











CLEF eHealth Evaluation Lab 2021

Lorraine Goeuriot¹ , Hanna Suominen² , Liadh Kelly³ ,
Laura Alonso Alemany⁴, Nicola Brew-Sam⁵ , Viviana Cotik⁶, Darío Filippo⁷,
Gabriela Gonzalez Saez¹, Franco Luque¹², Philippe Mulhem¹ ,
Gabiella Pasi⁸ , Roland Roller⁹ , Sandaru Seneviratne⁵, Jorge Vivaldi¹⁰,
Marco Viviani⁸ , and Chenchen Xu¹¹

¹ Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
{lorraine.goeuriot,gabriela.saez,philippe.mulhem}@univ-grenoble-alpes.fr

² The Australian National University (ANU), Data61/Commonwealth Scientific and
Industrial Research Organisation (CSIRO), and University of Turku,
Canberra, ACT, Australia

hanna.suominen@anu.edu.au

³ Maynooth University, Co., Kildare, Ireland

liadh.kelly@mu.ie

⁴ Universidad Nacional de Córdoba, Córdoba, Argentina

lauraalonsoalemany@unc.edu.ar

⁵ The ANU, Canberra ACT, Australia

{nicola.brew-sam,sandaru.seneviratne}@anu.edu.au

⁶ Universidad de Buenos Aires, CONICET, Buenos Aires, Argentina

vcotik@dc.uba.ar

⁷ Hospital de Pediatría ‘Prof. Dr. Juan P. Garrahan’, Buenos Aires, Argentina

⁸ University of Milano-Bicocca, Milan, Italy

{gabriella.pasi,marco.viviani}@unimib.it

⁹ German Research Center for Artificial Intelligence (DFKI),

Kaiserslautern, Germany

roland.roller@dfki.de

¹⁰ Institut de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona, Spain

jorge.vivaldi@upf.edu

¹¹ The ANU and Data61/CSIRO, Canberra, ACT, Australia

chenchen.xu@anu.edu.au

¹² Universidad Nacional de Córdoba, CONICET, Córdoba, Argentina

francolq@unc.edu.ar

Abstract. Motivated by the ever increasing difficulties faced by laypeople in retrieving and digesting valid and relevant information to make health-centred decisions, the CLEF eHealth lab series has offered shared tasks to the community in the fields of *Information Extraction* (IE), management, and *Information Retrieval* (IR) since 2013. These tasks have attracted large participation and led to statistically significant improvements in processing quality. In 2021, CLEF eHealth is calling for participants to contribute to the following two tasks: Task 1 on IE focuses on

LG, HS, & LK co-chair the CLEF eHealth lab and contributed equally to this paper. Task 1 is led by LAA and VC, and organized by DF, FL, RR, and JV; Task 2 is led by LG, GP, and HS, and organized by NB-S, GGS, LK, PM, SS, MV, and CX.

© Springer Nature Switzerland AG 2021

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, pp. 593–600, 2021.

https://doi.org/10.1007/978-3-030-72240-1_69

IE from noisy text. Participants will identify and classify Named Entities in written ultrasonography reports, containing misspellings and inconsistencies, from a major public hospital in Argentina. Identified entities will then have to be classified, which can be very challenging as it requires to handle lexical variations. Task 2 is a novel extension of the most popular and established task on consumer health search (CHS), aiming at retrieving relevant, understandable, and credible information for patients and their next-of-kins. In this paper we describe recent advances in the fields of IE and IR, and the subsequent offerings of this years CLEF eHealth lab challenges.

Keywords: eHealth · Medical informatics · Information extraction · Information storage and retrieval

1 Introduction

The requirement to ensure that patients¹ can understand their official, privacy-sensitive health information in their own *Electronic Health Records* (EHRs) is stipulated by policies and laws [16]. Patients' better abilities to understand their own EHR empowers them to take part in the related healthcare judgment, leading to their increased independence from healthcare providers, better healthcare decisions, and decreased healthcare costs [16]. Improving patients' ability to access and digest this content could mean paraphrasing the EHR-text, enriching it with hyperlinks to term definitions, care guidelines, and further supportive information on patient-friendly and reliable websites, helping them to discover good search queries to retrieve more contents, and allowing not only text but also speech as a query modality for example.

