1. Automated Extraction and Utilization of Clinical Data from Electronic Health Records of Cardiac Patients

The invasion of artificial intelligence (AI) in modern medicine could pave the way for the transformation of everyday clinical practice. The application of AI to extract useful clinical information from electronic health records (EHRs) is a valuable and feasible solution in the era of an unprecedented amount of clinical data. The goal of this thesis, in collaboration with the Cardiology Department of AHEPA hospital, is to automatically transform the clinical information of the EHRs from manually organized, raw text-based clinical notes to interpretable datasets by developing an AI model based on natural language processing (NLP) and machine learning (ML). Applying this model to EHRs from Cardiology wards will form the largest national dataset of cardiac patients. The thesis will use data from a retrospective, multi-center study based on large, longitudinal data obtained from the EHRs of the largest tertiary hospitals in Greece. The data extracted from the EHRs will include patient demographics, hospital administrative data, medical history, medications, lab tests, imaging notes and reports, therapeutic interventions, in-hospital management and post-discharge instructions. This information will be extracted both manually and by an AI model, integrating NLP to structuralize the acquired data. All this diverse collected data will be organized into a well-documented registry that will also be embellished with prognostic data aggregated centrally from the National Institute of Health, which will be refined to allow for the application of ML techniques.

Support

- Dimitris Papadopoulos, PhD student

Bibliography

- BERT-based Ranking for Biomedical Entity Normalization. (n.d.). Retrieved June 6, 2022, from https://pubmed.ncbi.nlm.nih.gov/32477646/
- Li, F., Jin, Y., Liu, W., Rawat, B. P. S., Cai, P., Yu, H. (2019). Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. JMIR Medical Informatics, 7(3), e14830. https://doi.org/10.2196/14830
- Gligic, L., Kormilitzin, A., Goldberg, P., Nevado-Holgado, A. (2019). Named Entity Recognition in Electronic Health Records Using Transfer Learning Bootstrapped Neural Networks. https://doi.org/10.48550/arxiv.1901.01592
- Nesterov, A., Umerenkov, D. (2022). Distantly supervised end-to-end medical entity extraction from electronic health records with human-level quality. https://doi.org/10.48550/arxiv.2201.10463
- Vashishth, S., Newman-Griffis, D., Joshi, R., Dutt, R., Rosé, C. P. (2021). Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. Journal of Biomedical Informatics, 121. https://doi.org/10.1016/j.jbi.2021.103880

2. Plain Language Summarization

In an ideal world, all scientific articles should be understandable by a broad audience, including people of both scientific and non-scientific backgrounds. This way, the potential impact of the article would be maximized, as a large number of individuals can benefit from the findings of the research. In contrast, for most scientific articles and domains, only a handful of people with a scientific background are able to understand their content. Frequently, even research scientists that are not well versed in that particular domain might have a hard time comprehending all of the information in an article. As a result, the majority of published scientific research is only accessible to a small subset of specialized domain experts, while the rest of the public cannot benefit from the findings of the research. This problem became particularly apparent during the COVID-19 crisis. Millions of scientific articles have been produced regarding different aspects of the topic but the majority of them are simply “too scientific" to be easily understood by the general public. As a result, the outcomes of these works are frequently overlooked or, even worse, misinterpreted to a point that gives rise to misinformation and fake news. Lately, the academic community started to realize that producing research that is not relatable to a broader audience is leading to scientific and health illiteracy. This led a number of research institutions and venues to take a different approach in order to make published research more accessible to the general public. For example, several publishing venues now require authors to also provide lay summaries that explain their work in simpler terms. A similar approach is also adopted by public funding agencies that require lay summaries for grant applications. Although this approach is definitely towards the right direction, still it has several major drawbacks. First, such an approach requires a significant amount of human effort on the author’s side and it is unlikely that it will be widely adopted. Second, the author of an article might not be the best person to describe their work in lay terms. Finally, it is simply not possible to apply this treatment to the massive amount of already published work. The goal of this thesis is to develop NLP techniques that will help the automation of producing plain language summaries of scientific articles.

