Semantic Interoperability Issues in the Master's Curriculum in Artificial Intelligence at Sofia University

Maria Nisheva-Pavlova[0000-0002-9917-9535]

Faculty of Mathematics and Informatica, Sofia University St. Kliment Ohridski

marian@fmi.uni-sofia.bg

Abstract. In today's world of digital transformation and big data the issues related to the semantic interoperability of different information systems that support the decision making activities in the same or similar areas are becoming increasingly important. It is therefore natural for such issues to be in the focus of the education at master's level in a variety of professional fields. The paper discusses the experience in this regard of the Master's program in Artificial Intelligence at the Faculty of Mathematics and Informatics at Sofia University, focusing on some good examples of student projects.

Keywords: Semantic Interoperability, Ontology, Ontology Matching, Semantic Enrichment, Semantic Search.

1 Introduction

In recent years, information systems have to process huge amounts of heterogeneous data coming from different sources and in various formats. The work with big heterogeneous datasets makes it necessary to use proper methods and tools for data integration and achievement of semantic interoperability of the respective software systems.

The requirement for semantic interoperability of two information systems supposes that each of them will understand the semantics of the information sent or requested by the other, as well as the semantics of its information sources. Currently ontologies underlie the only widely accepted paradigm for representing and managing open knowledge that can be shared and reused in a way that allows automatic interpretation and inference.

Therefore, the study of ontology design methodologies and formal methods for ontology matching as well as acquiring practical skills for using ontologies in building different types of intelligent software systems is one of the important goals of the university education at master's level in the field of Artificial Intelligence.

2 Overview of the Master's program in AI

The master's program in Artificial Intelligence (AI) has been operating successfully at the Faculty of Mathematics and Informatics at Sofia University for nearly 20 years. Its educational objectives include mastering of deep theoretical knowledge in the classical and some modern areas of Artificial Intelligence and acquisition of various practical skills needed for the application of AI methods and techniques in a wide range of fields of Informatics and Information Technologies. The curriculum includes courses in fundamentals of Artificial Intelligence, knowledge modeling and design of knowledge bases, machine learning (in particular deep learning), information retrieval, data mining and knowledge discovery in large datasets, natural language processing, image processing and pattern recognition, embedded and autonomous systems, neural networks and genetic algorithms, robot control, semantic technologies, recommender systems, legal and ethical aspects of the development and use of AI systems, etc.

The successful graduates of the master's program in AI are able to apply their knowledge and skills in research and educational organizations, as well as in leading software companies in the development of, for example:

- software for data analysis and knowledge discovery in big data;
- software for semantic web and semantic network services;
- intelligent search engines;
- intelligent user interfaces;
- expert systems, recommender systems, intelligent virtual assistants, intelligent learning environments and other types of knowledge-based software systems;
- smart databases;
- image processing and image recognition tools;
- different types of intelligent embedded systems: intelligent robots, smart home systems, etc.

The education in the master's program follows the good methodological practices of combining the accumulation of abstract theoretical knowledge through lectures and providing various opportunities for its understanding and acquiring skills for its application in real problem situations through workshops, homework assignments and especially in the development of appropriate course projects, many of which act as bridges between different subjects.

3 Theoretical aspects – knowledge modeling and reasoning

The curriculum of the master's program includes two compulsory courses, whose curricula consistently cover the theoretical foundations and some technological issues of semantic interoperability.

The *Knowledge Representation and Engineering* course has mostly abstract content and introduces the fundamental principles of functioning and the modern methods for creation of knowledge-based systems (KBS). Following the methodology proposed by Brachman and Levesque in [1], the course introduces the basic principles of functioning and some advanced methods for design and implementation of KBS. Special attention is paid to the problems of domain analysis and the conceptualization of domain knowledge. The most important theoretical and practical aspects of a set of classical and modern methods for knowledge representation and reasoning are discussed. Students who have successfully passed the course in Knowledge Representation and Engineering are expected to be able to analyze and construct conceptual models of knowledge and to design KBS aimed to solve complex tasks with various characteristics. The course syllabus covers the following topics:

- Key concepts: knowledge, knowledge representation and reasoning. Knowledge-based systems. Knowledge engineering;
- The language of first-order logic (FOL). Syntax, semantics and pragmatics of FOL;
- Resolution. Reasoning with Horn clauses;
- Production rule systems. RETE algorithm;
- Object-oriented knowledge representation. Frames;
- Structured description of knowledge. Computing entailments. Taxonomies and classification;
- Inheritance. Strict inheritance. Strategies for defeasible inheritance;
- Default reasoning. Closed-world reasoning. Circumscription;
- Knowledge representation and reasoning with KRL;
- Concepts and language tools for describing information resources with RDF/RDFS;
- Ontologies – definition, classification, basic characteristics and requirements, applications. Concepts and language tools for describing ontologies with OWL.

As basic readings we recommend the classic textbooks [1 – 3] as well as the W3C standard recommendations [4 – 6].

