Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210205

Patient and Graph Embeddings for Predictive Diagnosis of Drug Iatrogenesis

Lina F. SOUALMIA[a,b,1] Vincent LAFON[c] and Stéfan J. DARMONI[b,d]
[a]*Normandie Univ, UNIROUEN, LITIS EA 4108, F-76000 Rouen, France*
[b]*LIMICS U1142, Sorbonne Université, F-75015, Paris, France*
[c]*INSILIANCE, F-75014, Paris, France*
[d]*CHU Rouen, Department of Biomedical Informatics, F-76100 Rouen, France*

Abstract. In the context of the IA.TROMED project we intend to develop and evaluate original algorithmic methods that will rely on semantic enrichment of embeddings by combining new deep learning algorithms, such as models founded on transformers, and symbolic artificial intelligence. The documents' embeddings, the graphs' embeddings of biomedical concepts, and patients' embeddings, all of them semantically enriched with aligned formal ontologies and semantic networks, will constitute a layer that will play the role of a queryable and searchable knowledge base that will supply the IA.TROMED's clinical, predictive, and iatrogenic diagnosis support module.

Keywords. Knowledge Bases; Biomedical Ontologies; Neural networks.

1. Introduction

Since the few past years, the medical language processing community has adapted embeddings trained on general language to the clinical one [1]. These new models have to be explored and adapted to real life biomedical data. In France, several billion concepts have to be extracted from clinical data warehouses, that include electronic health records, and need to be linked to the French National System of Health Data (SNDS). Moreover, integrating complementary and heterogenous data, such as biomedical scientific literature, may improve the accuracy of the algorithms. Based on these hypotheses, we aim at developing original technologies in the context of the IA.TROMED project. It will combine machine learning, deep neural networks trained on heterogenous data sources, such as drug databases, scientific and grey literature, and real-life data, but also symbolic and semantic reasoning process on biomedical ontologies.

2. Methods

The IA.TROMED project (Fig.1) relies on several hypotheses: (i) new approaches (embeddings and transformer-based) are needed to be explored and adapted to real life biomedical data: several billion concepts have to be linked from EDSaN [2] to the SNDS; (ii) the new models need to be launched and trained on several complementary and heterogeneous data (scientific literature in English and French [3]); (iii) Enriching

[1] Corresp. Author : Avenue de l'Université, 76801 SER, France; E-mail: lina.soualmia@chu-rouen.fr.

semantically the deep learning models will allow to benefit from advanced rule-based processes and semantic web technologies; (iv) Mashing-up all the extracted knowledge from those heterogeneous resources will facilitate a better collaboration between clinicians and pharmacists, and researchers that will (v) adapt and optimize the algorithms, including technological trial.

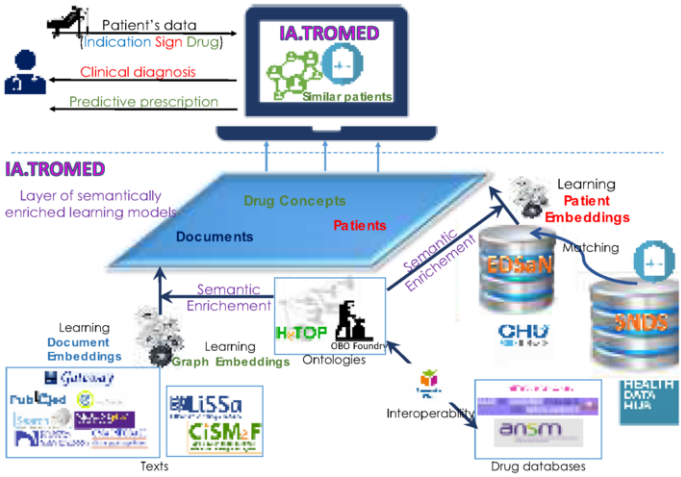

Figure 1. General architecture of IA.TROMED. The clinician updates Indication-Sign-Drug. Automatic analysis is performed on heterogeneous data sources, based on a layer of learnt/semantically enriched models (documents, graphs, patients embeddings). The mashup performs a diagnosis and adapt the prescription.

3. Preliminary Results

We have already developed a vectorial space trained on EDSaN and generated a hybrid semantic annotator [4,5], and document embeddings to create inter-scientific paper similarities in PubMed [6]. Moreover, our MeSH-gram neural network model extends word embedding vectors with MeSH concepts and improves semantic similarity and relatedness [7]. Researchers and clinicians will be associated all along the process – development-evaluation-exploitation-optimization – to get an efficient, effective, and easy-to-use application.

References

[1] Hahn U, Oleynik M. Medical information extraction in the age of deep learning. Yearb Med Inform 2020;29(1):208-20.
[2] Lelong R, et al. Building a semantic health data warehouse in the context of clinical trials: development and usability study. JMIR Med Inform. 2019;7(4):e13917.
[3] Cabot C, et al. Evaluation of the terminology coverage in the French corpus LiSSa. Stud Health Technol Inform. 2017;235:126-30.
[4] Cabot C, et al. Cimind: a phonetic-based tool for multilingual named entity recognition in biomedical texts. J Biomed Inform. 2019;94:103176.
[5] Dynomant E, et al. Word embedding for the French natural language in health care: comparative study. JMIR Med Inform. 2019;7(3):e12310.
[6] Dynomant E, et al. Doc2Vec on the PubMed corpus: study of a new approach to generate related articles.
[7] Abdeddaïm S, et al. The MeSH-gram neural network model: extending word embedding vectors with MeSH concepts for semantic similarity. Stud Health Technol Inform. 2019;264:5-9.