Proposal of the First International Workshop on Semantic Indexing and Information Retrieval for Health from Heterogeneous Content Types and Languages (SIIRH)

Francisco M. Couto[1(✉)] and Martin Krallinger[2]

[1] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
fcouto@di.fc.ul.pt
[2] Life Science Department, Barcelona Supercomputing Centre (BSC-CNS), C/Jordi Girona 29-31, 08034 Barcelona, Spain
martin.krallinger@bsc.es

Abstract. The application of Information Retrieval (IR) and deep learning strategies to explore the vast amount of rapidly growing health-related content is of utmost importance, but is also particularly challenging, due to the very specialized domain language, and implicit differences in language characteristics depending on the content type.

This workshop aims at presenting and discussing current and future directions for IR and machine learning approaches devoted to the retrieval and classification of different types of health-related documents ranging from layman or patient generated texts to highly specialized medical literature or clinical records. It includes a session on the MESINESP shared task, supported by the Spanish National Language Technology plan (Plan TL), in order to address the importance and impact of community evaluation efforts, in particular BioASQ, BioCreative, eHealth CLEF, MEDIQA and TREC, as scenarios for exploring evaluation settings and generate data collections of key importance for promoting the development and comparison of IR resources. Additionally, an open session will address IR technologies for heterogeneous health-related content open to multiple languages with a particular interest in the exploitation of structured controlled vocabularies and entity linking, covering the following topics: multilingual and non-English health-related IR, concept indexing, text categorization, generation of evaluation resources biomedical document IR strategies; scalability, robustness and reproducibility of health IR and text mining resources; use of specialized machine translation and advanced deep learning approaches for improving health related search results; medical Question Answering search tools; retrieval of multilingual health related web-content; and other related topics.

Supported by FCT through funding of the DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017, and LaSIGE Research Unit, ref. UIDB/00408/2020.

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 654–659, 2020.
https://doi.org/10.1007/978-3-030-45442-5_87

Keywords: Semantic indexing · Ontologies · Controlled vocabularies · Information Retrieval · Text mining · Natural language processing · Biomedical informatics

1 Introduction

There is an increasing interest in exploiting the vast amount of rapidly growing content related to health [7] by means of Information Retrieval [12] (IR) and deep learning strategies [14,18]. Health-related content is particularly challenging, due to the highly specialized domain language and implicit differences in language characteristics depending on the content type (patient-generated content like discussion forum [15], blogs [8], social media [17] and other Internet sources, healthcare documentation and clinical records [6], professional or scientific publications [9], clinical practice guidelines, clinical trials documentation, medical questionnaires, medical informed consent documents, etc.). Moreover, it is also critical to provide search solutions for non-English content as well as cross-language or multilingual IR solutions [4,10,16].

Efficient retrieval of biomedical documents is key for evidence-based medicine, preparing systematic reviews or retrieval of particular clinical case studies. Due to particular search conditions of caregivers and healthcare professionals (limited amount of time spent per patient), they are also in need of more sophisticated retrieval approaches applied to electronic health records [11], a type of content highly challenging due to its telegraphic and domain specific language and the presence of negations and abbreviations. There is also interest in processing patient-generated content like social media and patient fora, a key resource for rare disease research, clinical trials patient selection/stratification or for discovering new patient-reported symptoms and treatment-related adverse effects. In the health-domain, indexing strategies relying on structured controlled vocabularies, like MeSH/DeCS or SNOMED CT, represent a critical component for efficient biomedical search engines, enabling query expansion and refinement [2] and the improvement of recommender systems [3].

1.1 BioASQ MESINESP Session

Currently, most of the Biomedical NLP and IR research is being done on English documents [13], and only few tasks have been carried out on non-English texts [5]. Many structured controlled vocabularies are also available only in English [19]. Nonetheless, it is important to note that there is also a considerable amount of medically relevant content published in languages other than English and particularly clinical texts are entirely written in the native language of each country, with a few exceptions. The critical importance of semantic indexing with medical vocabularies motivated several-shared tasks in the past, in particular the BioASQ tracks[1], with a considerable number of participants and impact in the field. Following the outline of previous medical indexing efforts,

[1] http://bioasq.org/.

in particular the success of the BioASQ tracks centered on PubMed, the BioASQ MESINESP TASK[2], supported by the Spanish National Language Technology plan (Plan TL), proposes to carry out the first task on semantic indexing of Spanish medical texts.

