Project of PhD thesis in Computer Science: "Patient representation learning for analogical reasoning in healthcare"

Miguel Couceiro (Loria),
Adrien Coulet (Inria Paris) and
Nicolas Garcelon (IHU Imagine)

Sept. 2022

Keywords: AI in healthcare, Analogy reasoning, Electronic health records, supervised machine learning, diagnosis, prognosis, patient representation learning
Titre en français : Apprentissage de representation de patients pour le raisonnement par analogie en santé :

1 Motivation and context

1.1 Analogy Reasoning

Analogical reasoning (AR) is a capability of human thought that exploits parallels between situations of different nature to infer plausible conclusions, by relying simultaneously on similarities and dissimilarities. Artificial Intelligence (AI) has tried to develop AR by adopting the view of analogical proportions as statements of the form "A is to B as C is to D", usually denoted A:B::C:D. Such proportions are at the root of the analogical inference mechanism and they have been formalized by logical and algebraic approaches. Analogical inference has been used in machine learning tasks such as classification, preference prediction and recommendation with competitive results [13]. Moreover, analogical extrapolation can solve difficult reasoning tasks such as IQ tests and machine translation, and can support dataset augmentation (analogical extension) for model learning, especially in environments with few labeled examples [4]. Most recent algorithmic approaches to detecting and solving analogies rely on postulates inspired from the common view of analogy as a geometric or an arithmetic proportion. For instance, recent NLP approaches to solving morphological analogies formulated as statements A:f(A)::B:f(B) with f being a grammatical flection (*e.g.*, do:doing::make:making) , include those proposed by Fam *et al.* [14] and the Alea algorithm proposed by Langlais *et al.* [7]. A more empirical approach was proposed by Murena *et al.* [8], relaxing the formal definition of analogical proportion. Based on evidence that humans follow a simplicity principle when solving analogies, the authors proposed to solve analogical equations A:B::C:X by finding a function f of minimal Kolmogorov complexity (description length) such that B=f(A), and thus giving a solution X=f(C). More recent approaches use deep learning architectures together with tailor made embeddings [9, 10]. Despite the effectiveness of the latter, they remain task and domain specific, and strongly rely on ad-hoc representations of objects.

As mentioned above, most algorithmic approaches to reason with analogies remain task and domain specific, strongly relying on ad-hoc representations of objects. However, two recent deep learning approaches, sharing the same architecture on two distinct NLP tasks, highlighted the importance of a correct choice of a representation space: the approach in [10] with SOTA results on semantic tasks, whereas that in [9] achieves SOTA results on morphological analogies. Both rely on the same deep learning architecture, but differ on the representation space: semantic embeddings (Glove) for [10], and embeddings capable of capturing morphemes for [9]. This motivates the study and choice of representation spaces. We claim that suitable representations are the key to transfer the analogy-based framework to other settings and to handle complex objects, for which data are available, such as patients for which heterogeneous data are available in clinical data warehouses.

1.2 Electronic Health Records

Electronic Health Records (EHRs) enable secondary use of hospital historical data, and in particular the design and conduct of statistical clinical studies [6]. But, EHRs also demonstrated that they provide enough information on patients to train supervised algorithms with useful applications in healthcare, such as patient prioritization and diagnosis guidance [2, 11]. EHRs are complex data as they combine structured data (*e.g.*, diagnostic codes, drug prescriptions, laboratory results) and unstructured data (*e.g.,* clinical texts, images) in two dimensions: patient and time. These temporal and multidimensional aspects motivated with success the development of supervised approaches to learn *patient representations* from EHRs [12, 3]. This is similar to what is done in NLP with language representations, which embed language complex features *e.g.,* word contexts, order, sentence syntax. For instance, in the healthcare context, sequences of words can be replaced by sequences of patient-related events.

2 Objective

We will investigate healthcare applications of the analogy based framework to patient representations learned from EHRs [3]. In a first application, we will omit the temporal dimension of EHRs to build analogies within EHR data. Here, analogies will have the form *Alice's documents:Alice's lab results::Bob's* *documents:Bob's lab results*. We will explore various combinations of data types and modalities (*e.g.,* text, drug prescription, lab result) as well as of patient representation models. If successful, this would lead to an original study of relationships between different types of patient observations. The second application is more ambitious and will include the temporal dimension. We will investigate analogies of the form *Alice before intervention:Alice after intervention::Bob before intervention:Bob after intervention.* This would be particularly useful for prognostic purposes: given other patients' history, the idea is to find values for *Bob after intervention.* For both applications, we will produce data sets of valid and invalid analogies on the basis of real-world EHRs and expert knowledge.

