Knowledge Extraction for the Web of Things (KE4WoT)

[WWW 2018 Challenge Summary]

Amelie Gyrard*△, Manas Gaur*,
Swati Padhee*, Amit Sheth*
*Kno.e.sis, Wright State University, USA
{amelie,swati,manas,amit}@knoesis.org

Mihaela Juganaru-Mathieu△
△ MINES Saint-Etienne, Institut Henri Fayol, France
mathieu@emse.fr

ACM Reference Format:
Amelie Gyrard*△, Manas Gaur*, Swati Padhee*, Amit Sheth* and Mihaela Juganaru-Mathieu△. 2018. Knowledge Extraction for the Web of Things (KE4WoT): [WWW 2018 Challenge Summary]. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3184558.3192305

1 KE4WOT CHALLENGE OVERVIEW

The Web of Things (WoT)[1] is an extension of the Internet of Things (IoT) to ease the access to data generated by things/devices using the benefits of Web technologies [6]. Data is exploited by WoT applications to monitor healthcare or even control home automation devices. The purpose of the **Knowledge Extraction for the Web of Things (KE4WoT) challenge**[2] is to automatically extract the relevant knowledge from already designed smart WoT applications in various applicative domains. Those applications design and release Knowledge Bases (e.g., datasets and/or models) on the web.

Standardizations have a growing interest in designing models to represent devices and generated data. **W3C Semantic Sensor Networks (SSN)** [3] is the first initiative to address interoperability issues to describe sensor networks through an ontology. Sensors and devices are required to build WoT applications. The last release of the SSN ontology[3] became a W3C recommendation in October 2017. It is a joint contribution with the **Open Geospatial Consortium (OGC)** standard, extending and improving the SSN ontology published in 2011. **W3C Web of Things (WoT)** Interest Group is designing a vocabulary to describe interactions between objects through the Web, a potential implementation is the WoT ontology[4]. A "Remote health monitoring system"[5] scenario has been designed among several use cases. **OneM2M**, an international standard for Machine-to-Machine (M2M) designed the OneM2M ontology[6]. OneM2M extends the European **ETSI M2M** standard. At the current date of writing, neither WoT ontology nor OneM2M ontology are aligned with W3C ontologies.

The "MyOntoSens" ontology, based on SSN V1 is being standardized as a Technical Specification (TS) within the SmartBAN (Body Area Networks) Technical Committee of the ETSI standardization body [8]. This ontology is relevant to build health applications based on smart devices. **Smart Appliances REFerence (SAREF)**[7] is a European standard supported by ETSI M2M and SmartM2M. It mainly covers the smart building applicative domain. The SAREF ontology has been designed re-using SSN and oneM2M [7]. **Schema.org** is a well-known schema catalog to structure data on Web pages to describe the location, person, etc. [4]. The **IoT Schema.org** extension[8] is planned; nothing concrete has been developed yet, but discussions are ongoing. The **Haystack**[9] project aims to standardize semantic data models and web services. The Haystack Tagging Ontology[10] which employs SSN V1 ontology has been developed [2].

The purpose of this challenge is to automatically extract the redundant knowledge already designed within existing standardizations, WoT applications, and different communities. Most of the existing knowledge extraction techniques are frequently applied to text found within documents or social networks. The main novelty of this challenge is to apply web-based knowledge extraction techniques on models.

2 CHALLENGE DESCRIPTION

The challenge is open to a large audience and complementary web communities: (1) Knowledge extraction experts to detect common patterns within ontologies - Task 1.1. (2) WoT/IoT and healthcare communities willing to discover and study already designed models - Task 1.1. (3) Ontology matching experts to align existing ontologies - Task 1.2. (4) Natural Language Processing researchers - Tasks 2.1 and 2.2. (5) Semantic Web researchers. (6) Any developers and/or data scientists willing to implement innovative methods.

2.1 Task 1: Exploiting the WoT Knowledge Base

The LOV4IoT ontology catalog[11] [5] is referencing almost 400 WoT research projects in various domains. For both Task 1.1 and Task 1.2, we provide a set of 6 web services and dump files to easily query ontologies from the LOV4IoT ontology catalog (e.g., healthcare, IoT, agriculture, sensor networks, WoT and smart city ontologies).

[1] https://webofthings.org/
[2] http://wiki.knoesis.org/index.php/KE4WoTChallengeWWW2018
[3] https://www.w3.org/TR/vocab-ssn/
[4] http://iot.linkeddata.es/def/wot/index-en.html
[5] http://w3c.github.io/wot/wot-ucr.html#domain-healthcare_and_medical
[6] http://www.onem2m.org/technical/onem2m-ontologies

[7] https://ec.europa.eu/digital-single-market/en/blog/new-standard-smart-appliances-smart-home
[8] http://iot.schema.org/
[9] http://project-haystack.org/
[10] https://raw.githubusercontent.com/vcharpenay/hto/master/hto.xml
[11] http://lov4iot.appspot.com/

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3192305

Task 1.1: Extracting the most popular terms and properties. This task aims to extract the most popular concepts and properties from a set of ontologies[12]. For instance, if a set of health ontologies are queried, the most popular expected concepts are `Patient` and `Blood Pressure`. Regarding IoT ontologies, the most popular expected concepts are `Device` and `Temperature`. Approaches such as [10] can be taken into consideration.
Impact: Such algorithms would demonstrate the most relevant concepts and properties in a set of domain. The algorithms might be generic enough to be applied to any domains. Such algorithms would be relevant to assist in creating iot.schema.org for instance.

