Towards Knowledge Graph Creation from Greek Governmental Documents

Amalia Georgoudi[1], Nikolaos Stylianou[1][0000−0002−6396−5374], Ioannis Konstantinidis[2][0000−0002−4400−7074], Georgios Meditskos[1][0000−0003−4242−5245], Thanassis Mavropoulos[3][0000−0002−7326−5910], Stefanos Vrochidis[3][0000−0002−2505−9178], and Nick Bassiliades[1][0000−0001−6035−1038]

[1] School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece
[2] School of Science and Technology, International Hellenic University, Thessaloniki, 57001, Greece
[3] Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, 57001, Greece

Abstract. Documents contain textual information, which is of the utmost importance for all the organizations. Document management systems have been used to store vast amounts of unstructured textual data described with minimal metadata, a method that has several limitations. In order to convert hidden knowledge into machine-readable data with rich connections, this paper presents work in progress on the development of the first end-to-end guided approach to construct a Knowledge Graph from Greek government documents from the Greek open government portal. The resulted Knowledge graph consists a proof-of-concept graph, that illustrates the beneficial semantic relationships between the textual data.

Keywords: Knowledge Graph Construction · RDF Triples · Ontologies · Natural Language Processing · Knowledge Representation · Government Documents.

1 Introduction

With the rise of the terms open government and open data, several public sector Document Management Systems are publishing their documents to the Web as open data to increase transparency and accessibility. Yet, the problem with those document management systems are that their information is not machine-readable and they do not support data interconnection and linking. This is the situation with the Greek web portal, known as DIAVGEIA (in English: Clarity), which publishes all public sector administrative decisions and acts, as required by the law, resulting in a massive and rapidly growing collection of over 43 million documents.

Knowledge Graphs (KGs) [1] are a way of representing data in graphs, through which it is possible to describe the significance of the data as well

as the relationships between them. In order to utilize the advantages of using a KG, this paper has focused on transforming the retrieved information from the governmental documents in DIAVGEIA to Resource Description Framework (RDF) [2] triples in order to construct a KG.

In this paper, we present our findings, challenges faced and approaches used, as part of an ongoing work, focusing on guided triple exaction from Greek government documents.By organizing the retrieved information in a Knowledge Graph, semantic relationships and links interconnect the data, giving them a machine-readable form, which easily could lead to drawing conclusions and extract new information. Yet, Greek is a low-resource language with very limited available NLP tools and many open challenges. Our approach addresses these issues and constructs a Knowledge Graph, based on a widely used ontology schema, populated with the extracted triples.

The remainder of the paper is organized as follows. Section 2 presents the related work. In Section 3 the proposed method is introduced, while in Section 4 the resulted proof-of-concept KG is presented. Section 5 concludes and gives directions for future work.

2 Related Work

Our work was based on a previous study [9], in which the authors proposed a theoretical framework for KG construction in DIAVGEIA. In this paper, the problem will be analyzed in a more technical way, by taking into account the technical difficulties and the limited number of available NLP tools for the Greek language.

A very limited number of works have focused on end-to-end knowledge graph construction from a document repository. The majority of the studies concentrate on particular subtasks, such as entity recognition, entity disambiguation, entity linking, and relation extraction. T2KG [3] presents a hybrid method for mapping predicates in an existing KG that combines a rule-based approach with a similarity-based approach. Other methods approach Knowledge Graph Construction as a multi-label sequence labeling task and uses a deep learning neural network architecture to jointly learn to produce and extract triples from text ([4]).

Knowledge graph construction is a task that has been applied in text with different types of domains ([6]) and languages ([7]). Regarding the Greek language, there is no other work for Knowledge Graph construction from Greek unstructured documents, except [8]. In [8] they proposed to utilize transformer models for machine translation from Greek to English and backwards, in order to and apply existing triple extraction tools for English texts. Yet, this approach do not achieve satisfactory results with the domain-specific vocabulary of government documents.

3 Proposed Method

The proposed approach uses a guided triple extraction algorithm, that searches for pre-determined triples and relations. After a systematic analysis of the documents by domain experts, the approach is focused only on obtaining triples that contain the most important information.

The proposed guided triple extraction method consists of three phases. Figure 1 presents the architecture of the proposed approach. In the first phase, the PDF file is parsed into different machine-readable formats (text, HTML) in order to apply NLP techniques to extract the important information. In the second phase, a reusable ontology is created based on a widely-used schema (Party in Role), while in the third phase, the extracted information is transformed into RDF triples to construct the Knowledge Graph.

