Towards an Italian Healthcare Knowledge Graph*

Marco Postiglione[0000−0003−1470−8053]

University of Naples Federico II, 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.

Keywords: Knowledge Graphs · Electronic Health Records · Transformer Language Models.

1 Introduction

The Big Data paradigm has become a reality thanks to the availability of an ever-growing quantity of data and the synergetic progress in computing infrastructures and data analysis techniques [6]. To the current state, Big Data solutions are already in use to support us in our daily life for safety [26], entertainment [25] and healthcare [3] inter alia.

In the healthcare industry, the recent progress made in *Electronic Health Records (EHRs)* has enabled the collection of huge quantities of data related to the medical histories of patients (e.g. laboratory measurements, radiology imaging, clinical notes). Being closer to the actual practice of medicine as compared with the idealized information presented in textbooks and journals, EHRs provide the possibility to (1) identify possible causal relations between healthcare entities (e.g. symptoms, diseases, measurements) which are not even written in books [30] and (2) suggest personalized treatments.

The heterogeneous information of EHRs can be organized in graph data structures, a.k.a. *Knowledge Graphs (KGs)*, which capture the relationships between different entities by linking them through edges. Once the KG is constructed, not only can data be easily and interactively visualized and explored

* Supported by Oracle Labs

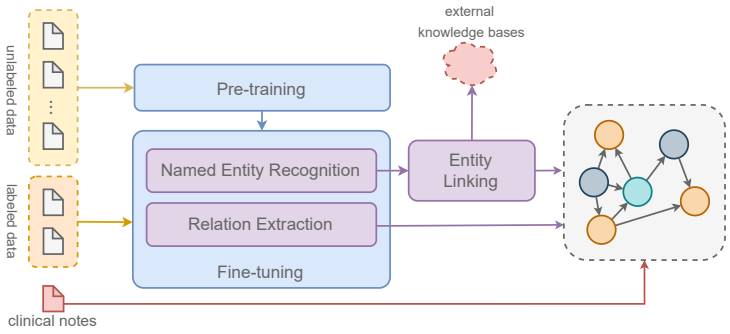

Fig. 1. Overview of the planned methodology.

by analysts and physicians, but they can also be analyzed with machine and deep learning techniques to solve complex tasks, such as providing personalized therapies. For example, KGs have been effectively used for the prediction of adverse drug reactions in patients [46] and for drug repurposing for the treatment of COVID-19 [40] especially thanks to embedding methods which allow to represent the KG entities in an Euclidean space and thus to exploit distance and similarity-based metrics to analyze relations between nodes.

In this work, we will try to pave the way towards the application of healthcare KGs in low-resource languages, where all the advances detailed above — in EHRs, KGs and analytics techniques — cannot be fully exploited due to the lack of data. Specifically, the overall project contributions are summarized as follows:

1. Pre-training of a *transformer* language model [4, 8, 29] based on Italian biomedical corpora
2. Definition of few-shot learning approaches to use the pre-trained model to recognize entities and relations from clinical notes
3. Entity linking to external knowledge bases with similarity-based approaches
4. Smart navigation and analysis of the constructed KG with deep learning similarity-based techniques.

2 Planned Methodology

Figure 1 shows an outline of the planned methodology, which will be detailed in the remainder of this section.

2.1 Language Model: pre-training & fine-tuning

The automatic understanding and processing of clinical notes is a challenging task due to several peculiarities, i.e. negations, synonyms, alternate spelling of entities, non-standard abbreviations, polysemous words [7, 43]. Thanks to their

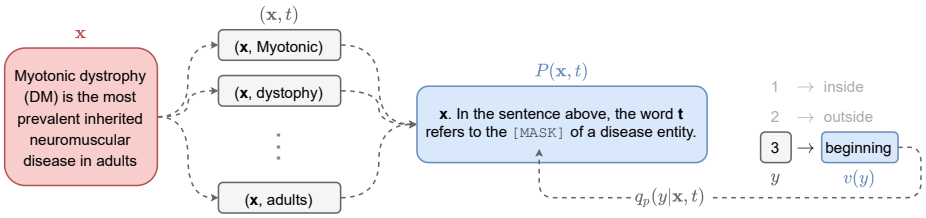

Fig. 2. Example of PETER application.

effectiveness in leveraging both words and their contexts, transformer language models (e.g. BERT [8], GPT-3 [4], T5 [29]) have proven to be a valuable solution for this challenge. They are first *pre-trained* on huge quantities of unlabeled text data and then *fine-tuned* with labeled data to solve downstream tasks (e.g. sentence classification, part-of-speech labeling). Inspired by recent works in biomedical language understanding which have shown that performance of downstream tasks can be strongly improved by pre-training on biomedical text data (e.g. papers, clinical notes) [2,20,22], we will pre-train an Italian biomedical language model and fine-tune it to build our KG by detecting entities (*Named Entity Recognition*, a.k.a. NER [14]) and extracting unknown relational facts (*Relation Extraction* [28,35,44,45]) from clinical notes.

