IARIA

SEMAPRO 2019

The Thirteenth International Conference on Advances in Semantic Processing

ISBN: 978-1-61208-738-2

September 22 - 26, 2019

Porto, Portugal

SEMAPRO 2019 Editors

Tim vor der Brück, Lucerne University of Applied Sciences and Arts, Switzerland

Efstratios Kontopoulos, Center for Research & Technology Hellas (CERTH), Greece

SEMAPRO 2019

Forward

The Thirteenth International Conference on Advances in Semantic Processing (SEMAPRO 2019), held between September 22-26, 2019 in Porto, Portugal, continued a series of events that were initiated considering the complexity of understanding and processing information. Semantic processing considers contextual dependencies and adds to the individually acquired knowledge emergent properties and understanding. Hardware and software support and platforms were developed for semantically enhanced information retrieval and interpretation. Searching for video, voice and speech [VVS] raises additional problems to specialized engines with respect to text search. Contextual searching and special patterns-based techniques are current solutions.

With the progress on ontology, web services, semantic social media, semantic web, deep web search /deep semantic web/, semantic deep web, semantic networking and semantic reasoning, SEMAPRO 2019 constituted the stage for the state-of-the-art on the most recent advances.

The conference had the following tracks:

- Basics on semantics
- Domain-oriented semantic applications
- Semantic applications/platforms/tools

We take here the opportunity to warmly thank all the members of the SEMAPRO 2019 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to SEMAPRO 2019. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also gratefully thank the members of the SEMAPRO 2019 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that SEMAPRO 2019 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the field of semantic processing. We also hope that Porto provided a pleasant environment during the conference and everyone saved some time to enjoy the historic charm of the city.

SEMAPRO 2019 Chairs

SEMAPRO Steering Committee

Fabio Grandi, University of Bologna, Italy
Tassilo Pellegrini, St. Pölten University of Applied Sciences, Austria

Sandra Lovrenčić, University of Zagreb, Croatia
Giuseppe Berio, Université de Bretagne Sud / IRISA, France
Takahiro Kawamura, Japan Science and Technology Agency (JST), Japan
Sule Yildirim Yayilgan, Gjøvik University College, Norway
Michele Melchiori, Università degli Studi di Brescia, Italy
Muhammad Javed, Cornell University, USA
Wladyslaw Homenda, Warsaw University of Technology, Poland

SEMAPRO Industry/Research Advisory Committee

Enrico Francesconi, ITTIG-CNR / Publications Office of the EU, Italy
Peera Pacharintanakul, TOT, Thailand
Raoul G. C. Schönhof, High Performance Computing Center Stuttgart (HLRS), Germany
Raghava Mutharaju, GE Global Research, Niskayuna, USA
Riccardo Albertoni, Istituto per la Matematica Applicata e Tecnologie Informatiche "Enrico Magenes" - Consiglio Nazionale delle Ricerche (IMATI-CNR), Italy
Sofia Athenikos, Morgan Stanley, USA
Shun Hattori, Muroran Institute of Technology, Japan

SEMAPRO 2019

COMMITTEE

SEMAPRO Steering Committee

Fabio Grandi, University of Bologna, Italy
Tassilo Pellegrini, St. Pölten University of Applied Sciences, Austria
Sandra Lovrenčić, University of Zagreb, Croatia
Giuseppe Berio, Université de Bretagne Sud / IRISA, France
Takahiro Kawamura, Japan Science and Technology Agency (JST), Japan
Sule Yildirim Yayilgan, Gjøvik University College, Norway
Michele Melchiori, Università degli Studi di Brescia, Italy
Muhammad Javed, Cornell University, USA
Wladyslaw Homenda, Warsaw University of Technology, Poland

SEMAPRO Industry/Research Advisory Committee

Enrico Francesconi, ITTIG-CNR / Publications Office of the EU, Italy
Peera Pacharintanakul, TOT, Thailand
Raoul G. C. Schönhof, High Performance Computing Center Stuttgart (HLRS), Germany
Raghava Mutharaju, GE Global Research, Niskayuna, USA
Riccardo Albertoni, Istituto per la Matematica Applicata e Tecnologie Informatiche "Enrico Magenes" - Consiglio Nazionale delle Ricerche (IMATI-CNR), Italy
Sofia Athenikos, Morgan Stanley, USA
Shun Hattori, Muroran Institute of Technology, Japan

SEMAPRO 2019 Technical Program Committee

Witold Abramowicz, Poznan University of Economics, Poland
Riccardo Albertoni, Istituto per la Matematica Applicata e Tecnologie Informatiche "Enrico Magenes" - Consiglio Nazionale delle Ricerche (IMATI-CNR), Italy
Jose María Alvarez Rodríguez, Carlos III University of Madrid, Spain
Ioannis Anagnostopoulos, University of Thessaly, Greece
Marta Andersson, Stockholm University, Sweden
Sofia Athenikos, Morgan Stanley, USA
Agnese Augello, ICAR - Istituto di Calcolo e Reti ad alte prestazioni | Consiglio Nazionale delle Ricerche, Palermo, Italy
Isabel Azevedo, Instituto Superior de Engenharia do Porto (ISEP), Porto, Portugal
Carlos Badenes-Olmedo, Universidad Politécnica de Madrid (UPM), Spain
Jarosław Bąk, Poznan University of Technology, Poland
Phạm The Bao, University of Science - Ho Chi Minh City, Vietnam
Giuseppe Berio, Université de Bretagne Sud / IRISA, France
Jorge Bernardino, Polytechnic of Coimbra - ISEC, Portugal
Stefano Bortoli, HUAWEI TECHNOLOGIES Duesseldorf GmbH - German Research Center - Munich Office,

Germany
Loris Bozzato, Fondazione Bruno Kessler, Trento, Italy
Zouhaier Brahmia, University of Sfax, Tunisia
Okan Bursa, Ege University, Turkey
Ozgu Can, Ege University, Turkey
Rodrigo Capobianco Guido, São Paulo State University (UNESP), Brazil
Elena Cardillo, Institute of Informatics and Telematics - National Research Council, Italy
Damir Cavar, Indiana University, USA
Julio Cesar Duarte, Military Institute of Engineering (IME), Brazil
Mingmin Chen, Uber Inc., USA
Muhao Chen, University of California Los Angeles, USA
Christian Chiarcos, Goethe Universität Frankfurt am Main, Germany
Christos Christodoulopoulos, Amazon Research Cambridge, UK
Ioannis Chrysakis, Foundation for Research and Technology-Hellas, Institute of Computer Science (FORTH-ICS), Greece
Francesco Corcoglioniti, Fondazione Bruno Kessler - Trento, Italy
Stefano Cresci, IIT-CNR, Italy
Valentin Cristea, University Politehnica Bucharest, Romania
Ademar Crotti Junior, Trinity College Dublin, Ireland
Zhihua Cui, Taiyuan University of Science and Technology, China
Mariana Damova, Mozajka Ltd, Bulgaria
Cláudio de Souza Baptista, Federal University of Campina Grande, Brazil
Maria Teresa Chiaravalloti, Institute of Informatics and Telematics | National Research Council of Italy, Italy
Amitava Das, Indian Institute of Information Technology, Andhra Pradesh, India
Monica De Martino, IMATI - National Research Council, Italy
Anastasia Dimou, Ghent University - IDLab – imec, Belgium
Melike Sah Direkoglu, Near East University, North Cyprus
Christian Dirschl, Wolters Kluwer Deutschland GmbH, Germany
Milan Dojchinovski, InfAI, Leipzig University, Germany / Czech Technical University in Prague, Czech Republic
Mauro Dragoni, Fondazione Bruno Kessler (FBK-IRST), Italy
Surya Durbha, Indian Institute of Technology Bombay (IITB), India
Ivan Ermilov, University of Leipzig, Germany
Vadim Ermolayev, Zaporozhye National University, Ukraine
Diego Esteves, University of Bonn, Germany
Muhammad Fahad, Centre Scientifique et Technique du Batiment CSTB (Sophia-Antipolis), France
Enrico Francesconi, ITTIG-CNR / Publications Office of the EU, Italy
Rolf Fricke, Condat AG, Germany
Panorea Gaitanou, Ionian University, Greece
Marcos Garcia, University of Coruña, Galiza, Spain
Chetana Gavankar, Cummins College of Engineering for Women, India
Joseph Giampapa, Carnegie Mellon University, USA
José M. Giménez-García, Université Jean Monnet, Saint-Étienne, France
Rafael Gonçalves, Stanford University, USA
Tatjana Gornostaja, Tilde, Latvia
Natalia Grabar, Université Lille 3, France
Fabio Grandi, University of Bologna, Italy

William Grosky, University of Michigan, USA
Ali Hasnain, National University of Ireland Galway, Ireland
Shun Hattori, Muroran Institute of Technology, Japan
Abdelati Hawwari, The George Washington University, USA
Benjamin Heitmann, Fraunhofer Institute for Applied Information Technology FIT, Germany
Gerald Hiebel, University of Innsbruck, Austria
Tobias Hellmund, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Germany
Armin Hoenen, Johann Wolfgang Goethe-Universität Frankfurt am Main, Germany
Jack Hodges, Siemens Corporate Technology, USA
Tracy Holloway King, Amazon, USA
Wladyslaw Homenda, Warsaw University of Technology, Poland
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Thomas Hubauer, Siemens AG Corporate Technology, Germany
Sergio Ilarri, University of Zaragoza, Spain
Agnieszka Jastrzębska, Warsaw University of Technology, Poland
Muhammad Javed, Cornell University, USA
Takahiro Kawamura, Japan Science and Technology Agency (JST), Japan
Haklae Kim, KISTI (Korea Institute of Science and Technology Information), Korea
Young-Gab Kim, Sejong University, Seoul, Korea
Gerda Koch, AIT Angewandte Informationstechnik Forschungsgesellschaft mbH, Austria
Mieczyslaw "Mitch" M. Kokar, Northeastern University, Boston, USA
Stasinos Konstantopoulos, Institute of Informatics and Telecommunications, NCSR "Demokritos", Greece
Efstratios Kontopoulos, Information Technologies Institute (ITI) / Center for Research & Technology Hellas (CERTH), Greece
Petr Kremen, Czech Technical University in Prague, Czech Republic
Jaroslav Kuchař, Czech Technical University in Prague, Czech Republic
Claudia Lanza, University of Calabria, Italy
Kyu-Chul Lee, Chungnam National University, Republic of Korea
Els Lefever, Ghent University, Belgium
Antonio Lieto, University of Turin and ICAR-CNR, Italy
Yunfei Long, Hong Kong Polytechnic University, Hong Kong
Giuseppe Loseto, Polytechnic University of Bari, Italy
Sandra Lovrenčić, University of Zagreb, Croatia
Wencan Luo, University of Pittsburgh, USA
Federica Mandreoli, Universita' di Modena e Reggio Emilia, Italy
Eugenio Martínez Cámara, Universidad de Granada, Spain
Béatrice Markhoff, Université de Tours, France
Miguel A. Martínez-Prieto, University of Valladolid, Spain
Miguel Felix Mata Rivera, UPIITA-IPN, Mexico
Brigitte Mathiak, GESIS - Leibniz institute for Social Sciences, Germany
John McCrae, Insight Centre for Data Analytics | National University of Ireland, Galway, Ireland
Imen Megdiche, Institut National Universitaire Champollion | IRIT, France
Michele Melchiori, Università degli Studi di Brescia, Italy
Panagiotis Mitzias, Information Technologies Institute (ITI) of the Centre for Research & Technology Hellas (CERTH), Greece
Luis Morgado da Costa, Nanyang Technological University, Singapore
Najmeh Mousavi Nejad, University of Bonn / Fraunhofer IAIS, Germany

Fadi Muheidat, University of the Pacific, USA
Raghava Mutharaju, GE Global Research, Niskayuna, USA
Yotaro Nakayama, Technology Research & Innovation, Nihon Unisys Ltd., Tokyo Japan
Sangha Nam, KAIST, South Korea
Fatemeh Nargesian, University of Toronto, Canada
Costanza Navarretta, University of Copenhagen, Denmark
Nikolay Nikolov, SINTEF Digital, Norway
Lyndon Nixon, MODUL Technology GmbH, Austria
Fabrizio Orlandi, University of Bonn and Fraunhofer IAIS, Germany
Peera Pacharintanakul, TOT, Thailand
Peteris Paikens, University of Latvia, Latvia
Panagiotis Papadakos, FORTH-ICS (Foundation for Research and Technology - Hellas, Institute of Computer Science), Greece
Tassilo Pellegrini, St. Pölten University of Applied Sciences, Austria
Rivindu Perera, Auckland University of Technology, New Zealand
Vassilios Peristeras, International Hellenic University, Greece
Silvia Piccini, Istituto di Linguistica Computazionale «A. Zampolli» - CNR, Italy
Octavian Popescu, IBM T. J. Watson Research Center, USA
Livia Predoiu, University of Oxford, UK
Behrang Qasemizadeh, Heinrich-Heine-Universität Düsseldorf, Germany
Joe Raad, Vrije Universiteit Amsterdam, Netherlands
Gerasimos Razis, School of Sciences - University of Thessaly, Greece
Livy Real, University of Sao Paulo, Brazil
Georg Rehm, DFKI GmbH, Germany
Irene Renau, Pontificia Universidad Católica de Valparaíso, Chile
German Rigau Claramunt, University of the Basque Country, Spain
Juergen Rilling, Concordia University, Montreal, Canada
Tarmo Robal, Tallinn University of Technology, Estonia
Christophe Roche, University Savoie Mont-Blanc, France
Alejandro Rodríguez González, Universidad Politécnica de Madrid, Spain
Federica Rollo, University of Modena and Reggio Emilia, Italy
Jacqueline Rosette, Swansea University, UK
Anisa Rula, University of Milano-Bicocca, Italy / University of Bonn, Germany
Michele Ruta, Technical University of Bari, Italy
Patrick Saint-Dizier, IRIT, France
Minoru Sasaki, Ibaraki University, Japan
Stefan Schlobach, Vrije Universiteit Amsterdam, Netherlands
Raoul G. C. Schönhof, High Performance Computing Center Stuttgart (HLRS), Germany
Anne-Kathrin Schumann, ProTechnology GmbH, Dresden, Germany
Wieland Schwinger, Johannes Kepler University Linz (JKU) | Inst. f. Telekooperation (TK), Linz, Austria
Floriano Scioscia, Polytechnic University of Bari, Italy
Emine Sezer, Ege Universitesi, Izmir, Turkey
Nuno Silva, School of Engineering - Polytechnic of Porto, Portugal
Vasco N. G. J. Soares, Instituto de Telecomunicações / Instituto Politécnico de Castelo Branco, Portugal
Ahmet Soylu, Norwegian University of Science and Technology / SINTEF Digital, Norway
Lars G. Svensson, Deutsche Nationalbibliothek, Germany
George Tambouratzis, Institute for Language and Speech Processing, Athena, Greece
WeldeRufael B. Tesfay, Deutsche Telekom Chair of Mobile Business and Multilateral Security | Goethe

University Frankfurt, Germany
Christos Tryfonopoulos, University of the Peloponnese, Greece
Jouni Tuominen, University of Helsinki, Finland
Murat Osman Ünalir, Ege University, Turkey
Jung-Ho Um, KISTI (Korea Institute of Science and Technology Information), Republic of Korea
L. Alfonso Ureña López, Universidad de Jaén, Spain
E. Lynn Usery, Center of Excellence for Geospatial Information Science | U.S. Geological Survey, USA
Taketoshi Ushiama, Kyushu University, Japan
Jack Verhoosel, Netherlands Organisation for Applied Scientific Research (TNO), Netherlands
Sirje Virkus, Tallinn University, Estonia
Daiva Vitkute-Adzgauskiene, Vytautas Magnus University, Lithuania
Tim vor der Brück, Lucerne University of Applied Sciences and Arts, Switzerland
Rita Zaharah Wan-Chik, Malaysian Institute of Information Technology (MIIT) - Universiti Kuala Lumpur, Malaysia
Yingying Wang, Snap Research, USA
Wai Lok Woo, Newcastle University, UK
Adam Zachary Wyner, Swansea University, UK
Sule Yildirim Yayilgan, Gjøvik University College, Norway
Alexander Yohan, University of Science and Technology, Taiwan
Roberto Yus, University of California, Irvine, USA
Fouad Zablith, Olayan School of Business | American University of Beirut, Lebanon
Stefan Zander, Hochschule Darmstadt, Germany
Martin Zelm, INTEROP-VLab, Brussels, Belgium
Xiaowang Zhang, Tianjin University, China
Qiang Zhu, University of Michigan - Dearborn, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Example Sentence Selection for Feedback on Preposition Usage

John Lee

Department of Linguistics and Translation, City University of Hong Kong
Email: jsylee@cityu.edu.hk

***Abstract*—While many writing assistance systems can automatically correct grammatical errors, most do not provide any explanation about their suggested corrections. This paper proposes an algorithm that selects example sentences to serve as feedback on preposition usage correction. This algorithm exploits the argument/adjunct distinction to select the most relevant example sentences. Evaluation shows that the use of argumenthood information improves the quality of the selected sentences.**

***Keywords*–computer-assisted language learning; example sentence selection; grammatical error correction feedback; preposition.**

I. Introduction

A Grammatical Error Correction (GEC) system detects and corrects grammatical errors in a learner text [1]. Given the input sentence "... go shopping **to* a store" [2], for example, the system may flag the preposition "to" and propose to replace it with "at". Studies in second language acquisition have shown that feedback from language teachers can be beneficial to foreign language pedagogy [3]. Most GEC systems, however, do not provide any feedback or explanation to complement their proposed corrections.

Research on automatic feedback generation has mostly focused on *explanatory feedback*, imitating the kind of comments traditionally composed by teachers (Section II-A). For the input sentence above, the system may elaborate on its correction with a feedback message such as "To mean traveling to a place in order to take part in an activity, *go* takes *at*, *in* or *on* depending on the activity ..." [2]. To date, most algorithms generate explanatory feedback by compiling comments from experts for different error types, and assigning them to unseen learner errors [2], [4]. Significant manual effort is required to cover the large variety of learner errors.

In contrast, *example-based feedback*, which presents example sentences to illustrate correct usage for the user, requires no manual composition (Section II-B). This approach can provide wider coverage, because it can address virtually any kind of usage issues, however idiosyncratic, as long as relevant examples can be found in the corpus. In addition, it supports data-driven learning by encouraging users to discover language patterns through observation of real-world example sentences, rather than through direct comments from the system or teacher [5]. Some existing systems can already provide example-based feedback by searching for similar sentences on the web [6] or in text corpora [7]. However, we are not aware of any reported evaluation on the quality or effectiveness of the retrieved example sentences.

This paper proposes an algorithm for generating example-based feedback aimed at preposition usage errors, a common error type for students of English as a Foreign Language [8]. This algorithm exploits the argument/adjunct distinction in prepositional phrases to help determine the most relevant example sentences. Evaluation shows that argumenthood information can help select higher-quality example sentences.

The rest of the paper is organized as follows. The next section presents previous work in feedback generation and argumenthood prediction. Section III presents our approach. Section IV describes our evaluation dataset and Section V discusses the results. Finally, Section VI concludes.

II. Previous Work

The feedback generated by existing writing assistance systems tends to fall into one of two types, *explanatory feedback* (Section II-A) and *example-based feedback* (Section II-B). After summarizing current approaches for generating these two types of feedback, we describe the argument/adjunct distinction in prepositional phrases (Section II-C), which will be exploited by our algorithm.

A. Generation of Explanatory Feedback

Among GEC systems that provide explanatory feedback, most rely on experts to manually compose the feedback or explanation for each error category. In the more coarse-grained approach, a "comment bank" [9] provides generic comments for broad error types such as "wrong preposition" or "wrong article". After correcting an error in the input text, the system delivers the comments associated with that error type to the user. While these hand-crafted comments can be comprehensive, they also tend to be generic and do not directly address the specific word usage in the input sentence.

In the more fine-grained approach, the feedback is associated not to broad error types, but rather to parse tree patterns [4], [10] or error case frames [2], which facilitate more in-context feedback. Case frames for preposition errors, for example, can be specific to the particular subject, verb, direct object, preposition and prepositional object in the sentence [2]. This approach still requires a significant amount of manual annotation, since error coverage is proportional to the number of frames for which comments are available.

B. Generation of Example-based Feedback

A GEC system can also offer example sentences as feedback to illustrate correct usage, either as an alternative or a supplement to explanatory feedback. This approach requires no hand-crafted messages. Further, given the size of contemporary text corpora, it can potentially cover a wider range of errors with corpus examples that more closely address the user's errors. The *ESL Assistant*, for example, automatically performs web search to retrieve sentences containing the original and corrected phrases [6]. A CALL tool for prepositions offers a review function, where users can request fill-in-the-blank items that are similar to those with which they previously experienced difficulties [7]. A sentence is considered "similar" if it contains the same preposition, prepositional object and lexical head, which can be identified in a parse tree such as the one in Figure 1.

Example sentence selection has primarily been investigated in the context of dictionary entries [11], [12], test item generation [13] and general language learning [14], typically using heuristics-based approaches. The kinds of example sentences required in these contexts share many similar characteristics with those for example-based feedback, such as well-formedness, simplicity of vocabulary, and ease of understanding. However, they target a larger variety of sentences, in order to provide a comprehensive portrait of the various aspects of the word's usage and collocational behavior. In contrast, example-based feedback aims at a narrower set of sentences that can precisely address the user's specific problem. Sentence selection for this purpose, therefore, often requires more syntactic and semantic analysis to determine the nature of the usage error. For preposition usage, this entails analyzing whether the preposition is used as an argument or adjunct.

C. Argumenthood

Arguments and adjuncts are linguistic concepts that have been intensively studied. In principle, "arguments depend on their lexical heads because they form an integral part of the phrase. Adjuncts do not." [15] An argument prepositional phrase (PP) is thus more closely related to the lexical head than an adjunct PP. For example, the phrase "to our topic" in sentence (1) in Table I is an argument PP, namely an argument of the lexical head "relevant". In contrast, "to some extent" in sentence (2) is an adjunct, serving as an adverbial to the lexical head "relevant". Argumenthood information has been shown to benefit a variety of natural language processing tasks, including PP attachment [15] and semantic role labeling [16]. It has not, however, been applied to sentence selection for feedback on grammatical errors.

There are a number of language resources that encode argument constructions, such as the verb subcategorization forms in VerbNet [17] and the grammar patterns in COBUILD [18]. Past work has attempted to distinguish between PP arguments and adjuncts with these resources, logical forms and formal grammars [19], as well as statistical models based on word embeddings and a variety of linguistic features [20].

III. Approach

Assuming that a grammatical error correction (GEC) system has corrected a preposition error in the input sentence, our task is to select the best example sentences from a corpus to explain and clarify the preposition usage. Similar to [21], we characterize preposition usage with three features: the corrected **preposition** (p'); the **prepositional object** (obj); and the **lexical head** (h). These features can be identified from a dependency parse tree. Figure 1 shows the tree for sentence (1) in Table I. Based on the dependency relations, we extract $p' =$"to", $obj =$"topic", and $h =$"relevant".

After defining the objective of the feedback (Section III-A), we discuss the types of example sentences to be considered (Section III-B), and the algorithms to be used for predicting the most suitable type (Section III-C).

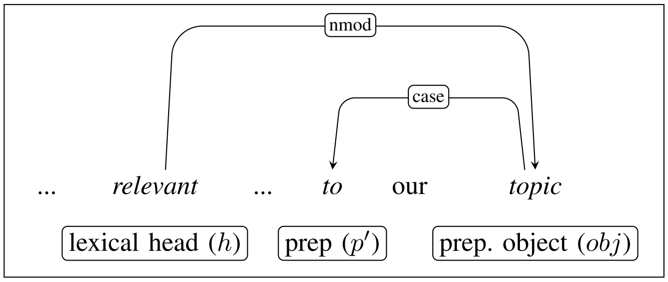


Figure 1. Extraction of the preposition, prepositional object and lexical head from a dependency parse tree, as derived by the Stanford parser [22].

A. Feedback Objective

When a GEC system corrects a preposition p to p', the user may not be able to discern the underlying reason:

> Is p' better than p because of (a) the lexical head, regardless of the choice of prepositional object? or because of (b) the prepositional object, regardless of the choice of lexical head?

For sentence (1), the answer is (a) because its PP is an argument. The ideal example sentence should make the point that the preposition "to" is required by the word "relevant", even when using other prepositional objects. In contrast, for sentence (2), the answer is (b) because its PP is an adjunct. A useful example should emphasize that "to ... extent" is the expected expression, even for lexical heads other than "relevant".

B. Types of example sentences

Table I lists some possible types of example sentences to provide feedback. An *Identical Example* is a sentence with the same p', obj and h as the input sentence. Sentences (3) and (4), for example, would serve as Identical Examples for sentences (1) and (2), respectively. An Identical Example seems useful in reinforcing the correction, since its content most closely matches the input sentence. However, by merely repeating the correction with the same h and obj, it gives no new insight and does not resolve the ambiguity noted in Section III-A: the user still would not be able to tell whether h or obj triggered the correction. We will therefore not give further considerations to Identical Examples. Instead, we focus on two kinds of example sentences:

1) Argument Example: We use the term "Argument Example" to refer to a sentence with the same p' and h as the input sentence. In Table I, sentence (5) serves as an Argument Example for (1) and (2). It gives useful feedback for sentence (1), where the *to*-PP is an argument. By using a different obj ("her"), it makes clear that the use of "to" is linked to the lexical head "relevant". In other words, it highlights the fact that *to* is an argument PP for the adjective "relevant".

In contrast, this example is less optimal for sentence (2). In reusing the lexical head "relevant", it fails to make the point that the expected expression is "to ... extent", and may even give the false impression that "**in* some extent" could be appropriate with other lexical heads.

TABLE I. TYPES OF EXAMPLE SENTENCES AS FEEDBACK ON PREPOSITION USAGE.

Type	Sentence	Lexical head (h)	Prep (p')	Prep. object (obj)	Remarks
(Corrected) input	(1) This book is **relevant** ~~with~~ *to* our **topic**. (2) This book is **relevant** ~~in~~ *to* some **extent**.	relevant relevant	to to	topic extent	*to*-PP is an argument *to*-PP is an adjunct
Identical Example	(3) This movie is **relevant** *to* the current **topic** (4) This movie is **relevant** *to* a large **extent**.	relevant relevant	to to	topic extent	Same h, p' and obj as (1) Same h, p' and obj as (2)
Argument Example	(5) This movie is **relevant** *to* **her**.	relevant	to	her	Same h and p' as (1), (2)
Adjunct Example	(6) This movie is **outdated** *to* a large **extent**.	outdated	to	extent	Same p' and obj as (2)

2) Adjunct Example: We use the term "Adjunct Example" to refer to a sentence with the same p' and obj as the input sentence. In Table I, sentence (6) serves as an Adjunct Example for (2). It gives useful feedback for sentence (2), where the *to*-PP is an adjunct. By using a different lexical head, "outdated", it clarifies that the choice of "to" as preposition is not tied to "relevant"; rather, it is required for the PP "to ... extent", even when under another lexical head.

C. Algorithm for Example Sentence Selection

To evaluate the effect of the argument/adjunct distinction on the quality of example-based feedback, we implemented the following algorithms for selecting example sentences. Given h, p' and obj, the algorithm is to determine whether Argument Examples or Adjunct Examples are more suitable as example sentences.

1) Majority Baseline: Ignoring the argument/adjunct distinction, this baseline always chooses the majority type in the evaluation dataset (Section IV).

2) COBUILD Grammar Patterns Baseline: These grammar patterns consist of phrases or clauses that are used with a verb [18], adjective or noun [23]. One pattern for the adjective *relevant*, for example, is the PP "`to n`". This baseline opts for Argument Examples as feedback if p' is listed among the patterns for h. Otherwise, it chooses Adjunct Examples.

3) Association Score Difference: Recent research suggested that the phenomenon of argumenthood is not binary, but gradient [20]. The grammar patterns define a boundary between argument and adjunct, but this boundary may not be the one at which Argument Examples become more useful than Adjunct Examples, or vice versa. This algorithm uses the logDice score [24], which measures word collocation strength based on the Dice Coefficient, as a proxy to learn this boundary from user data.

Let $logDice(h, p)$ represent the logDice score for the lexical head and the preposition, and let $logDice(obj, p)$ represent the score for the prepositional object and the preposition. We compute the difference between these scores, i.e., $\Delta logDice = logDice(h, p) - logDice(obj, p)$. We choose Argument Examples if $\Delta logDice > \theta$ and Adjunct Examples otherwise, with θ to be optimized on user data.

While there are many other approaches for predicting argumenthood (Section II-C), most concentrate on verbs as lexical heads and would have required non-trivial extension for nouns and adjectives. Since our goal is not to investigate the state-of-the-art in argumenthood prediction, we chose to use the logDice score for its simplicity and availability via Sketch Engine.

IV. EVALUATION DATASET

We extracted sentences containing preposition usage errors from Release 3.3 of the National University of Singapore (NUS) Corpus of Learner English (NUCLE) [25]. To construct an evaluation dataset that is balanced in terms of the part-of-speech (POS) of the lexical head and the argument/adjunct distinction, we randomly selected 24 sentences within the following constraints:

- **Lexical head POS**: 10 sentences have verbs as lexical head, 10 have nouns, and 4 have adjectives;
- **Argument vs. Adjunct**: Among sentences with lexical heads of each POS, one half have argument PPs and the other half have adjunct PPs, according to the COBUILD Grammar Patterns (Section III-C).

Since our research focus is on example sentence selection rather than grammatical error correction (GEC), we used the gold preposition in NUCLE to ensure that GEC accuracy would not be a confounding variable. A total of 8 prepositions ("at", "for", "from", "in", "of", "on", "through", "to", and "within") are represented in the dataset.

We retrieved example sentences in Sketch Engine with the collocation (h, p) to serve as Argument Examples, and sentences with the collocation (obj, p) to serve as Adjunct Examples. For each of the 24 input sentences, we collected the first three sentences returned by Good Dictionary EXamples (GDEX) [11] to create an Argument Example Set and an Adjunct Example Set. Table II shows an example item in the evaluation dataset.

For each item, we asked five human raters to decide whether the Argument or Adjunct Example Set was more useful. All five raters were advanced non-native speakers of English with a postgraduate degree in linguistics. The argument/adjunct distinction of the sets was not disclosed to the raters.

TABLE II. EXAMPLE ITEM IN EVALUATION DATASET

(Corrected) Input	The only way to satisfy the increasing demands ~~of~~ *for* space is by achieving a better usage ...
Adjunct Example Set	Canisters aren't the best option *for* big **spaces**. It's the perfect accent lamp *for* a small **space**. So that is a very practical use *for* **space**.
Argument Example Set	The **demand** *for* processed food items have increased ... They show no sign of scaling back their **demands** *for* human rights. Thus, **demand** *for* base metals will remain very strong.

V. EVALUATION RESULTS

We applied each algorithm in Section III-C to select either the Argument or Adjunct Example Set for each item in the evaluation dataset. For the Association Score Difference algorithm (Section III-C), we obtained the logDice scores in the English Web 2015 corpus on Sketch Engine, and tuned the value of θ using leave-one-out cross-validation in the evaluation dataset (Section IV).

TABLE III. ACCURACY IN SELECTING EXAMPLE SENTENCES FOR FEEDBACK ON PREPOSITION USAGE

Algorithm	Accuracy
COBUILD Grammar Patterns baseline	67.50%
Majority baseline	72.50%
Association Score Difference	**76.67%**

Table III shows the algorithms' accuracy in selecting the example set preferred by the rater. The Majority baseline achieved an accuracy of 72.50% by always choosing the Argument Example Set. Recall that only 50% of the items in the dataset have p' listed in the COBUILD Grammar Patterns as an argument marker (Section IV). This suggests a general preference among raters for example sentences illustrating argument usage. This preference holds regardless of the POS of the lexical head. For the rater with the strongest such preference, the Argument Example Set was deemed more useful in 8 out of the 12 adjunct items.

The COBUILD Grammar Patterns yielded an accuracy of 67.50%, below the Majority baseline. When it chose Argument Examples, the raters almost always agreed. Most errors occurred when it opted for Adjunct Examples, when the raters often preferred the Argument ones. This may reflect incomplete coverage in the grammar patterns, or could be the result of the gradient effect of the argumenthood phenomenon [20].

The Association Score Difference algorithm produced the best performance, at 76.67% accuracy. The improvement over the Majority baseline, at $p < 0.074$ by McNemar's Test, approaches statistical significance. The logDice score turned out to be a close proxy of the COBUILD Grammar Patterns, generally giving higher scores to (h, p') collocations where p' is listed in the patterns. Reflecting the raters' general preference for Argument Examples, the threshold θ was tuned to a relatively large negative value. This means that the algorithm selected Adjunct Example Sets only when the logDice score for (obj, p') enjoyed a large margin over the score for (h, p'). Experimental results thus show that the Dice Coefficient was effective in making more judicious selections for Adjunct Example Sets to cater to user preference on the argument-adjunct gradient for example sentences for preposition usage.

VI. CONCLUSION

We have presented a novel approach to select example sentences as feedback on preposition usage. This algorithm exploits the argument/adjunct distinction to determine the most useful examples. Evaluation shows that it can learn user preference on the argument-adjunct gradient to improve the quality of the selected example sentences.

ACKNOWLEDGMENT

The work reported in this paper is partially funded by an Applied Research Grant (#9667151) from City University of Hong Kong.

REFERENCES

[1] A. Rozovskaya and D. Roth, "Grammatical Error Correction: Machine Translation and Classifiers," in Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL), 2016.

[2] R. Nagata, M. Vilenius, and E. Whittaker, "Correcting Preposition Errors in Learner English Using Error Case Frames and Feedback Messages," in Proc. 52nd Annual Meeting of the Association for Computational Linguistics (ACL), 2014, pp. 754–764.

[3] J. Bitchener and S. Young, "The Effect of Different Types of Feedback on ESL Student Writing," Journal of Second Language Writing, vol. 14, no. 3, 2017, pp. 191–205.

[4] K. F. McCoy, C. A. Pennington, and L. Z. Suri, "English Error Correction: A Syntactic User Model based on Principled "Mal-rule" Scoring," in Proc. 5th International Conference on User Modeling, 1996.