Support

- Tatiana Passali, PhD student

Bibliography

- M.K. Chandrasekaran, G. Feigenblat, Hovy. E., A. Ravichander, M. Shmueli-Scheuer, and A De Waard. Overview and Insights from Scientific Document Summarization Shared Tasks 2020: {CL-SciSumm}, {LaySumm} and {LongSumm}. In Proceedings of the First Workshop on Scholarly Document Processing (SDP 2020), 2020.
- Seungwon Kim. Using Pre-Trained Transformer for Better Lay Summarization. 2020.
- Yea Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. SimpleScience: Lexical simplification of scientific terminology. In EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 2016

3. Τεχνικές Ανάλυσης Κειμένου για τον Υπολογισμό Συνάφειας Ερευνητών με Αντικείμενα (συνεργατικό θέμα για 2-3 φοιτητές)

Κατά τη διαδικασία εκλογής ενός καθηγητή σε ένα Πανεπιστήμιο, υπάρχουν διάφορα στάδια που θα μπορούσαν να υποστηριχθούν από υπολογιστικές τεχνικές υπολογισμού της συνάφειας ενός ερευνητή με ένα ερευνητικό αντικείμενο. Οι θέσεις καθηγητών προκηρύσσονται με βάση ένα γνωστικό αντικείμενο, το οποίο περιγράφεται με έναν τίτλο, αλλά και μια σύντομη περιγραφή. Π.χ. σε μια πρόσφατη εκλογή στο τμήμα μας ο τίτλος ήταν «Ευφυή Συστήματα-Συμβολική Τεχνητή Νοημοσύνη» και η περιγραφή ήταν «Ανάπτυξη ευφυών συστημάτων με την χρήση συνδυασμού μεθοδολογιών της συμβολικής Τεχνητής Νοημοσύνης, όπως Αναπαράστασης Γνώσης και Συλλογιστικής, Πολυπρακτορικών Συστημάτων, Μηχανικής Μάθησης, Ευφυών Αυτόνομων Συστημάτων, Σχεδιασμού και Χρονοπρογραμματισμού Ενεργειών, Ικανοποίησης Περιορισμών». Με βάση αυτήν την περιγραφή καταρχάς επιλέγονται οι πιο συναφείς καθηγητές από το τμήμα, αλλά και εκτός του τμήματος, απ' όλη την Ελλάδα ή/και το εξωτερικό μέσα από μια λίστα που καθορίζει το κάθε τμήμα, προκειμένου να καταρτιστεί η εξεταστική επιτροπή. Αυτό το βήμα κάποιες φορές είναι υποκειμενικό, καθώς μπορεί να στηρίζεται στο γνωστικό αντικείμενο των εκλεκτόρων, ενώ θα ήταν πιο ενδιαφέρον να στηρίζεται στην ερευνητική εμπειρία των εκλεκτόρων όπως αυτή τεκμηριώνεται σε βιβλιογραφικές βάσεις δεδομένων. Επιπλέον, ένα υποσύνολο της εξεταστικής επιτροπής, η τριμελής επιτροπή, οφείλει να κατατάξει τους υποψήφιους για τη θέση με βάση τη συνάφειας της. Και αυτό θα πρέπει να γίνει αντικειμενικά με βάση το δημοσιευμένο τους έργο.