3 Data

We plan to experiment with two sources of EHRs: the MIMIC database [1] and the Necker Hospital instance of Dr. Warehouse [5]. MIMIC ($40,000$ patients) is a shared anonymized dataset issued from a US hospital; Dr. Warehouse is a document-oriented clinical data warehouse installed at Necker Hospital ($800,000$ patients). The HeKA team has experience in the use of these data sources. The agreement to have access to MIMIC is acquired and a request has been initiated for Dr. Warehouse. Datasets of valid analogies and models built from Dr. Warehouse will stay on the local network of the AP-HP and will be shared internally, upon agreement by the associated CSE (*Comité Scientifique et Ethique*).

4 Context of the thesis

The candidate will be part of the HeKA research team (Inserm-Inria-Université Paris Cité) research team (`https://team.inria.fr/heka/`), and located at PariSanté Campus (`https://parisantecampus.fr/`). Weelky visits to the IHU Imagine are planed. The candidate will be co-supervised by Miguel Couceiro, Adrien Coulet and Nicolas Garcelon. The candidate will register to the Doctoral School ED386 at the Université Paris Cité (`http://www.math.univ-paris-diderot.fr/formations/doctorats/index`) under the direction of Adrien Coulet. Prof. Anita Burgun of the HeKA team will also participate in the supervision of the thesis.

The thesis is funded on the ANR project AT2TA (2022-2026).

Contacts

Miguel Couceiro, Université de Lorraine, Loria `miguel.couceiro@loria.fr`
Adrien Coulet, Inria, Inria Paris `adrien.coulet@inria.fr`
Nicolas Garcelon, IHU Imagine `nicolas.garcelon@institutimagine.org`

References

[1] A. Johnson et al. Mimic-iv (version 0.4). *PhysioNet*, 2020.

[2] A. Schuler et al. Performing an informatics consult: methods and challenges. *J. American College of Radiology*, 15(3):563–568, 2018.

[3] I. Landi et al. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ digital medicine*, 3(1):1–11, 2020.

[4] M. Couceiro et al. Analogy-preserving Functions: A Way to Extend Boolean Samples. In *26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*, pages 1–7, Melbourne, Australia, August 2017. International Joint Conferences on Artifical Intelligence (IJCAI), International Joint Conferences on Artifical Intelligence (IJCAI).

[5] N. Garcelon et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. warehouse. *J. biomedical informatics*, 80:52–63, 2018.

[6] P. B. Jensen et al. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.*, 13(6):395–405, 2012.

[7] P. Langlais et al. Improvements in analogical learning: Application to translating multi-terms of the medical domain. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 487–495, Athens, Greece, March 2009. Association for Computational Linguistics.

[8] P.A. Murena et al. Solving analogies on words based on minimal complexity transformation. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1848–1854. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[9] S. Alsaidi et al. A Neural Approach for Detecting Morphological Analogies. In *The 8th IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Porto/Online, Portugal, October 2021.

[10] Suryani Lim et al. Classifying and completing word analogies by machine learning. *Int. J. Approx. Reason.*, 132:1–25, 2021.

[11] Y. Gao et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in china: a retrospective, multicentre, diagnostic study. *The Lancet Digital Health*, 4(3):e179–e187, 2022.

[12] Y. Si et al. Deep representation learning of patient data from electronic health records (ehr): A systematic review. *J. Biomedical Informatics*, 115:103671, 2021.

[13] Z. Bouraoui et al. From Shallow to Deep Interactions Between Knowledge Representation, Reasoning and Machine Learning (Kay R. Amel group). 53 pages ; Kay R. Amel is the pen name of the working group “Apprentissage et Raisonnement” of the GDR (“Groupement De Recherche”)named “Aspects Formels et Algorithmiques de l’Intelligence Artificielle”, CNRS, France (https://www.gdria.fr/presentation/), 2019.

[14] R. Fam and Y. Lepage. Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).