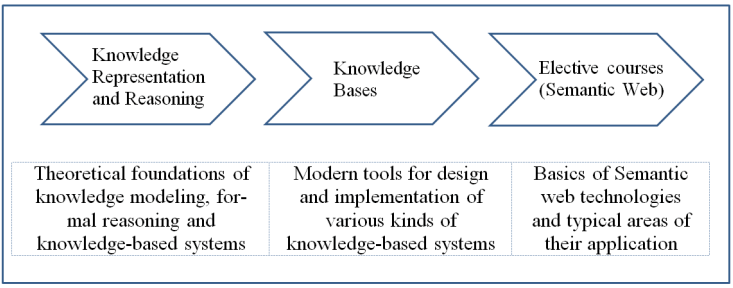

Fig. 1. Courses in the Master's in AI curriculum providing semantic interoperability knowledge and technical skills.

The *Knowledge Bases* course is the second one in the series shown in Fig. 1. It aims to acquaint students with the current state of research and practical developments in the field of knowledge bases, focusing primarily on the study of modern tools for creating knowledge bases with different characteristics and their application in the development of various types of KBS. Issues related to the basic principles and technologies of the Semantic Web, the semantic interoperability of information systems, the creation of semantic digital libraries, etc. are also studied. Here are some of the main topics covered by the course syllabus in the context of semantic interoperability:

- Semantic web and semantic technologies. Language standards for the Semantic web;
- Ontology engineering methodologies. Ontology mapping and merging;
- Cyc's knowledge base – the world's broadest and deepest commonsense knowledge base. Cyc's inference engines;
- Semantic databases. Tools for creating and using semantic databases;
- Semantic annotation. Semantic enhancement. Semantic search;
- Semantic digital libraries.

The knowledge and skills gained through these compulsory courses are upgraded by a number of elective courses, among which the most important in terms of the topic discussed is the Semantic Web one.

The Semantic Web course presents the basics of semantic technologies and the work with RDF data and linked open data. The main types of ontologies and examples of their application are discussed. Methods for implementation and work with semantic knowledge graphs and their applications are presented. The course has been taught by experts from the recognized leader in enterprise knowledge graph technology and semantic database engines Ontotext[1] and covers several specific topics including

[1] https://www.ontotext.com/

- Knowledge graphs;
- RDF data model and its serialization formats;
- Semantic integration of heterogeneous data;
- Linked open data;
- Data visualization.

The issues related to the semantic interoperability and the reusability of knowledge and data, in particular in unforeseen problem situations, connect these courses and form one of the thematic areas of study in the master's program in AI.

4 Practical aspects – course projects

A significant part of the students' independent work, aimed at understanding and mastering the abstract theory, is related to the preparation of homework assignments and the development of course projects.

Most homework assignments and course projects have generic topics and formulations that may be specified by the individual student depending on his/her interests and preferences.

Here are several examples in this regard.

Example 1

Design a knowledge base for a subject area of your choice. The concepts and their properties (roles) should be described in terms of description logic (the DL language [2]).

Based on these descriptions, build a corresponding ontology implemented with the means of Protégé/OWL [7] (concepts ↔ classes, roles ↔ properties, constants ↔ individuals/instances).

The knowledge base should include both atomic and different types of non-atomic concepts (constructed by the operators EXISTS, FILLS, ALL, AND). Provide appropriate statements of the three main types: $d \sqsubseteq e$, $d \doteq e$, $c \rightarrow e$.

Quantitative characteristics:

- number of concepts (classes): at least 20,
- number of constants (individuals/instances): at least 10,
- roles (properties): at least 10. Include properties with different characteristics (inverse, functional, transitive) and with appropriate and diverse domains and ranges.

Describe appropriate examples for automatic reasoning (using a reasoner of your choice) on the knowledge base – at least one inference of the type KB $\models (c \rightarrow e)$ and at least one inference of the type KB $\models (d \sqsubseteq e)$.

Describe at least one example of classification of the knowledge base. The same example should be illustrated with Protégé/OWL.

Example 2

Based on a series of paragraphs (one or more) in an article of your choice in Wikipedia, create an ontology that presents the concepts, objects, and relationships described in the text. Implement the ontology in Protégé/OWL or Apache Jena/OWL[2] and check its consistency.

The paragraphs in the article should be selected so that the ontology contains both primitive and defined classes.

Describe an appropriate example for performing automatic reasoning based on the created ontology.

Example 3

The project is aimed at in-memory implementation of the Web Annotation Data Model [8]. The task has a general formulation as follows.

Write a program that implements semantic annotation of text with support for the main elements of the annotation model (body, target, selector – for example text position selector).

Your program should read a short text (no longer than two paragraphs) and annotate it based on a publicly available ontology of your choice, tailored to the content of the text. The result of the work of the program should include finding the location in the text of at least three previously known (explicitly stated) concepts from the ontology.