This workshop will be a forum where the community can present and discuss current and future directions for the area based on the experience in participating at the MESINESP shared task or other medical IR, QA or text categorization evaluation campaigns, as well as the exploitation of evaluation settings and data collections generated through these kind of community evaluation efforts (both during and after the competition period).

1.2 Open Session

In addition to the MESINESP and shared task/evaluation campaign participation experience session, the workshop will include an Open Session covering IR technologies for heterogeneous health-related content open to multiple languages with a particular interest in the exploitation of structured controlled vocabularies and entity linking for document indexing and semantic search applications.

Among the proposed topics for the Open Session are: (1) multilingual and non-English health related IR, concept indexing and text categorization strategies, (2) generation of evaluation resources for biomedical document IR strategies, (3) scalability, robustness, reproducibility, utility and usability [1] of health IR and text mining resources, (4) use of specialized machine translation and advanced deep learning approaches for improving health related search results, (5) medical Question Answering search tools, (6) retrieval of multilingual health related web-content. Note that we will also consider other submissions related to innovative cutting-edge health and biomedical IR strategies, including evaluation and Gold Standard evaluation data set generation.

2 Planned Format and Structure

All the teams implementing systems for MESINESP will be invited to submit an article describing their participation strategy. The program committee will review the papers and select which of them will have a presentation slot at the workshop. For the Open Session we will invite researchers to submit novel IR approaches to process heterogeneous health-related content with particular interest in non-English content, novel content types as well as semantic indexing strategies exploiting structured controlled vocabularies and ontologies.

We expect that further investigation on the topics will continue after the workshop, based on new insights obtained through discussions during the event. As a venue to compile the results of the follow-up investigation, a journal special issue will be organized to be published a few months after the workshop.

[2] http://temu.bsc.es/mesinesp.

3 People Involved

3.1 Organizers

Martin Krallinger: head of the Text Mining unit at the Barcelona Supercomputing Center (BSC), Spain
Francisco M. Couto: LASIGE member and associate professor at the University of Lisbon, Portugal

3.2 Programme Committee

Alberto Lavelli: FBK, Trento, Italy
Alfonso Valencia: Barcelona Supercomputing Center, Spain
Analia Lourenco: Universidade de Vigo, Spain
Anastasios Nentidis: National Center for Scientific Research Demokritos, Greece
André Lamurias: LASIGE, Portugal
Anne: Lyse Minard - University of Orleans, France
Aron Henriksson: Stockholm University, Sweden
Bruno Martins: INESC-ID, Portugal
Carsten Eickhoff: Brown University, USA
Chih: Hsuan Wei - NCBI/NIH, National Library of Medicine, USA
Cyril Grouin: LIMSI, CNRS, Université Paris-Saclay, Orsay, France
Diana Sousa: LASIGE, Portugal
Dimitrios Kokkinakis: University of Gothenburg, Sweden
Eben Holderness: McLean Hosp., Harvard Med. School & Brandeis University, USA
Ellen Vorhees: National Institute of Standards and Technology (NIST), USA.
Fabio Rinaldi: IDSIA, University of Zurich, Switzerland & FBK, Trento, Italy
Fleur Mougin: University of Bordeaux, France
Georgeta Bordea: Université de Bordeaux, France
Georgios Paliouras: National Center for Scientific Research Demokritos, Greece
Goran Nenadic: University of Manchester, UK
Graciela Gonzalez: Hernandez - University of Pennsylvania, USA
Hanna Suominen: CSIRO, Australia
Henning Muller: University of Applied Sciences Western Switzerland, Switzerland
Hercules Dalianis: Stockholm University, Sweden
Hyeju Jang: University of British Columbia, Canada
James Pustejovsky: Brandeis University, USA
Jin: Dong Kim - Research Organization of Information and Systems, Japan
Jong C. Park: KAIST Computer Science, Korea
Kevin Bretonnel Cohen: University of Colorado School of Medicine, Colorado, USA
Maria Skeppstedt: Institute for Language and Folklore, Sweden