Στόχος αυτής της πτυχιακής είναι να αναπτυχθεί ένα σύστημα που θα υπολογίζει την συνάφεια των ερευνητών με κάποιο γνωστικό αντικείμενο και την περιγραφή του, τεκμηριωμένα με βάση το βιβλιογραφικό έργο σε βάσεις όπως Google Scholar, Scopus, DBLP, ORCID, Semantic Scholar. Για τον σκοπό αυτό απαιτείται ένας συνδυασμός (1) τεχνολογίας λογισμικού για την ανάπτυξη του συστήματος με μια διαδικτυακή ανοιχτή διεπαφή, (2) επικοινωνίας με τεράστιες βάσεις δεδομένων προκειμένου να αντληθεί το δημοσιευμένο έργο των εκλεκτόρων/υποψηφίων με αυτόματο τρόπο βάση του ονόματος τους (δεν είναι πάντα εύκολο/προφανές), και (3) επεξεργασίας φυσικής γλώσσας προκειμένου να παραχθεί μια διανυσματική αναπαράσταση του γνωστικού αντικειμένου και της περιγραφής του, π.χ. με νευρωνικά μοντέλα γλώσσας όπως το BERT, καθώς και των δημοσιεύσεων του κάθε εκλέκτορα/υποψηφίου. Με μια απλή συνάρτηση ομοιότητας (π.χ. συνημίτονου) το σύστημα μετά θα μπορεί να βρίσκει τους πιο σχετικούς ερευνητές με ένα δοθέν γνωστικό αντικείμενο και περιγραφή. Το σύστημα θα πρέπει να διατηρεί/επαυξάνει τις διανυσματικές αναπαραστάσεις των ερευνητών που υπάρχουν στη λίστα ενός τμήματος, για να μην εκτελεί κοστοβόρους υπολογισμούς και αναζητήσεις σε βάσεις δεδομένων από το μηδέν.

Βιβλιογραφία

- D. Luo, W. Cheng, J. Ni, W. Yu, X. Zhang, B. Zong, Y. Liu, Z. Chen, D. Song, H. Chen, X. Zhang. Unsupervised Document Embedding via Contrastive Augmentation. https://arxiv.org/pdf/2103.14542.pdf
- Β. Μοσχόπουλος, Κ. Νικηφορίδης. Υπολογισμός Συνάφειας Επιστημόνων με Ερευνητικά Αντικείμενα για Σύστημα Προτάσεων με Χρήση Τεχνικών Επεξεργασίας Φυσικής Γλώσσας, Διπλωματική Εργασία., https://ikee.lib.auth.gr/record/339729?ln=el

4. Explaining Transformer Models

Because it deals with socio-ethical phenomena such as discrimination, machine learning explainability is a popular research topic. Recent legislation also proposes stricter criteria for the deployment of machine learning systems in situations where end users are significantly impacted by them. Many techniques for explaining machine learning models have been suggested. In a recent research [1] attention information is suggested to be a good alternative to other interpretability techniques regarding the explainability of Transformer models on text classification, despite the heavy criticism [2, 3]. In this thesis, we aim to study the applicability of the research proposed in [1] on different transformer architectures.

Support

- Nikos Mylonas, PhD student
- Ioannis Mollas, PhD student

Bibliography

1. Mylonas Nikolaos, Ioannis Mollas, and Grigorios Tsoumakas. "Improving Attention-Based Interpretability of Text Classification Transformers." arXiv preprint arXiv:2209.10876. 2022.
2. Jain Sarthak, and Byron C. Wallace. "Attention is not Explanation." Proceedings of NAACL-HLT. 2019.
3. Bastings Jasmijn, and Katja Filippova. "The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?." Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. 2020.

Διαδικασία υποβολής αίτησης

Διευκρινήσεις επί των θεμάτων θα δοθούν την Παρασκευή 14/10 στις 11:00 στην παρακάτω τηλεδιάσκεψη:
https://authgr.zoom.us/j/95161623128?pwd=UWgrMkxnNFNhdm51SncvNE9yd09JQT09

Οι ενδιαφερόμενοι καλούνται να ανεβάσουν ένα αρχείο zip με (α) σύντομο βιογραφικό σημείωμα (β) αναλυτική βαθμολογία, (γ) πρόσφατη φωτογραφία, και (δ) λίστα με τα θέματα για τα οποία ενδιαφέρονται, ταξινομημένη σε σειρά προτίμησης στο URL https://www.dropbox.com/request/hfrhBeyE6EyrMYq0AEQV*.*

Καταληκτική ημερομηνία υποβολής αιτήσεων: 19/10

Ανακοίνωση αναθέσεων: 21/10