[5] T. Johns, "Should You be Persuaded — Two Samples of Data-driven Learning Materials," in Classroom Concordancing, T. Johns and P. King, Eds. ELR Journal (4), 1991, pp. 1–16.

[6] C. Leacock, M. Gamon, and C. Brockett, "User Input and Interactions on Microsoft Research ESL Assistant," in Proc. NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications, 2009, pp. 73–81.

[7] J. Lee and M. Luo, "Personalized Exercises for Preposition Learning," in Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL) — System Demonstrations, 2016, pp. 115–120.

[8] J. Tetreault and M. Chodorow, "The Ups and Downs of Preposition Error Detection in ESL Writing," in Proc. 22nd International Conference on Computational Linguistics (COLING), 2008.

[9] D. Wible, C.-H. Kuo, F.-Y. Chien, and A. Liu, "A Web-based EFL Writing Environment: Integrating Information for Learners, Teachers, and Researchers," Computers and Education, vol. 37, no. 3-4, 2001, pp. 297–315.

[10] J. Kakegawa, H. Kanda, E. Fujioka, M. Itami, and K. Itoh, "Diagnostic Processing of Japanese for Computer-Assisted Second Language Learning," in Proc. 38th Annual Meeting of the Association for Computational Linguistics (ACL), 2000.

[11] A. Kilgarriff, M. Husák, K. McAdam, M. Rundell, and P. Rychlý, "GDEX: Automatically Finding Good Dictionary Examples in a Corpus," in Proc. EURALEX, 2008.

[12] J. Didakowski, L. Lemnitzer, and A. Geyken, "Automatic Example Sentence Extraction for a Contemporary German Dictionary," in Proc. EURALEX, 2012.

[13] S. Smith, P. V. S. Avinesh, and A. Kilgarriff, "Gap-fill Tests for Language Learners: Corpus-Driven Item Generation," in Proc. 8th International Conference on Natural Language Processing (ICON), 2010.

[14] V. Baisa and V. Suchomel, "SkELL: Web Interface for English Language Learning," in Proc. Recent Advances in Slavonic Natural Language Processing, 2014.

[15] P. Merlo and E. E. Ferrer, "The Notion of Argument in Prepositional Phrase Attachment," Computational Linguistics, vol. 32, no. 3, 2006, pp. 341–378.

[16] L. He, K. Lee, M. Lewis, and L. Zettlemoyer, "Deep Semantic Role Labeling: What Works and What's Next," in Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017.

[17] M. McConville and M. O. Dzikovska, "Evaluating Complement-Modifier Distinctions in a Semantically Annotated Corpus," in Proc. LREC, 2008.

[18] G. Francis, S. Hunston, and E. Manning, Collins COBUILD Grammar Patterns 1: Verbs. London: HarperCollins, 1996.

[19] A. Villavicencio, "Learning to Distinguish PP Arguments from Adjuncts," in Proc. CoNLL, 2002.

[20] N. Kim, K. Rawlins, B. V. Durme, and P. Smolensky, "Predicting the Argumenthood of English Prepositional Phrases," in Proc. AAAI, 2018.

[21] J. Lee and S. Seneff, "Automatic generation of cloze items for prepositions," in Proc. Interspeech, 2007.

[22] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in Proc. ACL System Demonstrations, 2014, pp. 55–60.

[23] G. Francis, S. Hunston, and E. Manning, Collins COBUILD Grammar Patterns 2: Nouns and Adjectives. London: HarperCollins, 1998.

[24] P. Rychlý, "A Lexicographer-Friendly Association Score," in Proc. Recent Advances in Slavonic Natural Language Processing, 2008.

[25] D. Dahlmeier, H. T. Ng, and S. M. Wu, "Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English," in Proc. 8th Workshop on Innovative Use of NLP for Building Educational Applications, 2013.

Text Similarity Estimation for Targeted Marketing with Outlier Robust Centroids of GloVe Word Embeddings

Tim vor der Brück

School of Information Technology
Lucerne University of Applied Sciences and Arts
Rotkreuz, Switzerland
e-mail: tim.vorderbrueck@hslu.ch

***Abstract*—Customer segmentation is an important task for marketeers. It is a prerequisite for precise and successful marketing campaigns. The traditional way of conducting it is by clustering based on demographic, geographic and psychographic variables like sex, age, city, or profession. Such an approach has several drawbacks. First, some of these variables might be hard to obtain in practice. Second, deducing from them actual interests for certain products is very hard in practice. In this paper, we present a different approach, in which we use short text snippets provided by users in an online contest to come up with a much more precise user interest profile. In particular, these text snippets are matched to keyword lists representing several marketing target groups like *Freestyle Action Sportsmen*, *Young Performer*, etc. For that, we employed the cosine measure on outlier robust centroids of GloVe word embeddings. These centroids are determined in an iterative fashion that gives most focus on non-outlier vectors and tends to disregard vectors, which are far off from the others. The evaluation showed that we obtained superior results with our method than several baseline approaches including one alternative method of noise reduction based on tf-idf weights.**

***Keywords*–GloVe; Targeted marketing; Outlier robust centroid**

I. Introduction

Market segmentation is one of the key tasks of a marketer. Usually, it is accomplished by clustering over demographic variables, geographic variables, psychographic variables, and behaviors [1]. In this paper, we will describe an alternative approach based on unsupervised natural language processing. In particular, our business partner operates a commercial youth platform for the Swiss market, where registered members get access to third-party offers such as discounts and special events like concerts or castings. Actually, several hundred online contests per year are launched over this platform sponsored by other firms. An increasing number of them require the members to write short free-text snippets, e.g., to elaborate on a perfect holiday at a destination of their choice in case of a contest sponsored by a travel agency. Based on the results of a broad survey, the platform provider's marketers assume five different target groups called youth milieus. A sixth milieu called *Special groups* comprises all those who cannot be assigned to one of the upper five milieus. For each milieu (with the exception of *special groups*) a keyword list was manually created to describe its main characteristics. For triggering marketing campaigns, an algorithm shall be developed that automatically assigns each contest answer to the most likely target group: we propose the youth milieu as best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user answer is maximal.

The semantic similarity of the two documents is then estimated by computing the cosine measure on the two centroids. One typical issue in this approach is that text can contain noise in form of words irrelevant for the actual topic. Such words are often either function words or have a very general meaning and can partly be filtered out using stop word lists. Additionally, one can mitigate this problem by weighting the word vectors according to their associated tf-idf value. We follow in this paper an alternative method to tf-idf word vector weighting, which is the use of an outlier robust centroid. This centroid reduces the influence of outliers by an iterative approach that weights the individual word vectors by their distance to the current centroid.

The remainder of this paper is structured as follows. In Section II, we summarize existing work on semantic similarity estimation. In Section III we describe the process of obtaining the outlier robust centroid in detail. Section IV describes the application to targeted marketing. Section V contains the evaluation results obtained on two manually annotated contests including a discussion. Finally, the paper concludes with Section VI, which summarizes the obtained results and gives an outlook to possible future work.

II. Related Work in Semantic Similarity Estimation

Semantic similarity estimation is usually based on word or sentence vectors that are first aggregated document-wise to centroid vectors, which are afterwards compared by the cosine similarity. The most popular method to come up with word vectors is Word2Vec [2], which is based on a 3 layer neural network architecture in which the word vectors are obtained as the weights of the hidden layer. Alternatives to Word2Vec are GloVe [3], which is based on aggregated global word co-occurrence statistics and the Explicit Semantic Analysis (or shortly ESA) [4], in which each word is represented by the column vector in the tf-idf matrix over Wikipedia.

The idea of Word2Vec can be transferred to the level of sentences as well. In particular, the so-called Skip Thought Vector model (STV) [5] derives a vector representation of the current sentence by predicting the surrounding sentences.

Sond and Roth [6] propose an alternative approach to applying the cosine measure to the two word vector centroids for ESA word embeddings called Dense-ESA. In particular, they establish a bipartite graph consisting of the best matching vector components by solving a linear optimization problem. The similarity estimate for the documents is then given by the global optimum of the objective function. However, this method is only useful for sparse vector representations. In case of dense vectors, Mijangos et al. [7] suggested to apply the Frobenius kernel to the embedding matrices, which contain the embedding vectors for all document components (usually either sentences or words) (cf. also [8]). However, crucial limitations are that the Frobenius kernel is only applicable if the number of words (sentences respectively) in the compared documents coincide and that a word from the first document is only compared with its counterpart from the second document. Thus, an optimal matching has to be established already beforehand.

Another similarity estimate that employs the entire embedding matrix is the word mover‘s distance [9], which is a special case of the earth mover‘s distance, a well studied transportation problem. Basically, this approach determines the minimum effort (with respect to embedding vector changes) to transform the words of one text into the words of another text. The word mover‘s distance requires a linear optimization problem to be solved. Linear optimization is usually tackled by the simplex method, which has in the worst case, which rarely occurs however, exponential runtime complexity.

In two former papers we proposed for the task of customer segmentation two additional text similarity estimates, one based on an ontology [10] and the other on matrix norms [11] applied on the word similarity matrix over the two texts to compare.

A drawback of most conventional similarity estimates as described above is that slightly related word pairs can have in aggregate a considerable influence on their values, i.e., these estimates are sensitive to noise in the data.

III. Outlier Robust Centroids

The outlier robust centroid is illustrated in Figure 1. We basically use a variant of the Huber centroid but applied on ordinary vectors instead of covariance matrices [12]. The red dot denotes the ordinary centroid of the black dots, the blue dot is the outlier robust centroid. As one can perceive from the figure, the ordinary centroid is much more drawn in direction of the outlier (the black dot on the very bottom) than its outlier robust counterpart.

The procedure to obtain the latter is given in pseudocode (see algorithm 1). First, the word vector weights are initialized with 1 divided by the number of word vectors. In this way, each word vector is weighted identically at the beginning. Afterward (step 2), our initial centroid is computed as the weighted sum of all word vectors. Now we update the weight, where each vector is weighted by the reciprocal of its distance to the centroid. In bid to avoid weights of infinity, we add a tiny positive amount to the distance prior to building the reciprocal. Using this weighting procedure, very distant vectors, typical outliers, are weighted less than closeby ones. Now we repeat this process returning to step 2 until the coordinates of the centroid have sufficiently converged and remain basically unchanged.

Algorithm 1 Outlier Robust Centroid

```
 1: procedure ROBUST_OUTIER($vec$)
 2: $numit = size(vec)$
 3: $w \leftarrow [1/numit, \dots, 1/numit]$
 4: for ever:
 5:     $C := [0, \dots, 0]$
 6:     $w\_sum := 0$
 7:     $i := 0$
 8:     for vec in vecs:
 9:         $C+ = w[i++] \cdot vec$
10:     $i := 0$
11:     for vec in vecs :
12:         $w[i] = 1/(dist(vec, C) + 0.00001)$
13:         $w\_sum+ = w[i++]$
14:     for wi in w:
15:         $wi := wi/w\_sum$
16:     if $dist(C, last\_C) < threshold$ then
17:         break
18:     $last\_C := C$
19: return C
```

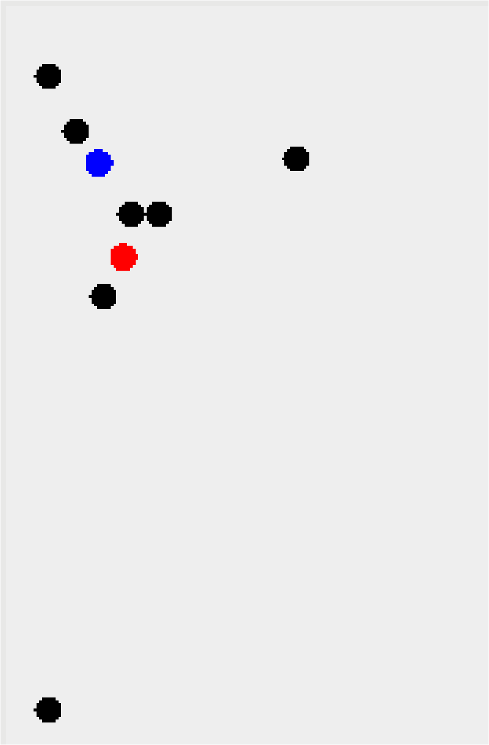


Figure 1. Outlier robust centroid (blue dot) vs ordinary centroid (red dot).

IV. Application to Targeted Marketing

Market segmentation is one of the key tasks of a marketer. Usually, it is accomplished by clustering over demographic variables, geographic variables, psychographic variables and behaviors [1]. In this paper, we will describe an alternative approach based on unsupervised natural language processing. In particular, our business partner Jaywalker GmbH operates a commercial youth platform for the Swiss market, where registered members get access to third-party offers such as discounts and special events like concerts or castings. Actually, several hundred online contests per year are launched over this platform sponsored by other firms. Some of them require the participants to write short free-text snippets. For instance, in one of our contests, the participants should specify their three preferred travel destination countries and elaborate on how a perfect holiday there would look like. An example of a participant's answer is given below:

1) Jordanien: Ritt durch die Wüste und Petra im Morgengrauen bestaunen bevor die Touristenbusse kommen
2) Cook Island: Schnorcheln mit Walhaien und die Seele baumeln lassen
3) USA: Eine abgespaceste Woche am Burning Man Festival erleben

English translation:

1) Jordan: Ride through the desert and marveling Petra during sunrise before the arrival of tourist buses
2) Cook Island: Snorkeling with whale sharks and relaxing
3) USA: Experience an awesome week at the Burning Man Festival

Based on the results of a broad survey, the platform provider's marketers assume five different target groups (called *milieus*) being present among the platform members: *Progressive Postmodern Youth* (people primarily interested in culture and arts), *Young Performers* (people striving for a high salary with a strong affinity to luxury goods), *Freestyle Action Sportsmen*, *Hedonists* (rather poorly educated people who enjoy partying and disco music) and *conservative youth* (traditional people with a strong concern for security). A sixth milieu called *Special Groups* comprises all those who cannot be assigned to one of the upper five milieus. For each milieu (with the exception of *Special Groups*) a keyword list was manually created to describe its main characteristics. For triggering marketing campaigns, an algorithm shall be developed that automatically assigns each contest answer to the most likely target group: we propose the youth milieu as best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user answer is maximal. In case the highest similarity estimate falls below the 10 percent quantile for the distribution of highest estimates, the special groups' milieu is selected. Since the keyword list typically consists of nouns (in the German language capitalized) and the user contest answers might contain a lot of adjectives and verbs as well, which do not match very well to nouns in the Word2Vec vector representation, we actually conduct two comparisons for our Word2Vec based measures, one with the unchanged user contest answers and one by capitalizing every word beforehand. The final similarity estimate is then given as the maximum value of both individual estimates.

Note that we apply the outlier robust centroid only to the word vectors derived from the user snippets since the keyword list is manually defined and should usually be free of noise.

V. Evaluation

For evaluation, we selected three online contests (language: German), where people elaborated on their favorite travel destination, speculated about potential experiences with a pair of fancy sneakers (contest 2) and explained why they emotionally prefer a certain product out of four available candidates. We experimented with different keyword list sizes (see Table IV) but obtained the best results with rather few, and therefore precise keywords (see Table V).

TABLE I. Obtained accuracy values for several similarity measures and for several baseline methods. (W)W2VC=(tf-idf-weighted) Word2Vec Embedding Centroids. GloVe,R.=GloVe using Outlier Robust Centroids.

Method	Contest			
	1	2	3	All
Random	0.167	0.167	0.167	0.167
ESA	0.357	0.254	0.288	0.335
ESA2	0.355	0.284	0.227	0.330
W2VC	0.347	**0.328**	0.227	0.330
WW2VC	0.347	0.299	0.197	0.322
GloVe	0.350	0.269	0.258	0.328
GloVe,R.	**0.365**	0.239	0.303	**0.342**
STV	0.157	0.313	0.258	0.189

TABLE II. Minimum and maximum average inter-annotator agreements (Cohen's kappa) / average inter-annotator agreement values for our automated matching method.

Method	Contest		
	1	2	3
Min kap.	0.123	0.295/0.030	0.110/0.101
Max. kap.	0.178	0.345/0.149	0.114/0.209
# Entr.	1544	100	100

TABLE III. Corpus sizes measured by number of words.

Corpus	# Words
German Wikipedia	651 880 623
Frankfurter Rundschau	34 325 073
News journal *20 Minutes*	8 629 955

In bid to provide a gold standard, three professional marketers from different youth marketing companies annotated independently the best matching youth milieus for every contest answer. We determined for each annotator individually his/her average inter-annotator agreement with the others (Cohen's kappa). The minimum and maximum of these average agreement values are given in Table II. Since for contest 2 and contest 3, some of the annotators annotated only the first 50 entries (last 50 entries respectively), we specified min/max average kappa values for both parts. We further compared the youth milieus proposed by our unsupervised matching algorithm with the majority votes over the human experts' answers (see Table I).

The Word2Vec and GloVe word embeddings were trained on the German Wikipedia (dump originating from 20 February 2017) merged with a Frankfurter Rundschau newspaper Corpus and 34 249 articles of the news journal *20 minutes* (see http://www.20min.ch), where the latter is targeted to the Swiss market and freely available at various Swiss train stations (see Table III for a comparison of corpus sizes). By employing articles from *20 minutes*, we want to ensure the reliability of word vectors for certain Switzerland specific expressions like *Velo* or *Glace*, which are underrepresented in the German

TABLE IV. ORIGINAL KEYWORD LISTS (TRANSLATED FROM GERMAN).

Youth Millieu	Keywords
Progressive Postmodern Youth	Trend, Trendsetter, Opinion Maker, Opinion Maker, Opinion, Opinion Leader, Individuality, Individual, Self-realization, Urban, Urbanity, City, Urban, Forward Thinker, Academic, Academic, Conscious, Design, Culture, Cultural, Choice, Freedom, Flexibility, Unbound, Progressive, Ecology, sustainability, new, discover, discovery, postmodern, hip, future, cultural journey, history, buildings, architecture, theatre, language travel, ipster, acting, instrument, musical, vegan, vegetarian, vegetarian, vegetarian, arthouse, independent, beard, criticism, rehearsal, band, books, literature, language, green, secondhand, fair, human right, human rights
Young Performer	Mobility, mobile, flexible, flexibility, performance, performance, performance-oriented, elite, elitist, risk, risk-averse, luxury, luxurious, income, self-realization, self-management, career, spontaneous consumption, education, educated, student, Status, bespoke, Brands, individual, Individuality, excessive, Success, materialistic, materialistic, possession, wealth, enjoyment, Enjoy, Wellness, City break, Gourmet, First class, Business, Opera, Metropole, Money, Account, MBA, CAS, MAS, business, Tailored, Individual
Freestyle Action Sportsman	Apprentice, Music, Sports, Sporty, New, New, Action, Action, Joy, Joy, Experience, Freestyle, Social, Joy, Just, Justice, Adventure, Adventurous, Optimism, Optimist, Extreme, Casual, Sportmania, Improvisation, Improvise, Freestyle, Freedom, Unbound, free, positive, celebrate, party, party, yolo, rap, rhythm, freeride, adventure, snow, mountains, bladen, skating, board, authenticity, interculturality, self-determination, left-liberal, curiosity, sea, nature, natural, video, film, nature, chill, group, homies, style, cool, go-pro
Hedonist	Mainstream, enjoy, enjoyment, intense, casual, unectritical, mass, communicative, entertainment, variety, inconspicuous, carefree, consumption, hedonist, materialistic, materialistic, joy, pleasure, lust, desire, painless, selfish, momentary, present, decadence, decadent, egoism, celebrate, party, party, lazy, lazy, all-inclusive, discount, cheap, last minute, beach, rock, pop, hits, new, current, charts, cinema, stadium, exit, club, drink, event, weed, grass, smoking, street parade, carnival, television, sofa, playstation, xbox
Conservative Youth	conservative, bourgeois, bourgeois, tradition, traditional, modest, modesty, community, common, down-to-earth, down-to-earth, associations, considered, orderly, Switzerland, future, middle class, virtue, virtuous, preserve, Existing, Stable, Stability, Preserve, Protect, Protection, Social, Craft, Democracy, Democratic, People, Hiking, Mountains, History, Homeland, Folklore, Popular, Carnival, Guugen, SVP, Work, Former, Quarter, Closing time, Stammtisch, Beiz

TABLE V. REDUCED KEYWORD LIST (TRANSLATED FROM GERMAN).

Youth milleu	Keywords
Progressive Postmodern Youth	clothing, music, art, freedom, culture, educated
Young Performer	rich, elite, luxury, luxurious
Freestyle Action Sportsmen	Sports, Fitness, Music
Hedonist	poor, communication, self-fulfilment, entertainment, party, music, disco
Conservative Youth	conservation of value, conservativity, citizenship, Switzerland

Wikipedia and the Frankfurter Rundschau corpus. ESA is usually trained on Wikipedia, since the authors of the original ESA paper suggest that the articles of the training corpus should represent disjoint concepts, which is only guaranteed for encyclopedias. However, Stein and Anerka [13] challenged this hypothesis and demonstrated that promising results can be obtained by applying ESA on other types of corpora like the popular Reuters newspaper corpus as well. Unfortunately, the implementation we use (Wikiprep-ESA, URL: https://github.com/faraday/wikiprep-esa) expects its training data to be a Wikipedia Dump. Furthermore, Wikiprep-ESA only indexes words that are connected by hyperlinks, which are usually lacking in ordinary newspaper articles. So we could train Wikiprep-ESA on Wikipedia only but additionally have developed a version of ESA that can be applied on arbitrary corpora (in the following referred to as ESA2) and which was trained on the full corpus (Wikipedia+Frankfurter Rundschau+20 minutes). The STVs were also trained on the same corpus as GloVe and Word2Vec embedding centroids. The actual document similarity estimation is accomplished by the usual centroid approach. An issue we were faced with is that STVs are not bag of word models but actually take the sequence of the words into account and therefore the obtained similarity estimate between milieu keyword list and contest answer would be dependent on the keyword ordering. However, this order could have arbitrarily been chosen by the marketers and might be completely random. A possible solution is to compare the contest answers with all possible permutation of keywords and determine the maximum value over all those comparisons. However, such an approach would be infeasible already for medium keyword list sizes. Therefore, we use a beam search approach instead, which extends the keyword list iteratively and keeps only the n-best performing permutations.

Finally, to verify the general applicability of our approach, we conducted a second experiment, where a novel by Edgar Allen Poe (The purloined letter) was independently translated by two different translators into German. We aim to match a sentence from the first translation to the associated sentence of the second by looking for the assignment with the highest semantic relatedness disregarding the sentence order. The obtained accuracy values based on the first 200 sentences of both translations are given in Table VI. To guarantee an 1:1 sentence mapping, periods were partly replaced by semicolons.

The evaluation showed that the use of outlier robust centroids leads to superior results on our evaluation set in terms of classification accuracy in comparison with its non-robust counterpart on 2 of the three contests and also for the overall comparison, for which all entries of the three contests are merged by concatenation as well as for the second task of translation matching. Furthermore, our method also

TABLE VI. ACCURACY VALUE OTBAINED FOR MATCHING A SENTENCE OF THE FIRST TO THE ASSOCIATED SENTENCE OF THE SECOND TRANSLATION.

Method	Accuracy
ESA	0.672
GloVe	0.706
GloVe,R.	**0.726**
STV	0.716
W2VC	**0.726**

clearly outperforms the use of centroids of tf-idf weighted embeddings, which is an alternative method for noise reduction in the data.

VI. CONCLUSION

We proposed a similarity measure to compare GloVe embeddings from different documents based on outlier robust centroids. This measure was evaluated on the task to assign users to the best matching marketing target groups. We obtained superior results compared to the usual non-robust centroid / cosine measure similarity estimation for contests 1 and 2 as well as overall (just appending the individual contests to form one large contest). As future work, we plan to evaluate additional similarity measures like Dense ESA or novel sentence embedding approaches on our dataset.

ACKNOWLEDGEMENT

Hereby we thank the Jaywalker GmbH as well as the Jaywalker Digital AG for their support regarding this publication and especially for annotating the contest data with the best-fitting youth milieus.

REFERENCES

[1] M. Lynn, “Segmenting and targeting your market: Strategies and limitations,” Cornell University, Tech. Rep., 2011, online: http://scholorship.sha.cornell.edu/articles/243.

[2] T. Mikolov, I. Sutskever, C. Ilya, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Proceedings of the Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, 2013, pp. 3111–3119.

[3] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Katar, 2014.

[4] E. Gabrilovic and S. Markovitch, “Wikipedia-based semantic interpretation for natural language processing,” Journal of Artificial Intelligence Research, vol. 34, 2009.

[5] R. Kiros, Y. Zhu, R. Salakhudinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fiedler, “Skip-thought vectors,” in Proceedings of the Conference on Neural Information Processing Systems (NIPS), Montréal, Canada, 2015.

[6] Y. Song and D. Roth, “Unsupervised sparse vector densification for short text similarity,” in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Denver, Colorado, 2015.

[7] V. Mijangos, G. Sierra, and A. Montes, “Sentence level matrix representation for document spectral clustering,” Pattern Recognition Letters, vol. 85, 2017.

[8] K.-J. Hong, G.-H. Lee, and H.-J. Kom, “Enhanced document clustering using wikipedia-based document representation,” in Proceedings of the 2015 International Conference on Applied System Innovation (ICASI), 2015.

[9] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, “From word embeddings to document distances,” in Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 957–966.

[10] T. vor der Brück, “Estimating semantic similarity for targeted marketing based on fuzzy sets and the odenet ontology,” in Proceedings of SemaPro, Athens, Greece, 2018.

[11] T. vor der Brück and M. Pouly, “Text similarity estimation based on word embeddings and matrix norms for targeted marketing,” in Proceedings of NAACL, Minneapolis, USA, 2019.

[12] I. Ilea, H. Hajiri, S. Said, L. Bombrun, C. Germain, and Y. Berthoumieu, “An m-estimator for robust centroid estimation on the manifold of covariance matrices: performance analysis and application to image classification,” in Proceedings of the 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 2016.

[13] T. Gottron, M. Anderka, and B. Stein, “Insights into explicit semantic analysis,” in Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, UK, 2011, pp. 1961–1964.

A Data Referencing Formalism for Information Exchange between Deafblind People and Databases

Carlos Seror
Independent researcher
Valencia, Spain
email: serorcarlos@gmail.com

***Abstract*—A categorizing feature of human language could be usefully exploited to facilitate an interactive access of deafblind people to database information. To that end, an interpretation of databases in terms of categories and instances is proposed, as well as the basic elements of a syntax consistent with it.**

Keywords-databases; categories; natural language; syntax; data referencing; deafblind people

I. INTRODUCTION

The access of DeafBlind (DB) people to information is generally limited to communication with other human beings through various, often non-standard, languages. Therefore, they require a human intermediary to query a database, or to be able to understand database reports expressed in some readable string form. Attempts to facilitate such communication have been described [1][2], including based on ontologies [16], sign languages [10], the Semantic Web [17], and even neural networks [9]. However, the formal aspect of data referencing has not been addressed in the literature in great depth. In this paper, based on a categorizing feature of natural languages [11], a formalism will be proposed that should make it possible to establish syntax correspondences between natural language expressions and databases.

In Section II, the concept of data aggregate will be introduced, to be subsequently endowed, in Section III, with a simple structure based on the logical connectives ∧, ∨. In Section IV, a string syntax will be derived from the resulting expressions and shown to be consistent with both natural language and databases. In Section V, the scope of the connectives will be enlarged to include categories of data items, and the resulting n-tuple structure will be seen to be, in Section VI, a particular case of more complex two-dimensional structures. Based on such general structures, a referencing and updating language will be developed in Section VII, and used in Section VIII to establish equivalences with simple database operations, including a few examples showing the structures' potential for semantic representation. Finally, Section IX will assess the potential of the proposed approach for communication with deafblind users.

II. DATA AGGREGATES

Databases have been around for a number of decades now. They are just a particular way to organize data [12]. From an abstract standpoint, databases could be described as symbol aggregates endowed with a particular structure, usually in the form of tables, but sometimes as graphs, objects or other ways of organization [6]. Natural Language (NL), being a means to deliver information, could also be argued to use data but, except in specific, explicitly structured subject areas, its users are usually unaware of the structure of such data. Describing a data structure consistent with NLs would therefore be a first step to establish a working correspondence between NLs and databases. This paper formally describes one such structure, based on a categorizing feature of NLs.

For a general approach to a diversity of data structures, the term 'data aggregate' will be adopted here. A data aggregate is defined as a number of data items that could be represented as points on a surface. This is arguably the minimal structure that can be conceived of, and it does not exclude other additional structures. Thus, a number of colors could be considered as a data aggregate, irrespective of whether they are located within a rainbow or on a painter's palette. In a data aggregate, items can be pointed to but do not have to be distinctly labelled —you may know nearly everything about a wood and not have a name for any of its trees. An item in a data aggregate could also be identified through a number of instructions on how to locate it. Also, a data aggregate could be indefinitely updated, as long as any new item could be represented as an additional point on the same surface. Goats in a herd, or data in a form, are simple examples of data aggregates.

III. INTENSION AND EXTENSION IN A DATA AGGREGATE

Items in a data aggregate could be discriminated by applying criteria to them. A criterion is a notion more general than a property, because it encompasses properties as well as fancy choices and algorithms. Two of the simplest criteria that could be applied to a data aggregate are the ones associated to intension and extension. Aggregations of items in a data aggregate do not have any extension or intension connotation *per se*, and are therefore objects more general than sets. Extension and intension could be implemented on them by means of resp. the connectives ∧, ∨, e.g.,

blue ∧ red ∧ yellow ∧ ...	extension [colors]
blue ∨ red ∨ yellow ∨ ...	intension [color]

In general, therefore, a number of items u_1, u_2, u_3, ... in a data aggregate could be discriminated in two alternate ways, i.e.,

$$u_1 \wedge u_2 \wedge u_3 \wedge \dots \tag{1}$$
$$u_1 \vee u_2 \vee u_3 \vee \dots \tag{2}$$

Because mathematical sets are said to be describable both in extensional and intensional terms, we shall rather not mix things up and separately discern either description instead. Hence, any expression in the form (1) will be referred to as a **combination**, while any expression in the form (2) will be referred to as a **category**. This difference reflects the use of plural resp. singular in human language. Thus, ‘colors’ could be associated to a combination, while ‘color’ could be associated to a category. When the scope of the criteria is not specified, the expressions (1) and (2) could also be interpreted as reflecting the difference between resp. ‘every’ and ‘any’.

Any item *u* encompassed by a category *C* —i.e., complying with the criteria that define *C*— will be referred to as an **instance** of *C*. A category *C* encompassing the instances u_1, u_2, u_3, ... will therefore be expressed as

$$C \equiv u_1 \vee u_2 \vee u_3 \vee \dots$$

The definitions of category and instance could be used as a means to locally refer to an item in a data aggregate. Indeed, in a data aggregate *E* where a category *C* has been identified, any instance of *C* could be expressed as a disambiguation of *C*. That is, if we denote a category as *C()* and an instance *u* of that category as *C(u)*, we could refer to *u* as

$$C() \rightarrow C(u)$$

The expression above may be interpreted as a path in *E*, i.e., “select *C*, then select the instance *u* of *C*”, where *u* could be identified by means of either a label or a number of instructions. In the following sections, an enlarged notion of disambiguation will be shown to be a powerful device to refer to data items in a data aggregate —arguably, the basic addressing device used by natural language users.

A data item in a data aggregate could also be referred to through a disambiguation of categories. For example, the word ‘green‘ may denote either a color or a political affiliation. In the absence of additional cues, they could be disambiguated resp. as either *color(green)* or *political_affiliation(green)*. In formal terms, if *H* is a category having the category *C* as an instance, then the instance *C(u)* could be referred to as

$$H(u) \rightarrow C(u)$$

IV. Symbol Syntax and String Syntax

The notation used thus far, based on symbols such as connectives or arrows, will be referred to as **symbol syntax**. An alternative syntax, which shall be referred to as **string syntax**, will express categories and instances as single words, and disambiguations as strings, as follows:

symbol syntax	string syntax
C()	C
C() → C(u)	C [δ u]

where the symbol δ denotes the fact that *u* complies with the criterion that defines *C*. The correspondence between symbol expressions and string expressions will be denoted as >>, e.g.,

color() >> color
color() → color(blue) >> color [δ blue]

Note that, in practice, if we deem it obvious that, e.g., the word ‘blue‘ refers to a color, we will not precede it with the word ‘color’, which will have to be guessed by the receiver. This data compression feature hinges on an implicit operation that pervades human language —and arguably also human thought—, i.e., spontaneous categorization.

V. Combined Categories

The connective ∧ could also be used to discriminate combinations of categories in a data aggregate. For example, from the categories

mass, electric charge, spin

a combined category could be derived, which in turn would give rise to a number of objects, e.g.,

mass ∧ electric_charge ∧ spin >> particle
mass ∨ electric charge ∨ spin >> observable
mass(9.1×10^{-31} kg) ∧ electric_charge(-1.6×10^{-19} C) ∧ spin(1/2) >>
particle [δ electron]

As the latter example shows, a combination of categories is itself a category, having as instances combinations of instances of its component categories. The latter example is a full disambiguation of the category ‘particle’. However, combined categories could also be partially disambiguated by specifying just some instances of its component categories, e.g.,

bearing() ∧ altitude(7000 ft) ∧ No._of_passengers(80)

Component categories in a combined category could also be discriminated by means of the connective ∨. The resulting object could be used to identify a path within the data aggregate leading to any of such component categories. For example, the category ‘color’ could also be construed as an attribute, i.e., an instance of the category

color ∨ shape ∨ size ∨ ...

Categories pervade language, a fact which is obscured by the spontaneous categorization mechanism, of which language users are usually unaware. In human languages, spontaneous categorization is a run-of-the-mill feature, associated not only to adjectives such as ‘blue’ or ‘big’, but to virtually any kind of meaningful component. Thus, the sentence ‘birds fly’ could be inaccurately categorized as implying that hens fly, or meaningfully categorized by instead interpreting ‘birds’ as an instance of some category, e.g., birds $\vee$ mosquitos $\vee$ bats $\vee$ ... having ‘fly’ as an attribute. Similarly, the meaning of ‘a through person’ could only be captured by evoking a category of concepts having ‘through’ as an instance.

