Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200113

A Semantic Similarity Evaluation for Healthcare Ontologies Matching to HL7 FHIR Resources

Athanasios KIOURTIS[a,1], Argyro MAVROGIORGOU[a] and Dimosthenis KYRIAZIS[a]
[a] *Department of Digital Systems, University of Piraeus, Greece*
{kiourtis,margy,dimos}@unipi.gr

Abstract. Healthcare 4.0 demands healthcare data to be shaped into a common standardized and interoperable format for achieving more efficient data exchange. Most of the techniques addressing this domain are dealing only with specific cases of data transformation through the translation of healthcare data into ontologies, which usually result in clinical misinterpretations. Currently, ontology alignment techniques are used to match different ontologies based on specific string and semantic similarity metrics, where very little systematic analysis has been performed on which semantic similarity techniques behave better. For that reason, in this paper we are investigating on finding the most efficient semantic similarity technique, based on an existing approach that can transform any healthcare dataset into HL7 FHIR, through the translation of the latter into ontologies, and their matching based on syntactic and semantic similarities.

Keywords. Healthcare, Ontology alignment, Semantic similarity, HL7 FHIR

1. Introduction

In the last decade, there has been a transition from a data-poor to a data-rich world, with the aim of improving the quality of transport, governance, environment, communication and health. Much of this unprecedented increase in data generation can be attributed to the abundance of thousands of mobile devices, wearables, and sensors. Most of these devices are typically characterized by a high degree of heterogeneity, in terms of having different characteristics, capabilities, or network specifications, needing to be easily manageable in particular with regard to the management of the heterogeneous data they generate. Especially for the Internet of Medical Things (IoMT) devices it is undeniable that vast amounts of heterogeneous medical data are becoming available, thus completely reshaping the Healthcare 4.0. In order for these heterogeneous medical data to be exchanged with multiple stakeholders, and to be a key driver of providing personalized and efficient medical care to patients and citizens, interoperability is the only way.

Taking into consideration these challenges, in [11] a holistic approach has been presented for achieving interoperability through the transformation of healthcare data into its corresponding HL7 FHIR structure. Shortly, the provided mechanism was building the healthcare ontologies that were stored into a triplestore, in order to identify

[1] Corresponding Author, Athanasios Kiourtis, 80, M. Karaoli & A. Dimitriou St., 18534 Piraeus, Greece E-mail: kiourtis@unipi.gr.

and compare their syntactic and semantic similarity with the HL7 FHIR Resources. According to the aggregation of the syntactic and semantic similarity results, the alignment and translation to the HL7 FHIR was taking place. In this paper, we are going to fill the gap of the semantic similarity of the aforementioned mechanism, as there has been little systematic analysis on which semantic similarity techniques perform well when applied to ontology alignment.

The remainder of this paper is organized as follows. In Section 2, some semantic similarity techniques are illustrated, while Section 3 depicts the overall approach. Section 4 includes the evaluation of the derived results, presenting our concluding remarks.

2. Semantic Similarity Techniques

Several methods for determining semantic similarity between terms have been proposed in the literature. Word2Vec [2] is a two-level neural network that processes text. Its input is a body of text and its output is a set of vectors. The purpose and utility of Word2Vec is to group similar word vectors together in the vector space, detecting mathematical similarities. Word2Vec creates vectors that are distributed arithmetic word representations, such as the single word context, without human intervention. Given enough data, usage, and the current environment, Word2Vec can make very accurate guesses about the meaning of a word based on impressions of the current word in other situations in the past. WordNet [3] is a great English based dictionary. Nouns, verbs, adjectives, and adverbs are grouped into sets of synsets, each of which expresses a separate meaning. Synsets are interrelated through conceptual-semantic and lexical relationships. The resulting network of related words and concepts can be navigated with the browser. The structure of WordNet makes it a useful tool for computational linguistics and NLP. WordNet looks superficially like a treasure trove because it groups words according to their meanings. Cortical.io [4] proposes a new way of understanding the natural language that transcends the limitations of other artificial intelligence approaches. Semantic Folding [5] is inspired by the most recent findings on how the brain processes information. This theory introduces the Semantic Fingerprint, which explicitly codifies the concept, including all concepts and frameworks. The system understands the affinity of two elements by measuring the overlap of their fingerprints.

