CLEF eHealth Evaluation Lab 2020

Hanna Suominen[1,2,3], Liadh Kelly[4], Lorraine Goeuriot[5(✉)], and Martin Krallinger[6]

[1] The Australian National University, Acton, ACT 2601, Australia
hanna.suominen@anu.edu.au
[2] Data61/Commonwealth Scientific and Industrial Research Organisation, Acton, ACT, Australia
[3] University of Turku, Turku, Finland
[4] Maynooth University, Co., Kildare, Ireland
liadh.kelly@mu.ie
[5] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
lorraine.Goeuriot@imag.fr
[6] Barcelona Supercomputing Center (BSC-CNS), 08034 Barcelona, Spain
martin.krallinger@bsc.es
https://researchers.anu.edu.au/researchers/suominen-h

Abstract. Laypeople's increasing difficulties to retrieve and digest valid and relevant information in their preferred language to make health-centred decisions has motivated CLEF eHealth to organize yearly labs since 2012. These 20 evaluation tasks on *Information Extraction* (IE), management, and *Information Retrieval* (IR) in 2013–2019 have been popular—as demonstrated by the large number of team registrations, submissions, papers, their included authors, and citations (748, 177, 184, 741, and 1299, respectively, up to and including 2018)—and achieved statistically significant improvements in the processing quality. In 2020, CLEF eHealth is calling for participants to contribute to the following two tasks: The 2020 Task 1 on IE focuses on term coding for clinical textual data in Spanish. The terms considered are extracted from clinical case records and they are mapped onto the Spanish version of the International Classification of Diseases, the 10th Revision, including also textual evidence spans for the clinical codes. The 2020 Task 2 is a novel extension of the most popular and established task in CLEF eHealth on CHS. This IR task uses the representative web corpus used in the 2018 challenge, but now also spoken queries, as well as textual transcripts of these queries, are offered to the participants. The task is structured into a number of optional subtasks, covering ad-hoc search using the spoken queries, textual transcripts of the spoken queries, or provided automatic speech-to-text conversions of the spoken queries. In this paper we describe the evolution of CLEF eHealth and this year's tasks. The

HS, LK & LG co-chair the CLEF eHealth lab and contributed equally to this paper. MK leads the 2020 IE task supported by the Spanish Plan TL while HS & LG lead the 2020 IR task. We gratefully acknowledge the contribution of the people and organizations involved in CLEF eHealth in 2012–2020. We thank the CLEF Initiative, Dr Benjamin Lecouteux (Université Grenoble Alpes), Dr João Palotti (Qatar Computing Research Institute), and Dr Guido Zuccon (University of Queensland).

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 587–594, 2020.
https://doi.org/10.1007/978-3-030-45442-5_76

substantial community interest in the tasks and their resources has led to CLEF eHealth maturing as a primary venue for all interdisciplinary actors of the ecosystem for producing, processing, and consuming electronic health information.

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

1 Introduction

Improving the legibility of *Electronic Health Record* (EHR) can contribute to patients' right to be informed about their health and health care. The requirement to ensure that patients can understand their own privacy-sensitive, official health information in their EHR are stipulated by policies and laws. For example, the *Declaration on the Promotion of Patients' Rights in Europe* by *World Health Organization* (WHO) from 1994 obligates health care workers to communicate in a way appropriate to each patient's capacity for understanding and give each patient a legible written summary of these care guidelines. This patient education must capture the patient's health status, condition, diagnosis, and prognosis, together with the proposed and alternative treatment/non-treatment with risks, benefits, and progress. Patients' better abilities to understand their own EHR empowers them to take part in the related health/care judgment, leading to their increased independence from health care providers, better health/care decisions, and decreased health care costs [11]. Improving patients' ability to digest this content could mean enriching the EHR-text with hyperlinks to term definitions, paraphrasing, care guidelines, and further supportive information on patient-friendly and reliable websites, and the enabling methods for such reading aids can also release health care workers' time from EHR-writing to, for example, longer patient-education discussions [14].

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 health care experts' information needs only [1,2,6]. Such exception, the *CLEF eHealth Evaluation-lab and Lab-workshop Series*[1] has been organized every year since 2012 as part of the *Conference and Labs of the Evaluation Forum* (CLEF) [4,5,8–10,13,16,17]. In 2012, the inaugural scientific CLEF workshop took place, and from 2013–2019 this annual workshop has been supplemented with a lead-up evaluation lab, consisting of, on average, three shared tasks each year (Fig. 1). Although the tasks have been centered around the patients and their families' needs in accessing and understanding eHealth information, also *Automatic Speech Recognition* (ASR) and IE to aid clinicians in IM were considered in 2015–2016 and in 2017–2019, tasks on technology assisted reviews to support health scientists and health care policymakers' information access were organized.