In general, therefore, a **combined category** G^{η} will be defined as the general expression

$$G^{\eta} = O1 \wedge O2 \wedge O3 \wedge ... \qquad (3)$$

where *O1*, *O2*, *O3*, ... are categories, whether they have been disambiguated or not, and η uniquely identifies that particular combination of categories. The definition (3) is consistent with a number of concept theories [13][14], that describe concepts as n-tuples of symbols representing attributes.

VI. Category Clusters

An n-tuple is just a one-dimensional combination of categories, and therefore a particular case of the more general concept of **category cluster**, where complex spatial relations could be incorporated as additional discrimination criteria in a data aggregate. A data form is a familiar example of data cluster. In general terms, therefore, a **representation** can now be defined as a data aggregate together with any number of category clusters. Now, given a category cluster *G* and one of its component categories *C*, any set of instructions *r* to uniquely identify *C* within *G* will be referred to as a **relation** *r(G, C)*.

As in the one-dimensional case, specific category clusters could also be referred to by specifying one or more of its component categories. For example, an employee’s record might be uniquely identified by specifying just the employee’s name, or his age and height. This will be formally described as follows. Let *G* be a category cluster, *r* a set of instructions to identify *C* within *G*, and G_k a copy of *G* where the data item *u* has been specified for the category *C*. The cluster G_k could therefore be referred to as

$$G_k = G \mid r(G, C(u)) \qquad (4)$$

i.e., G_k can be interpreted as a partial disambiguation of *G*. In string syntax, this will be expressed as

$$G \mid r(G, C(u)) >> G\ [r\ u]$$

For example,

$$ball = shape(round) \wedge color() \wedge size()$$

$$ball_k = shape(round) \wedge color(red) \wedge size(big) >> ball\ [r_2\ red]\ [r_3\ big]$$

where r_1, r_2, r_3 would represent resp. the sets of instructions to locally identify each of the component categories ‘shape’, ‘color’, ‘size’. In the general case, the disambiguation of a category cluster *G* will be expressed in string syntax as

$$G\ \Sigma[r_j\ u_j] \qquad (5)$$

where Σ denotes a string made up of $[r_j\ u_j]$ pairs, u_j denotes an instance of the component category C_j, and r_j denotes the relation $r_j(G, C_j)$. The possibility to uniquely identify a category cluster even when only some of its component categories have been specified is a feature heavily used by natural language users as a data compression device. Indeed, if there is only one red ball in the room, you would hardly want to refer to it as “the big red expensive air-filled ball on the sofa”.

Any set of rules to convert string syntax expressions into different strings will be referred to as a **conventional** syntax. For example,

String syntax	*Conventional syntax*	
ball [r_2 red] [r_3 big]	big red ball	(English)
boule [r_2 rouge]	boule rouge	(French)
bam [r ug]	bugam	(imaginary)
✌ [r_2 👌] [r_3 🖐]	👌 ✌ ♒ 🖐	(non-word)

The notion of category cluster, plus the relations it entails, endow data aggregates with powerful structures that could be used to semantically represent a vast number of concepts, e.g., ontologies, verbs, or semantic representations of space-time concepts, together with a local mechanism to refer to them [8][11]. An example of semantic cluster is described in Section VIII, *A.*

VII. Reference and Updating

The definition of the cluster-category relation implies that categories could also be referred to in terms of the cluster or clusters they are part of. This stems from the definition of the converse relation. Given a relation *r(G, C)*, the expression

$$C \rightarrow C \mid r(G, C) \qquad (6)$$

describes the constriction of the general category C to the range of instances allowed in G. In string syntax, (6) will be expressed as

$$C\ [r'\ G]$$

and r’ will be referred to as the **converse relation** of r. Example:

$$color \mid r_2(ball, color) >> color\ [r'_2\ ball] \qquad (7)$$

If we now incorporate (5) into the definition (6), the general expression will be

$$\mathrm{C}\ [\mathrm{r}'\ \mathrm{G}\ \Sigma[\mathrm{r_j}\ \mathrm{u_j}]] \qquad (8)$$

When G $\Sigma[\mathrm{r_j}\ \mathrm{u_j}]$ describes a full disambiguation, the expression (8) will point to the unique content of *C* in *G*, i.e., it could be used to indirectly refer to a specific component instance in a category cluster.

The expressions (5) and (8), used to refer to resp. category clusters and component categories, could therefore be used by a sender to update the receiver's presumed representation. For example, if the receiver is believed to ignore that there is a ball on the sofa, then that information could be sent by means of (5), i.e.,

$$\text{!ball}\ [\mathrm{r_4}\ \text{sofa}]$$

where the symbol ***!*** denotes a new category cluster to be included in the receiver's representation. Such updating is a commonplace device used in natural language exchanges, to indicate, e.g., that a new character has appeared in a film, or a new guest has arrived at a party. If the receiver were presumed to know that there is a ball on the sofa but not its size, then that fact could be conveyed by means of the expression

$$\text{ball}\ [\mathrm{r_4}\ \text{sofa}]\ [\mathrm{r_3}\ \text{big!}]$$

where ***!*** now denotes an instance intended to be included in a category presumed to be empty at the receiver. In a converse situation, where the sender ignores some information item supposedly known by the receiver, the expressions (5) and (8) could also be used for querying purposes, by pointing to the required item by means of a different symbol, e.g.,

$$?\ [\mathrm{r_4}\ \text{sofa}]$$
$$?\ \text{size}\ [\mathrm{r'_3}\ \text{ball}\ [\mathrm{r_4}\ \text{sofa}]]$$

where the symbol ***?*** points to resp. an category or instance unknown by the sender. The referencing and updating uses of (5) and (8) could be summed up as follows

reference	G [r u], C [r' G]	(9)
updating	!G [r u], G [r !u]	(10)
querying	? [r u], ? C [r' G]	(11)

VIII. Databases as Category Clusters

The expressions (10) and (11) provide a means to resp. update and query a communicating party, provided that the latter's representation is consistent with the sender's. Therefore, whatever the spatial configuration of a database and the language used by it to refer to its items, string syntax communication will be possible if the database's content could be interpreted in terms of categories and category clusters. In order to explicitly formalize how that could be done, we shall use two different conceptual frames for the same data, i.e., a database D, configured in the form of tables, and an associated representation R, configured in terms of category clusters.

In a database *D*, an empty table T_θ consisting of the columns C_1, C_2, ... could be interpreted as a combination of the associated categories C_1, C_2, ..., and any instantiation of that combined category would describe a row (or potential row) in T_θ. For example, let the table T_k consist of the columns 'name', 'age', and 'address'. Because each of these can take *any* value within its respective scope, they can also be construed as categories, i.e.,

$$\text{name()} \wedge \text{age()} \wedge \text{address()}\ \ >>\ \ \mathrm{G}_\theta$$

where G_θ would be a category cluster associated to T_θ. By specifying values for those columns, a number of rows would be obtained, e.g.,

$$\text{name(Oz)} \wedge \text{age(39)} \wedge \text{address(7th Av.)} >> \text{row}_1$$
$$\text{name(John)} \wedge \text{age(54)} \wedge \text{address(97 St.)} >> \text{row}_2$$
$$\text{name()} \wedge \text{age(33)} \wedge \text{address(221B St.)}\ \ >>\ \ \text{row}_3$$

Now, if we use the connective ∨ to link the rows above, then the table T_θ could be interpreted as a category ambiguously referring to any of its rows. If we use the connective ∧ instead, then T_θ could be interpreted as a combination of rows, i.e.,

$$\text{row}_1 \vee \text{row}_2 \vee \text{row}_3 \vee ...\ \ >>\ \ \text{employee}$$
$$\text{row}_1 \wedge \text{row}_2 \wedge \text{row}_3 \wedge ...\ \ >>\ \ \text{employees}$$

In string syntax, both rows and cells within a row could be referred to by means of (5) and (8), e.g.,

$$\text{employee}\ [\mathrm{r_2}\ \text{age(33)}]$$
$$\text{age}\ [\mathrm{r'_2}\ \text{employee}\ [\mathrm{r_3}\ \text{221B St.}]]$$

where the relations r_2, r'_2, r_3 would be defined according to (4) and (6). In the general case, communication between a database and a user could be established in either direction as follows:

A. User to database

By reversing the rules used to derive conventional syntaxes, messages sent by users to a database *D* for updating or querying purposes could be expressed in string syntax by means of resp. (10) or (11). Such messages could be processed at *D* insofar as its tables could be interpreted as category clusters and those clusters would be consistent with the sender's. When that is the case, updating and querying could be interpreted in *D* as follows

String syntax	Database operation	
$\mathrm{G}\ [\mathrm{r_m}\ \mathrm{!u}]$	$\mathrm{N_1(u_1)} \wedge ... \wedge \mathrm{N_m(u} \leftarrow)$	(12a)
$\mathrm{!G}\ [\mathrm{r_m}\ \mathrm{u}]$	$\mathrm{N_1(u_1)} \wedge ... \wedge \mathrm{N_m(u)} \leftarrow \mathrm{N_m(u)}$	(12b)
$\mathrm{?G}\ [\mathrm{r_m}\ \mathrm{u}]$	$\mathrm{N_m(u)} \rightarrow \mathrm{N_1(u_1)} \wedge ... \wedge \mathrm{N_m(u)}$	(12c)
$\mathrm{?C_m}\ [\mathrm{r'_m}\ \mathrm{G}]$	$\mathrm{N_1(u_1)} \wedge ... \wedge \mathrm{N_m}(\rightarrow \mathrm{u})$	(12d)

where the symbols ← and → denote resp. the incorporation of a new item and the identification of an extant item. The updating operation would add resp. a value or a row to D, while the query would prompt D to identify resp. an item or a table, and then send the result to the querying party. Therefore, to be able to process string syntax expressions, a database should be configured so that either (a) the column headers in its tables reflect categories potentially referred to by the user, or (b) a sub-table could be identified in D for each category cluster that might be referred to by a user.

This is not uncommon. Meteorological and geolocation databases usually record data expected to be of interest for the general user, and databases containing spatial/temporal data most often lend themselves to semantic interpretation. As an example, let us define the category cluster G as follows:

...	h_1	h_1	H	h_2	...
...	T	t_1	T	t_2	...

which could be interpreted as describing a stay at the location h_1 for an indefinite time T until the time t_1, then some movement along some distance H during an indefinite time span T, and then the presence on a fixed location h_2 at the time t_2. From that cluster, the subclusters

h_1	H
t_1	T

H	h_2
T	t_2

could be denoted resp. as G_{depart}, G_{arrive}, implying the relations

$\mathrm{from}(G_{\mathrm{depart}}, \mathrm{loc})$
$\mathrm{at}(G_{\mathrm{arrive}}, \mathrm{time})$

The above relations could be used to construct a number of useful string syntax expressions, e.g.,

G_{depart} [from Rome]
G_{arrive} [at 09:23]

and therefore also updating and querying expressions, e.g.,

?time [r'_2 G_{arrive} [to Rome]]

For a database to be able to interpret such expressions, the adjacency relations in the sub-table

origin	destination	departure	arrival
Bonn	Rome	20:15	22:30

should be reconfigured so as to reflect the semantic relations in G, e.g.,

[origin]	H	[destination]
[departure]	T	[arrival]

so that, e.g., the sub-table

H	h_2
T	t_2

could be associated to the category cluster

$G_{\mathrm{arrive}}(\mathrm{loc}, \mathrm{time})$

The reconfigured table in *D* is actually a three-dimensional table, where the original columns are now arranged differently, i.e., only the topology of the table has been changed.

B. Database to user

The correspondences (12a-b) could reciprocally be used by *D* to derive reports expressed in string syntax, i.e.,

Database operation	String syntax
$N_1(u_1) \wedge ... \wedge N_m(\rightarrow u)$	G [r_m !u]
$N_m(u) \rightarrow N_1(u_1) \wedge ... \wedge N_m(u)$	!G [r_m u]

that would prompt the receiver to update her representation in response to the query previously sent, or by, e.g., a geolocation algorithm intended to keep a user updated about his surroundings. An example would hopefully illustrate the reporting process. In a meteorological database *M*, the column headers 'temp', 'humidity', 'loc', and 'time' could be associated a combined category that a user would interpret as a number of variables describing different weather states, i.e.,

Column headers	Combined category
temp humidity loc time	temp ∧ humidity ∧ loc ∧ time

A query intended to find out, e.g., the temperature in Paris at 22:05 would be expressed in string syntax as

?temp [r_1 Paris] [r_2 22:05]

In response to that query, the database would locate the row *R* having 'Paris' under the header 'loc' and '22:05' under the header 'time'. It would then retrieve from that row the cell under the header 'temp', and express the resulting value in string syntax as

R [r_3 !33ºC] [r_1 Paris] [r_2 22:05] (13)

If we use English words for the subindices, then we can write

r_1	r_{in}
R	$R_{a\_row\_in\_this\_database}$
r_2	r_{at}

ISBN: 978-1-61208-738-2

A few translation rules, together with (8), would convert (13) into the conventional syntax expression

> the temperature from a row in this database in Paris at 22:05 is 33ºC

However, the receivers need not even know that the data has been retrieved from some table in the source database. They have chosen to ask the source because they trust it to output reliable data. Therefore, the source might safely decide to just translate

> the temperature in Paris at 22:05 is 33ºC

This omission might seem like a trick shrewdly devised to get the desired result. On the contrary, it is an information compression device routinely used by natural language speakers. Consider just a few examples.

- the kitchen [of our house] is in the ground floor
- I can see the airport [of Beijing] now
- the book [you expressed an interest to buy three minutes ago] is *Finnegan's Wake*

IX. Communication with Deafblind Users

By applying or reversing the rules that define a conventional syntax (cf. Section IV), communications with a database could be established in any conventional syntax, including haptic languages such as the ones used by DB people [10]. As to the possible implementations, a portable device, that could physically change hands to send and receive messages by other human parties, would arguably provide a higher degree of autonomy than garments or other wearable devices. At the same time, it could be programmed to cope with the wide variety of languages and dialects used by DB people, due to local learning environments and different degrees of sensory impairment. But it could also be a means, or at least provide a stimulus, for the users to simply learn the rules of string syntax as a universal language. Its three basic elements, i.e., categories, instances and relations, could be readily expressed by means of tactile icons, and its syntax rules are simplest and intuitive, and might help DB users to enhance their knowledge of the world [4][15]. The author has devised an interface that demonstrates this. However, such an interface is sufficiently specific and detailed to be reported in a separate paper.

X. Conclusion

The categorizing feature of natural languages provides a means to refer to items in a data aggregate that is consistent with both conventional languages and databases. This could be the basis for a communication interface connecting DB people to (a) databases, either through actively querying or updating the database or by passively receiving reports from it, or (b) other human partners, by providing a portable means to send and receive messages without the help of an assistant. Additionally, a portable interface could also provide a starting point for both DB and non-DB people to use string syntax as a universal language.

References

[1] I. Androutsopoulos, G. D. Ritchie, and P. Thanisch, “Natural language interfaces to databases - an introduction,” Natural Language Engineering, 1(1), pp. 29–81, 1995.

[2] N. Caporusso, et al., “Enabling touch-based communication in wearable devices for people with sensory and multisensory impairments,” in: 1st International Conference on Human Factors and Wearable Technologies, pp. 149-159, 2017.

[3] E. F. Codd, “A Relational Model of Data for Large Shared Data Banks,” IBM Research Laboratory, San Jose, California, 1970.

[4] W. S.Curtis, E. T. Donlon, and D. Tweedie, “Learning behavior of deaf-blind children. Education of the Visually Handicapped,” 7(2), American Psychological Association, pp. 40-48, 1975.

[5] T. Hachisu, M. Sato, S. Fukushima, and H. Kajimoto, “HaCHIStick: simulating haptic sensation on tablet PC for musical instruments application,” Adjunct Proceedings of the 24th annual ACM symposium on User interface software and technology, pp. 73-74, 2011.

[6] E. Horowitz and S. Sahni, “Fundamentals of Data Structures,” Computer Science Press, 1983.

[7] J. Kramer and L. Leifer, “The talking glove,” SIGCAPH Comput. Phys. Handicap 39, pp. 12–16, 1988.

[8] M. Petruck, "Frame semantics," Handbook of pragmatics, pp. 1-13, John Benjamins Publishing Company, 1996.

[9] P. Z. Revesz and R-R. K. Veera, "A Sign-To-Speech Translation System Using Matcher Neural Networks," https://cse.unl.edu/~revesz/papers/ANNIE93.pdf, 1993.

[10] L. O. Russo et al., “PARLOMA – A Novel Human-Robot Interaction System for Deaf-Blind Remote Communication,” https://journals.sagepub.com/doi/10.5772/60416, 2015.

[11] C. Seror, “Human language and the information process,” https://zenodo.org/record/3263753#.XRmuceszaUk, 2019.07.10.

[12] C. E. Shannon, “The lattice theory of information,” in Report of Proceedings, Symposium on Information Theory, London, Sept., 1950," Institute of Radio Engineers, Transactions on Information Theory, No. 1, February, 1953, pp. 105-107..

[13] R. Vigo, “The GIST of concepts,” Cognition, Vol. 129, Issue 1, Elsevier, pp. 138-162, 2013.

[14] R. Wille, “Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies,” Formal Concept Analysis: Foundations and Applications. Ganter, B., Stumme, G., Wille, R. Eds. Springer, pp. 1-33, 2005.

[15] K. Wolff Heller, P.A. Alberto, and J. Bowdin, “Interactions of communication partners and students who are deaf-blind: A model,” Journal of Visual Impairment & Blindness, 89(5), pp. 391-401, 1995.

[16] M. Zerkouk, A. Mhamed, and B. Messabih, “A user profile based access control model and architecture,” http://www.airccse.org/journal/cnc/0113cnc12.pdf, 2013.

[17] B. Munat, “The Lowercase Semantic Web: Using Semantics on the Existing World Wide Web,” http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.3445&rep=rep1&type=pdfMay 2004.

Static and Dynamic Haptograms to Communicate Semantic Content

Towards Enabling Face-to-Face Communication for People with Deafblindness

Sándor Darányi, Nasrine Olson
Swedish School of Library and Information Science
University of Borås
Borås, Sweden
e-mail: {sandor.daranyi, nasrine.olson}@hb.se

Marina Riga, Efstratios Kontopoulos, Ioannis Kompatsiaris
Information Technologies Institute (ITI/CERTH)
Thessaloniki, Greece
e-mail: {mriga, skontopo, ikom}@iti.gr

***Abstract*—Based on the ontology developed in the ongoing SUITCEYES EU-funded project to bridge visual analytics for situational awareness and navigation with semantic labelling of environmental cues, we designed a set of static and dynamic haptograms to represent concepts for two-way communication between deafblind and non-deafblind users. A haptogram corresponds to a tactile symbol drawn over a touchscreen, its dynamic nature referring to the act of writing or drawing, where the touchscreen can take several forms, including a smart textile screen designated for specific areas on the body. In its current version, our haptogram set is generated over a 4 x 4 matrix of cells and is displayed on the back of the user, tested for robustness at the receiving end. The concepts and concept sequences simulating simple questions and answers represented by haptograms are focused on ontology content for now but can be scaled up.**

Keywords-deafblind communication; conceptual haptograms; word and sentence semantics; ontology.

I. INTRODUCTION

Communication with and between users with deafblindness is constrained by the medical nature of this disability, ranging from congenital to acquired deafblindness, including worsening sight or worsening hearing or both over time, plus, ultimately, symptoms of ageing as well. This renders parties with and without this condition in a difficult position. Below we focus on the severest case, congenital deafblindness, and propose a novel solution for improving the communication between user and trainer, but with the hope in mind that it can be used in the future between two such users too. In this use case, a new model of mutual understanding between the partners must be developed practically from scratch.

To this end we took inspiration from Lahtinen [1] and Lahtinen et al. [2], whose approach, while being expanded over the decades, basically reproduces ideograms on different regions of the body by a combination of hand strokes, gestures, pressure, etc. Branded as the social-haptic mode of communication, by default this is a rich tactile language with its own syntax and vocabulary of so-called *haptemes*, built from phoneme-like *haptices*, and tailored to a range of situations and topics of high practical importance including environmental descriptions, different situations, behaviour, the arts and advertisements to sum up a quick sampling. At the same time, due to its consensual nature, it is idiosyncratic and in need of being applicable in distance mode as well. This constraint makes it an ideal candidate for testing within the framework of the SUITCEYES EU-funded project [3], which is aimed at improving the quality of life for people with deafblindness through intelligent haptic technologies [4].

Another important parallel is McDaniel's PhD thesis [5], where he describes a different approach. As in a situation of sensory overload, touch is a promising candidate for messaging given that it is our largest sensory organ with impressive spatial and temporal acuity, there is need for a theory that addresses the design of touch-based building blocks for expandable, efficient, rich and robust touch languages that are easy to learn and use; moreover, beyond design, there is a lack of implementation and evaluation theories for such languages. To overcome these limitations, he proposed a unified, theoretical framework, inspired by natural, spoken language, called *Somatic ABC's* for Articulating (designing), Building (developing) and Confirming (evaluating) touch-based languages. To evaluate the usefulness of Somatic ABC's, its design, implementation and evaluation theories were applied to create communication languages for two very unique application areas: audio-described movies and motor learning. It was found that Somatic ABC's aided the design, development and evaluation of rich somatic languages with distinct and natural communication units.

Because the mission of SUITCEYES is to deploy a prototype which is wearable, combines situational awareness, visual analytics and face-to-face communication by the same ontology, and works in distance mode by default, our below approach is conceptual. Instead of haptemes to reproduce phonemes by graphemes by a combination of consecutive dots, dashes and strokes as in [2], we propose to use haptograms where the limited size and resolution of a body part as screen is counterbalanced by evolving patterns, i.e., the dynamics of signs. Our effort is in line with the approach by Israr and Poupyrev [6], building on their Tactile Brush approach, but focusing on language design by means of an ontology-compliant vocabulary vs. grammar, where the latter implements relational contextualization and sign sequencing. Thus, it belongs to the category of a priori defined spatial-temporal patterns in the semiotic vein.

The rest of the paper is structured as follows: Section 2 starts with an account of related research approaches, followed by Section 3 that provides a background on haptograms. Section 4 then introduces the SUITCEYES ontology that will play the role of the unified model for semantic integration of information from the environment, while Section 5 presents our approach for designing the haptogram vocabulary. Finally, Section 6 concludes the paper and gives insight into our future work directions.

II. Related research

The phonemic approach to a haptic language by Lahtinen et al. [2] finds support from the study by Chen *et al*. [7], where they investigated that decomposing spoken or written language into phonemes and transcribing each phoneme into a unique vibrotactile pattern enables people to receive lexical messages on the arm. A potential barrier to adopting this new communication system is the time and effort required to learn the association between phonemes and vibrotactile patterns. However, their study was limited to the learning of 100 patterns by different methodologies, displayed on the arm, and the concepts were not connected to an ontology.

On the other hand, Reed *et al*. [8] experimented with a new tactile speech device based on the presentation of phonemic-based tactile codes. The device consisted of 24 tactors under independent control for stimulation at the elbow to wrist area. Using properties that included frequency and waveform of stimulation, amplitude, spatial location, and movement characteristics, unique tactile codes were designed for 39 consonant and vowel phonemes of the English language. The participants, 10 young adults, were then trained to identify sets of consonants and vowels, before being tested on the full set of 39 tactile codes.

Walker and Reed [9] investigated several haptic interfaces designed to reduce mistakes in Morse code reception of 12 characters. Results concluded that a bimanual setup, discriminating dots/dashes by left/right location, reduced the amount of errors to only 56.6% of the errors compared to a unimanual setup that used temporal discrimination to distinguish dots and dashes.

Very much in line with what we would like to achieve, Israr and Poupyrev in [6] proposed Tactile Brush, an algorithm that produces smooth, two-dimensional tactile moving strokes with varying frequency, intensity, velocity and direction of motion. The design of the algorithm was derived from the results of psychophysical investigations of two tactile illusions, *apparent tactile motion* and *phantom sensations.* Combined together they allowed for the design of high-density two-dimensional tactile displays using sparse vibrotactile arrays. In a series of experiments and evaluations, they demonstrated that Tactile Brush is robust and can reliably generate a wide variety of moving tactile sensations for a broad range of applications.

III. Background on haptograms

Haptograms as a concept were introduced by Korres and Eid [10]. In their approach, “Haptogram” is a system designed to provide point-cloud tactile display via acoustic radiation pressure. A tiled 2-D array of ultrasound transducers is used to produce a focal point that is animated to produce arbitrary 2-D and 3-D tactile shapes. The switching speed is very high, so that humans feel the distributed points simultaneously. The Haptogram system comprises a software component and a hardware component; the software component enables users to author and/or select a tactile object, create a point-cloud representation, and generate a sequence of focal points to drive the hardware.

Our haptograms, on the other hand, are conceptual, and correspond to ideograms and logograms in the tactile domain, using evolving dot patterns instead of tactile shapes. Further, we distinguish between *stable* vs. *changing* patterns and call them *static* vs. *dynamic* haptograms in a communication context. Their purpose in our framework is to implement an ontology-constrained messaging language to convey visual analytics results, situation awareness assessments, and everyday conversation raw material outside of the scope of the above two areas. As these haptograms are to be mapped to the back of a vest made of smart textile, i.e., use that body area to display semantic content, the resolution of this screen places a limit on the conceptual vocabulary. Since screen resolution goes back to the number of actuators in a rectangular grid, as a proof of concept, in the current arrangement we designed a test vocabulary of both static and dynamic dot patterns conveyed to the body by vibration, pressure, heat, stimulated position, their combinations, and combination sequences to map short messages from an external sender. This approach can be scaled up either by increasing actuator density, or by generating virtual actuators [6].

IV. The SUITCEYES ontology

The key aim of the SUITCEYES ontology is to semantically integrate information coming from the environment (via sensors), and from the system’s analysis components (e.g., visual analysis of camera feed). In this sense, the ontology is primarily focused on semantically representing aspects relevant to the users’ context, in order to provide them with enhanced situational awareness and augment their navigation and communication capabilities. More importantly, the proposed ontology also serves as the bridge between environmental cues and content communicated to the user via the haptograms described in the next section.

In ontology engineering, it is common practice to reuse existing third-party models and vocabularies during the development of a custom ontology. We also followed this approach, in order to rely on previously used and validated ontologies. We, thus, adopted the semantic representation of objects and activities from the Dem@Care ontology [11], [12], which contains a set of descriptions of every-day activities and common objects used in an every-day context that are highly relevant to our goals. Moreover, we are relying on SOSA/SSN [13] for representing sensors and the respective observations, and on the Friend-Of-A-Friend (FOAF) specification [14] for representing persons and social associations. Finally, we integrated the SEAS (Smart Energy Aware Systems) Building Ontology [15], which is a schema for describing the core topological concepts of a building, such as buildings, building spaces and rooms.

A. Ontology Conceptualisation

Figure 1 displays an overview of the core ontology classes based on the Grafoo ontology visualization notation [16]: the yellow rectangles represent classes, while the green ones represent data properties (i.e., properties that take a raw data value, like, e.g., integers and strings). The prefixes in front of some of the class names indicate the namespace of the respective third-party ontologies, as mentioned above. Classes and properties that have no prefix belong to the core SUITCEYES ontology.

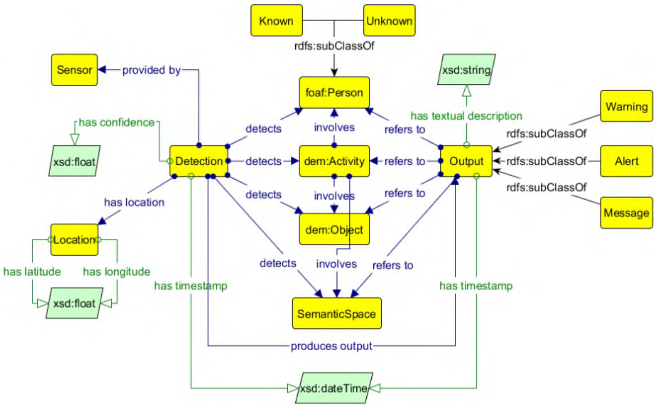


Figure 1. Overview of the core classes of the SUITCEYES ontology.

As indicated in the figure, class `Detection` is fundamental and refers to environmental cues (detected by the sensors) that have been instantiated in the ontology. A `Detection` instance may be associated with persons, objects, activities, and semantic spaces (more details on the latter follow next). The respective information is communicated to the user via class `Output` and its specializations: `Alert`, `Message`, `Warning`.

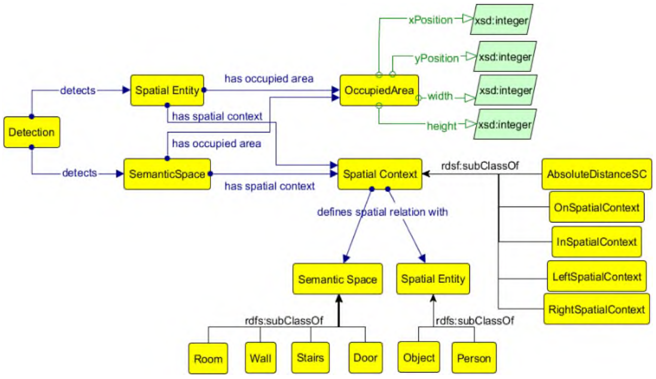


Figure 2. Semantic spaces and spatial contexts in the SUITCEYES ontology.

An entity that occupies space (e.g., persons, objects) is considered a `Spatial Entity` and the occupied space (e.g., a room or a location) belongs to the `Semantic Space` representation. These two aspects formulate the respective entity's `Spatial Context`, which provides information regarding the entity's relationship to the semantic space it is located in. Examples include: `in`, `on`, `left`, `right`, `far`, `close`, etc. The aforementioned concepts are depicted in Figure 2.

B. Sample Usage

Based on the ontological concepts presented above, Figure 3 illustrates a sample instantiation resembling an activity detected by the system's camera. The activity involves two people speaking to each other, one of them is known to the user (i.e., `john`) and the other is unknown. Moreover, these two people are currently located in the kitchen (i.e., `in_room_spatial_context`), and the respective message is communicated to the user via a textual description, which is then converted to haptograms as described in the next section.

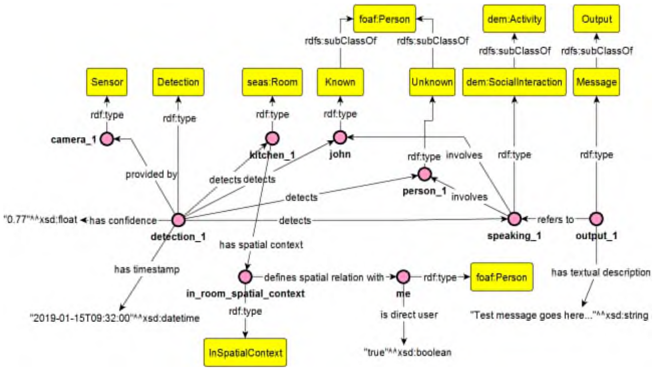


Figure 3. Sample instantiation of an activity involving two people discussing in the kitchen.

This flexible ontology-based representation described thus far allows the system to convey various types of information to the user. Below is an indicative list:

- Who is involved in an activity?
- Where is my mobile phone located?
- In which room am I now located?
- What objects are on the table?
- Which objects are observed on my left side?
- Which are the objects I am closer to/farther from?
- Alert! An obstacle (e.g., stairs) is in front of you!

V. HAPTOGRAM VOCABULARY DESIGN

Although in the next phase two-way communication will be our goal, in the current stage of the project our haptogramic vocabulary was designed for in-principle receiver testing over a 4 x 4 actuator grid. We were interested in finding out if the ontology and such a haptic conceptual vocabulary can be aligned, and how pattern sequences reminiscent of sentences can implement the transmission of more complex semantic content.

A. Examples

In our approach, haptograms can be static or dynamic, and can represent both word meaning and sentence meaning. Figure 4 illustrates the basic idea of the former version which was derived from the ASCII code table, where in our matrix cells, instead of characters, concepts are encoded.

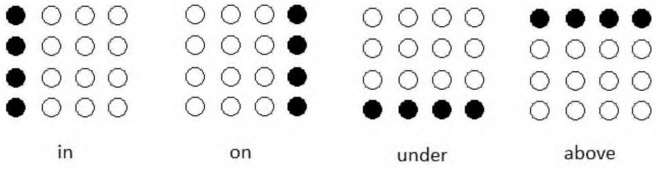


Figure 4. Sample static haptograms over a 4 x 4 actuator grid.

In Figure 5, we illustrate two sample dynamic haptograms. Above, in the matrix cells, the numbers indicate the firing sequence of the actuators for concepts (a) and (b), meaning "*stand*" and "*door*". Below the completed shape of the dynamic haptograms is indicated.

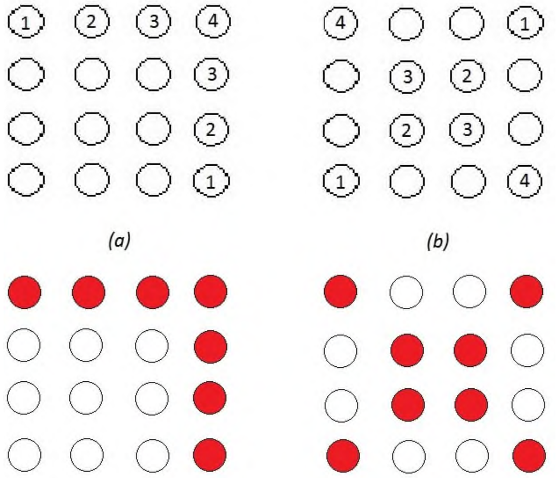


Figure 5. Unfolding sequences of two patterns over a 4 x 4 actuator grid, yielding different dynamic haptograms: (a) "*stand*"; (b) "*door*".