3. Proposed Approach

In the context of ontology alignment of healthcare data, several techniques are presented that lead to better decision making, and interoperable exchange and use of data, providing solutions that are adapted to deliver results under specific scenarios. To address this gap, in [1] a holistic approach is proposed that is capable of identifying common links between healthcare ontologies and the ontologies that represent the HL7 FHIR Resources. The preliminary steps of the developed approach were dealing with the automatic transformation of the healthcare dataset and the HL7 FHIR Resources into their ontological structure. As soon as the ontologies had been constructed, they were stored into a triplestore. Afterwards, two sub-mechanisms had occurred, providing the syntactic and the semantic similarity between two different ontologies, based on their syntactic representations and their semantic knowledge accordingly. The final stage of the mechanism implemented an additional sub-mechanism that aggregated and merged the

outcomes of the aforementioned sub-mechanisms, providing the overall alignment results, for the final transformation of the healthcare dataset into HL7 FHIR. Henceforth based on this developed mechanism, in our case, we assume that healthcare ontologies have already been constructed for both the ingested healthcare dataset and the different HL7 FHIR Resources. Consequently, the three described semantic similarity techniques are going to be implemented, for measuring the semantic similarity of the built ontologies.

3.1. Semantic Similarity Identifier

The objective of the Semantic Similarity Identifier is to provide the means for aligning and matching the different healthcare dataset and HL7 FHIR Resources ontologies, according to their semantic meaning (Fig. 1).

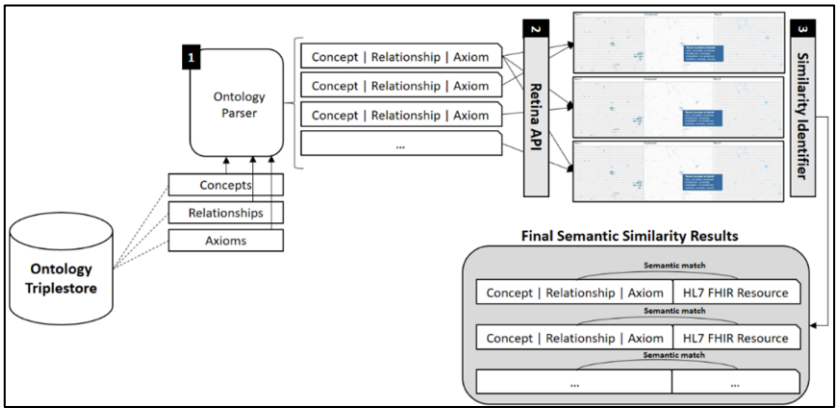

Figure 1. Semantic Similarity Identifier.

4. Evaluation

In our case, in order to evaluate the three different semantic similarity techniques, we are going to implement them under the same use case scenario where we will calculate their precision, recall, F-measure, and total errors [6], with respect to a UMLS-based reference alignment [7]. After this calculation, a comparison between the results will take place in order to conclude to the semantic similarity technique that is more suitable for aligning ontologies and calculating their semantic similarity in the healthcare domain.

4.1. Application of Semantic Similarity Techniques

The use case exploits a healthcare dataset structured in CSV format, covering all the previous steps, consisting of 5000 different instances of the personal information of certain citizens about their: (i) personal identifier (*subject*), (ii) gender (*gender*), (iii) date of birth (*dateOfBirth*), and (iv) cause of death (*causeOfDeath*). After performing the steps introduced into the Semantic Similarity Identifier, the following tables are being created (Table 1, Table 2, and Table 3), depicting the HL7 FHIR Resources with higher semantic similarity degrees, compared with the concept names of the use case dataset, through the three different semantic similarity techniques.

Table 1. Semantic Similarity Identifier Top Results for Cortical.io

Use Case Dataset Attribute	HL7 FHIR Resources	Cortical.io
subject	Patient.identifier	0.947 (~95%)
gender	Patient.gender	0.863 (~86%)
dateOfBirth	Patient.birthDate	0.971 (~97%)
causeOfDeath	Patient.deceased	0.458 (~46%)

Table 2. Semantic Similarity Identifier Top Results for Word2Vec

Use Case Dataset Attribute	HL7 FHIR Resources	Word2Vec
subject	Patient.contact.address	0.357 (~36%)
gender	Patient.gender	0.663 (~66%)
dateOfBirth	Patient.birthDate	0.891 (~89%)
causeOfDeath	-	0 (0%)