[1] http://clef-ehealth.org/.

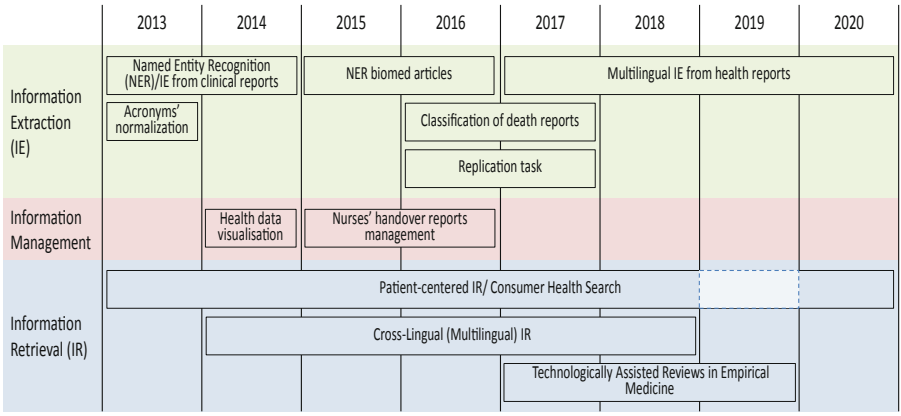

Fig. 1. Timeline of the CLEF eHealth tasks in 2013–2020

This paper presents first an overview of CLEF eHealth lab series from 2012 to 2019 and introduces its 2020 evaluation tasks. Then, it concludes by presenting our vision for CLEF eHealth beyond 2020.

2 CLEF eHealth Contributions and Growth in 2012–2019

CLEF eHealth tasks offered yearly from 2013 have brought together researchers working on related information access topics, provided them with resources to work with and validate their outcomes, and accelerated pathways from scientific ideas to societal impact. In 2013, 2014, 2015, 2016, 2017, 2018, and 2019 as many as 170, 220, 100, 116, 67, 70, and 67 teams have registered their expression of interest in the CLEF eHealth tasks, respectively, and the number of teams proceeding to the task submission stage has been 53, 24, 20, 20, 32, 28, and 9, respectively [4,5,8–10,16,17].[2]

According to our analysis of the impact of CLEF eHealth labs up to 2017 [15], the submitting teams have achieved statistically significant improvements in the processing quality in at least 1 out of the top-3 methods submitted to the following eight tasks:[3]

1. 2013 Task 1a on English disorder identification with $F1$ and random shuffling ($P = .009$) as the performance measure and statistical significance test, respectively, on independent sets of 200 and 100 annotated EHRs

[2] "Expressing an interest" for a CLEF task consists of filling in a form on the CLEF conference website with contact information, and tick boxes corresponding to the labs of interest. This is usually done several months before run submission, which explains the drop in the numbers.

[3] Some tasks have not presented a method ranking and/or statistical significance evaluation of this kind in the lab/task overviews. In other words, different kinds of improvements have been obtained in other tasks as well.

for training and testing. The top-3 submissions had on the test set $F1$ of 0.750, 0.737, and 0.707, whilst to illustrate the task difficulty, typically using a simple baseline method by the task organizers, the worst $F1$ was 0.428.

2. 2013 Task 1b on English disorder normalization with respect to the *Systematized Nomenclature of Medicine—Clinical Terms* (SNOMED-CT) codes with the accuracy and random shuffling ($P = .009$) as the performance measure and statistical significance test, respectively, on independent sets of 200 and 100 annotated EHRs for training and testing. The top-3 submissions had on the test set the accuracy of 0.589, 0.587, and 0.546, while the worst accuracy was 0.006.
3. 2013 Task 2 on English shorthand extension with respect to the *Unified Medical Language System* (UMLS) codes with the accuracy and random shuffling ($P = .009$) as the performance measure and statistical significance test, respectively, on independent sets of 200 and 100 annotated EHRs for training and testing. The top-3 submissions had on the test set the accuracy of 0.719, 0.683, and 0.664, while the worst accuracy was 0.426.
4. 2013 Task 3 on English IR with the *Precision at 10* (P@10) and Wilcoxon test ($P = .04$) as the performance measure and statistical significance test, respectively, on 50 test queries and the matching result set. The top-3 submissions had P@10 of 0.518, 0.504, and 0.484, while the worst P@10 was 0.006.
5. 2015 Task 1 on English nursing handover ASR with the error and Wilcoxon test ($P = .04$) as the performance measure and statistical significance test, respectively, on independent sets of 100 and 100 annotated EHRs for training and testing. The top-3 submissions had on the test set the error of 0.385, 0.523, and 0.528, while the worst error was 0.954.
6. 2016 Task 1 on English nursing handover IE with $F1$ and Wilcoxon test ($P = .04$) as the performance measure and statistical significance test, respectively, on independent sets of 200 and 100 annotated EHRs for training and testing. The top-3 submissions had on the test set $F1$ of 0.382, 0.374, and 0.345, while the worst $F1$ was 0.000.