Marcia Barros: LASIGE, Portugal
Mariana Lara: Neves - German Federal Institute for Risk Assessment, Germany
Marta Villegas: BSC, Spain
Pedro Ruas: LASIGE, Portugal
Rafael Berlanga Llavori: Universitat Jaume I, Spain
Rezarta Islamaj: Dogan - NIH/NLM/NCBI, USA
Sérgio Matos: University of Aveiro, Portugal
Shyamasree Saha: Europe PubMed Central, EMBL-EBI, UK
Suzanne Tamang: Stanford University School of Medicine, USA
Thierry Hamon: LIMSI, CNRS, Université Paris-Saclay & Université Paris 13, France
Thomas Brox Røst: Norwegian University of Science and Technology, Norway
Yifan Peng: NCBI/NIH, National Library of Medicine, USA
Yonghui Wu: University of Florida, USA
Yoshinobu Kano: Shizuoka University, Japan
Zhiyong Lu: NCBI/NIH, National Library of Medicine, USA
Zita Marinho: Priberam, Portugal

References

1. Arighi, C.N., et al.: BioCreative III interactive task: an overview. BMC Bioinformatics **12**(8), S4 (2011). https://doi.org/10.1186/1471-2105-12-S8-S4
2. Barros, M., Couto, F.M.: Knowledge representation and management: a linked data perspective. Yearb. Med. Inform. **25**(01), 178–183 (2016)
3. Barros, M., Moitinho, A., Couto, F.: Hybrid semantic recommender system for chemical compounds. In: European Conference on Information Retrieval. Springer (2020)
4. Bawden, R., et al.: Findings of the WMT 2019 biomedical translation shared task: evaluation for MEDLINE abstracts and biomedical terminologies. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pp. 29–53 (2019)
5. Campos, L., Pedro, V., Couto, F.: Impact of translation on named-entity recognition in radiology texts. Database **2017** (2017)
6. Costumero, R., García-Pedrero, Á., Gonzalo-Martín, C., Menasalvas, E., Millan, S.: Text analysis and information extraction from Spanish written documents. In: Ślęzak, D., Tan, A.-H., Peters, J.F., Schwabe, L. (eds.) BIH 2014. LNCS (LNAI), vol. 8609, pp. 188–197. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09891-3_18
7. Couto, F.M.: Data and Text Processing for Health and Life Sciences. AEMB, vol. 1137. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-13845-5
8. Denecke, K., Nejdl, W.: How valuable is medical social media data? Content analysis of the medical web. Inf. Sci. **179**(12), 1870–1880 (2009)
9. Intxaurrondo, A., et al.: Finding mentions of abbreviations and their definitions in Spanish clinical cases: the BARR2 shared task evaluation results. In: IberEval@ SEPLN, pp. 280–289 (2018)
10. Kelly, L., et al.: Overview of the CLEF eHealth evaluation lab 2019. In: Crestani, F., et al. (eds.) CLEF 2019. LNCS, vol. 11696, pp. 322–339. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28577-7_26

11. Koleck, T.A., Dreisbach, C., Bourne, P.E., Bakken, S.: Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J. Am. Med. Inform. Assoc.J. Am. Med. Inform. Assoc. **26**(4), 364–379 (2019)
12. Krallinger, M., Rabal, O., Lourenco, A., Oyarzabal, J., Valencia, A.: Information retrieval and text mining technologies for chemistry. Chem. Rev. **117**(12), 7673–7761 (2017)
13. Lamurias, A., Couto, F.M.: Text mining for bioinformatics using biomedical literature. In: Encyclopedia of Bioinformatics and Computational Biology, vol. 1 (2019)
14. Lee, J., et al.: BioBERT: pre-trained biomedical language representation model for biomedical textmining. Bioinformatics **36**, 1234–1240 (2019)
15. Liu, X., Chen, H.: AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. In: Zeng, D., et al. (eds.) ICSH 2013. LNCS, vol. 8040, pp. 134–150. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39844-5_16
16. Marimon, M., et al.: Automatic de-identification of medical texts in Spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), vol. TBA, p. TBA. CEUR Workshop Proceedings, Bilbao, Spain, September 2019, TBA. CEUR-WS. org (2019)
17. Segura-Bedmar, I., Revert, R., Martínez, P.: Detecting drugs and adverse events from Spanish social media streams. In: Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (LOUHI), pp. 106–115 (2014)
18. Sousa, D., Couto, F.: BiOnt: deep learning using multiple biomedical ontologies for relation extraction. In: European Conference on Information Retrieval. Springer (2020)
19. Villegas, M., Intxaurrondo, A., Gonzalez-Agirre, A., Marimon, M., Krallinger, M.: The MeSpEN resource for English-Spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. In: Proceedings of the LREC 2018 Workshop "MultilingualBIO: Multilingual Biomedical Text Processing", Paris, France. European Language Resources Association (ELRA) (2018)