Table 3. Semantic Similarity Identifier Top Results for Wordnet

Use Case Dataset Attribute	HL7 FHIR Resources	Wordnet
subject	Patient.gender	0.772 (~77%)
gender	Patient.gender	0.861 (~86%)
dateOfBirth	Patient.multipleBirth	0.561 (~56%)
causeOfDeath	Patient. deceased	0.619 (~62%)

4.2. Conclusions & Discussion of Results

In order to evaluate these results with regards to the metrics of precision, recall, F-measure, and errors, the specific use case dataset was translated manually in HL7 FHIR, so as to compare the results of the developed mechanism, with the actual outcomes. In Fig. 2, a visualization of the aforementioned results is depicted through a bar chart, showing how the three different semantic similarity techniques behave in the total semantic similarity identification, with regards to the manually provided results which did not contain any errors (0% of errors), while their precision, recall and F-measure had the value of 100%.

It is clear that there was not any semantic similarity technique that provided 100% accurate results. As it can be seen through Fig. 2, among the different techniques, the Cortical.io behave better since it provided less errors (13%) compared with the other semantic similarity techniques (31%, and 23% accordingly). Moreover, the F-measure (a result of both the precision and recall) of the Cortical.io had the greatest value (84%) in comparison with the other semantic similarity techniques (77%, and 65% accordingly). Consequently, this lead us to the conclusion that for the Semantic Similarity Identifier the most convenient semantic similarity technique to use is the Cortical.io, since it provides more efficient and reliable results.

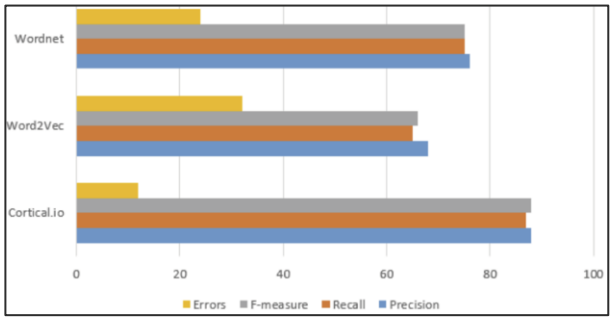

Figure 2. Visualization of Comparison Results.

In the field of healthcare interoperability, semantic similarity has to be identified, for which there has been little systematic analysis on which techniques perform well when applied to ontology alignment. For addressing this gap, in this paper a previous research was considered, about a healthcare ontology matching mechanism that has the ability of transforming healthcare data to HL7 FHIR format. Therefore, based upon this research, in order to identify the required semantic similarity, three of the most well-known semantic similarity techniques were used under the same scenario, in order to result to the most efficient technique of semantic similarity in the cases of healthcare ontologies alignment, and consequently to the most efficient HL7 FHIR transformation.

To this context, we will continue working on the evaluation of the Semantic Similarity Identifier mechanism with additional semantic similarity techniques, to identify the technique that is most suitable for the case of healthcare ontologies alignment. In addition, we will continue on this evaluation with datasets of different sizes, standards and formats, respecting privacy issues, based on the mechanism developed in [8].

Acknowledgment

The research leading to this result has received funding from the European Union's Horizon 2020 research and innovation programmes under grant agreement No 727560 (CrowdHEALTH project) and grant agreement No 826106 (InteropEHRate project).

References

[1] A. Kiourtis, A. Mavrogiorgou, D. Kyriazis, S. Nifakos, Aggregating the Syntactic and Semantic Similarity of Healthcare Data towards their Transformation to HL7 FHIR through Ontology Matching, Int. Journal of Medical Informatics, 132 (2019).

[2] Y. Goldberg, O. Levy, word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014).

[3] Wordnet, https://wordnet.princeton.edu/

[4] Coritcal.io, https://www.cortical.io/

[5] F. D. S. Webber, Semantic folding theory and its application in semantic fingerprinting. arXiv preprint arXiv:1511.08855 (2015).

[6] A. Kiourtis, A. Mavrogiorgou, A. Menychtas, I. Maglogiannis, D. Kyriazis, Structurally Mapping Healthcare Data to HL7 FHIR through Ontology Alignment. Journal of medical systems, 43(3) (2019).

[7] Ontology Alignment Evaluation Initiative, http://oaei.ontologymatching.org/2018/

[8] A. Kiourtis, A. Mavrogiorgou, D. Kyriazis, Towards a Secure Semantic Knowledge of Healthcare Data Through Structural Ontological Transformations. Joint Conference on Knowledge-Based Software Engineering, (2018) 178-188.