7 & 8. 2016 Task 2 on French IE, with entity recognition and cause of death subtasks. Both subtasks used $F1$ and t-test ($P \leq .001$) as the performance measure and statistical significance test, respectively, on 1,668 titles of scientific articles and 6 full text drug monographs for training and testing. The corpus was split evenly between training data supplied to the participants at the beginning of the lab, and an unseen test set used to evaluate participants' systems. In the entity recognition subtask, the top-3 submissions had on the test set $F1$ of 0.749, 0.702, and 0.699, while the worst $F1$ was 0.126. In the cause of death subtasks, the top-3 submissions had on the test set $F1$ of 0.848, 0.844, and 0.752, while the worst $F1$ was 0.554.

The 2012–2017 contributions have been reported by October 2018 in 184 papers for the 741 included authors from 33 countries across the world, and the

papers have attracted nearly 1,300 citations, creating h-index and i10-index of 18 and 35, respectively, on Google Scholar [14]. CLEF eHealth 2012 lab workshop has resulted in 16 papers and each year CLEF eHealth 2013–2017 evaluation labs have increased this number from 31 to 35. In accordance with the CLEF eHealth mission to foster teamwork, the number of co-authors per paper has been from 1 to 15 (the mean and standard deviation of 4 and 3, respectively). In about a quarter of the papers, this co-authoring collaboration has been international, and sometimes even intercontinental.

This substantial community interest in the CLEF eHealth tasks and their resources has led to the evaluation campaign maturing and establishing its presence over the years. In 2020, CLEF eHealth is one of the primary venues for all interdisciplinary actors of the ecosystem for producing, processing, and consuming eHealth information [1,2,6]. Its niche is addressing health information needs of laypeople—and not health care experts only—in retrieving and digesting valid and relevant eHealth information to make health-centered decisions.

3 CLEF eHealth 2020 Information Extraction and Retrieval Tasks

The 2020 CLEF eHealth Task 1 on IE, called *CodiEsp* supported by the *Spanish National Plan for the Advancement of Language Technology* (Plan TL), builds upon the five previous editions of the task in 2015–2019 [4,5,8,10,16] that have already addressed the analysis of biomedical text in English, French, Hungarian, Italian, and German. This year, the CodiEsp task, will focus on the *International Classification of Diseases, the 10th Revision* (ICD10) coding for clinical case data in Spanish using the *Spanish version of ICD10* (CIE10).

The CodiEsp task will explore the automatic assignment of CIE10 codes to clinical case documents in Spanish, namely of two categories: procedure and diagnosis (known as 'Procedimiento' and 'Diagnostico' in Spanish). The following three subtasks will be posed: (1) *CodiEsp Diagnosis Coding* will consist of automatically assigning diagnosis codes to clinical cases in Spanish. (2) *CodiEsp Procedure Coding* will focus on assigning procedure codes to clinical cases in Spanish. (3) *CodiEsp Explainable Artificial Intelligence* (AI) will evaluate the explainability/interpretability of the proposed systems, as well as their performance by requesting to return the text spans supporting the assignment of CIE10 codes.

The CodiEsp corpus used for this task consists of a total of 1,000 clinical cases that were manually annotated by clinical coding professionals with clinical procedure and diagnosis codes from the Spanish version of ICD10 together with the actual minimal text spans supporting the clinical codes. The CodiEsp corpus has around 18,000 sentences, and contains about 411,000 words and 19,000 clinical codes. Code annotations will be released in a separate file together with the respective document code and the span of text that leads to the codification (the evidence). Additional data resources including medical literature abstracts in Spanish indexed with ICD10 codes, linguistic resources, gazetteers, and a

background set of medical texts in Spanish will also be released to complement the CodiEsp corpus, together with annotation guidelines and details.