Information Extraction. In the first phase, our method focuses on extracting the important information from the documents.To extract those triples, first, the PDF was parsed in a machine-readable form. More specifically, during this phase, the PDF was initially parsed into a textual form. From the text was possible to extract information such as the unique ADA number of the document, which is a unique Internet Uploading Number that every uploaded document in Diavgeia has, using regular expressions. Then the PDF was parsed also into HTML, in order to obtain additional information, such as the location or the number of words of a part of the PDF.

The next step was to detect the main body of the document. The main text is usually located in document's text body and has the highest number of words, compared to other parts of texts. As already described, the documents

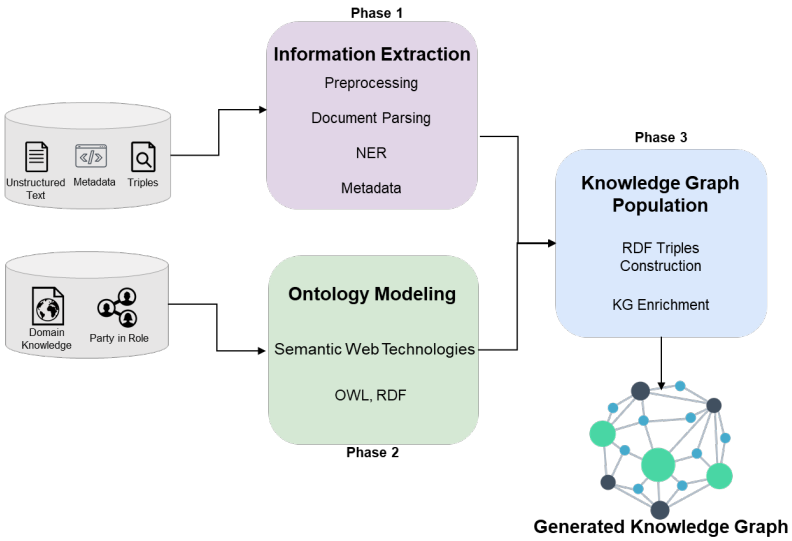

Fig. 1. Proposed Approach Architecture Overview

in DIAVGEIA usually contain a lot of noise. There are many numbers, symbols, words' initials, abbreviations, multiple tabs, etc. Furthermore, text inside parenthesis and quotation marks adds extra "noise" to text data, since they usually contain alphanumeric numbers of laws and decisions. For those reasons, a pipeline for text cleaning and normalization needed to be applied. As such, the text inside parentheses, quotation marks, numbers, punctuation, and multiple spaces were deleted so as to facilitate parsing the content. The cleaned text was then machine-readable and easier to parse. A Named Entity Recognition tool was then utilized to find the name of the individual who was appointed to the position.

Structural information was taken into account, i.e., the position of the text in the PDF. Finally, the GreekBERT-based NER tagger [11] was used to locate and classify named entities, such as the names of the signer and the organization. Based on preliminary experiments, on benchmark datasets and intrinsically on documents originating from DIAVGEIA, we found GreekBERT [10] to have the best performance in Greek language tasks

Ontology Modelling. During the second phase, a readable, scalable and reusable schema that describes the documents and the various parties (person, organization), was created. The ontology schema was based on the logic of a "Party in Role", which is a very common schema used by large organizations to describe people in different roles (someone may now be appointed in a position but later in another document, he could be the signatory). This pattern is also used in very large ontologies, such as FIBO (Financial Industry Business Ontology)[12]. Figure 2 shows the ontology's class hierarchy.

The class Document describes the Greek government documents from DIAVGEIA. The class Party represents all the participants/entities and has two subclasses: Person and Organization. The class Party_in_Role represents a participant who has a specific role. The class Role describes the roles that a person or an organization could have. In this case, the possible roles are the publisher, the signer and the assigned in. These roles are declared as instances of the class Role.

Knowledge Graph Population. The next step was to transform the extracted triples and their relationships into RDF triples, to populate the Knowl-

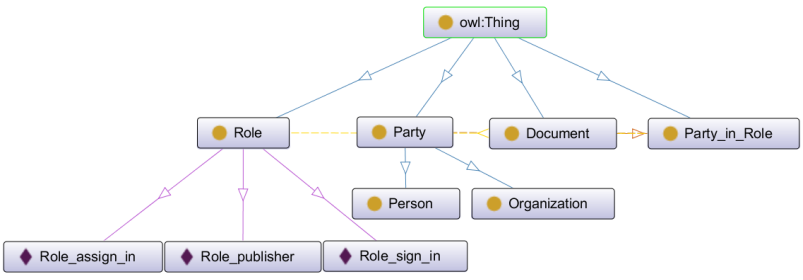

Fig. 2. Ontology: Party in Role

edge Graph.RDF stores information in the form of semantic triples. Each triple consists of a subject, a predicate, and an object. Each element in the triple is denoted by a unique Web URI. The resulted triples were parsed into an RDF file following the Turtle (Terse RDF Triple Language) format [13].