The lack of annotated data characterizing low-resource languages imposes the use of few-shot learning approaches [12,19,27,33,36,42]. *Pattern-Exploiting Training (PET)* [32] has been proved to be an effective technique to fine-tune language models for few-shot classification tasks. Hence, we developed PETER (**P**attern-**E**xploiting **T**raining for Named **E**ntity **R**ecognition), a slight adaptation which allows to use PET for NER and we intuitively describe it in Figure 2. Given a sequence of tokens $\mathbf{x}$ (i.e. a sentence) a "pattern" is applied to each token to generate input examples containing a *mask* token which will be replaced by the model with the appropriate label (which indicates if the token is at the *beginning*, *inside* or *outside* of an entity mention). In this way, the language model leverages the knowledge it has acquired during the pre-training phase to solve the downstream task, hence requiring few samples to obtain satisfactory performance.

2.2 Entity Linking

The knowledge retrieved from EHRs will be extended with external knowledge bases containing additional useful information. For example, WikiData [37] allows to link diseases with their corresponding *International Classification of Diseases (ICD)* code and to enrich nodes with suggested drugs, therapies, health specialty, and so on. The *Entity Linking* task consists in annotating mentions with their corresponding identifier in an external knowledge base. It involves candidate-entity *generation* and *ranking* [1], i.e. retrieving all the possible entities which may be linked to an entity mention and returning the most likely

one. Current literature shows the effectiveness of semantic similarity-based approaches which use neural networks and word embeddings [11,18] to capture the semantic correspondence between entity mentions and external entities. To this end, we plan to use word embeddings obtained with the pre-trained language model to compute the distance between the entities recognized in clinical notes and WikiData concepts, and thus link the most appropriate one.

2.3 Knowledge Graph Analysis

Despite the construction of the KG being an important step in our project, its smart navigation and analysis is essential to effectively use it as a supportive tool in the actual practice of medicine. To this end, we will leverage similarity measures which take count of not only the informative content of nodes (i.e. node properties) but also the topological graph structure (e.g. path length and depth). More specifically, we will leverage *knowledge representation learning* techniques, which aim to learn low-dimensional embedding of nodes and relations.

Embeddings have to guarantee the possibility to define *scoring functions* [10], which are used to measure the plausibility of facts, i.e. $(head, relation, tail)$ triples. As an example, a scoring function f can be exploited to return the probability that a patient suffers from a disease given all its attributes (e.g. age, sex, lifestyle) and laboratory exams $\rightarrow^{e.g.} f(Alice, hasDisease, Diabetes) = 0.6$.

We plan to employ similarity-based functions, which use semantic matching to calculate the semantic similarity between entities [39, 41, 47]. To this end, neural networks have been proven to effectively encode the semantic matching principle by feeding entities or relations or both into deep networks to compute a similarity score [17]. In particular, *Graph Convolutional Neural networks (GCNs)* have been proven to be effective in leveraging the attributes associated with nodes [15]. Node structures, attributes and relation types can be integrated in weighted GCN models [34], which treat multi-relational KGs as multiple single-relational graphs and learn weights when combining GCN embeddings for each subgraph and node. The output of the l-th layer for the node v_i can be thus computed as:

$$h_i^{l+1} = \sigma\Big(\sum_{j \in N_i} \alpha_t^l g(h_i^l, h_j^l) \Big), \tag{1}$$

where: h_i^l and h_j^l are the input vectors for nodes v_i and v_j, respectively, and v_j is a node in the neighborhood N_i of v_i; α_t is a learnable parameter specifying the strength of the relation type t between two adiacent nodes; σ is an activation function; g incorporates neighboring information with a coefficient matrix.