Moving over to sentence meaning, in Figure 6 we show how a simple statement, "*An unknown person is standing by/at the door*", can be made by concatenating static and dynamic haptograms. The statement begins with a single-blink sign, indicating the start of a new message, and finishes with a double-blink meaning end of transmission. It can be accompanied by a separate alert sign to add weight to the communicated content. Apart from this example, our test included questions and exclamations to enable a future dialogue between two users with deafblindness or a user with deafblindness and her/his trainer, family member, etc. Further, the vocabulary is both aligned with the ontology, and is including concepts and parts of speech not covered by the current version, i.e., indicates expansion opportunities. Likewise, e.g., logical operators, numbers, signs for operations etc. can be added following the same line of thought.

VI. Conclusion and future work

We plan to update the current pattern generator to a maximum of 9 x 9 actuator size matrices, subject to feasibility evaluation by psychophysics to make sure users are able to easily and consequently distinguish between the communicated haptograms, a prerequisite of noise-free or low-noise communication. This could include adding numbers and ways of calculation to the haptogram kit for example. Parallel to this, new concepts and relations from ongoing ontology development will be mapped to a more systematically designed, structured set of static and dynamic haptograms so that their semantics, including statements and limited argumentation, can be easier to follow by users. A mobile sender unit will be added to the receiving kit to enable two-way communication, and we aim to extend the framework to sending messages over a distance as well.

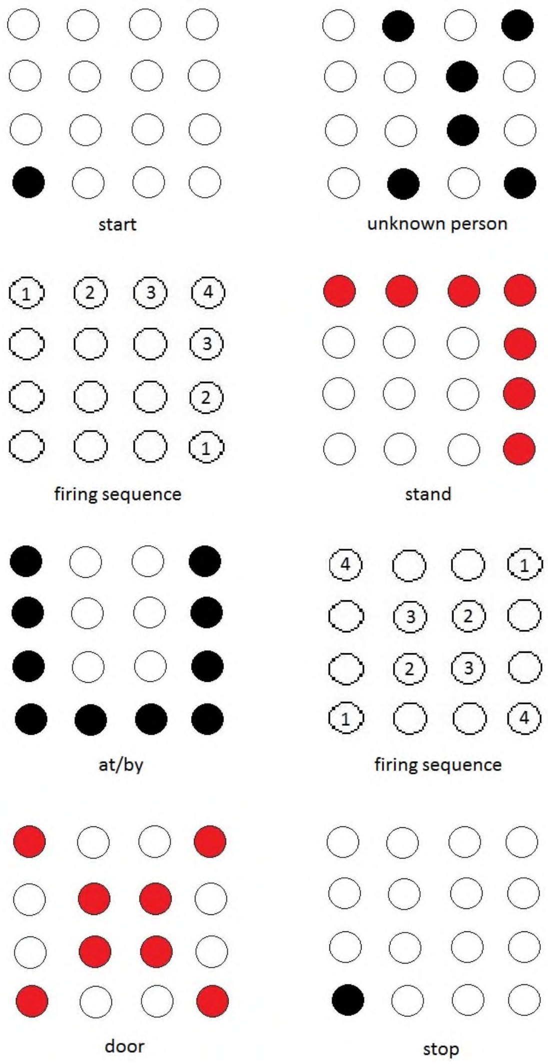


Figure 6. Sample statement constructed from static and dynamic haptograms over a 4 x 4 actuator grid: "Unknown person stand(ing) at/by (the) door".

In more detail, the current approach has its constraints by design, limiting incoming sensations and pattern recognition to the back of the individual. Given that this area is one of the less sensitive body interfaces for pattern recognition, an obvious way ahead will be to add more parts of the body as a screen, and combine pattern construction with diverse reception areas to extend the grammatical functionality of haptograms, while increasing the richness of the conceptual vocabulary. According to Lemmens et al. [17], for the torso, up to 62 sensors and actuators can be considered. This, combined with a more granular pattern generator, opens up new opportunities for a more systematic next effort, adding scalability to the approach, inviting knowledge graphs to replace ontologies, and increasing the number and complexity of situations to be described. One of the subsequent new challenges will be to match haptogram drawing on a mobile device by the sender over a much more sensitive surface, and its translation to the body.

ACKNOWLEDGMENTS

The authors are grateful to Sarah Woodin (University of Leeds) and Astrid Kappers (Eindhoven University of Technology) for helpful suggestions. This work has been partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 780814 SUITCEYES.

REFERENCES

[1] R. M. Lahtinen, Haptices and Haptemes: A Case Study of Developmental Process in Social-haptic Communication of Acquired Deafblind People. PhD dissertation, University of Helsinki, 2008.

[2] R. M. Lahtinen, R. Palmer, and M. Lahtinen, Environmental Description: For Visually and Dual Sensory Impaired People. A1 Management UK, 2010.

[3] http://suitceyes.eu/.

[4] O. Korn, R. Holt, E. Kontopoulos, A. M. Kappers, N. K. Persson, and N. Olson, "Empowering Persons with Deafblindness: Designing an Intelligent Assistive Wearable in the SUITCEYES Project," Proc. 11th PErvasive Technologies Related to Assistive Environments Conference, ACM, 2018, pp. 545-551.

[5] T. L. McDaniel, Somatic ABC's: A Theoretical Framework for Designing, Developing and Evaluating the Building Blocks of Touch-based Information Delivery. PhD dissertation, Arizona State University, 2012.

[6] A. Israr, and I. Poupyrev, "Tactile Brush: Drawing on Skin with a Tactile Grid Display," Proc. SIGCHI Conference on Human Factors in Computing Systems, ACM, 2011, pp. 2019-2028.

[7] J. Chen, R. Turcott, P. Castillo, W. Setiawan, F. Lau, and A. Israr, "Learning to Feel Words: A Comparison of Learning Approaches to Acquire Haptic Words," Proc. 15th ACM Symposium on Applied Perception, ACM, 2018, p. 11.

[8] C. M. Reed et al., "A Phonemic-Based Tactile Display for Speech Communication," IEEE Transactions on Haptics, 12(1), pp. 2-17, 2018.

[9] M. Walker, and K. B. Reed, "Tactile Morse Code Using Locational Stimulus Identification," IEEE Transactions on Haptics, 11(1), pp. 151-155, 2017.

[10] G. Korres, and M. Eid, "Haptogram: Ultrasonic Point-cloud Tactile Stimulation," IEEE Access, Vol 4, pp. 7758-7769, 2016.

[11] G. Meditskos, S. Dasiopoulou, and I. Kompatsiaris, "MetaQ: A Knowledge-driven Framework for Context-aware Activity Recognition Combining SPARQL and OWL 2 Activity Patterns," Pervasive and Mobile Computing, Vol 25, pp. 104-124, 2016.

[12] G. Meditskos, and I. Kompatsiaris, "iKnow: Ontology-driven Situational Awareness for the Recognition of Activities of Daily Living," Pervasive and Mobile Computing, Vol 40, pp. 17-41, 2017.

[13] K. Janowicz, A. Haller, S. J. Cox, D. Le Phuoc, and M. Lefrançois, "SOSA: A Lightweight Ontology for Sensors, Observations, Samples, and Actuators," Journal of Web Semantics, Vol 56, pp. 1-10, 2019.

[14] D. Brickley, and L. Miller. *FOAF Vocabulary Specification 0.99. Namespace Document*. [Online]. Available from: http://xmlns. com/foaf/spec/ 2019.08.13.

[15] M. Lefranois, "Planned ETSI SAREF Extensions based on the W3C&OGC SOSA/SSN-compatible SEAS Ontology Patterns," Workshop on Semantic Interoperability and Standardization in the IoT, SIS-IoT, Amsterdam, Netherlands, 2017.

[16] R. Falco, A. Gangemi, S. Peroni, D. Shotton, and F. Vitali, "Modelling OWL Ontologies with Graffoo," Proc. European Semantic Web Conference, Springer, Cham, 2014, pp. 320-325).

[17] P. Lemmens, F. Crompvoets, D. Brokken, J. van den Eerenbeemd, and G. J. de Vries, "A body-conforming Tactile Jacket to Enrich Movie Viewing," Proc. 3rd Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2009, p. 7.

A State of the Art Survey: Business Cases Based on Semantic Web Technologies in Healthcare

Vivi Ntrigkogia, Thanos G. Stavropoulos,
Ioannis Kompatsiaris
email: {vividrig, athstavr, ikom}@iti.gr
Centre for Research & Technology Hellas,
Information Technologies Institute
Thessaloniki, Greece

Maro Vlachopoulou
email: mavla@uom.gr
University of Macedonia,
Department of Applied Informatics
Thessaloniki, Greece

*Abstract***— Semantic web technologies promise to facilitate a long-awaited paradigm shift in the healthcare industry towards more efficiency, extended interoperability and intelligent analysis capabilities. This paper presents a critical review of current healthcare businesses incorporating semantic web technologies. Initially, it presents the potential of semantic web technologies in general and the challenges to apply them in business. State of the art businesses that use semantics to provide innovative healthcare services are pinpointed and critically reviewed. Through an analysis of their aspects, their semantic component, business models, target audience and value propositions, the competitive advantage and tangible business value of semantics in healthcare is revealed, serving as a reference for the growing number of emerging solutions in the near future.**

Keywords-semantics; semantic web; ontologies; eHealth; healthtech; business models; healthcare;

I. INTRODUCTION

Healthcare can largely benefit from technology throughout its lifecycle: from diagnosis, to hospitalization, prescription, treatment and prevention. However, complex governance structures, heterogeneity and lack of convergence in the healthcare industry are few of the reasons that such benefits of technological applications are slowly emerging. While data mining techniques can aid in certain sectors, such as diagnosis and prevention, it is the semantic web technologies that show great potential to resolve the interoperability problem, which, in turn, would benefit healthcare throughout its lifecycle.

The semantic web technologies [1] provide methods and tools to define the semantics, i.e., hierarchical models and relations between data of any particular domain in an interoperable, machine-interpretable format. The former property, interoperability, allows communications between disperse and vendor-agnostic systems on the basis of commonly established and agreed upon models, referred to as ontologies. Interoperability use case scenarios include patient data exchange between hospital and medical practice systems, exchange of clinical trial data and much more. The latter property, machine-interpretability, gives intelligent software applications the potential not only to read, but also to understand information in a way that humans do. The so-called reasoning process can utilize ontologies to deduce new information. The technology can be applied to any sector by domain experts designing a domain ontology, i.e., the taxonomy particular to the problems, peculiarities and terminology of the pertinent domain. This includes healthcare [2], where semantic web technologies have the potential to lift ailments in data sharing, Electronic Health Record exchange [3], reasoning on epidemic and patient record and drug database information [4] for prescription, prevention, care, etc. achieving cross-provider and even cross-border healthcare.

This paper presents an overview of semantic web technology applications in the healthcare industry so far, giving valuable insights, highlighting the benefits and paving the way to future achievements towards more interoperable healthcare with great clinical, economical and societal impact.

The next Section presents technological challenges and trends of semantic web technologies in several applications, including healthcare. Section 3 presents current business challenges for tech innovation and their relation to semantics. Section 4 presents a critical analysis of several concurrent businesses that offer healthcare services based on semantics. Section 5 presents conclusions from this analysis and future work.

II. TECHNOLOGICAL CHALLENGES AND TRENDS

A recent report by Gartner [5] highlights how semantic-based technologies add critical context to data. According to it, the proliferation of data poses huge challenges for businesses that want to leverage their data assets. Data scientists and business managers are advised to opt for a semantic approach in order to gain competitive advantage. Semantic technologies and knowledge graphs provide (a) the toolbox to facilitate decision making by actually making sense of rapidly growing pools of data, (b) the basis for artificial intelligence and machine learning applications (c) the possibility to make relationships and interconnections of this data with evident benefits on developing new tools and synergies [6].

Thus, Knowledge Management and semantic web technologies can be greatly beneficial to the knowledge society, i.e., a society in which the creation, dissemination, and utilization of information and knowledge has become the most important factor of production. In such an environment, knowledge assets are the most powerful producer of wealth, sidelining the importance of land, the volume of labour, and physical or financial capital. This vision requires an extensive analysis of factors and actions that promote the value of knowledge and specify critical prerequisites for the design and implementation of human centric information systems and services. However, this still pertains several challenges. Lytras et al. [5] state that as extensive communication and networking infrastructures are now implemented, a critical shift is required from the relevant verbalism to applied strategies and technologies. As literature supports “a key question within the context of the knowledge society is how we can redesign the basic structural models of the information provision to target effective models of support. Personalisation and Adaptation are only a few terms aiming to promote the need of multidimensional and non-monolithic approaches” [7].

Several trends emerge around technology to represent the semantics of data. Knowledge Graphs have gotten a lot of attention as a backbone for Machine Learning, Deep Learning, and AI business use cases a trend that is expected to evolve. From a business perspective, it looks like more and more industries (agriculture, healthcare, smart cities, finance, etc.) are pursuing Semantic Technologies, often relying on Knowledge Graphs. Among the semantic-driven AI ventures those related to the healthcare space are forecasted to blossom. Using semantics to drive chatbots is also emerging [8]. Nature Language Processing (NLP) is increasingly based on ontologies to represent the semantics and understanding of speech and dialogue [9]. Data governance procedures, finally, also pertain ontologies and structured semantics.

The digitisation of many industry sectors requires information models describing assets and information sources to enable the semantic integration and interoperable exchange of data. Although this vision has gained much traction lately in many sectors, e.g., creative industry, healthcare industry, manufacturing, etc., it is still not clear how it can actually be implemented in an interoperable way using concrete standards and technologies.

Initiatives for content representation and linking can be exploited in various domains where there is a need to aggregate and fuse information in different levels of abstraction.

A prominent example is the Healthcare domain, where there is a need to coupling profile, behavioural and health knowledge to achieve human awareness and assist clinical experts in assessing the health condition of patients and help adjust and update the care plan and interventions. Another example are intelligent virtual agents, where there is a need to fuse verbal and non-verbal information, e.g., utterances, gestures and emotions, to achieve conversational awareness and provide meaningful responses to the users.

Pervasive environments, often met in healthcare-at-home technological solutions, such as Active and Healthy Ageing (AHA) and Ambient Assisted Living (AAL), present the need to deploy, manage, collect and analyse multimodal sensor observations from the Internet of Things (IoT) [10]. Semantic integration and reasoning solutions have been implemented in home care, in systems such as the AAL solutions of Dem@Care [11], i-PROGNOSIS in occupational health, such as in Healthy@Work [12] Active@Work [13] and Fit4Work [14] smart environments, such as MARIO [15] and the AHA solutions of ACTIVAGE [16].

III. Business Challenges and Semantics

Marketing technology-based, innovative products and services is a far more complicated and challenging task compared to more conventional business cases. Even the launch of new high-tech products is challenging, as usually consumers are more skeptical. Moreover, high tech startups often find it very difficult to define target market and often focus solely on R&D and expect that their super product will sell by itself. To be a successful global brand in the 21st century, a hi-tech firm needs to be market oriented, agile and locally relevant. The proliferation of smart mobile devices and the mass adoption of social media have created an utterly new marketplace where a new consumer has emerged. Spurring innovation adoption can be hard. The high tech sector that is constantly evolving and rapidly changing may cause a multitude of complications.

The central benefit of the semantic web is that it enables the extraction of knowledge patterns and useful information from unstructured content. It further empowers interoperability, integration of multiple and differentiated data sources. It appears that by integrating semantic technologies in the business pipeline, business performance is likely to be improved while new business models emerge. Literature suggests that the impact of semantic web on business performance is related to: a) Less labour hours (20-80%); b) Less cycle time (20-90%); c) Less set up (25-80%); d) Quality gain (50-500%); e) Productivity gain (2-50X; f) Increased return of assets (2-25X; g) Revenue growth (2-30X); h) Reduction of total cost of ownership (20-80%); and i) positive ROI (Return On Investment) over 3 Year (2-300X) [17] [18].

Although the benefits are clear, creating business value using semantic technology is in many ways no different from creating business value with any type of new technology. Literature suggests that there are three important aspects to take into account [19]:

Customers: companies need to understand their customers, their target group and their specific needs. In today’s knowledge based economies, corporations cannot rely only on themselves to deliver innovation. Co-creation is the tool to open innovation. An emerging cross-sector model is to bring the outside in and bridge the gap between the consumers and the brand. If not the most, it is definitely one of the best ways to engage consumers in the company, increase brand recognition and have tangible results: an increase in sales. Co–creation is the path to have increased flow of quality ideas and concepts into a firm's development pipeline.

Business models: High-Tech organizations should carefully consider business models. Companies that succeed in coupling cost-reducing technologies with innovative business models to deliver increasingly affordable and accessible products and services will gain competitive advantage [17] [18].

Technology: it is important to have a differentiated product or service that actually adds value to the end consumer. To this end, the product offering must not be easily replicated and IPR (Intellectual Property Rights) should be carefully managed.

There are, however, additional challenges to the market adoption of semantic technologies. Although significant amounts of money have been invested in the development of novel semantic technologies, industry uptake appears to have not reached its full potential. This is partly due to the fact that enterprises are unaware of the benefits that semantics can bring about to their offerings. There are specific aspects of Semantic Technologies that may explain why it is hard for these technologies to be adopted by enterprises in the mainstream market [2]:

(1) Semantic technologies are hard to explain
(2) It is not easy to describe how Semantic Technologies might fit within a business
(3) There is a lack of innovation in semantic business models. These challenges impose obstacles to the market adoption of semantic technologies.

The next Section presents instances of how those challenges are met in real applications of semantic web technologies for healthcare technology.

IV. BUSINESS CASES BASED ON SEMANTIC WEB TECHNOLOGIES FOR HEALTHCARE

Recent advances in pervasive computing and sensor technologies have enabled the contextualised enrichment of business processes capitalising on the ability to sense, process, combine and interpret data of different modalities. Nowadays, it is getting harder to extract useful information from the enormous amount of data that is being collected in the medical information systems or eHealth systems due to the distributed and very complex nature of this data [20].

A common question is why - given the advent of IoT, AI and multiple sophisticated medical technologies introduced each year - healthcare has not been disrupted to a significant degree up to now. A reason is that healthcare remains expensive and inaccessible to many because of the lack of business model innovation. In healthcare, most technological enablers have failed to bring about lower costs, higher quality and greater accessibility. It is believed that semantic technologies can play a pivotal role in guiding growth among healthcare businesses [19].

This Section initially presents prominent examples of semantic web technology applications in the healthcare industry, along with their pertinent value propositions, business models and target audience. It then presents an analysis over them, regarding the competitive advantage and tangible business value of semantics for a healthcare business.

A. *Healthcare Businesses using Semantics*

This Section presents companies or startups whose core component and business model relies on semantic technologies. These business cases, prominent to the best of our knowledge, have been identified via searching the web and specialized startup databases. Our objective is to create a listing of existing "semantic" healthtech companies and identify their value proposition, their business model, as well as the means of how semantic technologies add value to the corporation and the end consumer. Table 1 provides an overview of the aforementioned listing, while each business case is examined in detail below.

In-JeT [21] is a research company and service provider in the area of internet-based healthcare services such as telemedicine. The company offers LinkSmart® a set of middleware *telemonitoring applications able to interconnect devices, people, terminals, buildings, etc*. The Service-Oriented Architecture (SOA) and its related standards provide interoperability at a syntactic level. On top of it, the LinkSmart® middleware provides interoperability also at the semantic level. This is achieved through a semantic model-driven infrastructure, whereby services exposed by devices can be described and consumed by various applications in Ambient Intelligence, Pervasive Computing, Ubiquitous Computing, Mobile Computing and Cloud Technologies. The company's various software and hardware assets are sold together with consultancy services.

Life Semantics Corp [22], a Korean health-tech startup, has developed a platform that integrates *health record data scattered among hospitals, governments, and corporations* to create a Personalized Health Management (PHM) platform that can prepare itself for the upcoming diseases based on a disease prognosis prediction algorithm. Life Semantics developed the first commercial PHR-based data platform called LifeRecord. It develops Hospital Information Systems (HIS) and semantic web technology for application in the area of life sciences.

Hi3 Solutions [23] provides HIT products, education, and consulting services that enable clients to engage effectively in *health information exchange, health data integration, and health care quality measurement* required to establish and comply with evidence-based best practices in health care. The mission of Hi3 Solutions is to accelerate widespread adoption and compliance with emerging HIT standards by offering the information integration infrastructure necessary to enable the use of health information exchange standards, meaningful health care quality and performance measures, and standardized clinical decision support capabilities.

Intrepid Analytics [24] is using text analytics and custom medical ontologies, to be able to analyze online posts and reports to *anonymously track disease spread in near real-time,* as well as attributes, such as the medicines reported to be taken and subsequent reactions. This helps healthcare and regulatory bodies to stay on top of the quickly changing healthcare landscape.

Ontoforce [25] developed Disqover, a *semantic search platform that integrates various life-sciences data*. The platform uses semantic web technologies including ontologies in RDF and LinkedData, additionally supported by an indexing engine. The platform integrates private, public, and third-party data resources, all searchable via a single interface. Search results are enhanced by predefined data types. Ontoforce provides an integrated search of 80+ databases. The company also provides customizable visualizations: graphics, plots, tables, charts and maps.

Mendelian's [26] online technology addresses the needs of patients, physicians, providers, payers, and pharma. They provide for the best tools to *get the right diagnosis with speed and accuracy*. By continuously adding, curating and analysing conditions, symptoms, and genes along with clinical tests they aim to build the most comprehensive Rare Disease Knowledge Base. The patient can fill in a questionnaire with their signs and symptoms via an online form. The Mendelian engine processes then the information and provides a link to a detailed report to share with the doctor. The process is similar for physicians. Doctors enter patient's symptoms and clinical features. The input is processed semantically and provides an output with likely causative diseases genes and mutations.

SemanticMD [27] enables customers to *find, connect and license medical imaging data with expert annotations*. Customers can automate their data collection as well as use NLP to annotate radiological and clinical reports for search and analysis. SemanticMD Annotate enables teams to organize medical image annotation projects in a fun, flexible way and output the results for easy analysis by machine learning algorithms.

Teamarrayo [28] leverages the value of existing data sources, both internal and external, to transform them into ontologies and appropriate data models. In turn, it enables *data management and added value from data processing tools*. They provide bioinformatics services that include data curation, informatics ontologies (i.e., Gene Ontology, ChEBI, etc.). Their know-how also pertains to loading and utilizing large public data sources such as 1000 Genomes, TCGA, CCLE, and others.

Ontotext [29] is a company that utilizes semantic medical coding of patient records to help transform the raw patient data into structured knowledge. Its pipelines are designed to process large volumes of *patient records and to extract and semantically index data about patient diagnoses, treatments, medications and events timing*. All extracted medical data is normalized to resolvable instances from the medical Knowledge Graph. Thus, the extracted information is ready to be semantically fused with the LinkedData generated from multiple references public dataset (covering disease and symptoms, anatomical structures, generic drugs and products and much more).

Pangaea [30] is a domain expert in *bioinformatics, molecular biology, data engineering and machine learning*. The company aspires to help life science companies determine 'what data exists' and organize it in specific scientific or clinical contexts. Thereby, they are able to analyze and interpret it effectively, making the most from their investment in such data.

Seminte [31] provides assistance in preparing products for new markets by the use of international terminologies. SemInte assists in mapping an existing product's interface language to international terminology or in the development of a new product. The *process ensures that products are compliant to standards, data can be reused, compared and exchanged across third-party systems*, e.g., for EHR. SemInte identifies data required in specific documentation, creating datasets based on terminology e.g., SNOMED-CT, ICD-10, etc. needed, facilitates quality review process and helps with the technical dialogue with the vendors who shall implement the exchange standards (HL7-CDA and IHE-XDS) and data sets.

Healx [31] is a biotechnology company integrating artificial intelligence with expert pharmacology to *discover treatments for rare diseases, to share assets and to accelerate their uptake by clinical trials* [32] within as soon as two-years time. To achieve this, Healx has developed a comprehensive AI-based drug discovery platform for rare diseases, named HealNet. Their revenue model is asset sharing (e.g., clinical trial databases) across individuals and groups.

B. Discussion

TABLE I. HEALTHCAE AND TECHNOLOGY COMPANIES BASED IN SEMANTIC WEB TECHNOLOGIES AND THEIR PERTAINING ATTRIBUTES.

Company	**BUSINESS MODEL**				
	Semantic Component	***Value Proposition***	***Domain***	***Revenue Model***	***Target Customer***
In-JeT	Ontologies, Semantic Annotation, Semantic Middleware	Support patients in managing their chronic diseases efficiently and help healthcare professionals provide better care with more frequent, reliable and relevant data about health status.	Health Management, Telemedicine/AAL	Asset Sale, Consultancy	B2C, B2B
Life Semantics Corp	HL7 FHIR	Offer a total health management solution through a platform by utilising collaboration models with various healthcare related industries like insurance, finance, food and fitness.	Health Management, EHR	Asset Sale	B2B
Hi3 Solutions	HL7/v2/CDA/FHIR, HIT Standards	Health Information Technology vendor. They provide HIT products, education, and consulting services that enable their clients to engage effectively in health information exchange, health data integration, and health care quality measurement	Health Information Exchange, Healthcare Quality Measurement	Asset Sale, Consultancy	B2B
Intrepid Analytics	Data Mining, Medical Ontologies	Offer an AI platform focused on the healthcare industry- specifically for the biotech industry and patients. The platform supports the ingestion and organization of molecular and drug information. Home-grown medical ontologies support the integration and classification of data sets.	Health Information Exchange, Disease Tracking	Asset Sale, Consultancy	
Ontoforce	Semantic Search, Ongologies, LinkedData	Effortlessly extract "information" from public, third party and private big data and present them in a way they can be easily interpreted and used to support smart decisions.	Health Data Discovery, Health Data Visualization	License-based	B2B
Mendelian	Rare Disease Knowledge Base	Rare Disease Diagnosis, Faster - A search engine for rare diseases.	Health Data Discovery, Rare Disease Diagnosis	Freemium	B2B, B2C
Semantic MD	Semantic Annotation, Semantic Search, NLP, Ontologies (SNOMED, ICD-9/10)	SemanticMD provides a SaaS-based platform that enables the rapid training of medical image analysis applications and classifiers.	Health Data Discovery, EHR	SaaS	B2B
Teamarrayo	Data Mining, Ontologies (Gene Ontology, ChEBI)	Accelerate scientific research by providing solutions for data consolidation, management and visualization to scientists and clinicians.	Health Data Discovery, Health Data Visualization	Data as a Service	B2B
Ontotext	Data Mining, Semantic Annotation, LinkedData	To transform how organizations identify meaning across diverse databases and massive amounts of unstructured data by combining a semantic graph database with text mining, and machine learning.	Health Data Discovery, EHR	License-based	B2B
Pangaea	Data Mining, Ontologies	Pangaea's value proposition is that it helps end users such as scientists, clinicians and researchers with little or no IT experience to find 'what data exists' and execute their analysis from a single web portal regardless of underlying tools and applications.	Health Data Discovery	Asset Sale	B2B
Seminte	Ontologies (SNOMED-CT, ICD-10), HL7-CDA, IHE-XDS	Making Healthdata sharable and compareable	Health Information Exchange, EHR	Asset Sale, Consultancy	B2B, B2G
Healx	Data Mining, Ontologies	Healx's value proposition is about *asset sharing* (for example, making available clinical-trial databases that record the effectiveness of most drugs across therapeutic areas and diseases, including rare ones). Healx promises more *personalization* by revealing drugs with high potential for treating the rare diseases covered.	Health Data Discovery, Rare Disease Diagnosis	Asset Sharing	B2B

After examining real-world business cases of healthcare and technology, the benefits of semantic web technologies to them are clear. We may conclude that the findings from this state-of-the-art survey regarding those benefits are in line with previous studies in literature [17] [20], outlining the advantages and performance boosts in business due to semantic web technologies. Specifically, our review pinpoints the following benefits and advantages to the respective business cases examined:

- **The maximization of the value of information:** data online and offline is in abundance. As proved by the cases of Interpid Analytics, Teamarrayo, Mendelian, Pangaea and Healx semantic technologies can assist in making sense of these data, extracting knowledge patterns or detecting previously unknown trends or details, that could be leveraged for disease management, new clinical trials and rare disease diagnosis. Health management solutions like In-Jet and Life Semantics Corp further prove how a researcher, medical practitioner or patient can maximize the value of information (for example coming from a variety of sensors that monitor activity, bio-signals etc) for effective self-monitoring.

- **Facilitated information diffusion:** Semantic search bridges the gap between humans and machines, and takes us further on a quest for meaningful information and knowledge discovery. The business cases of Hi3 Solutions, Ontoforce, SemanticMD, Ontotext that all constitute successful corporate examples that add value through knowledge modeling and flexible information sharing. When data is released from individual applications the diffusion of knowledge is empowered.
- **Greater level of future-proofing and re-use**: to illustrate how business performance is strengthened by utilizing semantics for future-proofing and re-use, we will utilize the example of rare diseases. Rare diseases can take many years to diagnose. This represents an odyssey for patients, a challenge for physicians, a headache for care providers, a waste of resources for payers and missed opportunities for pharma. Diagnosing Rare Diseases is no small feat. Indeed, according to Mendelian, it takes on average 8 years and 4 specialists, often involving misdiagnoses. The fact is that there are over 8,000 rare conditions, the information on them is scattered across multiple sources and new research is published every day. By leveraging knowledge graphs and proprietary semantic web technologies, healthcare technology providers like Mendelian and Healx have the opportunity to extract new phenotypes from recent publications, access results from past studies so as to guide clinical investigations and assist in diagnosis. Such a knowledge structure can have another side effect. It is not rare in clinical trials that researchers discover that a drug is more effective on treating a completely different symptom (the case of Viagra constitutes a well-known example in this respect). Providing medical researchers and clinicians the ability to search and identify such cases easily and in a meaningful context by leveraging knowledge graphs and ontologies can evidently facilitate decision support, re-use of knowledge and clinical interventions.

From a business perspective, semantic adoption is still in its infancy, though the potential is huge. Most of the companies in our study are startups, which means that they are still developing their business models and their viability depends on funding. However, the fact that many of them, such as Healx, have raised millions of euros to scale up shows that investors and industry experts are eager to invest in such initiatives and believe in their sustainability.

As in healthcare technology knowledge extraction and information integration is pivotal for success, startups should consider adding a semantic component to their product suite and develop a business model based on a strong competitive advantage. Business models from the domain of e-retailers and electronic stores are the most common among high tech market providers. However, new semantic business models need to be different so as to address new customer needs and add value across the buyer's journey.

V. Conclusion and Future Work

To meet the challenge for high quality and efficient care, highly specialized and distributed healthcare establishments have to communicate and co-operate in a semantically interoperable way. Despite the complexity of current semantic web technologies, several businesses have realized the vision of bringing research to the industry and applied these technologies for profit. After examining real business cases and their pertaining technological and business aspects the benefits of this practice are clear. Technologies such as Data Mining, Semantic Annotation and Search, Ontologies and LinkedData already provide tangible solutions to problems such as Health Management, Telemedicine, Health Information Exchange, EHR and Health Data Discovery, servicing not only healthcare but also the life science research.

As for future research directions, we consider expanding the survey but also diving deeper into categorizations and analysis of criteria. Many more business cases are emerging and have to be added to future more in-depth reviews. In parallel, this review has only scratched the surface it terms of criteria, categories and clustering of the various semantic components and domains, i.e., problems the companies solve, from a technological perspective, as well as the value propositions, revenue models and customer bases from a business perspective. Finding the pertaining groups and strategies of companies in a more in-depth analysis survey can reveal significant trends and methods for applied semantic web technologies in healthcare in the coming future, with real, tangible business value.

Acknowledgment

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-02668, support2LIVE - ypostiriZO), RADAR-AD - IMI2 - Grant Agreement No. 806999 and H2020-IOT-732679 ACTIVAGE.