Information access conferences have organized evaluation labs on related *Electronic Health* (eHealth) *Information Extraction* (IE), *Information Management* (IM), and *Information Retrieval* (IR) tasks for almost 20 years. Yet, with rare exception, they have targeted the healthcare experts' information needs only [4, 5, 11]. The *CLEF eHealth Evaluation-lab and Lab-workshop Series*² has been organized every year since 2012 as part of the *Conference and Labs of the Evaluation Forum* (CLEF) [7, 8, 10, 12–14, 19, 22, 23] with the primary goal of supporting laypersons, and their next-of-kin, access to medical information. This year, the lab proposes two tasks: one centered on Information Extraction (identify and classify Named Entities in written ultrasonography reports); one centered on Information Retrieval (*Consumer Health Search* (CHS)).

In this paper we overview the interest in the CLEF eHealth evaluation lab series to-date. We then consider recent advances in IE and IR which inform the offered CLEF eHealth 2021 IE and IR tasks. These IE and IR evaluation lab challenge tasks are also described. The paper concludes with a vision for CLEF eHealth beyond 2021.

¹ In the paper, we consider *patients*, *layperson* or *consumer*, to be system users with no or little medical background.

² <http://clefehealth.imag.fr>.

2 CLEF eHealth in 2012–2020

The CLEF and other information access conferences have organized evaluation labs and shared tasks on eHealth IE, IR, and Information Management for approximately two decades. Yet, their primary focus has been on healthcare experts' information needs, with limited consideration of laypersons' difficulties to retrieve and digest credible, topical, and easy-to-understand contents in their preferred language to make health-centred decisions [4, 5, 11].

This niche of addressing patients, their families, health scientists, health-care policy makers, and other laypersons' health information needs in a range of languages in order to make health-centered decisions began stimulating the annual CLEF eHealth Evaluation-lab and Lab-workshop Series in 2012. Its first workshop took place in 2012 with an aim to organize an evaluation lab, and in 2013–2021, this lab with up to three shared tasks annually has preceded each campaign-concluding CLEF eHealth workshop [7, 8, 10, 12–14, 19, 22, 23].

3 CLEF eHealth 2021 Information Extraction Task

3.1 Preceding Efforts

In 2020, the CodiEsp task of the CLEF eHealth evaluation lab mastered the challenge of building a publicly available automatic clinical coding system for Spanish documents, which is a step towards the final application of *natural language processing* (NLP) technologies in non-English speaking countries [10]. In contrast to previous clinical coding tasks using death certificates and non-technical summaries of animal experimentations [14, 20, 21], the 2020 task was able to use a collection of clinical case reports from a variety of medical disciplines chosen to constitute a corpus of *electronic health records* (EHRs; 1,000 documents from the *Spanish clinical case reports* (SPACCC) corpus). CodiEsp shared tasks attracted participants from both Spanish and non-Spanish speaking countries, with different backgrounds in the 51 teams registered for the tasks. Thus, CodiEsp was able to prove that the language barrier (languages other than English) does not necessarily make the tasks more restrictive, but presents an opportunity to adapt well-known techniques to language-specific features. The diversity in profiles led to the development of heterogeneous resources, with a development of 167 novel clinical coding systems achieved. Finally, the 2020 task organizers' showed that individual task results could be combined, leading to further performance gains.

The 2020 task on Spanish resources was popular to the extent that it set the ground for the 2021 SpRadIE (Spanish Radiology Information Extraction) task focusing on further sub-aspects of the Spanish language: text in the radiology domain, image reports written under time constraints, resulting in misspellings and inconsistencies, coming from a public hospital in South America, as elaborated in the next subsection. These particularities pose an interesting challenge of domain and register adaptation for systems trained for general Spanish eHealth, in their application to a specific setting. With this objective, we are calling for submissions from hospitals and private companies to supplement academic participants.

3.2 The Task in 2021: Multilingual Information Extraction

In 2021, the SpRadIE task will target Named Entity Recognition and Classification in the domain of radiological image reports, more concretely, pediatric ultrasonographies. These reports are written in haste, under time pressure in a public Argentinean hospital. They tend to be repetitive, probably due to an extensive use of copy and paste. Nevertheless, these are actual free text reports with no pre-determined structure, which results in great variations in size and content. No element is mandatory in the report except the age of the patient. Also, there are misspellings and inconsistencies in the usage of abbreviations, punctuation and line breaks.