3 Early results

Our research activities have so far been focused on the first steps of our research plan, i.e. data collection, pre-training and the definition of few-shot learning techniques. In particular, the hospital *Azienda Ospedaliera Universitaria (AOU)*

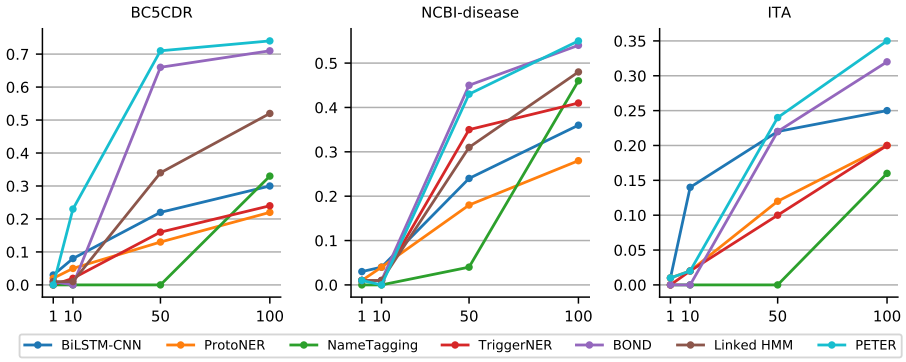

Fig. 3. PETER comparison with the state-of-the-art in terms of F1 scores.

Federico II has provided a database with information about hospitalizations in their cardiological departments. With reference to the pre-training phase, all the clinical notes included in the above-mentioned database (646, 774 sentences) have been collected and integrated with information collected from the forum *Medicitalia*[1] (14, 484, 684 sentences) and *DBpedia* [21] (7, 129 sentences).

Furthermore, a team of 8 biomedical engineers has labeled a subset of the clinical notes to allow the fine-tuning of NER models. More in detail, the dataset contains 6186 *disease* and 4918 *symptom* mentions annotated.

Finally, we compared PETER with state-of-the-art few-shot NER techniques [5, 13, 16, 23, 24, 31] on three datasets: BC5CDR [38], NCBI-disease [9], and the Italian NER dataset described above. Results in Figure 3 show that PETER obtains higher results w.r.t. the other techniques in terms of F1 scores (y-axis) in several few-shot contexts (x-axis). While models fine-tuned on BC5CDR and NCBI-Disease are initialized with BioBERT [20], i.e. a biomedical transformer, the Italian model has been initialized with GilBERTo[2], which is trained on general text corpora. The resulting poor performance shows the crucial importance of the pre-training step we are currently working on.

4 Conclusion & Future Work

In this paper, we have described the planned research methodology for the construction and analysis of an Italian KG in healthcare. We have so far (1) collected the required data to train the language model and (2) defined the few-shot approach to be used. In future work, we will link medical entities to external knowledge bases and analyze the developed KG with similarity-based techniques which take count of both the topological graph structure and the information content of nodes.

[1] https://www.medicitalia.it/
[2] https://github.com/idb-ita/GilBERTo

References

1. Al-Moslmi, T., Ocaña, M.G., Opdahl, A.L., Veres, C.: Named entity extraction for knowledge graphs: A literature overview. IEEE Access **8**, 32862–32881 (2020)
2. Alsentzer, E., Murphy, J., Boag, W., Weng, W., Jin, D., Naumann, T., McDermott, M.B.A.: Publicly available clinical bert embeddings. ArXiv **abs/1904.03323** (2019)
3. Amendola, S., Lodato, R., Manzari, S., Occhiuzzi, C., Marrocco, G.: Rfid technology for iot-based personal healthcare in smart spaces. IEEE Internet of Things Journal **1**, 144–152 (2014)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. ArXiv **abs/2005.14165** (2020)
5. Cao, Y., Hu, Z., Chua, T.S., Liu, Z., Ji, H.: Low-resource name tagging learned with weakly labeled data. ArXiv **abs/1908.09659** (2019)
6. Chen, M., Mao, S., Liu, Y.: Big data: A survey. Mobile networks and applications **19**(2), 171–209 (2014)
7. Dalloux, C., Claveau, V., Grabar, N., Oliveira, L.E.S., Moro, C., Gumiel, Y.B., Carvalho, D.: Supervised learning for the detection of negation and of its scope in french and brazilian portuguese biomedical corpora. Natural Language Engineering **27**, 181 – 201 (2020)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019)
9. Dogan, R., Leaman, R., Lu, Z.: Ncbi disease corpus: A resource for disease name recognition and concept normalization. Journal of biomedical informatics **47**, 1–10 (2014)
10. Ebisu, T., Ichise, R.: Toruse: Knowledge graph embedding on a lie group. In: AAAI (2018)
11. Francis-Landau, M., Durrett, G., Klein, D.: Capturing semantic similarity for entity linking with convolutional neural networks (2016)
12. Fries, J., Wu, S., Ratner, A., Ré, C.: SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data. arXiv:1704.06360 [cs] (Apr 2017), http://arxiv.org/abs/1704.06360, arXiv: 1704.06360
13. Fritzler, A., Logacheva, V., Kretov, M.: Few-shot classification in named entity recognition task. Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (2019)
14. Grishman, R., Sundheim, B.: Message understanding conference- 6: A brief history. In: COLING (1996)
15. Hamilton, W.L., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: NIPS (2017)
16. Hofer, M., Kormilitzin, A., Goldberg, P., Nevado-Holgado, A.: Few-shot learning for named entity recognition in medical text. ArXiv **abs/1811.05468** (2018)
17. Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S.: A survey on knowledge graphs: Representation, acquisition and applications. IEEE transactions on neural networks and learning systems **PP** (2021)
18. Karadeniz, I., Özgür, A.: Linking entities through an ontology using word embeddings and syntactic re-ranking. BMC Bioinformatics **20** (2019)