REFERENCES

1. Berners-Lee T, et al., (2001) The semantic web. Sci Am 284:28–37
2. Natoli J., Kermanshahche K. and Painter J. (2014) Healthcare semantic interoperability platform. US Pat App 14/
3. Garde S., et al (2007) Towards Semantic Interoperability for Electronic Health Records. Methods Inf Med 46:332–343. https://doi.org/10.1160/ME5001
4. Wild D., et al Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research. Elsevier
5. How to Use Semantics to Drive the Business Value of Your Data. https://www.gartner.com/en/documents/3894095/how-to-use-semantics-to-drive-the-business-value-of-your. Accessed 15 Jul 2019
6. 7 Ways Semantic Technologies Make Data Make Sense - InformationWeek. https://www.informationweek.com/big-data/big-data-analytics/7-ways-semantic-technologies-make-data-make-sense/d/d-id/1323580. Accessed 18 Jul 2019
7. Lytras M.D., Sakkopoulos E. and De Pablos P.O. (2009) Semantic web and knowledge management for the health domain: State of the art and challenges for the seventh framework programme (FP7) of the European Union (2007-2013). Int J Technol Manag 47:239–249
8. Semantic Web and Semantic Technology Trends in 2019 - DATAVERSITY
9. Zhang J. and El-Gohary N.M. (2017) Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking. Autom Constr 73:45–57. https://doi.org/10.1016/j.autcon.2016.08.027
10. Stavropoulos T.G., Meditskos G. and Kompatsiaris I. (2017) DemaWare2: Integrating sensors, multimedia and semantic analysis for the ambient care of dementia. Pervasive Mob Comput 34:. https://doi.org/10.1016/j.pmcj.2016.06.006
11. The Dem@Care Project. http://www.demcare.eu/. Accessed 25 Jul 2019
12. healthy@Work AAL Programme. http://www.aal-europe.eu/projects/healthywork/. Accessed 20 Jul 2019
13. Active@Work AAL programme. http://www.aal-europe.eu/projects/activework/. Accessed 21 Jul 2019
14. Fit4Work AAL
15. The Mario Project. http://www.mario-project.eu/portal/. Accessed 21 Jul 2019
16. ACTIVAGE project. https://www.activageproject.eu/. Accessed 21 Jul 2019
17. Benjamins VR, Radoff M. and Davis M, et al (2011) Semantic Technology Adoption: A Business Perspective. In: Handbook of Semantic Web Technologies. pp 619–657
18. Cifliganec Dimitar and Trajanov B. (2011), Semantic Web Business Models, The 8th International Conference for Informatics and Information Technology (CIIT 2011)
19. Hwang J. and Christensen C.M. (2008) Disruptive innovation in health care delivery: A framework for business-model innovation. Health Aff.
20. Dogdu E. (2009) Semantic web in eHealth. In: Proceedings of the 47th Annual Southeast Regional Conference, ACM-SE 47
21. In-JeT ApS. https://www.in-jet.dk. Accessed 15 Jul 2019
22. Life Semantics Corp. https://www.lifesemantics.kr. Accessed 15 Jul 2019
23. hi3Solutions. http://www.hi3solutions.com. Accessed 15 Jul 2019
24. Intrepid Analytics. https://www.intrepid-analytics.com/. Accessed 15 Jul 2019
25. Ontoforce. https://www.ontoforce.com. Accessed 15 Jul 2019
26. Mendelian. https://www.mendelian.co/. Accessed 15 Jul 2019
27. Semantic MD. https://semantic.md/nlp.html. Accessed 15 Jul 2019
28. Arrayo. https://www.teamarrayo.com. Accessed 15 Jul 2019
29. Ontotext. https://www.ontotext.com. Accessed 15 Jul 2019
30. Pangaea Entreprises. https://www.pangaeaenterprises.co.uk. Accessed 15 Jul 2019
31. Seminte. http://www.seminte.dk. Accessed 15 Jul 2019
32. The 6 Elements of Truly Transformative Business Models. https://hbr.org/2016/10/the-transformative-business-model. Accessed 25 Jul 2019

Knowledge-based Intelligence and Strategy Learning for Personalised Virtual Assistance in the Healthcare Domain

Eleni Kamateri*, Georgios Meditskos*, Spyridon Symeonidis*, Stefanos Vrochidis*,
Ioannis Kompatsiaris* and Wolfgang Minker†
*Information Technologies Institute
Centre for Research and Technology Hellas, 6th Km Charilaou-Thermi Road, Thessaloniki, Greece
Email: {ekamater, gmeditsk, spyridons, stefanos, ikom}@iti.gr
†Institute of Communications Engineering
Ulm University, 89081 Ulm, Germany
Email: wolfgang.minker@uni-ulm.de

***Abstract*—This paper introduces a virtual assistant framework that combines knowledge-based and statistical techniques to produce meaningful task-oriented conversations that are enhanced by ”chatty” style dialogues in order to increase system's naturalness and user engagement. The paper describes how appropriate ontologies, semantic reasoning, dialogue management and policy learning techniques can be linked together and integrated through the dialogue process to enable a) the internal representation of the conversational state, b) the conversational awareness that drives the retrieval of appropriate information from the Knowledge Base (KB) and the inference of unrelated system actions with the current conversational state, and c) the dynamic selection of the most appropriate strategy at each dialogue turn, tackling both informational and social-related needs of individuals. The framework is exemplified by a use case from the healthcare domain where companionship and supportive care-related services are prerequisites for an efficient human-system interaction through a conversational agent.**

Keywords–Dialogue management; Knowledge representations; Reasoning; Strategy learning; Virtual assistance.

I. Introduction

Nowadays, there is an increasing demand for intelligent agents. A challenging domain includes personalised virtual assistants that carry out human-like conversations taking into account the latest user's utterance, the dialogue history, as well as the background knowledge about the user. The development of such personalised systems requires a knowledge representation model for describing the semantics of various contexts and structuring the background knowledge about individuals.

Current task-oriented dialogue systems focus on one task at a time using frame-based [1] or agenda-based [2] mechanisms, while it was only recently, when some ontology-based dialogue systems (such as [3] and [4]) have been proposed using semantic models for the representation of user's utterance and the generation of the system's response. Access to a rich domain model and the conversation memory can deal with complex task-oriented dialogues. However, the typical problem of task-oriented dialogue solutions remains that is the difficulty of tackling user utterances that go beyond the agent's representational model and the smooth transition between task-oriented and ”chatty” style dialogues.

To succeed this, we propose a hybrid dialogue framework that can be placed at the heart of any personalised virtual assistant to enhance its model-driven operation by ”chatty” style responses. The proposed approach, which is an on-going work, combines knowledge representation and reasoning with statistical learning for the smooth transition between strategies, discussion topics and available knowledge with the aim to impose social skills in the personalised virtual assistants in order to efficiently realise meaningful task-oriented conversations, recover breakdowns in a natural way, and increase user engagement.

Our major contributions are summarised as follows:

1) **a domain and a dialogue representation model** are proposed and populated with local semantics coming from the language analysis of the user's utterance by means of semantic similarity and disambiguation techniques,
2) **a dialogue history representation model** is proposed and populated with global semantics of the entire dialogue session at each dialogue turn,
3) **semantic reasoning techniques** are applied on top of the semantically structured data with the aim to generate dynamically-inferred insights and actions,
4) **a dialogue management technique** analyses the system's confidence regarding the task-oriented response and produces a set of social-oriented action candidates, and
5) **a strategy selection technique** is used to select the appropriate strategy, i.e., action.

Such personalised virtual assistants can have many applications in the healthcare domain and provide a mixture of companionship and supportive care-related services, improving the quality of life of individuals. We selected to apply our framework in a rehabilitation setting, which involves people with motor, cognitive and behavioural disorders being in a clinical environment or after returning home.

The rest of the paper is structured as follows: Section II presents related work on dialogue systems. Section III describes the specifics of the proposed framework, elaborating on the representation, reasoning and dialogue management capabilities. Section IV presents an example use case in the rehabilitation domain, where the framework is currently being used. Finally, Section V concludes our work, mentioning future research directions.

II. Related Work

First conversational systems were mainly task-oriented (e.g., [5] realises restaurant reservations) lacking social competences. More recent personal assistants, such as the commercial platforms of Alexa, Siri, Google Assistant and Contana, have started to incorporate social features and support non-task-oriented dialogues as well, where users do not have a clear goal or intention. However, these systems are usually model-less, constrained to accessing the parameters of the last users utterance and thus, they are acceptable only for simple tasks that do not need to sustain the whole conversation memory.

On the other hand, non-task-oriented dialogue systems do not have a specific goal and are capable of addressing a wide range of topics. To succeed this, they are based on data-driven methods, such as the retrieval-based response selection [6] and the sequence-to-sequence recurrent neural networks [7]. Like most data-driven systems, they produce utterances that are incoherent or inappropriate from time to time and they require a big volume of data that may not be always available.

The combination of the two types of dialogue systems has only recently studied. Zu et al. [8] address the problems of task-oriented dialogue systems when the user's intention is not clear with a framework that incorporates non-task-oriented strategies to keep users interest in the conversation. Similarly, Papaioannou et al. [9] propose a system that combines task-oriented and chat-style dialogues. Both systems apply a reinforcement learning mechanism for selecting the appropriate strategy at each dialogue turn. Coronado et al. [10] propose a hybrid dialogue system that combines a Question Answering system with a conversational agent dealing with rest (small talk) phrases giving a social aspect to the system.

Although current works introduce social aspects through non-task-oriented strategies, we noticed that they mainly use retrieval-based methods with only exception the [11], which incorporates an extension of OwlSpeak dialogue manager [12] and decides whether to consult a knowledge-based module or react on its own. To the best of our knowledge, this is the first approach to combine knowledge-based and statistical techniques to produce task-oriented dialogues that will be used interchangeably with chatty style dialogues exploiting a rich domain model and sustaining the whole conversation memory.

III. Framework Overview

Our framework has four major components: (a) a Contextual Modelling and Representation (CMR) module, (b) a Semantic Intelligence (SI) module, (c) a Dialogue Management (DM) module and (d) a Strategy Selection (SS) module. Figure 1 shows the information flow among these components.

A user utterance is sent to the language understanding module that extracts useful information to help the CMR represent the parsed key entities and identify the discussion topic. Based on the CMR outcome, the SI updates the system's conversational picture, correlates it with background knowledge (e.g., the dialogue history) and infer unrelated insights and actions. Simultaneously, the DM accesses the discussion topic and produces topic-oriented action(s) along with a set of social-oriented actions. Finally, the SS selects among all the actions the most appropriate one and forwards it (along with relevant information from KB, if needed) to the language generation module to produce a system response.

A. *Contextual Modelling and Representation*

The module semantically represents and interlinks the user utterance against the system's cognitive models considering the information passed from the language understanding module.

To achieve this, the module employs existing ontologies and vocabularies. Existing ontologies form the basis of our domain model extended with application-specific aspects. Although there is a significant number of ontologies representing the domain knowledge, we found only few examples of respective ontologies for capturing the different features of the dialogue process. From these, we selected to reuse the well-established OwlSpeak ontology [12] extending it with domain-retrieved knowledge communicated within the user's utterance, exploiting the framework proposed in [4]. The dialogue turn, which is modelled by the *Move* concept, was extended with two new subclasses, the *UserMove* and the *SystemMove*, and each of them is broken down into a set of ”generic” actions, which are common for both edges. For these actions, we used the list of typical actions for multi-agent dialogues presented in [13], including: Open/Greeting, Close/Goodbye, Pause, Resume, Ask, Inform, Affirm, Assert, Remind, and Alert, and extended them with ”Repeat” and the ”Recommend” action.

Each action is further specialised by a set of topic-oriented actions, which constitute the ”discussion topics” that can be covered by the agent. Each topic might be associated with domain knowledge by means of a dialogue entity (*dialogueEntity*) which consist the target entity of each discussion topic. Additional entities extracted from the user's utterance might be associated with the dialogueEntity to further specify the requested entity.

The module semantically represents a user utterance using state-of-the-art disambiguation tools (e.g., UKB [14] or Babelfy [15]) that assign key entities extracted from the language understanding module to resource categories (i.e., synsets). These resource categories are then used to identify entities (synonyms) and topics against the domain and the dialogue ontology, respectively.

With respect to domain-driven mapping, we assume that $label(r)$, is the label of resource $r \epsilon KB$, $syn(k)$ is the synset of key entity $k \epsilon K$ and σ is a similarity function, the set $S(k)$ of all the relevant resources to k is defined as:

$$S(k) = argmax_{k \epsilon K} \sigma(k, label(r)) \tag{1}$$

The UMBC Semantic Similarity Service [16] is used to calculate the semantic similarity σ between k and $label(r)$ combining Latent Semantic Analysis (LSA) word similarity and WordNet knowledge.

With respect to dialogue-driven mapping, a simple classification algorithm calculates the conditional probability of each discussion topic for all parsed resources, given that each discussion topic is described by means of a set of similar resources:

$$P(Topic_i \mid t_x) = \frac{P(Topic_i \cap t_x)}{P(t_x)} \tag{2}$$

where $Topic_i$ is a topic defined by a set of resources $t_1, t_2, ...t_k$, while t_x is assumed to be a parsed resource from user utterance. This probability is then multiplied with respective probabilities for all parsed resources.

When a discussion topic is identified, the dialogue session is informed with the dialogue details including the dialogue topic, the dialogueEntity and associated entities populated with knowledge coming from the analysis of user utterance.

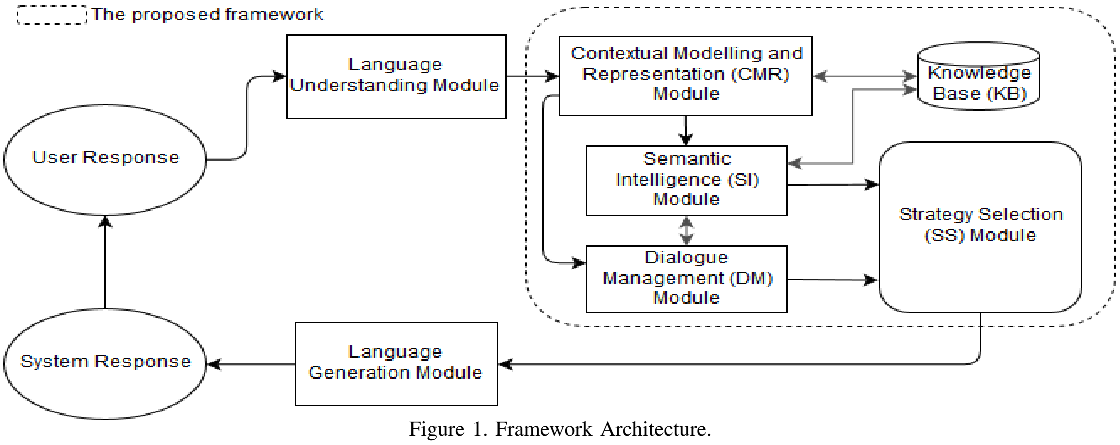

The proposed framework
Language Understanding Module
Contextual Modelling and Representation (CMR) Module
Knowledge Base (KB)
User Response
Semantic Intelligence (SI) Module
Strategy Selection (SS) Module
Dialogue Management (DM) Module
System Response
Language Generation Module

Figure 1. Framework Architecture.

B. Semantic Intelligence

The module utilises pattern-based models [17] to update domain models with new information communicated through the human-system interaction and inform the dialogue history with identified entities and topics at each dialogue turn. Moreover, it translates the system actions into actionable rules (SPARQL queries), which are then used to retrieve pertinent information from the underlying KB.

SPARQL Inferencing Notation (SPIN) are also applied to generate alerts, reminders and recommendations, which are triggered by the knowledge of the preceding discourse and the specific user profile. By this way, motivational or interventional actions are forwarded to the SS module, which might interrupt the usual flow and impose situation-oriented system responses. These actions consists of: (1) alert, (2) remind, (3) recommend and (4) repeat action.

C. Dialogue Management

The module processes the outcome of topic identification and decides the topic-oriented action to follow, selecting among: (5) predefined topic-based (re-)action, when the matching score of a topic exceeds a specific threshold, (6) clarification action, in case of partial topic identification with more than one topics receiving a significant matching score, and (7) say-again action, in case of incomplete topic identification.

Simultaneously, the module formulates a set of social-oriented action candidates considering the information received from the CMR and supportive information extracted from the KB. The social-oriented actions include (8) switch topic (a new topic is suggested based on user's preferences), (9) initiate a relevant topic, (10) end current topic and make an open question, (11) suggest to provide more info about the current topic, and (12) elicit more information.

D. Strategy Selection

This module chooses among all action candidates the most appropriate one with the aim to optimise the conversational flow towards natural and meaningful interaction. Different learning algorithms can be applied to train the strategy selection, such as Q-learning [8] and policy gradient [18]. Our strategy selection was implemented based on a simplified version of the reinforcement learning algorithm presented in [8]. The algorithm has a function that calculates the quantity of a state-action combination $Q : SxA- > R$, called Q table.

$$
\begin{aligned}
Q_{t+1}\left(s_t, a_t\right) \leftarrow Q_t\left(s_t, a_t\right) + a_t\left(s_t, a_t\right) \cdot \\
\left(R_{t+1} + \gamma max Q_t\left(s_{t+1}, a\right) Q_t\left(s_t, a_t\right)\right)
\end{aligned}
$$

For the reward function, we used domain experts' knowledge provided in [19] and [8]. According to them, the reward is calculated based on: turn index, number of times each strategy executed, sentiment polarity of previous utterances, most recently used strategy and coherence confidence of the response.

IV. A Use Case Example in Rehabilitation

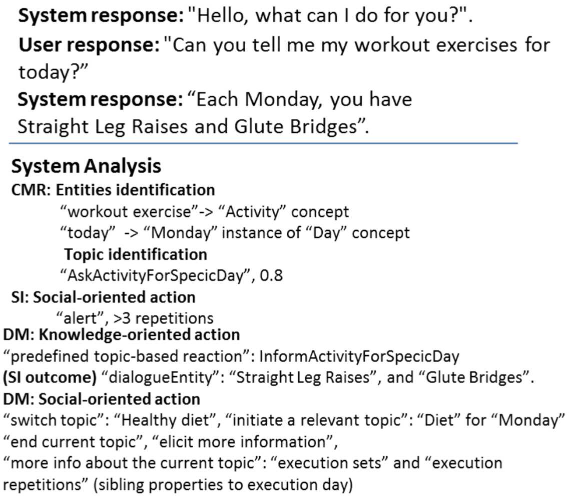
As depicted in Figure 2, the system starts a conversation saying ”Hello, what can I do for you?”. Let us assume that the user replying ”Can you tell me my workout exercises for today?”.

For domain modelling, we reused COPDology [20], an ontology which was designed to facilitate the systematic monitoring of Chronic Obstructive Pulmonary Disease (COPD) patients, containing concepts pertinent to an individual's profile, the conditions they suffer from, and the medications/workout exercises they receive. We extended it with new properties, such as the *hasExecutionDay*, *hasExecutionSets* and *hasExecutionRepetitions*, to describe the execution guidelines for the scheduled workout exercises. Moreover, we assume that there is a *AskActivityForSpecificDay* topic, with the *Activity* being the target entity and the *Day* specifying the topic receiving a specific value, e.g., Monday.

The CMR annotates the key entities parsed from the language understanding module and identifies the ”discussion topic”. The incoming information ”workout exercises” and ”today” are associated with the *Activity* concept and the *Monday* instance of *Day* concept, while the *AskActivityForSpecificDay* topic is identified with a matching score of 0.8.

The SI module updates the dialogue history and enforces predefined rules. Emergency situations can be detected, for example, if the user asks more than a couple of times about the same topic, the system initially conceives it as repetition but if it happens more than a predefined amount of times (e.g., three times) the system enforces an emergency situation.

The DM evaluates the matching score of identified discussion topic and decides that a ”predefined topic-based (re-)action” will be followed. This means that the *InformActivityForSpecificDay* system action, which is one-by-one associated



System response: "Hello, what can I do for you?".

User response: "Can you tell me my workout exercises for today?"

System response: "Each Monday, you have Straight Leg Raises and Glute Bridges".

System Analysis

CMR: Entities identification

"workout exercise"-> "Activity" concept

"today" -> "Monday" instance of "Day" concept

Topic identification

"AskActivityForSpecicDay", 0.8

SI: Social-oriented action

"alert", >3 repetitions

DM: Knowledge-oriented action

"predefined topic-based reaction": InformActivityForSpecicDay

(SI outcome) "dialogueEntity": "Straight Leg Raises", and "Glute Bridges".

DM: Social-oriented action

"switch topic": "Healthy diet", "initiate a relevant topic": "Diet" for "Monday"

"end current topic", "elicit more information",

"more info about the current topic": "execution sets" and "execution repetitions" (sibling properties to execution day)

Figure 2. Use case example.

with the user's action, will be enforced. In the meantime, the SI (upon DM's request) translates the system action and dialogue entities into SPARQL queries to retrieve instances of the "Activity" concept for Monday. Simultaneously, the module formulates a set of social-oriented action candidates.

Based on the learned Q table, the SS selects the most appropriate action and forwards it to language generation to produce the system response content.

V. Conclusion

The proposed framework combines dynamic knowledge-based features with social competences which are orchestrated by the means of a statistical policy learning that selects among action candidates the most appropriate one to optimise conversational effectiveness. The framework is currently validated in a running project involving clinicians and staff of a rehabilitation clinic. Our next steps is to establish an experimental set-up and evaluate it with real data. In addition, we plan to enrich the context understanding capabilities of the agent by integrating and fusing multimodal information, such as home activities and gestures, increasing the situational awareness of the agent.

Acknowledgment

This research has been cofinanced by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-00686)

References

[1] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd, "Gus, a frame-driven dialog system," Artificial intelligence, vol. 8, no. 2, 1977, pp. 155–173.

[2] A. Rudnicky and W. Xu, "An agenda-based dialog management architecture for spoken language systems," in IEEE Automatic Speech Recognition and Understanding Workshop, vol. 13, no. 4, 1999.

[3] D. Altinok, "An ontology-based dialogue management system for banking and finance dialogue systems," CoRR, vol. abs/1804.04838, 2018. [Online]. Available: http://arxiv.org/abs/1804.04838

[4] M. Wessel, G. Acharya, J. Carpenter, and M. Yin, OntoVPA—An Ontology-Based Dialogue Management System for Virtual Personal Assistants. Cham: Springer International Publishing, 2019, pp. 219–233.

[5] F. Jurcícek, S. Keizer, M. Gasic, F. Mairesse, B. Thomson, K. Yu, and S. J. Young, "Real user evaluation of spoken dialogue systems using amazon mechanical turk," in INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, 2011, pp. 3061–3064.

[6] R. E. Banchs and H. Li, "Iris: a chat-oriented dialogue system based on the vector space model," in Proc. of the ACL 2012 System Demonstrations, 2012, pp. 37–42.

[7] O. Vinyals and Q. V. Le, "A neural conversational model," CoRR, vol. abs/1506.05869, 2015. [Online]. Available: http://arxiv.org/abs/1506.05869

[8] Z. Yu, Z. Xu, A. W. Black, and A. Rudnicky, "Strategy and policy learning for non-task-oriented conversational systems," in Proc. of the 17th annual meeting of the special interest group on discourse and dialogue, 2016, pp. 404–412.

[9] I. Papaioannou, C. Dondrup, J. Novikova, and O. Lemon, "Hybrid chat and task dialogue for more engaging hri using reinforcement learning," in 26th IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN), 2017, pp. 593–598.

[10] M. Coronado, C. A. Iglesias, and A. Mardomingo, "A personal agents hybrid architecture for question answering featuring social dialog," in Int. Symposium on Innovations in Intelligent SysTems and Applications (INISTA), 2015, pp. 1–8.

[11] L. Pragst, J. Miehle, W. Minker, and S. Ultes, "Challenges for adaptive dialogue management in the kristina project," in Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents, ser. ISIAA 2017. New York, NY, USA: ACM, 2017, pp. 11–14. [Online]. Available: http://doi.acm.org/10.1145/3139491.3139508

[12] S. Ultes and W. Minker, "Managing adaptive spoken dialogue for intelligent environments," Journal of Ambient Intelligence and Smart Environments, vol. 6, no. 5, 2014, pp. 523–539.

[13] J. Baskar and H. Lindgren, "Human-agent dialogues and their purposes," in Proceedings of the European Conference on Cognitive Ergonomics 2017, ser. ECCE 2017. New York, NY, USA: ACM, 2017, pp. 101–104.

[14] [Online]. Available: http://ixa2.si.ehu.es/ukb/

[15] [Online]. Available: http://babelfy.org

[16] L. Han, A. L. Kashyap, T. Finin, J. Mayfield, and J. Weese, "Umbc_ebiquity-core: Semantic textual similarity systems," in 2nd Joint Conf. on Lexical and Computational Semantics (* SEM), Volume 1: Proc. of the Main Conf. and the Shared Task: Semantic Textual Similarity, 2013, pp. 44–52.

[17] G. Meditskos, S. Dasiopoulou, S. Vrochidis, L. Wanner, and I. Kompatsiaris, "Question answering over pattern-based user models," in Proc. of the 12th Int Conf. on Semantic Systems, 2016, pp. 153–160.

[18] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, "Deep reinforcement learning for dialogue generation," arXiv preprint arXiv:1606.01541, 2016.

[19] F. Sukno and other, "A multimodal annotation schema for non-verbal affective analysis in the health-care domain," in Proc. of the 1st Int. Workshop on Multimedia Analysis and Retrieval for Multimodal Interaction, 2016, pp. 9–14.

[20] H. Ajami and H. Mcheick, "Ontology-based model to support ubiquitous healthcare systems for copd patients," Electronics, vol. 7, no. 12, 2018, p. 371.

A Semantic Model for the Validation of ePassport Certificate Chain of Trust

Elwaleed Elmana, Hind Zantout, Hani Ragab Hassen
School of Mathematical & Computer Sciences
Heriot-Watt University
Dubai, UAE
Email: Eke1@hw.ac.uk
Email: H.Zantout@hw.ac.uk
Email: H.Ragabhassen@hw.ac.uk

***Abstract*—Chip-enabled passport (ePassport) data is secured by Public Key Infrastructure (PKI) Digital Certificates to validate that the digitally signed data has not been tampered with, thus creating trust. Border ePassport verification processes in place are diverse; each country defines its own rules taking into account the International Civil Aviation Organization (ICAO) published recommendations. This project attempted to represent the ePassport PKI domain and its related policies using semantic technologies based on the Resources Description Framework (RDF) and the Web Ontology Language (OWL). The objective is to help border authorities rely on a standardised and unified trust classification process. The ontology was built using Protege following the Ontology Development 101 Methodology. The results show that not only can the PKI certificate chain be represented, but also the related certificate policy and practice statement. Semantic Web Rule Language (SWRL) rules successfully managed to represent essential aspects of the borders validation policy. The pilot demonstrates that a reliable implementation to automate the trust level classification process is achievable.**

Keywords-ePassport; PKI; Border Control; Semantic Technologies.

I. INTRODUCTION

The introduction of chip technology into the identification document domain enabled passport and national ID documents with increased security features. The chip contains all the information printed on the document data page, and the relevant data is stored on the chip using encryption, making the document tamper-proof. The encryption methods applied to use Public Key Infrastructure (PKI) digital certificates, assuring that the document is not forged.

When travelers pass through a border checkpoint, a personal information and identification process takes place to verify that the passport holder information matches the data on the chip. This matching or authentication process utilises the biometric information stored on the chip, such as fingerprints or iris scans. Biometric authentication, however, is outside the scope of this paper. The validation process being considered here is the application of decryption mechanisms to read the data from the chip, something that includes managing a complex PKI system.

This border control validation process also has a political aspect to it as it depends on the general practice of a country's Certification Authority (CA) sharing the distributing digital certificates with the relevant authority in another country. The Country Signer Certificate Authority (CSCA) and the Document Signer (DS) certificates are crucial as they form a chain of trust. ICAO plays an important advisory role through its suggested roadmap and Public Key Directory (PKD) [1]. To date, the validation policies still vary from a country to a country despite the various recommendations and technical reports that aim to regulate how to trust a chip-enabled document, using protocols like passive authentication [1] [2]. However, the actual implementation on the ground will vary because verifying an electronic document involves not only checking the data on the chip against the real documents, but it also includes the verification of the trust level of the PKI system behind it. Therefore, a comprehensive solution must have a check of personal information and validate the trust level of the country's digital certificate. Also, it must automatically process both the Certificate Policy (CP) and Certificate Practice Statement (CPS) of the relevant CA. The CP and CPS will indicate how the CA is performing its duties.

We propose a solution that will use semantic technologies to create a system that can process both the digital certificates as well as the PKI policies relating to a travel document. The resulting decision support system will enhance the ability of the border control officer to determine the trust level of a travel document.

The rest of the paper is organised as follows. Section II reviews the literature on Machine Readable Travel Document (MRTD) and policies that govern the validation process with proposed solutions. Section III describes the objectives, requirements, and validation methods for the project. Section IV introduces the model, the design process, and discusses system capability. Implementation details are included in Section V, and the results are discussed in Section VI, followed by the conclusion and future work.

II. RELATED WORK

The ICAO recommendations are published in document 9303 [1], and several other regulators like The German Federal Office for Information Security (BSI) publish related technical reports [2]. Such publications advocate a general framework for the validation process and policies that include guidelines for trusting PKI certificates issued by other countries. These certificates should be distributed through verification means on the ICAO own PKD portal, or through bilateral exchange agreements between countries. Currently, the details of checking a travel document depend on the practices in place in each state as well as existing collaborations between countries. Several studies tried to address the gap between the validation result and the trust decision of a travel document, by proposing a centralised service with frameworks that utilise the certificate path validation as a tool to achieve trust, along with other PKI elements like the CP, the CPS and the Certificate Revocation List (CRL). However, CP and CPS documents are written in a natural language like English or German, which means the involvement of a human interpreter is an essential part of the validation process.

We start by discussing the attempts to include the quality of the CP and the commitment through CPS during the PKI certificate validation process. We then review the work related to the semantic representation of the policies which is needed for an automated system.

Sato and Kubo [3] in their patent application classified CA policies based on their level of assurance, and the paper proposes a dynamic chain or trust validation using a single certificate policy service provider. It manages the CP lifecycle independent of its corresponding CA, by pre-registering CA based on their compliance with a regularly published CP/CPS and classifying the trust level based on their CP/CPS level of assurance. In a multi-country situation, this will require all countries to share their CP/CPS with the single certificate policy provider. Currently, this ideal scenario of all countries around the globe sharing this information is not in place and unlikely to be in place in the foreseeable future.

Roh et al. [4] provide a solution that involves a server which upon receipt of the object certificate to be validated, the certificate of a trusted certification authority and the certificate policy proceeds to create a certification path for the object certificate as a first stage. If it is valid, it continues to the next step of validating the certificate path itself. This method was applied for as a patent in 2004.

Another ongoing research track investigates how to represent PKI CP and CPS in a machine-readable format. As described earlier, the CP defines the applicability of the CA certificate and the rules that govern it. The CPS describes in detail how the CA certificate has been managed and includes specifics of the issuing, the distribution and the revocation of a CA certificate [12]. The representation of the underlying rules is an essential step towards an automated system that can process both the PKI certificates as well as their policies.

Smith [5] worked on a Computational Framework for Certificate Policy Operations, using a machine-readable language to represent the CP elements as an object identifier. It based the CP representation on an encoding technique called “Canonical Text Services Uniform Resource Name (CTS-URN)", which provides the advantage of a validation system to read a semi-machine-readable CP without human interaction.

Grill [6] modelled X.509 Certificate Policies using Description Logics, his paper divided their approach, which used an ontology to represent policies into three stages.

1) Defining the domain schema classification or the taxonomy.
2) Having a reference ontology for usability purposes.
3) Working on the specific policy elements with an approach to compare CPs rather than to infer from them.

However, there were no proposals to include functionality that supports both the processing of the PKI certificate and their respective policies. Grill’s use of descriptive logic shows the potential role that semantic technologies can play in representing the PKI domain. The fact that the semantic technologies stack is built with security in mind and uses digital certificates as a means of trust can be leveraged to that end.

In our proposed solution, the RDF representation gives us the advantage to keep writing CP/CPS in a natural language while having rich metadata about the document that can be used by machines to evaluate the policy. Furthermore, OWL, coupled with rule-based reasoners, can provide a decision to trust or not to trust an MRTD based on predefined rules that reflect the actual practice in the real world.

The first step towards such a system is to build a knowledge-base that incorporates all the must-have elements of MRTD, PKI components, as well as the CP/CPS definition, and the border validation policy. Once the ontology that is comprehensive in nature is defined, it can be coupled with valuable inference rules and applied to specific instances. To do so, we followed the Ontology Development 101 Methodology [7] which enables the building of ontologies based on existing ones and uses the Certificate Ontology specification as outlined in the W3C standard as a baseline [8]. The domain knowledge is taken from MRTD regulator’s publications such as the ICAO Machine Readable Travel Documents Doc 9303 -part-11 [9] and Part-12 [1], as well as the BSI Technical Guideline BSI TR-03135 Machine Authentication of MRTDs for Public Sector Applications [10].

III. The Model requirements and Design

The primary objective of the proposed system is to answer questions related to how countries can develop an MRTD local border verification policy. The solution will have to incorporate the root certificate CSCA, document signers, together with their policies and practices statement. This can be achieved by building an ontology-based model that captures elements of the border validation process based on the current recommendation and best practice of border control validation policies and procedures.

The knowledge-base will represent the CSCA certificate policy along with DS certificates and ePassport chip Document Security Object (SOD) elements using OWL coupled with SWRL rules, a combination that provides rich vocabulary and a full inference capability [11]. The Protégé reasoner will be used to verify the ability of the rules in creating a model that can deliver a reliable trust decision capability.

In the design phase, we recap what we highlighted in Section II, the need for a system that is capable of processing PKI certificates and their respective policies. Figure 1 depicts the general framework design. In the first stage, the passport document Security Object SOD that contains a hash of all the data groups and the associated DS is processed. In the second stage, the data is prepared in a format that is compatible with the knowledge base. The preparation process is not within the scope of this paper. However, we assume that the data is RDF/OWL compatible. The third stage consists of applying an inference engine like Protégé DRool with the capability to run SWRL rules that will deliver the decision.

In Figure 2, we identify the concepts, properties, and relationships using the 101 Methodology. In that structure, the properties of the MRTD, CA, DS, and Policies were defined. For example, the main properties of the CP and CPS were listed based on Request for Comment (RFC) 3647 [12].

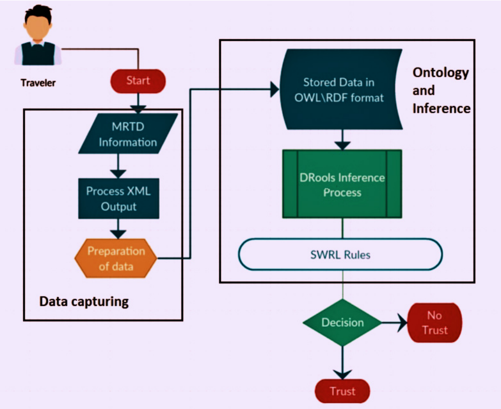


Figure 1. The Framework Design

In our work, the ontology is focused on answering the following four questions:

1) What is the type of the Document: is it an ID or passport?
2) Does the passport have a valid certificate chain or trusted path?
3) Does the passport root CA or CSCA have a trusted Policy?
4) Is the root CA or DS Trusted?

IV. Ontology Implementation

We used Protégé [13] as a primary tool to develop and validate the ontology. The tool provides a framework that has many add-on tabs that serve different functions such as the Entities tab where classes and their corresponding properties and individuals can be defined. In addition, Protégé provides integrated SWRL rules processing using the DRool extension as well as the option of using different reasoners.

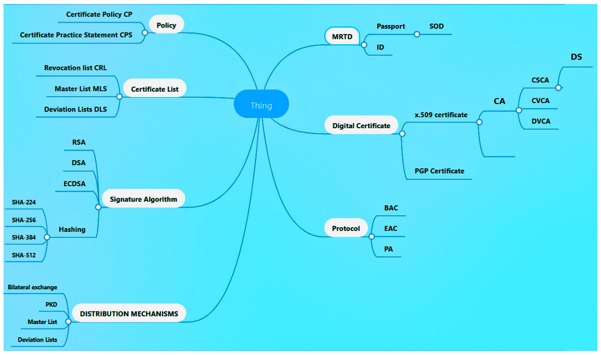


Figure 2. Concepts and Terms

For the system to answer the questions mentioned above, the knowledge base must include a sufficiently rich representation of concepts and their relationships to infer the required result correctly. A top-down approach is used to define the classes, object properties, and individuals. Figure 3 shows the main classes which are identified. An object property captures the relationship between classes and individuals [14] and can be used to specify the domain and the range [15]. In Table 1, the main pillars of the PKI are described and linked.