Task 1.2: Ontology Matching algorithms and software. This task is mainly designed for ontology matching experts to run their tools on the LOV4IoT benchmark to align WoT-related ontologies (instead of the popular OAEI benchmark). An evaluation has been designed[13] to align IoT ontologies with a specific ontology matching tool.
Impact: Ontology matching experts would observe that ontologies referenced are not structured in the same way. For instance, no labels or comments are provided within the ontology which is a huge problem since most of the methods are using this hypothesis. This would lead to the design of new ontology matching tools relevant for WoT.

2.2 Task 2: Named Entity Recognition and Question Answering in Healthcare Unstructured Text

Named entity recognition (NER) is considered as an essential natural language processing task [1] for entity extraction. The state-of-the-art NER systems lacks focus on structured healthcare text due to the complexity to capture context in medical domain. Good quality NER is crucial for accurate Question Answering (Q/A) systems and hence we have designed the following two sequential tasks.

Task 2.1: Extracting named entities using ontologies. In this challenge, we encourage the idea of NER from unstructured healthcare text obtained from tweets. We categorize the named entities as follows: (1) *Disease Entity*: It is the name of the disease that is stated explicitly in the text. (2) *Severity Entity* (a severe form of disease entity): It is a disease entity of etiological origin from a relatively mild disease entity. (3) *Trigger Entity*: It refers to the cause (entity/substance/environmental condition) of the disease. For example, `pollen`, `weather`, `cough` can cause `asthma`. (4) *Location Entity*: It refers to the location affected in human anatomy. For instance, `bones`, `muscles`, `nose`, `lungs`, etc. (5) *Procedure/Treatment/Device*: These are entities that define a procedure, treatment or device used by the patient or clinician to treat the disease entity stated in the text. For example, an `inhaler` is a device to treat asthma. (6) *Control*: It is a dichotomous concept whose value is "yes" when the tweet talks about disease control, reduction in severity or reduced frequency of asthmatic attacks. This category is created for supporting the question answering task.
Impact: This task motivates the participants to develop a quality named entity extraction technique for unstructured healthcare text. It can be leveraged by a question answering system for healthcare domain.

Task 2.2: Question Answering (Q/A) System. In this task, the participants will be required to develop a Q/A system for healthcare tweet questions using their developed NER module developed in Task 2.1. They are encouraged to utilize existing domain knowledge sources (e.g., SNOMED, DBpedia, etc.) to enhance the efficiency of their model. We provide 25 test questions to evaluate their performance[14].
Impact: Q/A systems have been an apogee of research in linked open data [9]. This task encourages research in Q/A for a less explored, complex and critical healthcare domain.
An illustrative example of this task: *Tweet*: Patients with severely uncontrolled asthma derive the most benefit from dupilumab. *Question*: Does dupilumab control asthma? *Answer*: Yes

3 LESSONS LEARNT

The KE4WoT challenge has attracted significant interest from the research community[15]. Interested participants gave feedback regarding time constraint to submit papers. We plan to pursue this research challenge idea and extend it.

4 ACKNOWLEDGMENTS

This work is partially funded by a bilateral research convention with ENGIE Research & Development, the ANR 14-CE24-0029 OpenSensingCity project, Hazards SEES NSF Award EAR 1520870, and KHealth NIH 1 R01 HD087132-01. Thanks to Dr. Antoine Zimmermann for his feedback.

REFERENCES

[1] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*, 2017.
[2] V. Charpenay, S. Käbisch, and H. Kosch. Introducing Thing Descriptions and Interactions: An Ontology for the Web of Things. In *SR+ SWIT@ ISWC*, 2016.
[3] M. Compton, P. Barnaghi, L. Bermudez, R. Garcia-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, et al. The ssn ontology of the w3c semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2012. http://www.w3.org/2005/Incubator/ssn/ssnx/ssn.
[4] R. V. Guha, D. Brickley, and S. Macbeth. Schema.org: Evolution of structured data on the web. *Communications of the ACM*, 2016.
[5] A. Gyrard, G. Atemezing, C. Bonnet, K. Boudaoud, and M.Serrano. Reusing and Unifying Background Knowledge for Internet of Things with LOV4IoT. In *International Conference on Future Internet of Things and Cloud (FiCloud)*. IEEE, 2016.
[6] A. Gyrard, P. Patel, A. Sheth, and M. Serrano. Building the web of knowledge with smart iot applications. *IEEE Intelligent Systems Magazine*, 2016.
[7] J. e. a. Moreira. Towards IoT platforms' integration: Semantic Translations between W3C SSN and ETSI SAREF. In *SIS-IoT: Semantic Interoperability and Standardization in the IoT Workshop at Semantics Conference*, 2017.
[8] L. Nachabe, M. Girod-Genet, and B. El Hassan. Unified data model for wireless sensor network. *IEEE Sensors Journal*, 2015.
[9] C. Unger, C. Forascu, V. Lopez, A.-C. N. Ngomo, E. Cabrio, P. Cimiano, and S. Walter. Question answering over linked data (qald-4). In *Working Notes for CLEF 2014 Conference*, 2014.
[10] G. Wu, J. Li, L. Feng, and K. Wang. Identifying potentially important concepts and relations in an ontology. In *International Semantic Web Conference*. Springer, 2008.

[12] http://lov4iot.appspot.com/?p=OntologyExtractionKE4WoTChallengeWWW2018
[13] http://lov4iot.appspot.com/?p=OntologyAlignmentKE4WoTChallengeWWW2018
[14] https://github.com/gyrard/KE4WoT_Challenge_WWW2018/tree/master/Challenge_Dataset
[15] https://goo.gl/WKciMV