The corpus consists of a total of 513 sonography reports, with over 17,000 annotated named entities with some class imbalance (the smallest class is a sixth of the majority class). Reports were manually annotated by clinical experts and then revised by linguists. Annotation guidelines and training were provided for both rounds of annotation. Interannotator (dis)agreement, detailed for each type of entity, will be used to better assess the performance of automatic annotators. Automatic annotators will be expected to perform well in those cases where human annotators have strong agreement, and worse in cases that are difficult for human annotators to identify consistently.

Five different classes of entities are distinguished: *Finding*, *Anatomical Entity*, *Location*, *Measure*, *Degree*, *Type of Measure* and *Abbreviation*. Hedges are also identified, distinguishing *Negation*, *Uncertainty*, *Condition* and *Conditional Temporal*. Entities can be embedded within other entities of different types. Moreover, entities can be discontinuous, and can span over sentence boundaries. The entity type *Finding* is particularly challenging, as it presents great variability in its textual forms. It ranges from a single word to more than ten words in some cases, and comprising all kinds of phrases. However, this is also the most informative type of entity for the potential users of these annotations. Other challenging phenomena are the regular polysemy observed between *Anatomical entities* and *Locations*, and the irregular uses of *Abbreviations*. In the manual annotation process, we have found that human annotators differ more on those categories than on the others, thus we expect automatic annotators will also have difficulties to consistently classify those as well.

For the SpRadIE 2021 task, submissions will be evaluated with different metrics, including exact and lenient match. The lenient evaluation will be carried out using a Jaccard Index, similarly as used in the 2013 BioNLP shared task [1]:

$$J_{(ref,pred)} = \frac{overlap_{(ref,pred)}}{length_{ref} + length_{pred} - overlap_{(ref,pred)}}$$

It takes the length (offsets) of the annotated reference concept, the predicted concept, as well as the overlap between them. This index amounts to 1 in the case of perfect match and 0 if there is no overlap between reference and prediction.

The official evaluation measures for the task are Slot Error Rate (SER) [15] with the Jaccard index as primary metric for entity match, and F1 for classification of matching entities within each type of entity.

4 CLEF eHealth 2021 Information Retrieval Task

4.1 Preceding Efforts

In 2020, the CHS task of CLEF eHealth consisted of an extension of the 2018 task. The use case was similar to previous years: helping patients and their next-of-kins find relevant health information online. The topics were extracted from query logs from the Health on the Net website and were representative of real information needs. The organizers oversaw the generation of spoken queries for these topics, and transcription of these spoken queries. Participants could submit their runs to two subtasks: one adhoc IR subtask using the textual queries; one spoken IR subtask using the spoken queries or their transcriptions. In each subtask, the effectiveness of the participants systems were evaluated considering three dimensions of relevance: topical relevance, understandability, and credibility. Three teams took part in the challenge, and all of them submitted runs to the 2 subtasks. However, none of them adapted the IR models used for each subtask – only the input query changed (textual query or transcription). This tendency was also observed in the previous multilingual tasks (running from 2014 until 2018), where only a few teams went further than adding a translation layer before the IR pipeline. Given the workload necessary to record and transcribe the topics, the organizers have decided not to carry on this task that failed to bring together several communities, and in the end did not really address the challenge of varying input type for IR models.

A constant effort has been made in the task since 2014 to integrate relevance dimensions. This has led to many interesting publications in order to adapt IR models to these dimensions, as well as the evaluation framework itself. Since 2020, the credibility dimension has been considered too. Integrating a dimension that, in itself, is already challenging to define, assess, and measure, led to a variety of interesting and exciting research questions. The 2021 CHS tasks reflect these new challenges.

4.2 The Task in 2021: Consumer Health Search

The 2018 CLEF eHealth CHS document collection will be used in the 2021 IR task. This collection consists of Web pages acquired from Common Crawl,³ which is augmented with additional pages collected from a number of known reliable health Websites and other known unreliable health Websites [9]. The topics for 2021 are manually created by medical professionals from realistic scenarios. Participants are challenged in the 2021 Task with retrieving the relevant documents from the provided document collection. A number of distinct subtasks can be completed using the considered queries and the provided labeled dataset: *ad-hoc search*, *credibility assessment*, and *personalized search based on multi-dimensional relevance assessment*.