For the implementation of the project students may use technology of their choice. A preferred option is the choice of the DBpedia ontology[3], programming in Java and presentation of the annotation in an open format, which facilitates its use for various purposes – for example, presentation based on JSON-LD [9]. Thus, in the process of developing their course project, students learn and master the work with new technologies and software platforms.

Usually many students take the opportunity to work on course projects on topics suggested by themselves (and agreed with the professor) and often such course projects grow into valuable master's theses.

5 Research and development – good practices of master's theses

Traditionally, graduates of our master's program in AI develop excellent diploma projects in modern and complex areas. The results achieved in most of them have been published by graduates and their supervisors in authoritative specialized scientific journals. In most cases, reusability and semantic interoperability with other systems are part of the characteristics of the developed software products.

As an example of good practice in this regard, we can consider the master's thesis entitled "Virtual health assistant" [10], defended in March 2021.

---

[2] https://jena.apache.org/documentation/inference/#owl

[3] https://wiki.dbpedia.org/

The aim of the thesis is to create a web-based virtual medical assistant that provides users with fast and convenient health information related to the symptoms and treatment of various socially significant diseases. For this purpose, the healthcare assistant uses an appropriate knowledge base. It supports functionality for automatic collection of data from verified sources on the Internet, processing this data, and building and extending a knowledge graph, which is the main component of the knowledge base.

The health assistant provides a user-friendly interface and receives as input a set of symptoms related to the condition and sufferings of the user. As a result, the assistant generates an appropriate answer, indicating probable diagnoses and detailed information about each of them, including a description of the disease, its synonyms and symptoms, as well as medications that help to treat it.

In addition to the information provided, the healthcare assistant asks questions about more symptoms that the user may have missed. If the user indicates other symptoms, the result of the assistant's work can be updated. All components of the health assistant can be easily expanded with additional functionalities.

Another good example of a master's thesis that successfully addresses most issues of semantic interoperability is the one on "Intelligent system for answering specialized questions about COVID-19". It is aimed at development of a question answering system based on information retrieval, natural language processing and text mining techniques. For the implementation of the system and the conduct of the planned experiments with it, the freely available COVID-19 Open Research Dataset (CORD-19) presented in one of the Kaggle competitions[4] has been used. CORD-19 is prepared as a reliable set of more than 500,000 scholarly resources about COVID-19, SARS-CoV-2 and related coronaviruses in order to assist the medical community in preparing answers to as many high-priority questions related to COVID-19 as possible.

After the necessary data processing, the system uses BERT [11] – a pre-trained model for recognizing the context of the words in the question and the possible answers. BERT uses neural networks to find the most accurate answer to a given question.

6 Conclusion

The analysis of the experience of the master's program in Artificial Intelligence at Sofia University for nearly 20 years shows that the chosen approach of combining lecture courses, providing deep theoretical knowledge at an abstract level, with the challenge for students to acquire modern technological skills in the process of developing proper course and diploma projects, gives very good results in the education of specialists in AI at master's level. It is particularly suitable for

[4] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

training in a number of areas that underlie the acquisition of knowledge and skills to create semantically interoperable information systems that can work flexibly with large heterogeneous datasets.

7 Acknowledgements

The presented work has been supported by Project BG05M2P001-1.001-0004 "Universities for Science, Informatics and Technologies in the e-Society (UNITe)" funded by Operational Program "Science and Education for Smart Growth" co-funded by European Regional Development Fund.

References

1. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach (3rd ed.). Pearson Education Ltd. (2010).
2. Brachman, R., Levesque, H.: Knowledge Representation and Reasoning. Elsevier (2004).
3. Brachman, R., Levesque, H.: Readings in Knowledge Representation. Morgan Kaufmann (1985).
4. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation 25 February 2014. http://www.w3.org/TR/rdf11-concepts/, last accessed 2021/03/30.
5. RDF Schema 1.1. W3C Recommendation 25 February 2014. http://www.w3.org/TR/rdf-schema, last accessed 2021/03/30.
6. OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation 11 December 2012. http://www.w3.org/TR/owl2-overview/, last accessed 2021/03/30.
7. Horridge, M., Brandt, S.: A Practical Guide to Building OWL Ontologies Using Protégé 4 and CO-ODE Tools, Edition 1.3. University of Manchester (2011). http://owl.cs.manchester.ac.uk/research/co-ode, last accessed 2021/03/30.
8. Sanderson, R., Ciccarese, P., Young, B.: Web Annotation Data Model (W3C Recommendation 23 February 2017). https://www.w3.org/TR/annotation-model/, last accessed 2021/03/30.
9. JSON for Linking Data. https://json-ld.org/, last accessed 2021/03/30.
10. Tsanova, R.: Virtual Health Assistant. Master Thesis. Faculty of Mathematics and Informatics, Sofia University St. Kliment Ohridski (2021).
11. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 4171–4186. Association for Computational Linguistics (2019). https://www.aclweb.org/anthology/N19-1423.pdf, last accessed 2021/03/30.