A. Use Cases Scenarios and SWRL Rules

We build the use cases to show that the ontology can simulate the current border validation scenario summarised below [10]:

1) The reader captures EMRTD information and uses BAC or PEAC protocols to access the chip.
2) Based on the document type information, it determines if it is a passport or ID.
3) Using the Passive Authentication protocol, it checks the digital signature of the DS.
4) The path validation checks if the DS has a valid CSCA signer or not.

When we add a new individual eMTRD instance to the system, the reasoner will be able to identify and classify it.

For example, the first primary use case will answer Question 1 above. Figure 4 shows the introduction of an individual with name Pass124 and has datatype property "hasPassportType" with value 3. The reasoner was able to identify that this individual is of class passport.

A more advanced use case is one where the reasoner had to process more than two classes with their various properties, to infer a result. In this complicated case, the system was able to answer Question 2 above.

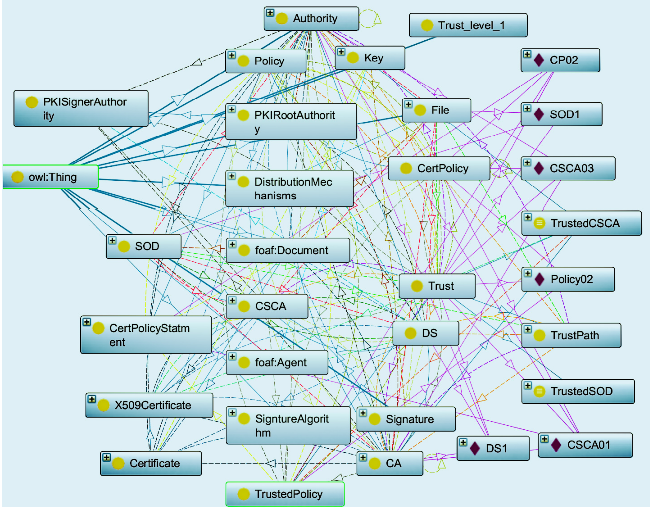


Figure 3. Main Classes

Here, an additional instance and its properties were identified as follow:

1) Add the instances Pass124 of class Passport, SOD1 of class SOD, DS1 of class DS, and CSCA01 of class CSCA to the knowledge base.

TABLE.1 OBJECT PROPERTIES

Object Properties		
Domain Class	**Object Property**	**Range Class**
Passport	AssociatedwithA	SOD
Digital Signature	CreatedBy	Private Key
Certificate	HasAKey Private Key and	Private Key and Public Key
CSCA OR DS	HasCertificateType	X.509Certificate
CSCA	RootCertificateType	X.509Certificate
DS	SignerCertificateType	X.509Certificate
SOD	HasSodIn	TrustPath
TrustPath	HasValidPathfrom SCATo	Policy
Domain Class	Object Property	Range Class
TrustPath	HasValidPathfrom DSTo	CSCA
TrustPath	HasValidPathfromSODTo	DS
SDO OR Certificate	Holds	Digital Signature
DS	IsKindOf	PKI SignerAuthoriy
CSCA	IsTypeOf	PKI RootAuthority
CSCA	Sign	DS
SOD	SignedBy	DS

2) Determine the instance to have general object properties CSCA01 Sign DS1, and SOD1 is Signed by DS1, and Pass124 AssociatedWith SOD1.

3) Define the main class called Trust, and a Subclass called Trusted path with Axiom:

(**HasValidPathfromCSCATo** Some Policy, and **HasValidPathfromDSTo** Only CSCA, and **HasValidPathFromSodTo** DS).

In Figure 5, the reasoner inferred that only the individuals SOD1, DS1, CSCA1 are part of the trusted path, although there were other individuals within the same domain.

The result of this use case as an example to prove that normal Protégé reasoner like HermiT and Pellet can give valuable outcome. Nevertheless, they were limited in that they cannot infer further results based on previously inferred results. Any result that is needed for further processing must be added as a new assertion to the knowledge base first.

B. SWRL Rules

The results obtained by the reasoned can also be reached using SWRL Rules. The SWRLAPI uses the DRool rule engine for inference purposes based on OWL 2 RL [13]. It uses the ontology as input, applies the rules, and returns inferred and asserted results.

Four rules have been developed using assumptions based on industry best practice. In a fully mature system, it is expected to have a much larger number of rules based on a formal written border validation policy.

1) Rule 1:

If a Document Signer signs a passport SOD, and a CSCA signs that Document Signer, then this passport component belongs to a Trusted Path class. Rule (1) shows the SWRL representation.

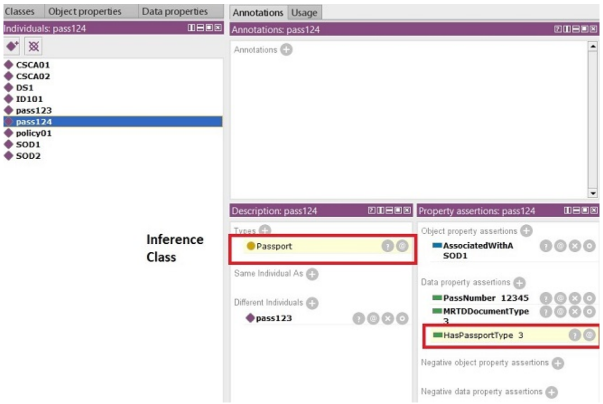
$$Passport(?P) \wedge SOD(?S) \wedge SignedBy(?D, ?S) \wedge CSCA(?C) \wedge Sign(?C, ?D) \Rightarrow TrustedPath(?P) \qquad (1)$$

2) Rule 2

If a CSCA certificate was distributed through a mechanism such as ICAO PKD and found to have some properties like a Trusted Policy, a signature algorithm of type ECDSA, and a signature hash algorithm of type SHA 256, then this CSCA certificate can be classified as Trusted CSCA.

Figure 6 shows the SWRL representation and result of Rule (2).

$$CSCA(?C) \wedge TrustedPolicy(?TPO) \wedge HasLinkFromCSCATo(?TPO, ?C) \wedge HasCSCASignAlgorithm(?SA, ?"ECDSA") \wedge HasCSCADistributionMechanism(?DM, \text{"PKD"}) \wedge CSCASigntureHashAlgorithms(?SHA, \text{sha256}) \Rightarrow TrustedCSCA(?C) \qquad (2)$$



CSCA01
CSCA02
DS1
ID101
pass123
pass124
policy01
SOD1
SOD2

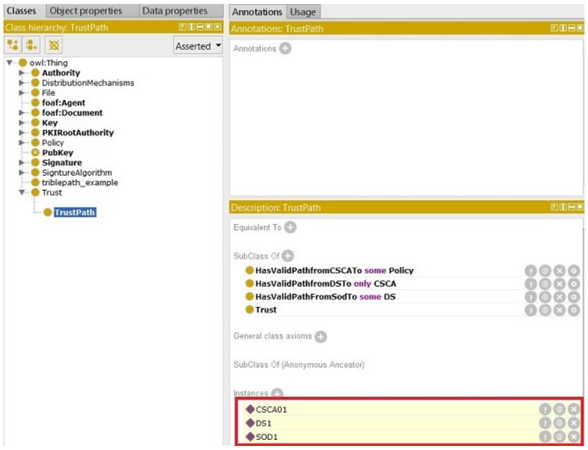
Passport

pass123

AssociatedWithA SOD1

PassNumber 12345
MRTDDocumentType 3
HasPassportType 3

Figure 4. Inference class of use case one



owl:Thing
Authority
DistributionMechanisms
File
foaf:Agent
foaf:Document
Key
PKIRootAuthority
Policy
PubKey
Signature
SigntureAlgorithm
triblepath_example
Trust
TrustPath

HasValidPathfromCSCATo some Policy
HasValidPathfromDSTo only CSCA
HasValidPathFromSodTo some DS
Trust

CSCA01
DS1
SOD1

Figure 5. Inferred Trusted Path Class members

3) Rule 3

If a policy CP or CPS is found to have a CRL Issuing Frequency of 2 weeks, and a rigorous Certificate Rekey process, as well as a publication frequency of 3 months, then it can be classified as a Trusted Policy.
Rule (3) shows the SWRL representation.

$$\mathrm{Policy}(?PO) \wedge \mathrm{PolicyPropCRLIssuingFreq}(?CRLIF,\text{"2weeks"}) \wedge \mathrm{PolicyPropCertRekeyProcess}(?CRKP,yes) \wedge \mathrm{PolicyPropPublicationFreq}(?PF,3months) \Rightarrow TrustedPolicy(?TP) \tag{3}$$

4) Rule 4

This rule depends on the result of previous rules. The aim here is to classify the CSCA of a country based on their trust level. If a CSCA certificate is a member of a Trusted path class, and a Trusted CSCA class, in addition to having a Trusted Policy class and a trusted SOD class, then this CSCA belongs to a Trust level 1 class.
Rule (4) shows the SWRL representation.

$$\mathrm{TrustPath}(?P) \wedge \mathrm{TrustedCSCA}(?C) \wedge \mathrm{TrustedPolicy}(?PO) \wedge \mathrm{TrustedSOD}(?TSOD) \Rightarrow \mathrm{TrustLevel1}(?C) \tag{4}$$

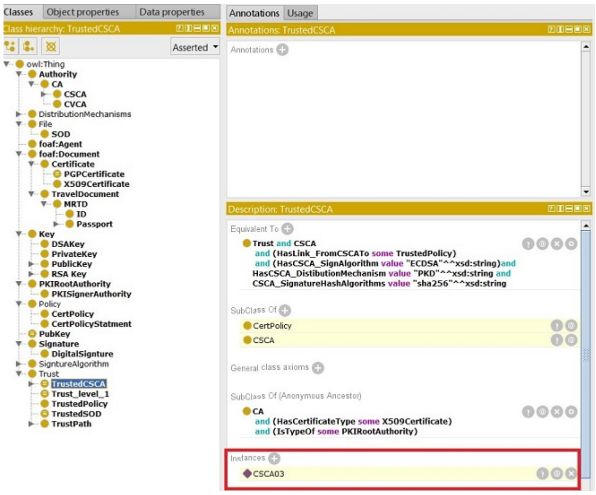
V. RESULTS AND DISCUSSION

The evaluation of the work is based on cross-checking the ontology against the most important criteria such as consistency and coherence, clarity and modularity and reusability. These are defined by the Ontology Quality Evaluation and Requirements Framework (OQuaRE) [16].

A) Consistency and Coherence

1) Protégé has set of reasoners and Debugger tools, which run through the Ontology axioms, object properties, and data properties to infer result. The Debugger run over 837 axiom and the result is “The ontology is consistent and coherent”.

2) We used The Ontology Pitfall Scanner developed by the Ontology Engineering group [17], as a comprehensive online tool that checks the consistency. The result showed the existence of critical cases related to using multiple domains or ranges in properties, and some crucial cases due to the use of recursive definitions, which refer to the use of a class name within its equivalent class axiom. As this was only detectable after the DRool inference result, we believe that it is due to Protégé internal ontology processes, and it should not harm the original ontology structure.



owl:Thing
Authority
CA
CSCA
CVCA
DistributionMechanisms
File
SOD
foaf:Agent
foaf:Document
Certificate
PGPCertificate
X509Certificate
TravelDocument
MRTD
ID
Passport
Key
DSAKey
PrivateKey
PublicKey
RSA Key
PKIRootAuthority
PKISignerAuthority
Policy
CertPolicy
CertPolicyStatment
PubKey
Signature
DigitalSignture
SigntureAlgorithm
Trust
TrustedCSCA
Trust_level_1
TrustedPolicy
TrustedSOD
TrustPath

```
Trust and CSCA
 and (HasLink_FromCSCATo some TrustedPolicy)
 and (HasCSCA_SignAlgorithm value "ECDSA"^^xsd:string)and
HasCSCA_DistibutionMechanism value "PKD"^^xsd:string and
CSCA_SignatureHashAlgorithms value "sha256"^^xsd:string
```

CertPolicy
CSCA

```
CA
 and (HasCertificateType some X509Certificate)
 and (IsTypeOf some PKIRootAuthority)
```

CSCA03

Figure 6. Inference of trusted CSCA

B) Clarity

The OntOlogy Pitfall Scanner result for clarity shows only minor remarks, suggesting more annotation and a unified naming convention should be used. Further clarification of the annotation definition can be discussed with the domain experts.

C) Modularity and Reusability

The extendibility or modulatory criteria check depicts the level of change in the ontology that can be introduced without affecting the overall function. We used the following OQuaRE metrics:

The Weighted Count Method (WMCOnto) is a metric, which can be measured by calculating the average number of properties and relationship per class.

$$\mathrm{WMCOnto} = (\Sigma|\mathrm{PCi}| + \Sigma|\mathrm{RCi}|/\Sigma|\mathrm{Ci}|$$

Our ontology scored 0.45, which is considered very low comparing to well-defined Ontologies that scores between 5-11 [16].

The DITOnto is a reusability metric, which counts the maximum length of the path from the leaf to the ontology root point “Thing”.

$$\mathrm{DITOnto} = \mathrm{Max}\,\Sigma|\mathrm{Ci}|$$

The NOMOnto is another reusability metric that considers the number of properties per class

$$\mathrm{NOMOnto} = \Sigma|\mathrm{PCi} \,/\, |\Sigma|\mathrm{Ci}|$$

The result of the DITOnto is 5. Moreover, NOMOnto is 0.36.

Comparing to the result of other well-defined ontologies that score between 2-8 on DITOnto and NOMOnto, the above result is an indication that the Ontology has its limitations concerning reusability.

VI. Conclusion and future work

With the model that we proposed and the ontology described, we were able to demonstrate that that ePassport PKI elements can be semantically represented, linked to relevant policies and classified based on Trust rules. Thus, a precise border control ontology-based validation procedure can be achieved. The ontology within the model can be considered as a core to an industry-ready solution, customizable to suit each border control authority rules and procedures. The initial knowledge base will need to be expanded with other countries’ certificates. Combined with the semantic representation of the CP and CPS we believe it will make border classification process more transparent, in addition to helping border control authorities build an ICAO recommended Master List [1] through the PKD portal.

Although the ontology did not score highly in the technical evaluation process, however, we were able to answer all key four questions and reach the goal of having a decision to trust or not to trust a given eMRTD. The taxonomies captured were modest, and the border validation elements and rules were not comprehensive. Nevertheless, within the defined scope, the ontology was able to demonstrate the validity of the concept of CP and CPS representation using ontologies.

Finally, this approach closes a severe gap in providing a meaningful border control solution. The issue of how countries are maintaining their PKI CA and the issuing of DS certificates needs to be addressed in a structured way as proposed by this project.

This project can be considered as a base for the following future work:

1) The semantic representation of the CP and CPS elements, having both Policies entirely written in RDF/OWL means they can be processed by a system without the need of a human expert and can make ePassport PKI classification an automated process.

2) The current model using SWRL rules is only intended as a proof of concept. In a real-world situation, we expect a comperhensive list of rules that covers the ePassport border validation process and procedures.

References

[1] ICAO Recommendation 9303, “Machine Readable Travel Documents - Part 12: Public Key Infrastructure for MRTDs”, Seventh Edition, 2015, INTERNATIONAL CIVIL AVIATION ORGANIZATION,[Online].Available:www.icao.int/Security/FAL/TRIP [Accessed: 06-Aug-2019].

[2] German Federal Office for Information Security, “Machine Authentication of MRTDs for Public Sector Applications Part 1: Overview and Functional Requirements,” Bonn, 2017.

[3] H. Sato and A. Kubo, “Graded Trust of Certificates and Its Management with Extended Path Validation,” Inf. Media Technol. J. Inf. Process., vol. 6, no. 19, pp. 980–990, 2011.

[4] Jong Hyuk Roh et al., “Method of Validating Certificate By Certificate Validation Server Using Certificate Policies And Certificate Policy Mapping In Public Key Infrastructure,” " ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUT" 21-May-2002.

[5] G. A. Weaver, S. Rea, and S. W. Smith, “for Certificate Policy Operations”, Public Key Infrastructures, Serv. Appl., vol. EuroPKI, 20, no. 2006, pp. 17–33, 2010.

[6] S. W Grill, “Modeling X.509 Certificate Policies Using Description Logics - Semantic Scholar,” 2007. [Online]. Available: https://www.semanticscholar.org/paper/Certificate-Policies-Using-Description-Logics Grill/37fab24f8de082c7e3ff2e23879cf2979a610a99. [Accessed: 06-Aug-2019].

[7] N. F. Noy and D. L. McGuinness, “A Guide to Creating Your First Ontology,” in Biomedical Informatics Reseach, 2001.

[8] W3C WebID Incubator Group, “The Cert Ontology Specification,” 2008. [Online]. Available: https://www.w3.org/ns/auth/cert#PublicKey. [Accessed: 06-Aug-2019].

[9] ICAO Recommendation, “Machine Readable Travel Documents - 9303 Part 11,” Montréal, 2015, INTERNATIONAL CIVIL AVIATION ORGANIZATION, [Online]. Available: www.icao.int/Security/FAL/TRIP [Accessed: 06-Aug-2019].

[10] BSI, “Advanced Security Mechanisms for Machine Readable Travel Documents and eIDAS Token-Part 1,” 2015, BSI Publications [Online].Available:https://www.bsi.bund.de/EN/Publications/TechnicalGuidelines/TR03110/BSITR03110.html.

[11] D. S. R. P.Hitzler, and M.Krötzsch, "Foundations Of Semantic Web Technologies", Taylor and Francis Group, 2010.

[12] S. Chokhani, W. Ford, R. Sabett, C. Merrill, and S. Wu, “Internet X.509 Public Key Infrastructure Certificate Policy and Certification Practices Framework,” 2003.

[13] stanford.edu, “Protege 3.5 Release Notes - Protege Wiki.” [Online]. Available: https://protegewiki.stanford.edu/wiki/Protege_3.5_Release_Notes. [Accessed: 06-Aug-2019].

[14] J. D. Allen and Unicode Consortium., "The Unicode standard 5.0", Addison-Wesley, 2007.

[15] U. Prot et al., “Tutorial Protege OWL,” Copyr. C Univ. Manchester, March 24, 2011.

[16] A. Duque-Ramos, J. T. Fernández-Breis, R. Stevens, and N. Aussenac-Gilles, “OQuaRE: A square-based approach for evaluating the quality of ontologies”, J. Res. Pract. Inf. Technol., vol. 43, no. 2, pp. 159–176, 2011.

[17] Ontology Engineering group, “OOPS! - OntOlogy Pitfall Scanner!Results.”[Online].Available:http://oops.linkeddata.es/response.jsp. [Accessed: 21-Nov-2018].

Semantic Queries Supporting Crisis Management Systems

Manfred Schenk, Tobias Hellmund,
Philipp Hertweck and Jürgen Moßgraber

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB
Karlsruhe, Germany
Email: {manfred.schenk, tobias.hellmund,
philipp.hertweck, juergen.mossgraber}@iosb.fraunhofer.de

***Abstract*—In crisis management, it is crucial to have up-to-date data available to assess an ongoing situation correctly. This information originates from different sources, such as human observations, various sensors or simulation algorithms with heterogeneous geographic scope. The aggregation of such data is conducted by decision support systems to disburden the end-users from automatable tasks. By using semantic technologies for the integration, these systems can benefit from the expressional power of semantic queries. Therefore, all data that is available in the system should also be accessible for these queries. In the following, we present an approach how sensor data can be accessed through semantic queries and how geospatial knowledge can be integrated in a Decision Support System.**

***Keywords*–decision support; semantic access to sensor observations; GeoSPARQL; geospatial semantic queries.**

I. Introduction

Being a specialized sub-category of a Decision Support System (DSS), crisis management systems are designed to support authorities in handling a crisis by providing collected and analyzed information as well as simulation results. While in the past the lack of information was one of the biggest challenges, the situation has changed now. Since the advent of the Internet of Things (IoT) placing great numbers of connected sensors the amount of available data has increased [1]. This results in new challenges: The decision makers have to be protected from information overflow [2] and the input from different heterogeneous sources has to be integrated and automatically analyzed.

Why is geospatial knowledge important in crisis management? A crisis situation usually will be limited to a certain area in most cases. When a high level of water in a river causes flooding of the land along the riverbanks, this flooding will be limited to a certain area surrounding the river. The knowledge about this area allows reducing the amount of data which has to be considered in the crisis management workflows. Thus, the location of a sensor or its observation can be used for filtering data that is of no value for the current crisis; the location of a human observer can be used to estimate the validity of a given statement; the geometry of an urban district can be used to determine if the district is affected by some event.

Our approach is to use semantic technologies to integrate information from different sources and then apply reasoning to draw conclusions from the gathered information focusing on the semantics of geospatial data. While Kontopoulos et al. presented the overall approach in [3] and [4], this paper concentrates on two parts of the approach: firstly, accessing of time series data, e.g. sensor observations and, secondly, exploiting geospatial knowledge.

The paper is structured as follows: Related Work introduces preliminaries and gives a brief overview of recent work. In Section 3 - Using geospatial knowledge, the motivation for the semantic integration of geospatial information in a crisis situation given. It is described how geospatial semantics can be integrated in a machine-understandable manner. Further, geoSPARQL [5] with possible applications is introduced. Section 4 - Semantic Access to Sensor Data - gives different approaches how sensors and their geospatial semantics can be integrated into a DSS. We present our approach for the integration and retrieval of such data with the standard SensorThings API [6] and ONT-D2RQ [7]. Section 5 - Use-Case Application describes the application of the developed functionalities within the project beAWARE [8]. We give an exemplaric query and analyze its architecture. Section 6 evaluates the approach on a qualitative base. Section 7 concludes our findings and gives an outlook on future tasks.

II. Related Work

Parts of the work presented here are based on the Open Geospatial Consortium (OGC) standard GeoSPARQL [5]. Battle and Kolas [9] provide a good introduction into GeoSPARQL. Zhang et al. [10] have already considered GeoSPARQL useful in the area of crisis management. While they concentrated on performance improvements of the GeoSPARQL implementations, our focus is how to bring all parts together. Nishanbaev et al. provided a survey of geospatial semantic web for cultural heritage [11]. While the survey focused on cultural heritage, major parts of it are also valid for the topic of crisis management.

The integration of databases into semantic systems has been adressed by Bizer and Seaborne [12] in their description of the D2RQ project. Hert et al. [13] also provided a comparison of several relational database to RDF mapping languages. Santipantakis et al. [14] also described the use of database to ontology mapping systems for the maritime domain. A discussion on the integration of sensors into a crisis management system can be found in [15]. The beAWARE ontology, which was used for the evaluation of our approach has been presented in [3]. The ontology integrates aspects of the domain sensors and observations, as well as metadata for geospatial information. Hence, we continue using this ontology.

III. Using geospatial knowledge

As already stated in the introduction, locations of events, places, buildings, creatures, etc., play an important role in crisis management. Locations can be expressed in different ways: Technical documentation will most likely provide coordinates of objects like sensors or buildings. In contrast, people tend to use symbolic locations, e.g. ”The fire at the cathedral in Paris” or ”the traffic accident at the crossing of St. James Street and Independance Avenue”. Sometimes, these symbolic locations are also called well-known places. Therefore, all those different types of location descriptions should be supported and there should be some automatic mapping between them. The resolving of symbolic locations is done with the help of gazetteers. A gazetteer is a geographical dictionary providing several information about the recorded well-known places, e.g. the location or geometry, the population, etc. A popular example available on the internet is GeoNames [16]. Some of those gazetteers can even be accessed via SPARQL and therefore can be integrated into a semantic crisis management system.

In some use-cases, the amount of data can be further restricted to increase the usability of the system. If some events are visualized on a map, only the events that are located inside the visual part of the map are of interest. Therefore, the semantic query used for populating the map should utilize the information about the map’s viewport. In this case, we have two geospatial restrictions for the data: process only data which is located inside the area which is affected by the crisis situation, process only data which is located inside the area visualized by the current viewport.

There are two preconditions for the use of such geospatial knowledge as part of semantic queries: First, the collected data has to be in relation with a location or geometry. This can either be explicit, e.g. the documented location of a sensor, or the geospatial information can be inferred, e.g. some textual statement mentions a well-known place. The second requirement is geospatial support of the semantic query engine. The GeoSPARQL extension from the OGC addresses this requirement. It provides SPARQL extension functions for geographic information. According to the standard document [5], GeoSPARQL provides the following features:

- An RDF/OWL vocabulary for representing spatial information consistent with the Simple Features model [17]
- A set of SPARQL extension functions for spatial computations
- A set of RIF (Rule Interchange Format [18]) rules for query transformation (not in the scope of this paper)

The vocabulary defines top-level spatial vocabulary components, as well as geometry vocabulary and topological relation vocabulary. Since the definition of relations can follow different approaches, GeoSPARQL supports three of these *relation families*: The *Simple Features* family follows the OpenGIS Simple Features specification [17], the *Egenhofer* family follows the formal definition of binary topological relationships by Egenhofer [19] and the *RCC8* family follows the Region Connection Calculus by Randell et al. [20]. The SPARQL extension functions can be divided into topological and non-topological query functions. The topological functions contain relations like *equals*, *intersects*, *touches*, *contains*, *overlaps*, etc. and are defined for each of the different relation families mentioned above. The non-topological functions contain relations like *distance*, *buffer*, *convexHull*, etc.

Geometries can have different numbers of dimensions: points (0-dimensional), lines (1-dimensional) and areas (2-dimensional). The function *equals* can be applied to all geometry types and expresses that two instances of SpatialObject are topologically equal, i.e. their interiors intersect and no part of the interior or boundary of one geometry intersects the exterior of the other. The function *intersects* can be also be applied to all geometries and states that both geometries have at least one point in common. The function *touches* can be applied to all geometries with a dimension greater than zero and expresses that both geometries have at least one boundary point in common, but no interior points. The function *overlaps* can be applied only to geometries with the same dimension. It states that they have some but not all points in common and the intersection of their interiors has the same dimension as the geometries themselves. The function *distance* returns the shortest distance in units between any two points in the two geometries as calculated in the spatial reference system of the first geometry. The function *buffer* returns a geometric object that represents all points whose distances from the given geometry are less than or equal to the given radius value measured in the given units. The calculations are made within the spatial reference system of the given geometry. Finally, the function *convexHull* returns a geometric object that represents all points in the convex hull of the given geometry.

The following enumeration lists examples of competency questions from a crisis management system, which could benefit from the integration of geospatial knowledge:

1) Which is the area with most people involved in an incident?
2) Which is the area with the highest density of incidents?
3) Will the playground be affected by the estimated flood zone?
4) Is there enough shelter capacity for all people affected by a specific incident type within a certain radius?
5) Where are the nearest sensors for a specific incident?

How can GeoSPARQL be used for answering those questions? In the following examples, the functions from the *Simple Features* relation family are used. The snippets are simplified to show only the GeoSPARQL part of the whole query to keep it short. For the first question, the locations of all incidents have to be collected and then the number of persons involved has to be aggregated for the different areas. In this case, the *contains* respectively the *sfContains* function determines which locations belong to which area as shown in the following SPARQL snippet.

```
SELECT ?area ?incidentLocation
WHERE {
?area a geo:wktLiteral;
geo:sfContains ?locLiteral.
?incidentLocation geo:asWKT ?locLiteral.
}
```

The second question is similar to the first one. If the geometries of the playground and the estimated flood zone are known, the *intersects* or *overlaps* functions will help answering question three.

```
SELECT ?playground ?estFloodZone
WHERE {
?playGround geo:hasGeometry ?playGeom.
?estFloodZone geo:hasGeometry
    ?estFloodGeom.
?playGeom geo:asWKT ?playWkt.
?estFloodGeom geo:asWKT ?estFloodWkt.
FILTER(geof:overlaps(?playWkt,
    ?estFloodWkt)
}
```

Question four is more complex. At first, the affected area has to be determined. With the help of the *buffer* function the area for potential shelters is calculated. After that, the matching shelters can be associated with those areas by using the *contains* function and their capacity can be summed up per area.

```
SELECT ?affectedZone ?incidentLocation
WHERE {
?incidentLocation geo:asWKT ?wktIncident.
BIND (geof:buffer(?wktIncident,
    uom:metre) AS ?affectedZone)
}
```

```
SELECT ?affectedZone ?shelterLocation
WHERE {
?shelterLocation geo:asWKT ?wktShelter.
?affectedZone geo:sfContains ?wktShelter.
}
```

The *distance* function will help answering the last question if the location of the sensors and the incident location is known.

```
SELECT ?sensorLocation ?incidentLocation
    ?distance
WHERE {
?sensorLocation geo:asWKT ?wktSensor.
?incidentLocation geo:asWKT ?wktIncident.
BIND (geof:distance( ?wktSensor,
    ?wktIncident, uom:metre) as ?distance)
}
ORDER BY DESC(?distance)
```

By using the functions mentioned above together with the standard SPARQL features, queries that are more complex can be built for the given questions. If some locations are provided as well-known places, they will have to be mapped to coordinates by some gazetteer services before they can be used together with the GeoSPARQL functions. If those gazetteers provide a SPARQL interface, the mapping can be done as part of a federated query, meaning that some parts of a query are sent to remote SPARQL endpoints and will be executed there.

IV. Semantic Access to Sensor Data

Changes in the environment can be observed by sensors through a large number of parameters. Since these sensors are usually connected to the internet, a large amount of information is generated, which can be used as possible input for decision support systems. Since we focus on the semantics of the available data, the information should be semantically integrated to make it accessible through semantic queries.

This integration could be done in several ways (in the following sections, the term *sensor management system* is used with the following meaning: *A system that manages sensor measurements and provides standardized access to the data measured by them.*):

- Sensors store their data directly into an ontology or use some generic adapter for this task
- An existing sensor management system is enhanced by a SPARQL interface
- The SPARQL queries are mapped to some query interface of an existing sensor management system
- The database of an existing sensor management system is accessed via some adaption layer which provides a SPARQL interface

The first option is enhancing each sensor in a way it stores its data as described by the used ontology. However, since this approach would require changes to each sensor it would counteract the idea to use existing sensors as input. Even if there were some generic adapters available for this task, another problem would still exist: While some raw observation data might not be suitable for the user of a Decision Support System, the results of some simulations or forecasts based on these observations are of greater value. Adding such simulation and processing functionality to these adapters also would make them less generic and increase their complexity. Furthermore, it could lead to massive concurrent write access to the ontology, causing a permanent index rebuild according to Pan et. al [21]. Therefore, this approach was not continued any further.

Another solution would be to integrate a SPARQL interface into an existing sensor management system. This would leave the task of collecting and storing the sensor data to an existing software, which follows a popular standard and therefore enables the use of a great number of already existing sensors. The OGC SensorThings API [6] is such a standard from the OGC, which provides a unified way to interconnect Internet of Things devices, data, and applications over the web. From the list of available implementations, the FROST-Server [22] is used for the current research. It uses a PostgreSQL relational database for the actual storage of the sensor data. Since the FROST-Server is an open-source implementation, integrating an additional SPARQL interface should be feasible. The second part of the SensorThings API standard already specifies how simulation and processing tasks can interact with an implementation of the standard.

Some small changes to this second solution leads to the third possibility: While the SPARQL interface of the previous solution would have direct access to the internals of the FROST-Server, in this approach it would be decoupled by only using the existing query interface of the server. This approach would not require changes to the implementation of the FROST-Server, but the mapping between SPARQL and the query interface would have to be implemented from scratch.

A fourth approach would be to access the underlying database of the FROST-Server via some adaption layer instead of enhancing the server itself. This would keep the server

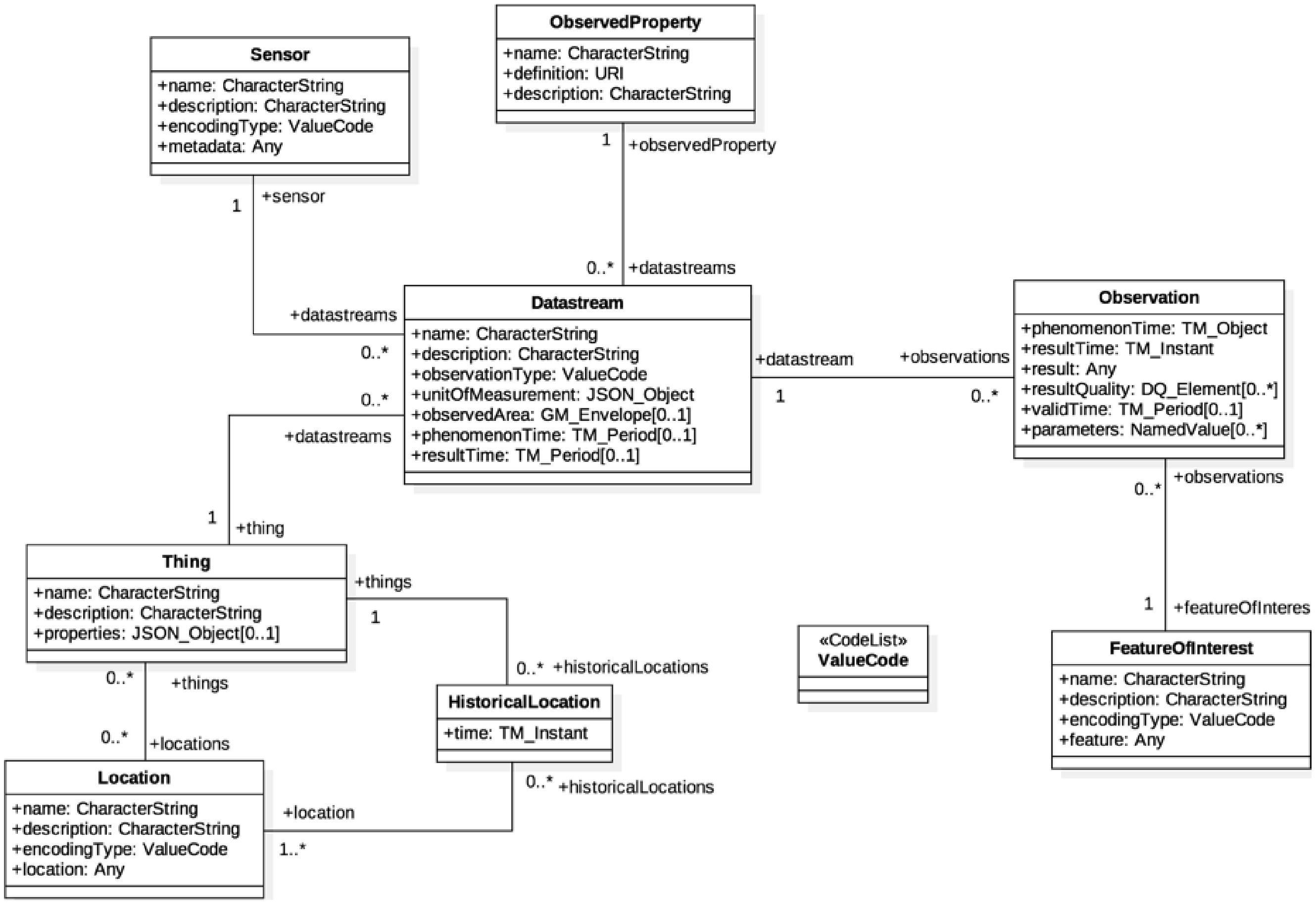


Figure 1. SensorThings Datamodel (from the OGC SensorThings API standard)

implementation simple by keeping out functionality, which is only used in some use cases. By using this approach, the original problem *Semantic access to sensor data* is transformed into the problem *Semantic access to relational databases*. For the transformed problem, there are already multiple solutions available: one of them, the D2RQ project [23] provides a generic implementation for accessing relational databases as virtual, read-only RDF graphs [12] [13]. The original project has been extended in the meantime by the community as part of the project ONT-D2RQ [7] to support OWL in addition to RDF.