For the CodiEsp Diagnosis and Procedure Coding subtasks, participants will submit their coding predictions returning ranked results. For every document, a list of possible codes will be submitted, ordered by confidence or relevance. Since these subtasks are designed to be ranking competitions, they will be evaluated on a standard ranking metric: Mean Average Precision. For the CodiEsp Explainable AI subtask, explainability of the systems will be considered, in addition to their performance on the test set. Systems have to provide textual evidence from the clinical case documents that supports the code assignment and thus can be interpreted by humans. This automatically returned evidence will be evaluated against manually annotated text spans. True positive evidence texts are those that consist in a sub-match of the manual annotations. $F1$ will be used as the primary evaluation metric.

The 2020 CLEF eHealth Task 2 on IR builds on the tasks that have run at CLEF eHealth since its inception in 2012. This *Consumer Health Search* (CHS) task follows a standard IR shared challenge paradigm from the perspective that it provides participants with a test collection consisting of a set of documents and a set of topics to develop IR techniques for. Runs submitted by participants are pooled, and manual relevance assessments conducted. Performance measures are then returned to participants.

In the 2017 CLEF eHealth CHS task, similarly to 2016, we used the ClueWeb 12 B13[4] document collection [12,18]. This consisted of a collection of 52.3 million medically related web pages. Given the scale of this document collection participants reported that it was difficult to store and manipulate the document collection. In response, the 2018 CHS task introduced a new document collection, named *clefehealth2018*. This collection consists of over 5 million medical webpages from selected domains acquired from the CommonCrawl [7]. Given the positive feedback received for this document collection, it will be used again in the 2020 CHS task.

Historically the CLEF eHealth IR task has released text queries representative of layperson information needs in various scenarios. In recent years, query variations issued by multiple laypeople for the same information need have been offered. In this year's task we extend this to spoken queries. These spoken queries are generated by 6 individuals using the information needs derived for the 2018 challenge [7]. We also provide textual transcripts of these spoken queries and ASR translations.

Given the query variants for an information need, participants are challenged in the 2020 task with retrieving the relevant documents from the provided document collection. This is divided into a number of subtasks which can be completed using the spoken queries or their textual transcripts by hand or ASR. Similar to the 2018 CHS tasks, subtasks explored this year are: ad-hoc/personalized search, query variations, and search intent with Binary Preference, Mean Reciprocal Rank, Normalized Discounted Cumulative Gain@1–10,

[4] http://lemurproject.org/clueweb12/index.php.

and (Understandability-biased) Rank-biased Precision as subtask-dependent evaluation measures. Participants can submit multiple runs for each subtask.

4 A Vision for CLEF eHealth Beyond 2020

The general purpose of CLEF eHealth throughout the years, as its 2020 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. Text has been our major media of interest, but speech has been, and continues to be, included in tasks as a major new way of interacting with systems. 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, popular IR tasks have considered multilingual, contextualized, and/or spoken queries and query variants. However, further exploration of query construction, aiming at a better understanding of CHS are still needed. The task into the future will also further explore relevance dimensions, and work toward a better assessment of readability and reliability, as well as methods to take these dimensions into consideration. 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 analyses of the influence of the CLEF eHealth evaluation series from the perspectives of publications and data/software releases [3,14,15].

References

1. 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)
2. 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)
3. Goeuriot, L., et al.: An analysis of evaluation campaigns in ad-hoc medical information retrieval: CLEF eHealth 2013 and 2014. Inf. Retrieval J. **21**(6), 507–540 (2018). https://doi.org/10.1007/s10791-018-9331-4
4. 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
5. 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
6. Huang, C.C., Lu, Z.: Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief. Bioinf. **17**(1), 132–144 (2016)

7. Jimmy, Zuccon, G., Palotti, J.: Overview of the CLEF 2018 consumer health search task. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2018)
8. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: 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
9. 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
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. 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)
12. Palotti, J., et al.: CLEF 2017 task overview: the IR task at the eHealth evaluation lab. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2017)
13. 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)
14. 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. Protoc. **7**(7), e10961 (2018)
15. Suominen, H., Kelly, L., Goeuriot, L.: The scholarly impact and strategic intent of CLEF eHealth labs from 2012 to 2017. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF, pp. 333–363. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22948-1_14
16. Suominen, H., et al.: Overview of the CLEF eHealth evaluation lab 2018. In: Bellot, P., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 286–301. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_26
17. 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
18. Zuccon, G., et al.: The IR task at the CLEF eHealth evaluation lab 2016: user-centred health information retrieval. In: CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS, September 2016