Towards an Italian Healthcare Knowledge Graph

Marco Postiglione
University of Naples Federico II
Department of Electrical and Information Technology (DIETI)
Naples, Italy

marco.postiglione@unina.it

Abstract

Electronic Health Records (EHRs), Big Data, Knowledge Graphs (KGs) and machine learning can potentially be a great step towards the technological shift from the *one-size-fit-all* medicine, where treatments are based on an equal protocol for all the patients, to the *precision* medicine, which takes count of all their individual information: lifestyle, preferences, health history, genomics, and so on. However, the lack of data which characterizes low-resource languages is a huge limitation for the application of the above-mentioned technologies. In this work, we will try to fill this gap by means of transformer language models and few-shot approaches and we will apply similarity-based deep learning techniques on the constructed KG for downstream applications. The proposed architecture is general and thus applicable to any low-resource language.

Enabling Factors

Data. Availability of an ever-increasing number of publicly available dataset and increased willingness of institutions to provide anonymized data.

EHRs. Collections of data related to the medical history of patients (e.g., laboratory measurements, radiology imaging, clinical notes).

Transformers. Deep learning-based language models which have extremely improved the state-of-the-art in natural language understanding tasks.

Representation Learning. Low-dimensional distributed embeddings which allow entities and relations in a Knowledge Graph to be represented in a Euclidean space and thus similarity-based metrics to be used.

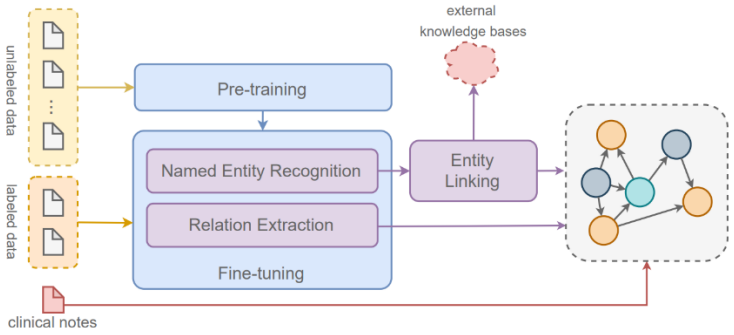
Challenges

Italian biomedical transformer. Lack of an Italian language model specialized in the biomedical field, which would enhance text understanding capabilities on EHR clinical notes.

Low resources. There is a lack of unlabeled text data for the *pretraining* phase and a lack of publicly available Italian datasets for NER, RE and EL tasks.

Few-shot Learning. Since the annotation process is time-consuming and requires domain knowledge, there is a need for few-shot learning techniques, i.e. high-performing methods even with few labeled samples.

Planned Methodology



Early Results

Data Collection

- Clinical notes provided by the hospital *Azienda Ospedaliera Universitaria (AOU) Federico II* – then labeled by 8 biomedical students with disease and symptom entity mentions.
- Collection of pre-training text: clinical notes (646,774 sentences), topics on the forum *Medicitalia* (14,484,684 sentences) and information from DBpedia (7,129 sentences)

Pattern-Exploiting Training for Named Entity Recognition