4 Produced Knowledge Graph

The produced Knowledge graph consists a proof-of-concept graph, that converts the textual data into machine-readable data and illustrates the beneficial semantic relationships between them.

Currently, there isn't any annotated dataset that could be used to test and validate the proposed approach. For this reason, the approach was tested on a manually annotated dataset consisting of 30 documents, created by domain experts. Table 1 illustrates the accuracy scores for the extracted information. The ADA number has the highest accuracy score, while the organization has the lowest. This was anticipated since the unique ADA number has a uniform format and a specific position in a document. On the other hand, extracting the organization from a document could be very challenging since sometimes it can be represented in different ways, such as pictures or logos. The results were also evaluated by humans, working in the public sector, who were very satisfied with the performance of the method.

5 Discussion and Conclusion

This paper presents an end-to-end guided approach to construct a Knowledge Graph from the Greek government documents, by extracting specific triples from the hidden textual data. Organizing and semantically linking the textual information in a Knowledge graph, which is included among the state-of-the-art technologies and has been acknowledged as the most efficient way of organizing and describing data as well as retrieving information, will enable and substantially improved semantically-based searching, transparency, and reusability.

Some limitations of our work that could encourage for future research are that we focused only on a specific type of decision while our approach only works for documents that follow a widely-accepted format. In future work, we plan to focus more on the triple extraction task for the Greek language and scale

Table 1. Accuracy scores for the retrieved information.

Extracted Information	Accuracy
Organization	0.70
ADA	1.00
Referred ADA	0.80
Appointed Person	0.85
Signer	0.80

our method to larger datasets, including more types of decisions and document formats. Yet, there is a lot of room for contributions regarding NLP tools for low resource languages, such as Greek. Having at our disposal more NLP tools would benefit the Knowledge Graph construction task.

Acknowledgements The research for this paper is partially funded by the Horizon Europe project ENCRYPT (Grant Agreement no. 101070670).

References

1. Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J. & Wahler, A. Introduction: what is a knowledge graph?. *Knowledge Graphs*. pp. 1-10 (2020)
2. Miller, E. An introduction to the resource description framework. *Journal Of Library Administration*. **34**, 245-255 (2001)
3. Kertkeidkachorn, N. & Ichise, R. T2kg: An end-to-end system for creating knowledge graph from unstructured text. *Workshops At The Thirty-First AAAI Conference On Artificial Intelligence*. (2017)
4. Stewart, M. & Liu, W. Seq2KG: An End-to-End Neural Model for Domain Agnostic Knowledge Graph (not Text Graph) Construction from Text. *International Conference On Principles Of Knowledge Representation And Reasoning*. **17**, 748-757 (2020)
5. Clancy, R., Ilyas, I., Lin, J. & Cheriton, D. Knowledge Graph Construction from Unstructured Text with Applications to Fact Verification and Beyond. *Second Workshop On Fact Extraction And VERification, Hong Kong, China*. pp. 3-7 (2019)
6. Rossanez, A., Dos Reis, J., Silva Torres, R. & Ribaupierre, H. KGen: a knowledge graph generator from biomedical scientific literature. *BMC Medical Informatics And Decision Making*. **20**, 1-24 (2020)
7. To, H. & Do, P. Extracting triples from vietnamese text to create knowledge graph. *12th International Conference On Knowledge And Systems Engineering (KSE)*. pp. 219-223 (2020)
8. Papadopoulos, D., Papadakis, N. & Matsatsinis, N. PENELOPIE: Enabling open information extraction for the greek language through machine translation. *ArXiv Preprint ArXiv:2103.15075*. (2021)
9. Stylianou, N., Vlachava, D., Konstantinidis, I., Bassiliades, N. & Peristeras, V. Doc2KG: Transforming Document Repositories to Knowledge Graphs. *International Journal On Semantic Web And Information Systems (IJSWIS)*. **18**, 1-20 (2022)
10. Koutsikakis, J., Chalkidis, I., Malakasiotis, P. & Androutsopoulos, I. Greek-bert: The greeks visiting sesame street. *11th Hellenic Conference On Artificial Intelligence*. pp. 110-117 (2020)
11. Smyrnioudis, N. & Koutsikakis, J. A Transformer-based Natural Language Processing Toolkit for Greek–Named Entity Recognition and Multi-task Learning. (2021)
12. Bennett, M. The financial industry business ontology: Best practice for big data. *Journal Of Banking Regulation*. **14**, 255-268 (2013)
13. Consortium, W. & Others RDF 1.1 Turtle: terse RDF triple language. (World Wide Web Consortium,2014)