³ <https://commoncrawl.org/>.

Like in the 2020 IR task, the pool of documents to be assessed will be labelled with respect to three relevance dimensions: *topicality*, *understandability*, and *credibility*. The assessment guidelines will follow up on 2020 guidelines: assessors will be asked to assess if the documents are on the same topic as the query, how readable/understandable the document is to a layperson, and how credible it is. Credibility has been introduced in the 2020 IR task. When assessing the credibility of online information, we consider credibility as an objective characteristic of an information item (either it is true, false, or partially true/false) [25], which is subjectively perceived by individuals [18]. Hence, the assessors are required to consider distinct aspects related to [24]: the *source* that disseminates information (e.g., its *trustworthiness* [3]), some characteristics associated with the *message* diffused (e.g., syntactic, semantic, and stylistic aspects [17]), and some *social aspects* if the information is disseminated through a virtual community (e.g., to be part of an *echo chamber* [2]).

The official evaluation measures include classic IR measures such as Binary Preference, Mean Reciprocal Rank, or Normalized Discounted Cumulative Gain @ 1–10, measuring how well systems retrieve relevant documents at low ranks (which is in line with the CHS use case). In order to measure how well systems can adapt the retrieved content to the consumers knowledge, understandability and credibility Rank-biased Precision will also be considered as official metrics. For the credibility assessment subtask, reference will be made to measures such as Accuracy and F-measure to establish the goodness of the classification between credible information or not.

5 A Vision for CLEF eHealth Beyond 2021

The general purpose of our lab throughout the years, as its 2021 IE and IR tasks demonstrate, has been to assist laypeople in finding and understanding health information in order to make enlightened decisions. Breaking language barriers has been our priority over the years, and this will continue in our multilingual tasks. Each year of the labs has enabled the identification of difficulties and challenges in IE, IM, and IR which have shaped our tasks. For example, our IR tasks have considered multilingual, contextualized, spoken queries, and query variants. However, further exploration of query construction, search scenario definition, aiming at a better understanding and management of CHS are still needed. The task will also further explore relevance dimensions, and work toward a better assessment of understandability and credibility, as well as methods to take these dimensions into consideration. Moreover, by better defining the search scenarios, the topics, and considering a document relevance in all its various aspects, the task will progress towards personalized and effective health search engines. As lab organizers, our purpose is to increase the impact and the value of the resources, methods and the community built by CLEF eHealth. Examining the quality and stability of the lab contributions will help the CLEF eHealth series to better understand where it should be improved and how. As future work, we intend continuing our analysis of the influence of the CLEF

eHealth evaluation series from the perspectives of publications and data/software releases [6, 20, 21].

Acknowledgements. The lab has been supported in part by the CLEF Initiative and the Our Health in Our Hands (OHIOH) initiative of The Australian National University (ANU). OHIOH is a strategic initiative of The ANU which aims to transform healthcare by developing new personalised health technologies and solutions in collaboration with patients, clinicians, and healthcare providers. The lab has been supported in part by the bi-lateral Kodicare (Knowledge Delta based improvement and continuous evaluation of retrieval engines) project funded by the french ANR (ANR-19-CE23-0029) and Austrian FWF.