Since the fourth approach promises to reach the goal with only a minimum of software development, it has been chosen and the idea of integrating a SPARQL interface directly into a sensor management system has been postponed.

The D2RQ-Server requires a mapping between the database tables and an ontology. As a first attempt, a simple ontology is created, representing the data model of the SensorThings API (see Figure 1). Parts of this ontology are integrated into the beAWARE Ontology, presented in [3]. The entities of this model are directly mapped to tables in the implementation of the FROST-Server. Therefore, table mapping means entity-mapping in our case. Each of the main entities is represented by an OWL class. The relations between entities are modeled as ObjectProperties. Finally, DatatypeProperties are used for the attributes of an entity. Since the relations of the SensorThings data model are not directed, there has to be an inverse ObjectProperty defined for each ObjectProperty. The actual mapping for ONT-D2RQ is defined in a configuration file.

Using the new modular Semantic Sensor Network Ontology (SSN, [24] would have been an alternative to the creation of an own simple ontology, but the complexity of the mapping would have been higher and some aspects of the SensorThings API which are not covered by the SSN would have required some additional extensions as well.

V. Use-Case Application

The integration of geospatial knowledge and sensor observation data with the help of ontologies (as introduced in Section 3 and 4) has been evaluated as part of the research project beAWARE: “Enhancing decision support and management services in extreme weather climate events” [8]. The goal is to develop an integrated solution to manage climate-related crises in all phases, starting with early warning before, managing during and recovery after the event.

A great number of sensors has been connected to the system for the surveillance of water levels, temperatures,

```
1  PREFIX webgenesis: <http://arqext.webgenesis.de/>
2  PREFIX beaw: <http://beaware-project.eu/beAWARE/#>
3  PREFIX sth: <http://www.iosb.fraunhofer.de/frost#>
4  PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
5  PREFIX uom: <http://www.opengis.net/def/uom/OGC/1.0/>
6
7  SELECT ?location ?jsonLoc  ?wktLocInc ?wktLoc ?distance
8  WHERE {
9           {SELECT ?location ?jsonLoc ?wktLocInc WHERE {
10          SERVICE <http://localhost:2020/sparql/> {
11              ?location a sth:Location;
12                        sth:LOCATIONS_LOCATION ?jsonLoc.
13          }
14                  ?wktLocInc webgenesis:convertGeoJSON2WKT ?jsonLoc.
15          }}
16      {{
17            ?incidentReportLocation beaw:latitude  _:lat;
18                                    beaw:longitude _:lon.
19            ?wktLoc webgenesis:convertLatLon2WKT (_:lat _:lon) .
20      }}
21    BIND ( geof:distance( ?wktLoc, ?wktLocInc, uom:metre) AS ?distance) .
22 }
```

Figure 2. A geoSPARQL-enhanced SPARQL query retrieving geospatial information

humidity, etc. The sensor observation data is stored inside the FROST-Server via the SensorThingsAPI interface. For the SPARQL access to these observations, the D2RServer from ONT-D2RQ has been chosen. It also supports OWL, whereas the original D2RQ implementation only supported RDF.

Since the different sources for crisis management system do not necessarily use the same representation for geospatial location data, we have implemented additional custom SPARQL extensions for converting between those representations, e.g. the *webgenesis:convertGeoJSON2WKT* property function converts the GeoJSON geometries delivered by the FROST-Server to a WKT representation required by GeoSPARQL.

Figure 2 shows a SPARQL query combining GeoSPARQL with information from within the FROST-Server. This example is constructed from several subqueries and uses the federated queries feature of SPARQL:

1) The subquery shown in lines 10 to 13 is sent to the SPARQL-endpoint of the ONT-D2RQ Server and processed there. This endpoint will then return all locations which have the DatatypeProperty sth:LOCATIONS_LOCATION set.
2) In the next step, those results are converted from GeoJSON to WKTLiterals by the use of custom property functions. This is the purpose of the subquery shown in lines 9 to 15.
3) For the incident locations, there is also a previous conversion step to have the data available as WKTLiterals. This conversation step is shown in lines 16 to 20.
4) Finally, the converted locations are used as part of a GeoSPARQL query to determine the distance between those observation locations and incident locations from the beAWARE ontology.

Similar SPARQL queries have been developed for the competency questions of the beAWARE decision support system. The query results are presented to the decision maker by means of different visualizations ranging from a simple list to geographical maps. The applicability of the approach used for the beAWARE decision support system was validated by two large-scale trials up to now an in one up-coming third trial. The evaluation report [25] of the first trial is publicly available on the project website [8].

VI. Evaluation

The feasibility of our approach could be shown, but also some problems were identified which have to be addressed before the implementation can be used in production: The performance of the mapping between the relational database and the semantic representation is not fast enough. This causes network timeouts of the SPARQL requests and degrades the usability of the system, if the user has to wait several minutes for the results of queries which were expected to be simple queries. Since the author of the SPARQL queries does not know how these queries will be mapped to SQL queries by the D2RQ implementation, the mapping may produce non-optimal SQL statements. The original D2RQ implementation already provided some optimizations to address this deficiency when handling large datasets, but it seems that not all of them are still present in the ONT-D2RQ implementation. Since both implementations provide different feature sets and the common features of both are not sufficient for our system, a direct comparison with a defined dataset was not possible.

The use of federated queries has been identified as another bottleneck: Since the variable bindings of such a query have to be transferred to the remote endpoint via the network and after the execution, the results will traverse the network again, some delay caused by the network has to be considered. In particular, the missing support of *LIMIT* and *OFFSET* restrictions for the federated queries makes the situation even more badly. Even in the cases where only a few results are needed from the remote SPARQL endpoint, that endpoint has to process the whole dataset and return a possibly large number of results, despite the fact that most of these results will then be thrown away by some *LIMIT* or *OFFSET* clause of the surrounding query.

In November 2019, the last beAWARE pilot will be conducted. Here, a quantitative benchmark will be performed. With these results, measures to improve the performance will be developed.

VII. Conclusion

In this paper, we presented the usage of GeoSPARQL for the integration of geospatial knowledge into a Decision Support System for crisis management based on semantic technologies. Competency questions that are of interest for such DSS were introduced, as well as their formal representations as geoSPARQL enhanced SPARQL queries. To access sensor data on a semantic base, the data must be semantically integrated. We introduced four possible ways how this could be achieved and how sensor observation data could be accessed from within such a system. In the Use-Case Application section, we discussed how these parts were combined within the EU-funded project beAWARE. As further task, performance optimization has been identified. In November, the last beAWARE pilot will take place in Valencia, where quantitative measures will be implemented to evaluate the systems performance. Finally, the evaluation of the approach within the project has been presented.

Acknowledgment

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 700395 and Grant Agreement No 700475.

References

[1] T. Usländer et al., "The trend towards the internet of things: what does it help in disaster and risk management?" Planet@ Risk, vol. 3, no. 1, 2015, pp. 140–145.

[2] M. van den Homberg, R. Monné, and M. R. Spruit, "Bridging the information gap: mapping data sets on information needs in the preparedness and response phase," Technologies for Development, 2018, p. 213.

[3] E. Kontopoulos, P. Mitzias, J. Moßgraber, P. Hertweck, H. van der Schaaf, D. Hilbring, F. Lombardo, D. Norbiato, M. Ferri, A. Karakostas et al., "Ontology-based representation of crisis management procedures for climate events." in ISCRAM, 2018, pp. 1064–1073.

[4] E. Kontopoulos, P. Mitzias, S. Dasiopoulou, M. J., S. Mille, P. Hertweck, T. Hellmund, A. Karakostas, S. Vrochidis, L. Wanner, and I. Kompatsiaris, "Applying semantic web technologies for decision support in climate-related crisis management," in Proceedings Citizen Observatories for natural hazards and Water management. 2nd Int. Conf. on Citizen Observatories for natural hazards and Water Management (COWM 2018). COWM, 2018.

[5] M. Perry and J. Herring, "Ogc geosparql-a geographic query language for rdf data," OGC implementation standard., 2012, last access date: 2019-09-05. [Online]. Available: https://www.opengeospatial.org/standards/geosparql

[6] S. Liang, C.-Y. Huang, T. Khalafbeigi et al., "Ogc sensorthings api-part 1: Sensing," OGC R Implementation Standard. Available online: http://docs. opengeospatial. org/is/15-078r6/15-078r6. html (accessed on 14 April 2018), 2016.

[7] A database to owl mapper. Last access date: 2019-09-05. [Online]. Available: https://github.com/avicomp/ont-d2rq

[8] beware project homepage. Last access date: 2019-09-05. [Online]. Available: https://beaware-project.eu/

[9] R. Battle and D. Kolas, "Geosparql: enabling a geospatial semantic web," Semantic Web Journal, vol. 3, no. 4, 2011, pp. 355–370.

[10] C. Zhang, T. Zhao, L. Anselin, W. Li, and K. Chen, "A map-reduce based parallel approach for improving query performance in a geospatial semantic web for disaster response," Earth Science Informatics, vol. 8, no. 3, 2015, pp. 499–509.

[11] I. Nishanbaev, E. Champion, and D. A. McMeekin, "A survey of geospatial semantic web for cultural heritage," Heritage, vol. 2, no. 2, 2019, pp. 1471–1498.

[12] C. Bizer and A. Seaborne, "D2rq-treating non-rdf databases as virtual rdf graphs," in Proceedings of the 3rd international semantic web conference (ISWC2004), vol. 2004. Proceedings of ISWC2004, 2004.

[13] M. Hert, G. Reif, and H. C. Gall, "A comparison of rdb-to-rdf mapping languages," in Proceedings of the 7th International Conference on Semantic Systems. ACM, 2011, pp. 25–32.

[14] G. Santipantakis, K. I. Kotis, and G. A. Vouros, "Ontology-based data integration for event recognition in the maritime domain," in Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics. ACM, 2015, p. 6.

[15] J. Moßgraber, D. Hilbring, H. van der Schaaf, P. Hertweck, E. Kontopoulos, P. Mitzias, I. Kompatsiaris, S. Vrochidis, and A. Karakostas, "The sensor to decision chain in crisis management." in ISCRAM, 2018, pp. 754–763.

[16] Geonames gazetteer. Last access date: 2019-09-05. [Online]. Available: https://www.geonames.org

[17] Opengis simple features specification for sql. Open GIS Consortium and others. Last access date: 2019-09-05. [Online]. Available: http://www.opengeospatial.org/standards/sfs (1999)

[18] M. Kifer and H. Boley, "Rif overview," W3C working draft, W3C, 2013, last access date: 2019-09-05. [Online]. Available: http://www.w3.org/TR/rif-overview

[19] M. J. Egenhofer, "A formal definition of binary topological relationships," in International conference on foundations of data organization and algorithms. Springer, 1989, pp. 457–472.

[20] D. A. Randell, Z. Cui, and A. G. Cohn, "A spatial logic based on regions and connection." KR, vol. 92, 1992, pp. 165–176.

[21] Z. Pan, T. Zhu, H. Liu, and H. Ning, "A survey of rdf management technologies and benchmark datasets," Journal of Ambient Intelligence and Humanized Computing, vol. 9, no. 5, 2018, pp. 1693–1704.

[22] Fraunhofer open source sensorthings api server. Last access date: 2019-09-05. [Online]. Available: https://github.com/FraunhoferIOSB/FROST-Server

[23] Accessing relational databases as virtual rdf graphs. Last access date: 2019-09-05. [Online]. Available: https://github.com/d2rq/d2rq

[24] Semantic sensor network ontology. W3C. Last access date: 2019-09-05. [Online]. Available: https://www.w3.org/TR/vocab-ssn/ (2017)

[25] F. Lombardo, D. Norbiato, M. Ferri, I. Vourvachis, M. Meliadis, A. Karakostas, I. Koulalis, T. Hellmund, and I. Koren, "D2.4 evaluation report of the 1st prototype," in beAWARE H2020 project deliverable, last access date: 2019-09-05. [Online]. Available: https://beaware-project.eu/wp-content/uploads/2019/01/D2.4_-beAWARE_Evaluation-report-of-P1_v0.5.pdf

Knowledge Base Completion With Analogical Inference on Context Graphs

Nada Mimouni and Jean-Claude Moissinac and Anh Tuan Vu

LTCI, Télécom Paris
Institut Polytechnique de Paris
France
Email: nada.mimouni, jean-claude.moissinac, anh.vu @telecom-paris.fr

***Abstract*—Knowledge base completion refers to the task of adding new, missing, links between entities. In this work we are interested in the problem of knowledge Graph (KG) incompleteness in general purpose knowledge bases like DBpedia and Wikidata. We propose an approach for discovering implicit triples using observed ones in the incomplete graph leveraging analogy structures deducted from a KG embedding model. We use a language modelling approach where semantic regularities between words are preserved which we adapt to entities and relations. We consider excerpts from large input graphs as a reduced and meaningful context for a set of entities of a given domain. The first results show that analogical inferences in the projected vector space is relevant to a link prediction task.**

Keywords–Knowledge Base; Context graph; Language embedding model; Analogy structure; Link discovery.

I. Introduction

General purpose knowledge Bases (KB), such as Yago, Wikidata and DBpedia, are valuable background resources for various AI tasks, for example recommendation [1], web search [2] and question answering [3]. However, using these resources bring to light several problems which are mainly due to their substantial size and high incompleteness [4] due to the extremely big amount of real world facts to be encoded. Recently, vector-space embedding models for KB completion have been extensively studied for their efficiency and scalability and proven to achieve state-of-the-art link prediction performance [5], [6], [7], [8]. Numerous KB completion approaches have also been employed which aim at predicting whether or not a relationship not in the KG is likely to be correct. An overview of these models with the results for link prediction and triple classification is given in [9]. KG embedding models learn distributed representations for entities and relations, which are represented as low-dimensional dense vectors, or matrices, in continuous vector spaces. These representations are intended to preserve the information in the KG namely interactions between entities like similarity, relatedness and neighbourhood for different domains.

In this work, we are particularly interested in adapting the language modelling approach proposed by [10] where relational similarities or linguistic regularities between pairs of words are captured. They are represented as translations in the projected vector space where similar words appear close to each other and allow for arithmetic operations on vectors of relations between word pairs. For instance, the vector translation $v(Germany) - v(Berlin) \approx v(France) - v(Paris)$ shows relational similarity between countries and capital cities. It highlights clear-cut the analogical properties between the embedded words expressed by the analogy "$Berlin$ is to $Germany$ as $Paris$ is to $France$". We propose to apply this property to entities and relations in KGs as represented by diagrams (a) and (b) in Figure 1. The vector translation example is likely to capture the $capital$ relationship that we could represent by a translation vector $v(capital)$ verifying the following compositionality [10]: $v(France)+v(capital)-v(Paris) \approx 0$. We use the analogical property for KB completion and show that it is particularly relevant for this task. Our intuition is illustrated by diagrams (b) and (c) in Figure 1 where an unobserved triple can be inferred by mirroring its counterpart in the parallelogram. To the best of our knowledge, leveraging analogy structure of linguistic regularities for KB completion has never been investigated prior to this work. We consider to apply such properties on excerpts from large KGs, we call context graphs, guided by representative entities of a given domain where interactions between entities are more significant. Context graphs show to be bearer of meaning for the considered domain and easier to handle because of their reduced size compared to source graphs.

In the following, Section II gives an overview of related work, Section III describes our approach to build context graphs and learn features for link prediction and Section IV gives the initial results.

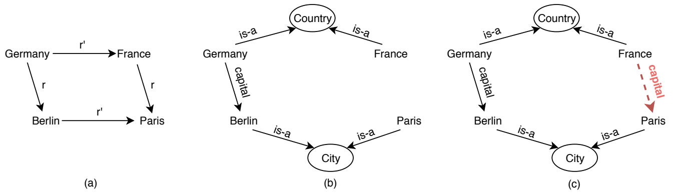


Figure 1. (a) Analogy relation diagram (parallelogram) between countries and capital cities. In KGs (b) and (c), r corresponds to the relation $capital$ and r' is decomposed into two type relations (*is-a*) to concepts $Country$ and $City$.

II. Related work

A closely related approach to our work is described in [11]. The RDF2vec approach uses the neural language model to generate embeddings on entities from walks on two general knowledge bases namely DBpedia and Wikidata. Short random walks are created for the whole set of entities in an image of the KB at a given date. Walks using RDF graph kernels are also used on small test datasets due to scalability limitation. The trained models are made available for reuse. The approach we propose here differs in several aspects. First, we consider undirected labelled edges in the RDF graph to adapt the neural language model, compared to directed graph. Second, we use biased walks guided by the application domain to generate sequences compared to random walks. Third, beside using object properties to build the sequences, we consider DataType properties and literals because we assume that they hold useful information for our application domain (e.g., dates, textual descriptions). Last, we propose to handle scalability issues by contextualizing the input graphs assuming that more relevant information is centralized within a perimeter of α hops around our main entities (α is defined later).

A more general technique called Node2vec is proposed in [12]. It aims to create embeddings for nodes in a (un)directed (a)cyclic (un)weighted graph $G(V, E, W)$ where V is the set of vertices, E the set of edges with weights W. The embeddings are learnt using the Skip-Gram model [10] trained on a corpus of sequences of nodes generated using the sampling strategy. The input graph is turned into a set of directed acyclic sub-graphs with a maximum out degree of 1 using two hyper-parameters for sampling: Return R (probability to go back to the previous node) and Inout Q (probability to explore new parts of the graph).

III. Approach

Here, we define a context graph and show how to build it, then we present how to create a model from our context graph.

A. Building Context Graphs

We define a Context Graph (CG) as a sub-graph of a general KG (e.g. DBpedia) representative of a domain D. The first step to build CG is to identify a list of seeds defining the domain. A seed is an entity from KG corresponding to a concept which is considered relevant for D. For example, if the concept is '*Musée du Louvre*', the corresponding entity in DBpedia is <http://dbpedia.org/resource/Louvre>. In some domains this list is obvious as for museums, hotels or restaurants. In general case, the common practice is to rely on a reference dataset (such as IMDB for cinema).

The second step extracts from KG the neighbourhood for each seed within a given depth filtering useless entities (not informative for D) and returns the final CG as the union of elementary contexts. We use CG in the following as basis for the embedding model.

We create the algorithm CONTEXT BUILDER (Algorithm 1) to build a context graph $context$ from a knowledge graph $\mathcal{KG}$ for a given domain D. For a set of seeds ($seedsEntities$), findNeighbors(s) extracts a neighbouring context C_v from a knowledge graph $\mathcal{KG}$ for each seed s. The final context, $context$, is updated adding C_v. A list of new seeds, $newSeeds$, is updated with the new collected entities after filtering the terminal nodes with the EntityFilter method. The exploration depth $level$ is incremented by 1 at each step up to the desired $radius$ limit. At the end of process, the resulting context $context$ is expanded with the classes of entities extracted from $\mathcal{KG}$ by the methods AddClasses and Entities.

Algorithm 1: $CONTEXT\ BUILDER$

```
Function ContextBuilder($KG$, $seedsEntities$, $radius$, $filteredEntities$)
    Input  : A knowledge graph $KG$
    A neighbourhood depth to reach $radius$
    A set of entities which are used as seeds $seedsEntities$
    A set of entities which are excluded from the seeds $filteredEntities$
    Output : Context Graph $context$
    $level \leftarrow 0$
    $context \leftarrow \emptyset$
    while $level$ < $radius$ do
        $newSeeds \leftarrow \emptyset$
        foreach $s \in seedsEntities$ do
            $C_v \leftarrow$ FindNeighbors($KG$, $s$)
            $context \leftarrow context \cup C_v$
            $newSeeds \leftarrow newSeeds \cup$
              EntityFilter($C_v$,$filteredEntities$)
        end
        $level \leftarrow level +1$
        $seedsEntities \leftarrow newSeeds$
    end
    $context \leftarrow context \cup$ AddClasses($KG$,
      Entities($context$))
    return $context$
```

B. Feature Learning

First, we adapt the language modelling approach to KG embedding. We transform the entities and relations in the CG as paths that are considered as sequences of words in natural language. To extract RDF graph sub-structures, we use the breadth-first algorithm to get all the graph walks or random walks for a limited number N. Let $G = (V, E)$ be an RDF graph where V is the set of vertices and E is the set of directed edges. For each vertex v, we generate all or N graph walks P_v of depth d rooted in the vertex v by exploring direct outgoing and incoming edges of v and iteratively direct edges of its neighbours v_i until depth d is reached. The paths after the first iteration follow this pattern $v \rightarrow e_i \rightarrow v_i$ where $e_i \in E$. The final set of sequences for G is the union of the sequences of all the vertices $\bigcup_{v \in V} P_v$.

Next, we train a neural language model which estimates the likelihood of a sequence of entities and relations appearing in the graph and represents them as vectors of latent numerical features. To do this, we use the continuous bag of words (CBOW) and Skip-Gram models as described in [10]. CBOW predicts target words w_t from context words within a context window c while Skip-Gram does the inverse and attempts to predict the context words from the target word. The probability $p(w_t|w_{t-c}...w_{t+c})$ is calculated using the $softmax$ function.

Finally, we extract analogical properties from the feature space to estimate the existence of new relationships between entities. We use the following arithmetic operation on the feature vectors (entities of Figure 1): $v(Berlin) - v(Germany) + v(France) = v(x)$ which we consider is solved correctly if $v(x)$ is most similar to $v(Paris)$. On the left-hand side of the equation, entities contribute positively or negatively to the similarity according to the corresponding sign. For example, $Germany$ and $France$ having the same type $Country$ contribute with different signs, $Berlin$, of a different $City$ type, contribute with the opposite sign of the corresponding Country. The right-hand side of the equation contains the missing corner of the diagram which remains to be predicted. We then use cosine similarity measures between the resulting vector $v(x)$ and vectors of all other entities of the same type in the embedding space (discarding the original ones of the equation) in order to rank the results.

IV. Experimental Evaluation

A. Case Study

We test our approach on a sub-graph of DBpedia representing a target domain: here we chose museums of Paris. We propose to address the scalability issue by contextualizing the input graphs assuming that more relevant information is centralized within a perimeter of α hops around main entities of this domain (we used $\alpha = 2$ as suggested by [13]). We build our KG as the union of individual contextual graphs of all entities representing the input data from the cultural institution *Paris Musées* (12 sites). We identify each site by its URI on DBpedia-fr after an entity resolution task (in the following, we denote the URI http://fr.dbpedia.org/resource/entity shortly as dbr:entity). The final graph contains 448309 entities, 2285 relations and 5122879 triples. To generate sequences of entities and relations we use random graph walks with $N = 1000$ for depth $d = \{4, 8\}$. We also consider for each entity all walks of depth $d = 2$ (direct neighbours).

We then train the Skip-Gram word2vec model on the corpus of sequences with the following parameters: window size = 5, number of iterations = 10, negative samples = 25 (for the purpose of optimisation) and dimension of the entities' vectors = 200. We use gensim implementation of word2vec [14]. We also trained our model with CBOW method and with larger vector dimension (500). We notice in general better performance with Skip-Gram method, but cannot do any assertion about vector dimension. Our method can't be evaluated against other ones using standard datasets such as FB15K, WN18 [6], [7]. It requires to define a context and extracts a subgraph from it, none of the other methods uses such a context in the available evaluations.

B. Evaluation Protocol

Existing benchmarks for testing analogy task in the literature are designed for words from text corpora. To the best of our knowledge, using language model driven analogy for link prediction in knowledge graphs has not been investigated yet. To evaluate our approach, we build a ground-truth for analogy between entities in the KG. Each entry corresponds to a parallelogram as described in Figure 1 with one unobserved triple in the KG. For each entity, corresponding to a museum site in our application, we collect a list of well-known artists for this site as follows: find in DBpedia-fr the list of artists (dbo:Artist) or otherwise, individuals (dbo: Person) who are associated with the site. For some sites, we manually create the list, for example by searching for well-known artists for a museum on the website [15].

The evaluation test aims at discovering $artist_a$ for $museum_a$ considering a known triple <$museum_b$, $artist_b$> while varying b and measuring the mean of the returned results. We use conventional metrics: Mean Reciprocal Rank (MRR) and the number of correct responses at a fixed rate (Hits@).

The evaluation protocol is as follows: for each $Muri_i$, URI of a museum, let $Auri_i$ be the URI of the first artist identified for $Muri_i$, consider all $Muri_j \mid j \neq i$, find the top most similar entities of the predicted vectors with positives = [$Auri_i$,$Muri_j$] and negative =[$Muri_i$]. In the list of results, we filter by type $Artist$, we then examine the intersection with artists $Auri_l$ associated with $Muri_j$. It is worth noticing that we frequently find loosely defined links between museums and artists; such links are very common in DBpedia and use the property `wikiPageWikiLink` representing an untyped link. Subsequent work is required to qualify them.

C. Results

Table I shows results of MRR and Hits@$\{3, 5, 10\}$ (%) for $d = \{4, 8\}$ and $N = 1000$. The final row of the table with columns $d = 8$ shows the impact of considering longer paths on the performances of the approach. In fact, longer paths capture richer contexts for entities and results in better vectors estimation by the neural language model.

We compared our approach with the one presented in [11] which creates a model, modelDB, for all entities in DBpedia. For each entity in our ground-truth built on DBpedia-fr, we look for its equivalent in DBpedia and verify that it is contained in the vocabulary of modelDB built with $d = 4$. Only 7 out of 12 museum entities are in modelDB, as well as their first associated artist among others. The analogy tests return globally poor results. ModelDB were unable to retrieve relevant entities in top 100 returned answers, as for our model trained on the CG, without any improvement even if extended to top 5000. This is not a surprising result if we look at the following table which shows that our CG has a better coverage of the ground-truth domain entities, mainly artists, compared to DBpedia.

TABLE II. Ground-truth entities in DBpedia and DDBpedia-fr.

	dbo:Person	dbo:Artist	No type	dbo:Museum
DBpedia	272	190	44	7
DBpedia-fr	272	327	6	12

The first row of Table II shows that not all dbo:Artist are linked to dbo:Person (ex: dbr:Sonia_Delaunay). With 12 museums and 334 artists in the reference list, 97.90% can be identified as an artist in our context graph vs. 56.88% in DBPedia, which partly explains the poor results. As we filter the returned results by type Artist (or more generally by Person), several relevant answers are filtered.

We also compared our approach with a random selection of entities of type dbo:Artist in the vocabulary of the model. The results, given in columns $d = 4R$ of table I, show a great benefit of leveraging the regularities in the vocabulary space to extract relationships between entities.

While analysing values on Table I, we noticed wide discrepancies between results of different museums. For example, Hits@10 values for dbr:Musée_d'art_moderne and dbr:Musée_de_Grenoble are respectively: 0.83 and 0.33. This impacts the global performance of all museums (last row of table I). The result means for the second value that the system was not able to retrieve the corresponding artist for dbr:Musée_de_Grenoble in top returned results. We argue this

TABLE I. MRR AND HITS@$\{3, 5, 10\}$ (%) OF A SUBSET OF REPRESENTATIVE EXAMPLES OF *Paris Musées* DATA FOR $d = \{4, 8\}$ AND $N = 1000$ WITH ANALOGY AND RANDOM FOR $d = 4$ (D=4R).

	MRR			Hits@3			Hits@5			Hits@10		
Entity	d=4R	d=4	d=8	d=4R	d=4	d=8	d=4R	d=4	d=8	d=4R	d=4	d=8
dbr:Musée_Bourdelle	0,05	0,39	0,43	0,09	0,50	0,42	0,18	0,50	0,42	0,18	0,66	0,50
dbr:Musée_Carnavalet	0,01	0,43	0,59	0,00	0,58	0,67	0,09	0,66	0,75	0,09	0,83	0,75
dbr:Musée_Zadkine	0,00	0,43	0,44	0,00	0,41	0,42	0,00	0,50	0,50	0,00	0,50	0,50
dbr:Musée_Cernuschi	0,01	0,42	0,50	0,00	0,50	0,58	0,00	0,58	0,67	0,09	0,75	0,67
dbr:Petit_Palais	0,04	0,38	0,63	0,09	0,50	0,75	0,09	0,66	0,75	0,09	0,66	0,75
dbr:Maison_de_Balzac	0,03	0,23	0,44	0,09	0,25	0,58	0,09	0,41	0,58	0,09	0,41	0,58
dbr:Musée_Cognacq-Jay	0,09	0,33	0,49	0,09	0,33	0,58	0,09	0,33	0,58	0,09	0,33	0,58
dbr:Musée_d'art_moderne	0,04	0,36	0,71	0,09	0,41	0,75	0,09	0,50	0,83	0,09	0,58	0,83
dbr:Musée_Romantique	0,03	0,34	0,48	0,09	0,41	0,50	0,09	0,41	0,58	0,09	0,50	0,58
dbr:Palais_Galliera	0,00	0,36	0,48	0,00	0,50	0,50	0,00	0,50	0,58	0,00	0,50	0,58
dbr:Maison_de_Victor_Hugo	0,01	0,38	0,55	0,00	0,50	0,58	0,00	0,58	0,58	0,18	0,58	0,67
dbr:Musée_de_Grenoble	0,00	0,34	0,33	0,00	0,41	0,33	0,00	0,50	0,33	0,00	0,50	0,33
All entities in *Paris Musées*	0,02	**0,37**	**0,52**	0,04	**0,44**	**0,58**	0,06	**0,51**	**0,62**	0,09	**0,57**	**0,64**

is mostly related to the representativeness of this artist's entity in the KG and how it is linked to the museum's entity; less interlinked entities (directly or indirectly through neighbours) have less chance to be related with the analogy structure in the embedding space. To explain this, we run another evaluation as follows: for each $Auri_i$, URI of an artist, consider all known triples $<Muri_j$, $Auri_j >$ | $j \neq i$, find the top most similar entities of the predicted vectors ranked by similarity. In the list of results, we filter by type $Museum$, we then examine the intersection with museums $Muri_l$ associated with $Auri_i$.

Table III shows results of MRR and Hits@$\{3, 5, 10\}$ (%) for $d = 4$ and $N = 1000$. The wide differences between artists' results in the last column of the table (Hits@10) (ex. dbr:Victor_Hugo and dbr:Geer_Van_Velde) reveals the impact of the triple interlinkage in the graph on the analogy prediction test. Thus, good prediction performance of new triples could be achieved with a good representativeness of known triples by the context graph.

TABLE III. MRR AND HITS@$\{3, 5, 10\}$ (%) OF REPRESENTATIVE EXAMPLES OF ARTISTS EXHIBITED IN MUSEUMS OF *Paris Musées* FOR $d = 4$ AND $N = 1000$

Entity	MRR	Hits@3	Hits@5	Hits@10
dbr:Antoine_Bourdelle	0,61	0,72	0,81	0,81
dbr:Israël_Silvestre	0,08	0,09	0,13	0,13
dbr:Gustave_Courbet	0,38	0,45	0,45	0,72
dbr:Ossip_Zadkine	0,67	0,81	0,90	0,91
dbr:Xu_Beihong	0,74	1,0	1,0	1,0
dbr:Honoré_de_Balzac	0,53	0,72	0,81	0,81
dbr:François_Boucher	0,65	0,72	0,81	0,81
dbr:Geer_Van_Velde	0,09	0,09	0,09	0,09
dbr:Ary_Scheffer	0,62	0,72	0,81	0,91
dbr:Jacques_Heim	0,18	0,18	0,45	0,72
dbr:Victor_Hugo	0,53	0,63	0,72	0,91

V. CONCLUSION

In this paper, we presented an approach for link discovery in KBs based on the neural language embedding of contextualized RDF graphs and leveraging analogical structures extracted from relational similarities which could be used to infer new unobserved triples from the observed ones. The test of our approach on a domain-specific ground-truth shows promising results. We will continue to expand upon the research and compare it with state-of-the-art approaches for KB completion on the standard baselines.

ACKNOWLEDGMENT

This work is conducted as part of the Data & Musée project selected in the $23rd$ call for projects of the Fonds Unique Interministériel (FUI) and certified by Cap Digital and Imaginove. http://datamusee.fr/le-projet/. This work is an extension of a poster paper scheduled for publication in the ESWC 2019 supplementary proceedings.

REFERENCES

[1] M. Al-Ghossein, T. Abdessalem, and A. Barré, "Open data in the hotel industry: leveraging forthcoming events for hotel recommendation," J. of IT & Tourism, vol. 20, no. 1-4, 2018, pp. 191–216.

[2] S. Szumlanski and F. Gomez, "Automatically acquiring a semantic network of related concepts," in Proceedings of the 19th ACM CIKM, 2010, pp. 19–28.