19. Kim, S., Toutanova, K., Yu, H.: Multilingual named entity recognition using parallel data and metadata from wikipedia. In: ACL (2012)
20. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**, 1234 – 1240 (2020)
21. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**, 167–195 (2015)
22. Lewis, P., Ott, M., Du, J., Stoyanov, V.: Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In: CLINICALNLP (2020)
23. Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., Zhang, C.: Bond: Bert-assisted open-domain named entity recognition with distant supervision. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020)
24. Lin, B.Y., Lee, D.H., Shen, M., Moreno, R.R., Huang, X., Shiralkar, P., Ren, X.: Triggerner: Learning with entity triggers as explanations for named entity recognition. In: ACL (2020)
25. Lippell, H.: Big data in the media and entertainment sectors. In: New Horizons for a Data-Driven Economy (2016)
26. Mannering, F., Bhat, C., Shankar, V., Abdel-Aty, M.: Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Analytic Methods in Accident Research **25**, 100113 (2020)
27. Mintz, M.D., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: ACL/IJCNLP (2009)
28. Nguyen, T., Grishman, R.: Relation extraction: Perspective from convolutional neural networks. In: VS@HLT-NAACL (2015)
29. Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. ArXiv **abs/1910.10683** (2020)
30. Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., Sontag, D.: Learning a health knowledge graph from electronic medical records. Scientific Reports **7** (2017)
31. Safranchik, E., Luo, S., Bach, S.H.: Weakly supervised sequence tagging from noisy rules. In: AAAI (2020)
32. Schick, T., Schütze, H.: Exploiting cloze-questions for few-shot text classification and natural language inference. In: EACL (2021)
33. Schmidhuber, J.: On learning how to learn learning strategies (1994)
34. Shang, C., Tang, Y., Huang, J., Bi, J., He, X., Zhou, B.: End-to-end structure-aware convolutional networks for knowledge base completion (2018)
35. Shen, Y., Huang, X.: Attention-based convolutional neural network for semantic relation extraction. In: COLING (2016)
36. Thrun, S., Pratt, L.Y.: Learning to learn (1998)
37. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**, 78–85 (2014)
38. Wei, C.H., Peng, Y., Leaman, R., Davis, A.P., Mattingly, C., Li, J., Wiegers, T.C., Lu, Z.: Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. Database: The Journal of Biological Databases and Curation **2016** (2016)
39. Xue, Y., Yuan, Y., Xu, Z., Sabharwal, A.: Expanding holographic embeddings for knowledge completion. In: NeurIPS (2018)

40. Yan, V.K.C., Li, X., Ye, X., Ou, M., Luo, R., Zhang, Q., Tang, B., Cowling, B., Hung, I., Siu, C., Wong, I., Cheng, R.C.K., Chan, E.: Drug repurposing for the treatment of covid-19: A knowledge graph approach. Advanced Therapeutics (2021)
41. Yang, B., tau Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. CoRR **abs/1412.6575** (2015)
42. Yarowsky, D., Ngai, G.: Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In: NAACL (2001)
43. Yoon, W., So, C.H., Lee, J., Kang, J.: Collabonet: collaboration of deep neural networks for biomedical named entity recognition. BMC Bioinformatics **20**(S10) (May 2019). https://doi.org/10.1186/s12859-019-2813-6, http://dx.doi.org/10.1186/s12859-019-2813-6
44. Zeng, D., Zhao, C., Quan, Z.: Cid-gcn: An effective graph convolutional networks for chemical-induced disease relation extraction. Frontiers in Genetics **12** (2021)
45. Zeng, X., He, S., Liu, K., Zhao, J.: Large scaled relation extraction with reinforcement learning. In: AAAI (2018)
46. Zhang, F., Sun, B., Diao, X., Zhao, W., Shu, T.: Prediction of adverse drug reactions based on knowledge graph embedding. BMC Medical Informatics and Decision Making **21** (2021)
47. Zhang, W., Paudel, B., Zhang, W., Bernstein, A., Chen, H.: Interaction embeddings for prediction and explanation in knowledge graphs. Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (2019)