach to enrich the knowledge graph is to include additional information that can be discovered from the texts that appear in open access datasets.

More recently, the research community has started exploring how to leverage deep learning to build linguistic features that were traditionally built by humans. There are papers[12] in which the main idea behind is to minimize the involvement of humans in the process of creating knowledge graphs from textual data. The authors hypothesize that transformer-based models like BERT or GPT2/3 have the capacity to learn and store domain knowledge, which can be converted into a structured knowledge graph. A transformer is a deep learning model that adopts the mechanism of attention, differentially weighing the significance of each part of the input data. In this context.

We have explored research on the building of knowledge graph from text by leveraging transformer-based language models. The knowledge graph represents a collection of interlinked descriptions of entities — real-world objects and events, or abstract concepts. It allows us to model the real world into a graph, which can then be reasoned over to assert facts about the world. The steps followed to enrich the IDS Model with textual corpus can be seen in the figure below.

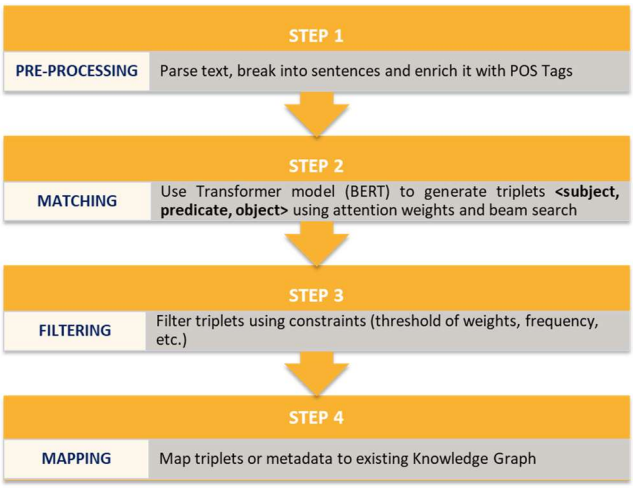

Figure 2 – Steps of textual semantic approach

[12] Language Models are Open Knowledge Graphs (https://arxiv.org/pdf/2010.11967.pdf)

The process starts with parsing, breaking into sentences and enriching the text with metadata such as POS tags (a process of pre-processing textual information). The next step is to use of attention mechanisms to infer candidate triplets in the text, taking advantage of transformer-based linguistic models such as BERT. The attention mechanism is also an attempt to implement the same action of selectively concentrating on a few relevant things, while ignoring others in deep neural networks. Attention takes two sentences, converts them into a matrix in which words from one sentence form the columns, and words from another sentence form the rows, and then it makes matches, identifying the relevant context. Attention weights, which models based on the pre-trained transformer have learnt during training, provide us with insights about the relationship between various terms in the text. These weights provide possible candidate facts (subject, predicate, object), which can be filtered using constraints (threshold, frequency, etc.) and then enrich the knowledge graph.

2.5 Conclusions and Future Work

In recent years, the exponential growth of data production has made us realise the importance and impact that data have in the economy, industry and society. In addition, social trends towards openness and sharing are drivers that are changing the global economy and society. How to provide methodologies or mechanisms for the harmonization of these data coming from different regions, a wide range of application areas (open, industrial, personal, research…), multiple sources with multiple data types, formats and licenses and using different metadata or schemas to describe data became a critical and key element to address interoperability issues. In this paper, we presented our work-in-progress related to semantic approaches for facilitating interoperability between data driven sources. In this context, the EuHubs4Data project, which links and establishes collaboration among existing initiatives in the domain of big data from different DIHs can leverage the capabilities of knowledge graphs based on the IDS model.

Our next steps involve the implementation of these enhanced semantic approaches within a single process, testing the generated algorithms with the real catalogue data provided in the EuHubs4Data project, providing an interface to enable visualization of the enriched knowledge graph, feeding the core knowledge graph with other vocabularies or ontologies to unify, integrate and provide all content and resources more easily. And analyze how to integrate with other components of the project and expose results. As future work, we would like to evaluate the quality of the generated knowledge graph with the suggestion of new metrics to foster higher quality and better outcomes in the analyzed context.

3. THE NAME, THE ROLE AND THE AFFILIATION OF THE AUTHORS

Paula Peña, Luis García, Rafael del Hoyo, María del Carmen Rodríguez-Hernández and Rosa Montañés
Technological Institute of Aragon (ITAINNOVA), María de Luna 7, Zaragoza, Spain
R&D technicians in the Artificial Intelligence, Big Data and Cognitive Systems group
ppena, lggarcia, rdelhoyo, mcrodriguez, rmontanes@itainnova.es