[3] J. Yin, X. Jiang, Z. Lu, L. Shang, H. Li, and X. Li, "Neural generative question answering," in Proceedings of IJCAI. AAAI Press, 2016, pp. 2972–2978.

[4] B. Min, R. Grishman, L. Wan, C. Wang, and D. Gondek, "Distant supervision for relation extraction with an incomplete knowledge base," in Proceedings of NAACL-HLT, 2013, pp. 777–782.

[5] M. Nickel, L. Rosasco, and T. A. Poggio, "Holographic embeddings of knowledge graphs," in AAAI, 2016.

[6] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in NIPS, 2013.

[7] H. Liu, Y. Wu, and Y. Yang, "Analogical inference for multi-relational embeddings," in Proceedings of ICML, 2017, pp. 2168–2178.

[8] A. García-Durán, A. Bordes, N. Usunier, and Y. Grandvalet, "Combining two and three-way embedding models for link prediction in knowledge bases," J. Artif. Intell. Res., vol. 55, 2016, pp. 715–742.

[9] D. Q. Nguyen, "An overview of embedding models of entities and relationships for knowledge base completion," CoRR, vol. abs/1703.08098, 2017.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in NIPS, 2013.

[11] P. Ristoski and H. Paulheim, "Rdf2vec: Rdf graph embeddings for data mining," in Proceedings of ISWC, 2016, p. 498 – 514.

[12] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in ACM SIGKDD, 2016.

[13] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using dbpedia," in Proceedings of WSDM, 2013, pp. 465–474.

[14] "Gensim implementation of word2vec," https://radimrehurek.com/gensim/models/word2vec.html, accessed: 2019-08.

[15] "Paris musée collection website." http://parismuseescollections.paris.fr/fr/recherche, accessed: 2019-03.

Knowledge Representation Frameworks for Terminology Management in Cybersecurity: The OCS Project Use Case

Claudia Lanza*, Elena Cardillo†, Maria Taverniti†, Roberto Guarasci*

*University of Calabria, Rende, Italy
Email: c.lanza@dimes.unical.it; roberto.guarasci@unical.it
†Institute of Informatics and Telematics, National Research Council, Rende, Italy
Email: {elena.cardillo;maria.taverniti}@iit.cnr.it

***Abstract*—Managing technical terms proper to specialized languages, represents one of the main tasks of Knowledge Organization Systems (KOSs). Cybersercurity domain contains a plethora of such terms, with a constant growth of new terms , which still need to be structured and organized from a semantic point of view. This paper aims at providing a presentation of KOSs for organizing specialized terminologies, specifically related to Cybersecurity, starting from a comparison between semantic resources presenting a higher level of semantic representation, i.e., thesauri and ontologies. To show their potentiality in the management of the Cybersecurity technical terminology, an outline of their application within a project carried out at the Institute of Informatics and Telematics of the National Research Council is described, and the distinction between them detailed in the conclusive discussion. A specific focus will be given to the more accurate description that ontologies are able to provide due to the way semantic relationships existing among terms and concepts belonging to a specific field of knowledge are formalized.**

Keywords- Cybersecurity; KOS; Thesauri; Ontologies; Specialized language; Knowledge Representation.

I. INTRODUCTION

Managing technical terms proper to specialized languages, represents one of the main tasks of Knowledge Organization Systems (KOSs). In the context of KOSs, semantic resources, such as thesauri and ontologies, are useful to index documents and to help people during the information searching and retrieval from all types of information resources related to specialized domains, where semantic ambiguity between terms should be avoided. In this scenario, the paper is focused on presenting some of the main differences existing in the way of organizing and representing the information related to highly technical domains, in particular that of Cybersecurity. Amongst the KOSs [1] the comparison will focus on the two mentioned means of semantic knowledge configuration: thesauri and ontologies. The reason why these two types of resources have been selected among the others basically relies on one of the objectives of the *OCS Project* Cyber Security Observatory of the CNR Institute of Informatics and Telematics (IIT-CNR) [2], that will be presented in detail in Section IV. The project concerns the development of an Italian controlled vocabulary, in other words a thesaurus, for the Cybersecurity domain, and the enhancement of semantic connections and representation by exploiting a more interoperable and formal language, i.e., the Web Ontology Language (OWL) [3] the recommended Semantic Web language for authoring ontologies.

Thesauri's main scope is that of structuring information and organizing it in a layered network of semantic connections, and its management and usability is piloted by KOSs functionalities [4][5]. As Soergel affirms in his work, ”A thesaurus is a structured collection of concepts and terms for the purpose of improving the retrieval of information.A thesaurus should help the searcher to find good search terms, whether they be descriptors from a controlled vocabulary or the manifold terms needed for a comprehensive free-text search all the various terms that are used in texts to express the search concept” [6]. In managing information represented by terms proper to specialized language, a thesaurus should provide a reliable and a well structured semantic means to guide the understanding of technical terms representing concepts belonging to a specific field of knowledge. Its indexing function proves to be helpful in the way the users are able to analyze documents according to an informative organization of descriptors. In other words, the abstraction of knowledge occurs indirectly by exploiting terminological units that take on the status of descriptor or indexing unit. The latter is the element that language uses to describe, synthesize and extract information from documents [7].

Another relevant work to understand the aims and the methods for building a thesaurus is that of Broughton [8]. In this work, the author gives light to the main guidelines to develop a semantic tool through which technical concepts can be organized by means of hierarchical, equivalence and associative relations between the terms that represent them [9][10].

The way thesauri are structured follows standardized rules that should be respected, as the ones included in the ISO standards [11][12]. The interoperability of semantic resources, such as thesauri and ontologies, is given by the principle of linked open data [13][14][15], which guarantees a shareable knowledge organization system that can facilitate the coordination among several users for different terminological tasks. On the basis of the idea of generating a language that can guarantee a higher form of interaction between informative systems, without losing the exact meaning of the shared information, the ontology seems to route towards a constant reuse of the managed information by providing conceptual representations of a domain [16][17]. The methods followed for building ontologies observe basic principles that can be found in the guidelines published by Noy and Mcguinness [18] or Bourigault [19].

The paper is structured as follows: Section II shortly gives an overview of main existing resources for Cybersecurity information management, both in English and Italian language. Section III includes related works focused on the construction of KOSs and on Cybersecurity. Section IV describes the construction of the Italian thesaurus for Cybersecurity and its enhancement through an ontological representation. Section V will provide a discussion about the main advantages derived from exploiting thesauri and the ontologies. Finally, Section VI sums up the key issues underlined in the paper giving some conclusions and providing some future perspectives.

II. State of the art

One of the main purposes of this research activity is related to the creation of a semantic resource, a thesaurus, that can be considered as a reliable knowledge organization system that structures the information related to Cybersecurity in Italian language. Indeed, the basis from which the activity has taken inspiration was connected to the absence in Italian language of a highly semantically structured way to manage the terminology of this field of study. Some of the resources that have been taken into account to build a source corpus to be processed in order to obtain a list of representative terms are hereafter summarized.These terms synthesize the concepts belonging to a specific domain and they represented the starting model to realize the ontology for Cybersecurity based on the structure created for the Italian thesaurus. The ontology has been developed with the goal of representing the classes linked to each other through more precise properties that could, at times, specify the interconnections between them better than a flat visualization that belongs to a thesaural organization of terms.

Among the examples of Cybersecurity glossaries and vocabularies, of great importance are: for English, the ones contained in the NIST 7298 [20] and ISO 27000:2016 [21] standards for Information and Communication Technologies (ICT) security, and, for Italian, the Italian book ”*Libro Bianco*” (White Book for Cybersecurity) realized by the National Laboratory of Cybersecurity of the Consorzio Interuniversitario Nazionale per l’Informatica (CINI) [22], which thoroughly sheds light on the key issues related to Cybersecurity guidelines and on the latest related episodes that have changed the mode of conduct to defend informative systems and to conceive some specific concepts proper to Cybersecurity. Another relevant existing resource for Italian is the Italian ”Glossario Intelligence” [23], a technical glossary published by the Presidency of the Council of Ministers, which contains several terms belonging to the Cybersecurity domain and which has been used as basis for the creation of the Italian thesaurus and the ontology for Cybersecurity under investigation.

With respect to ontologies, it is worth mentioning the works targeted at the creation of ontology models for Cybersecurity, i.e., [24][25][26][27], and the studies focused on the approaches for developing an architecture for Cybesecurity standards [28] and enterprise’s Cybersecurity metrics [29]. In particular, in [25] an ontology has been presented, which has been designed to integrate data from different heterogeneous sources, in the absence of a common terminology, offering a sufficiently complete knowledge on the possible threats, thus allowing Organizations to perform reasoning and support decision-making processes related to security.

III. Related works

Processing the information belonging to specific domains of interest involves the analysis of those documents which semantically tend to represent concepts through a technical language [30]. The creation of terminological databases [30] follows some given criteria linked to gathering the related documents that have to constitute the reference corpus from which terms can be retrieved. To achieve this first informative structure, the corpus firstly aims at including documents that can represent the domain in an official way [31], i.e., the gold standards, [32] collecting a terminological standardized repository made up of terms that are meant to be closely specific to the technical field of knowledge under review [33]. To obtain a matching system between the terminology shared by a community of experts from a particular domain and the terms contained in a list derived from the processing of a reference corpus, the documents gathered in the corpus undergo a process of terminology extraction, which shall compare the equivalence between the representative terms of a domain with the ones of the gold standards [34].

This last step is usually implemented by exploiting semi-automatic term extraction tools. Nazarenko *et al.*. [35] and Loginova [36] gave in their works detailed lists of several tools for extracting terminologies from texts. With regards to the Cybersecurity domain and the research activity treated in this paper, various existing sources, both in English and in Italian, have been analyzed in order to retrieve an accurate terminological basis from which to build a more sophisticated semantic resource to guide the knowledge representation process. The intent of this project task, as aforementioned, is to provide an Italian structure, firstly conceived as a thesaurus, to configure the terminology of Cybersecurity in a network of semantic relations that can better orientate to a lexical understanding of specialized concepts represented by terms belonging to this field. The goal of this research activity is also based on the reuse of the terms contained in the thesaurus to realize in a consequential way an ontology system that could support the inclusion of customized properties between classes and more comprehensively clarify the associative relationships of the thesaurus. This represents the reason why ontologies can be usually considered as resources that can provide a more exhaustive and explicit frame for knowledge representation.

IV. THE OCS PROJECT

In this Section the project case will be presented. The activity regarded the creation of a thesaurus in Italian language as a semantic tool to organize the terminology on the Cybersecurity domain. The thesaurus has been inserted amongst the services of the online platform Cyber Security Observatory (OCS) [37].

A. The Cybersecurity context

As previously mentioned, the Cybersecurity domain is mainly characterized by a technical terminology. Given that Cybersecurity is a synergy of different sub-fields, the schematising of this specialized field reflected this high level of heterogeneity. Cybersecurity is permeated by its multidisciplinary nature that involves Information and Communication Technologies (ICT) and its sub-areas, such as, audiovisual techniques, computer software, electronics, by its specificity with respect to technical and standardized terms, and by its cross-fielding thematic coverage, i.e., computer science field, legislative systems, regulations. Given these premises, the treatment of its internal language, that derives from the textual content extracted from the source corpus documents, is meant to be managed by formal semantic systems in order to obtain shareable standardized lists of the domain's representative terms organized according to their semantic relations, which, in turn, will orientate the understanding of the conceptual model of the domain [38].

B. The Italian thesaurus for Cybersecurity

The main focus of this paper is the creation of a semantic tool for the Italian project *Cyber Security Observatory* (OCS) [37], carried out in collaboration with the Institute of Informatics and Telematics of the National Research Council. During this task, while seeking a resource that could represent the Cybersecurity terminological framework and could be used as a service for experts and common users, some of the key differences between thesauri and ontologies in the management and organization of highly technical information and language arose.

Firstly, the choice to privilege a thesaurus structure instead of other semantic resources, such as glossaries or taxonomies, relies on its peculiarity of managing the representative terms of a specific domain as an entangled network of semantic relations that guide the comprehension of a conceptual model proper of a field of knowledge to be studied [8]. In order to obtain the knowledge organization with respect to a structuring system as provided by a controlled vocabulary, i.e., a thesaurus, several guidelines need to be observed [11][12]. These aforementioned standards depict the way the terms, that represent the concepts of a specialized domain, should indicate a unique and an unambiguous meaning (through the use of scope note, SN) and should be connected to each other. As mentioned in Section I, three main basic forms of connections are generated for structuring the information under the basis of thesaurus's modelling [39]:

1) Equivalence relation, marked with the tags *Used (USE)* and *Used For (UF)*
2) Hierarchical relation, marked with the tags *Broader Term (BT)* and *Narrower Term (NT)*;
3) Associative relation, marked with the tag *Related Term (RT)*.

The methodology followed for the realization of the Italian thesaurus for Cybersecurity covered classical sequences. As primary step, the terminology contained in the thesaurus has been extracted from reliable sources which made up the corpus characterized by documents distinctively selected for their content oriented to Cybersecuritity issues [31]. This collection of texts made the information retrieval highly oriented to the domain to be represented [40], and covered different types of documents, such as standards and laws [41], Cybersecurity-related magazines or guidelines and certifications. The conceptual content of these documents was meant to be processed to obtain lists of terms (a glossary) sorted according to statistical measurements able to provide a first semantic schematization [42]. Indeed, the second phase concerned the semi-automatic processing of the information included in the source corpus by exploiting a term extractor software [36] (more specifically the Italian native tool, Text to Knowledge (T2K)) [43] that provided, as outputs, lists of terms ranked according to their occurrence's value in the texts.

Only once having received the validation by the experts of the domain, i.e., the third phase of the methodology, the terms have been selected as candidate terms to be integrated in the thesaurus and their semantic relations with other terms belonging to the domain and deriving from the corpus have been created. The current thesaurus in Italian language contains 245 candidate terms, already validated and mapped to the taxonomies contained in the main gold standards for Cybersecurity, i.e., NIST 7298 [20] and ISO 27000:2016 [21] together with domain experts collaborating on the project. The alignment with the terms contained in the standards for ICT security granted a coordination between the knowledge shared by an international Cybersecurity community of experts and the one represented in the structured thesaurus, which are preferred terms selected amongst those extracted by T2K as the most frequent. In order to carry out a matching configuration with the standards as predictable and stable as possible, the terms included in the standards, and selected with the support of domain experts as key elements representing the domain, have been translated using the Interactive Terminology for Europe (IATE) term banks [44]. This is considered an important step given the instructive purposes of the application that the thesaurus would have had in the web portal of the Cybersecurity Observatory. The first main entries in the Italian thesaurus for Cybersecurity are four categories finely selected from the glossary including the frequency of terms and from the mapping with the standards alongside the approval by the domain experts. These macro categories are:

- Cybersecurity;
- Cyberdefence;
- Cyberbullism;
- Cybercriminality.

Almost each of the candidate terms included in the thesaurus network, generated by the semantic relations among the terms, are accompanied by their definitions, i.e., *Scope Note (SN)*, which helps in understanding the terms in their specific contexts giving their definition taken from the source documents [45].

For a better understanding of the terms in the Italian Thesaurus for Cybersecurity, Table I gives a metrics of the numbers of terms, as well as of the semantic relations:

TABLE I. FEATURES OF THE ITALIAN THESAURUS FOR CYBESECURITY

	Terms	Semantic Relations	Non-preferred Terms	Scope Notes
Total	246	280	33	74

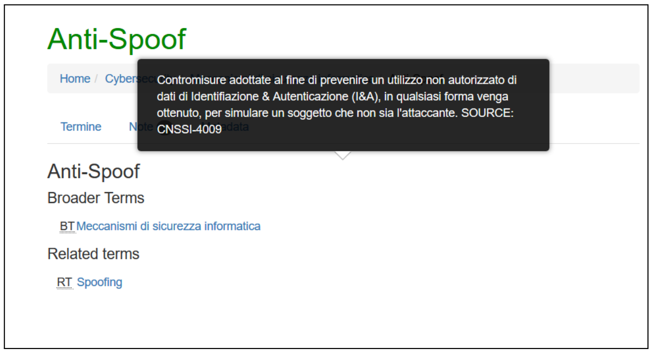
C. *Ontology enhancement*

Another activity of the OCS project has also been focused on the migration of the thesaurus into a more formal semantic resource, i.e., an ontology, to better organize and represent the information about Cybesecurity and addressed to users who want to get closer to this field of knowledge. The formalization of a thesaurus into an ontology is a task that in the last ten years has attracted much interest. In fact, in the literature different approaches are proposed for reusing thesaurus semantic content to build ontology meta-models and to populate knowledge bases in different domains, see for example [46][47][48].

The need for migrating the content included in the thesaurus to an ontology lies in the decision to better clarify the associative relationships between the terms of the thesaurus. In particular, the flat modality in which associative relationship between terms is represented in the thesaurus, i.e., via the RT relation, turned out to be not fully satisfactory in the seek of getting a complete terminological outline for Cybersecurity.

As shown in Figure 1 and Figure 2, there is a clear distinction between the two systems used to organize and represent the terminology belonging to the Cybersecurity domain. The example taken into account to represent the differences is referred to the semantic relationship linked to the idea of opposition, i.e., *Spoof* and *Antispoof*: in the thesaurus, even though a definition is present (within the black square), which corresponds to the *SN*, proper to thesauri, giving many details on the context from which terms come from, the ”opposition” is not so well represented because it is only shown through the associative relation (*RT*) between these aforementioned terms without giving other explications on the way the two terms are related as the OWL language does.
the other hand, in the ontology, these two concepts are connected through the *ObjectProperty* ”HasAsContrary” that helps in considering the *Domain* and the *Range* as linked by a precise relationship. 38304480

Another representative case is depicted by Figure 3 and Figure 4 that show how a thesaurus sometimes gives a weak visualization of some attributes associated to a concept.


Anti-Spoof

Home / Cyber

Contromisure adottate al fine di prevenire un utilizzo non autorizzato di dati di Identifiazione & Autenticazione (I&A), in qualsiasi forma venga ottenuto, per simulare un soggetto che non sia l'attaccante. SOURCE: CNSSI-4009

Termine   No

Anti-Spoof

Broader Terms

BT Meccanismi di sicurezza informatica

Related terms

RT Spoofing

Fig. 1. Thesaurus representation of the semantic relationship that describes opposition

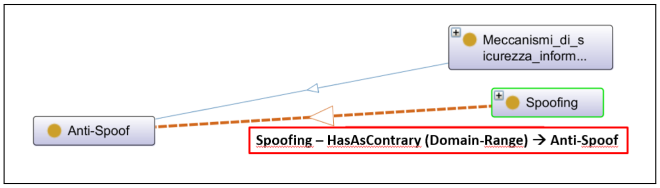

Meccanismi_di_s icurezza_inform...
Spoofing
Anti-Spoof
Spoofing – HasAsContrary (Domain-Range) → Anti-Spoof

Fig. 2. Protégé representation of the semantic relationship that describes opposition

In the following case, the relation that had to be provided was related to several attributes that security properties proper to informative systems own. For this specific purpose, the ontology resource gives more advantages in the visualization of the informative structure allowing a higher accurate organization and representation of the attributes related to the concepts. In detail, the main difference that makes ontologies a good semantic means to represent the conceptual model connected to certain semantic classes is related to the fact that, in this case, the security properties, i.e., integrity, authenticity, confidentiality, availability, reliability, non-repudiation, and privacy, are represented as *Data Properties* and are conceived as attributes. In the thesaurus, as shown in Figure 4, they are related to the *BT* ”Data” and are represented as its specific terms, i.e., the *NT* [11].

To give an idea of the content of the ontology derived from the Italian Cybersecuirty Thesaurus, Table II above shows some metrics and highlights the changes in the number of the relationships and concepts and the number of axioms with respect to the results shown for the thesaurus in Table I.

V. Discussion

Although thesauri and ontologies belong to the same family of knowledge organization systems and some of their functionalities are the same (e.g., their use for improving information retrieval, indexing, and knowledge organization) they are built for different purposes. In fact, it has been demonstrated in this contribution that ontologies allow for a more formal representation of knowledge for a given domain, by providing explicit relationships between concepts, disjunctions, applying data properties for each concept or instance and by providing

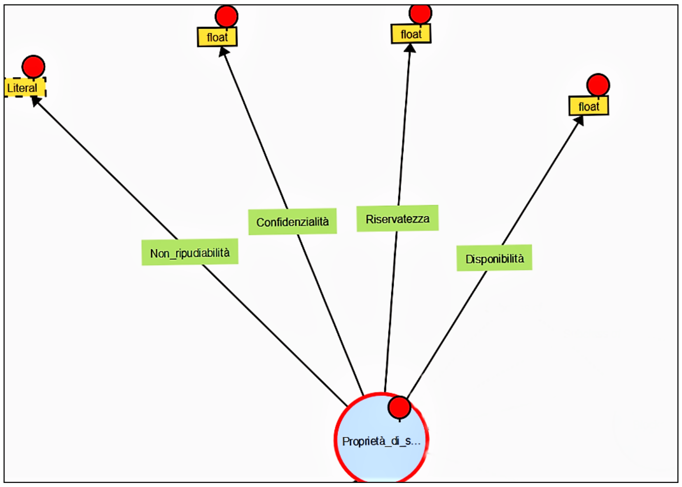
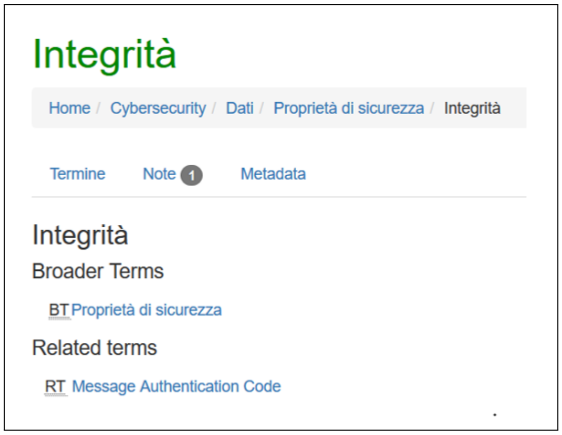


Fig. 3. WebVOWL representation of Security properties as *Data Properties*



Integrità

Home / Cybersecurity / Dati / Proprietà di sicurezza / Integrità

Termine   Note 1   Metadata

Integrità

Broader Terms

BT Proprietà di sicurezza

Related terms

RT Message Authentication Code

Fig. 4. Thesaurus representation of Security properties as hierarchical relations

restrictions that avoid ambiguity in the representation of the meaning and the context of use of a concept and their terms in the domain of reference. Nevertheless, the two semantic resources can be used together or, as widely demonstrated both in this paper and in the literature, one can be reused to build or populate the other, thus they complement each other, improving the end user's search experience.

The natural structural rigidity of thesauri, given by the use of *a priori* defined semantic relationships (hierarchical, associative and equivalence), seems to be a point against these type of controlled vocabularies; by contrast, such weakness seems to be overcome by the flexibility, scalability and reusability of ontologies that, as stressed by the semantic staircase of Blumauer and Pellegrini [49], compared to other KOSs, bring to a highest level of semantic richness thanks to an internal formal description of concepts. This latter combines a system of relations and properties of the concepts themselves.

Despite this, one of the strengths of the thesaurus compared to the ontology, when used in a specialized domain, is its greater capacity to eliminate ambiguity between the terms through the use of synonymy control [1] and the choice of preferred terms, compared to non-preferred terms for representing the concepts. This guarantees a standardization of technical terms in specialized domains, which can help in the process of unifying, and thus sharing, a specific field of knowledge's terminology.

TABLE II. CYBERSECURITY ONTOLOGY METRICS

Metric	**Total**
Axiom	640
Logical axiom count	316
Declaration axioms count	233
Class count	157
Object property count	37
Data property count	7
Individual count	31
Annotation Property count	5
CLASS AXIOMS	
SubClassOf	58
EquivalentClasses	0
DisjointClasses	24
OBJECT PROPERTY AXIOMS	
SubObjectPropertyOf	7
InverseObjectProperties	1
FunctionalObjectProperty	1
TransitiveObjectProperty	0
SymmetricObjectProperty	1
AsymmetricObjectProperty	0
ObjectPropertyDomain	40
ObjectPropertyRange	39
DATA PROPERTY AXIOMS	
SubDataPropertyOf	1
DataPropertyDomain	8
DataPropertyRange	5
INDIVIDUAL AND ANNOTATION AXIOMS	
ClassAssertion	31
AnnotationAssertion	89

VI. Conclusion

This paper aimed at presenting two different types of KOSs, i.e., thesauri and ontologies, exploiting their use and feasibility to organize and manage the specialized terminology proper to the Cybersecurity domain. Beginning from a general overview of Knowledge organization and representation systems, the analysis focused on the way the thesaurus, in particular, has proved to be a reliable system to structure the information derived from heterogenous sources belonging to the Cybersecurity domain, which is full of technical terms. Concurrently, attention has also been given to the comparison between the modality of representing in the thesaurus some of the relationships existing among terms, that represent the relevant concepts of the domain, with the ones proper to ontologies and the OWL language. The perspective has been oriented to provide a demonstrative outline of ontology peculiarities and advantages when using an existing thesaurus, like the one created in the Italian OCS project framework, as a basis for building the meta-model and populating the knowledge base. Being the presented activity a work in progress, in the near future both the thesaurus and the knowledge base in OWL will be extended with more terms, relationships and restrictions where needed, and a new evaluation will be executed. Among the future works there will be a translation in another language (firstly English) to allow, within the OCS project team, the recognition of threats even from non-Italian sources and improve the thesaurus/ontology usability and sharing also at an international level. Moreover, the remainder of this work targets also at taking into account the insertion of several other types of documents to be part of the source corpus. In particular, following the perspective of getting updated on the changes related to the Cybersecurity domain, documents shall be taken from the social media world, adjusting all the analysis related to the processing of information to the treatment of texts written in a specialized form.

References

[1] M. Zeng, "Knowledge organization systems (kos)," Knowledge Organization, vol. 35, pp. 160–182, 01 2008.

[2] Cybesecurity osservatorio. https://www.cybersecurityosservatorio.it\. Accessed: 2019-08-08.

[3] W3C Web Ontology Language (OWL). https://www.w3.org/OWL/. Accessed: 2019-08-08.

[4] R. Davis, H. Shrobe, and P. Szolovits, "What is a knowledge representation?" AI Magazine, vol. 14, p. 17, 03 2002.

[5] A. Miles and S. Bechhofer, SKOS Simple Knowledge Organization System Reference, ser. W3C Recommendation. United States: World Wide Web Consortium, 8 2009.

[6] D. Soergel, "The art and architecture thesaurus (aat): A critical appraisal," Visual Resources, vol. 10, pp. 369–400, 01 1995.

[7] M. Taverniti, "Tra terminologia e documentazione: estrazione automatica di voci indice da corpora documentali della pubblica amministrazione," Ainformazioni, vol. 1-2/2018, pp. 227–238, 2008.

[8] V. Broughton, Essential Thesaurus Construction. Facet, 2006.

[9] E. Morin and C. Jacquemin, "Projecting corpus-based semantic links on a thesaurus," in Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 1999, p. 389396.

[10] J. Morris and G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," Comput. Linguist., vol. 17, no. 1, pp. 21–48, Mar. 1991.

[11] ISO , Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies.

[12] ISO, Information and documentation — Thesauri and interoperability with other vocabularies Part 1: Thesauri for information retrieval, International Organization for Standardization, August 2011.

[13] A. A. Shiri and C. Revie, "Thesauri on the web: current developments and trends," Online Information Review, vol. 24, no. 4, pp. 273–280, 2000.

[14] D. Soergel, "The art and architecture thesaurus (aat): A critical appraisal," Visual Resources, vol. 10, pp. 369–400, 01 1995.

[15] M. van Assem, V. Malaisé, A. Miles, and G. Schreiber, "A method to convert thesauri to skos," 06 2006, pp. 95–109.

[16] N. Guarino, D. Oberle, and S. Staab, "What is an ontology?" Springer, Berlin, Heidelberg, 05 2009, pp. 1–17.

[17] D. W. Embley, S. W. Liddle, D. W. Lonsdale, and Y. A. Tijerino, "Multilingual ontologies for cross-language information extraction and semantic search," in *ER*, 2011.

[18] N. F. Noy and D. L. Mcguinness, "Ontology development 101: A guide to creating your first ontology," Tech. Rep., 2001.

[19] D. Bourigault and N. Aussenac-Gilles, "Construction d'ontologies á partir de textes," pp. 11–14, 01 2003.

[20] R. Kisserl, Glossary of Key Information Security Terms, National Institute of Standards and Technology, May 2013, NISTIR 7298 Revision 2.

[21] ISO/IEC 27000, Information technology — Security techniques — Information security management systems — Overview and vocabulary, International Standard, February 2016.

[22] R. Baldoni, R. De Nicola, and P. Prinetto, Il Futuro della Cybersecurity in Italia: Ambiti Progettuali Strategici Progetti e Azioni per difendere al meglio il Paese dagli attacchi informatici. Laboratorio Nazionale di Cybersecurity (CINI) - Consorzio Interuniversitario Nazionale per lInformatica, 2018.

[23] Presidenza del Consiglio dei Ministri - Sistema di informazione per la sicurezza della Repubblica, Il linguaggio degli organismi informativi, Glossario intelligence. https://www.sicurezzanazionale.gov.it/sisr.nsf/quaderni-di-intelligence/glossario-intelligence.html\. Accessed: 2019-08-08.

[24] B. Barnett and A. Crapo, "A semantic model for cyber security," 2011.

[25] A. Aviad, K. Wcel, and W. Abramowicz, "The semantic approach to cyber security. towards ontology based body of knowledge," vol. 2015, 01 2015, pp. 328–336.

[26] L. Obrst, P. Chase, and R. Markeloff, "Developing an ontology of the cyber security domain," in STIDS, 2012.

[27] A. Oltramari, L. Cranor, R. Walls, and P. McDaniel, "Building an ontology of cyber security," CEUR Workshop Proceedings, vol. 1304, pp. 54–61, 01 2014.

[28] M. C. Parmelee, "Toward an ontology architecture for cyber-security standards."

[29] A. Singhal and D. Wijesekera, "Ontologies for modeling enterprise level security metrics," ACM International Conference Proceeding Series, 01 2010.

[30] A. Condamines, "Sémantique et corpus spécialisés : Constitution de Bases de Connaissances Terminologiques," Habilitation à diriger des recherches, Université Toulouse Le Mirail, Jun. 2003. [Online]. Available: https://halshs.archives-ouvertes.fr/tel-01321042

[31] G. Leech, The state of the art in corpus linguistics, K. Aijmer and B. Altenberg, Eds. London: Longman, 1991.

[32] G. Bernier-Colborne, "Defining a gold standard for the evaluation of term extractors," in Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 15–18.

[33] J. Pearson, Terms in Context. John Benjamins, Amsterdam, 1998.

[34] A. Rigouts Terryn, V. Hoste, and E. Lefever, "A gold standard for multilingual automatic term extraction from comparable corpora : term structure and translation equivalents," in Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), 2018, pp. 1803–1808. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2018/index.html

[35] A. Nazarenko, Zargayouna, O. H.; Hamon, and van Puymbrouck, "Evaluation des outils terminologiques : enjeux, difficultés et propositions," Traitement Automatique de la Langue (TAL), vol. 50, no. 1, pp. 257–281, 2009.

[36] E. L. et al., "Reference Lists for the Evaluation of Term Extraction Tools," in Terminology and Knowledge Engineering Conference (TKE), Madrid, Spain, 2012.

[37] Cybesecurity Osservatorio - Thesaurus. https://www.cybersecurityosservatorio.it/it/Services/thesaurus.jspt\. Accessed: 2019-08-08.

[38] J. E. Rowley, J. E. Rowley, and R. J. Hartley, Organizing knowledge: an introduction to managing access to information / Jennifer Rowley and Richard Hartley , 4th ed. Ashgate Aldershot, England ; Burlington, VT, 2008.

[39] M. Hudon and D. Ménillet, Guide pratique pour l'élaboration d'un thésaurus documentaire /. [Montréal, QC] :: éditions ASTED,, 2009., publ. antérieurement sous le titre: Le thésaurus. 1995.

[40] C. Barrière, "Semi-automatic corpus construction from informative texts," in Text-Based Studies in honour of Ingrid Meyer, ser. Lexicography, Terminology and Translation, L. Bowkes, Ed. University of Ottawa Press, January 2006, ch. 5.

[41] G. Zagrebelsky, Il sistema costituzionale delle fonti del diritto, EGES, Ed. Turin: UTET, 1984.

[42] A. Condamines, "L'interprètation en sémantique de corpus : le cas de la construction de terminologies," Revue française de linguistique appliquée, vol. Vol. XII, no. 2007/1, pp. 39–52, 2007.

[43] F. Dell'Orletta, G. Venturi, A. Cimino, and S. Montemagni, "T2K: a system for automatically extracting and organizing knowledge from texts," in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[44] IATE European Union Terminology. https://iate.europa.eu/home\. Accessed: 2019-08-08.

[45] C. Lanza, “Italian domain-specific thesaurus as a means of semantic control for cybersecurity terminology,” in The Twelfth International Conference on Advances in Semantic Processing (SEMAPRO 2018), U. o. A. S. G. P. L. U. o. H. A. F. Michael Spranger, Hochschule Mittweida, Ed., Athens, Greece, November 2018.

[46] E. Cardillo, A. Folino, R. Trunfio, and R. Guarasci, “Towards the reuse of standardized thesauri into ontologies,” in Proceedings of the 5th International Conference on Ontology and Semantic Web Patterns - Volume 1302, ser. WOP’14, 2014, pp. 26–37.

[47] M. Nowroozi, M. Mirzabeigi, and H. Sotudeh, “The comparison of thesaurus and ontology: Case of asist web-based thesaurus and designed ontology,” Library Hi Tech, vol. 36, 01 2018.

[48] J. L. D. Kless, L. Jansen and J. Wiebensohn, “A method for re-engineering a thesaurus into an ontology,” in Proceedings of International Conference on Formal Ontology in Information Systems (FOIS 2012), 2012, pp. 133–146.

[49] A. Blumauer and T. Pellegrini, Semantic Web und semantische Technologien: Zentrale Begriffe und Unterscheidungen. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 9–25.