References

1. Bossy, R., Golik, W., Ratkovic, Z., Bessières, P., Nédellec, C.: BioNLP shared task 2013 - an overview of the bacteria biotope task. In: Proceedings of the BioNLP Shared Task 2013 Workshop. pp. 161–169. Association for Computational Linguistics, Sofia, Bulgaria, August 2013. <https://www.aclweb.org/anthology/W13-2024>
2. Bruns, A.: Echo chamber? What echo chamber? Reviewing the evidence (2017)
3. Ceravolo, P., Damiani, E., Viviani, M.: Adding a trust layer to semantic web metadata. In: Herrera-Viedma, E., Pasi, G., Crestani, F. (eds.) *Soft Computing in Web Information Retrieval. Studies in Fuzziness and Soft Computing*, vol. 197, pp. 87–104. Springer, Heidelberg (2006). https://doi.org/10.1007/3-540-31590-X_5
4. Demner-Fushman, D., Elhadad, N.: Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. *Yearb. Med. Inform.* **1**, 224–233 (2016)
5. Filannino, M., Uzuner, Ö.: Advancing the state of the art in clinical natural language processing through shared tasks. *Yearb. Med. Inform.* **27**(01), 184–192 (2018)
6. Goeuriot, L., et al.: An analysis of evaluation campaigns in ad-hoc medical information retrieval: CLEF eHealth 2013 and 2014. *Inf. Retr. J.* **21**, 507–540 (2018). <https://doi.org/10.1007/s10791-018-9331-4>
7. Goeuriot, L., et al.: Overview of the CLEF eHealth evaluation lab 2015. In: Mothe, J., et al. (eds.) *CLEF 2015. LNCS*, vol. 9283, pp. 429–443. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24027-5_44
8. Goeuriot, L., et al.: CLEF 2017 eHealth evaluation lab overview. In: Jones, G.J.F., et al. (eds.) *CLEF 2017. LNCS*, vol. 10456, pp. 291–303. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_26
9. Goeuriot, L., Liu, Z., Pasi, G., Saez, G.G., Viviani, M., Xu, C.: Overview of the CLEF eHealth 2020 task 2: consumer health search with ad hoc and spoken queries. In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings* (2020)
10. Goeuriot, L., et al.: Overview of the CLEF eHealth evaluation lab 2020. In: Aramatzis, A., et al. (eds.) *CLEF 2020. LNCS*, vol. 12260, pp. 255–271. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_19
11. Huang, C.C., Lu, Z.: Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief. Bioinform.* **17**(1), 132–144 (2016)
12. Kelly, L., et al.: Overview of the CLEF eHealth evaluation lab 2016. In: Fuhr, N., et al. (eds.) *CLEF 2016. LNCS*, vol. 9822, pp. 255–266. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_24

13. Kelly, L., et al.: Overview of the ShARe/CLEF eHealth evaluation lab 2014. In: Kanoulas, E., et al. (eds.) CLEF 2014. LNCS, vol. 8685, pp. 172–191. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11382-1_17
14. 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
15. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop, pp. 249–252 (1999)
16. McAllister, M., Dunn, G., Payne, K., Davies, L., Todd, C.: Patient empowerment: the need to consider it as a measurable patient-reported outcome for chronic conditions. *BMC Health Serv. Res.* **12**, 157 (2012)
17. Mukherjee, S., Weikum, G., Danescu-Niculescu-Mizil, C.: People on drugs: credibility of user statements in health communities. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 65–74 (2014)
18. Self, C.C.: *Credibility*. In: *An Integrated Approach to Communication Theory and Research*, pp. 449–470. Routledge (2014)
19. Suominen, H.: CLEFeHealth2012 – the CLEF 2012 workshop on cross-language evaluation of methods, applications, and resources for eHealth document analysis. In: Forner, P., Karlgren, J., Womser-Hacker, C., Ferro, N. (eds.) CLEF 2012 Working Notes. vol. 1178. CEUR Workshop Proceedings (CEUR-WS.org) (2012)
20. Suominen, H., Kelly, L., Goeuriot, L.: Scholarly influence of the conference and labs of the evaluation forum eHealth initiative: review and bibliometric study of the 2012 to 2017 outcomes. *JMIR Res. Prot.* **7**(7), e10961 (2018)
21. Suominen, H., Kelly, L., Goeuriot, L.: The scholarly impact and strategic intent of CLEF eHealth labs from 2012 to 2017. *Information Retrieval Evaluation in a Changing World*. TIRS, vol. 41, pp. 333–363. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22948-1_14
22. Suominen, H., et al.: Overview of the CLEF eHealth evaluation lab 2018. In: Bellet, P., et al. (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. LNCS, vol. 11018, pp. 286–301. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_26
23. Suominen, H., et al.: Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 212–231. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40802-1_24
24. Viviani, M., Pasi, G.: A multi-criteria decision making approach for the assessment of information credibility in social media. In: Petrosino, A., Loia, V., Pedrycz, W. (eds.) WILF 2016. LNCS (LNAI), vol. 10147, pp. 197–207. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52962-2_17
25. Viviani, M., Pasi, G.: *Credibility in social media: opinions, news, and health information—a survey*. *Wiley Interdisc. Rev. Data Min. Knowl. Discov.* **7**(5), e1209 (2017)