CLEF eHealth 2019 Evaluation Lab

Liadh Kelly[1], Lorraine Goeuriot[2], Hanna Suominen[3], Mariana Neves[4], Evangelos Kanoulas[5], Rene Spijker[6], Leif Azzopardi[7], Dan Li[5], Jimmy[8], João Palotti[9,10], and Guido Zuccon[8] *

[1] Maynooth University, Ireland liadh.kelly@mu.ie
[2] Univ. Grenoble Alpes, France Lorraine.Goeuriot@imag.fr
[3] The Australian National University (ANU), Data61/Commonwealth Scientific and Industrial Research Organisation (CSIRO), University of Canberra, and University of Turku, Canberra, ACT, Australia, hanna.suominen@anu.edu.au
[4] German Federal Institute for Risk Assessment (BfR), Germany, mariana.lara-neves@bfr.bund.de
[5] Informatics Institute, University of Amsterdam, Netherlands, E.Kanoulas@uva.nl
[6] Cochrane Netherlands and UMC Utrecht, Julius Center for Health Sciences and Primary Care, Netherlands, R.Spijker-2@umcutrecht.nl
[7] Computer and Information Sciences, University of Strathclyde, UK leif.azzopardi@strath.ac.uk
[8] Queensland University of Technology, Brisbane, QLD, Australia, g.zuccon@qut.edu.au
[9] Vienna University of Technology, Austria, palotti@ifs.tuwien.ac.at
[10] Qatar Computing Research Institute, Doha, Qatar jpalotti@hbku.edu.qa

Abstract. Since 2012 CLEF eHealth has focused on evaluation resource building efforts around the easing and support of patients, their next-of-kins, clinical staff, and health scientists in understanding, accessing, and authoring eHealth information in a multilingual setting. This year's lab offers three tasks: Task 1 on multilingual information extraction; Task 2 on technology assisted reviews in empirical medicine; and Task 3 on consumer health search in mono- and multilingual settings. Herein, we describe the CLEF eHealth evaluation series to-date and then present the 2019 tasks, evaluation methodology, and resources.

Keywords: eHealth · Medical Informatics · Information Extraction · Information Storage and Retrieval · Information Management · Systematic Reviews.

1 Introduction

In today's information overloaded society it is increasingly difficult to retrieve and digest valid and relevant information to make health-centered decisions. *Electronic Health* (eHealth) content is becoming available in a variety of forms ranging from patient records and medical dossiers, scientific publications, and

* LK, LG & HS co-chair the CLEF eHealth lab and contributed equally to this paper. MN, EK & RS & LA & DL, and J & JP & GZ lead 2019 lab Tasks 1–3, respectively.

health-related websites to medical-related topics shared across social networks. Laypeople, clinicians, and policy-makers need to easily retrieve, and make sense of this content to support their decision making.

Information retrieval (IR) systems have been commonly used as a means to access health information available online. However, the reliability, quality, and suitability of the information for the target audience varies greatly while high recall or coverage, that is finding all relevant information about a topic, is often as important as high precision, if not more. Furthermore, the information seekers in the health domain also experience difficulties in expressing their information needs as search queries.

CLEF eHealth[11], established as a lab workshop in 2012 as part of the Conference and Labs of the Evaluation Forum (CLEF), has offered since 2013 evaluation labs in the fields of layperson and professional health information extraction, management, and retrieval with the aims of bringing together researchers working on related information access topics and providing them with datasets to work with and validate the outcomes. More specifically, these labs and their subsequent workshops target

1. developing processing methods and resources (e.g., dictionaries, abbreviation mappings, and data with model solutions for method development and evaluation) in a multilingual setting to enrich difficult-to-understand eHealth texts and provide personalized reliable access to medical information, and provide valuable documentation;
2. developing an evaluation setting and releasing evaluation results for these methods and resources;
3. contributing to the participants and organizers' professional networks and interaction with all interdisciplinary actors of the ecosystem for producing, processing, and consuming eHealth information.

In this paper we overview the CLEF eHealth evaluation lab series to-date [20,11,5,10,6,19] and present this year's evaluation lab challenges.

2 CLEF eHealth — Past and Future

In 2012, the CLEF eHealth workshop was organized to prepare for evaluation labs. Its outcome was the identification of the need for an evaluation lab focusing on patient-centric health language processing. The subsequent CLEF eHealth tasks offered from 2013–2018 can be broadly categorized as information extraction, management and retrieval focused. In 2019 we offer information extraction and retrieval challenges (described in Section 3). Here we describe the growth path of these challenges.

2.1 Information Extraction from Clinical Text

The CLEF eHealth tasks on *information extraction* (IE) began in 2013 by considering English only but evolved by 2018 to considering more and more languages.

[11] `https://sites.google.com/site/clefehealth/` (last accessed on 19 October 2018)

In 2013, the focus of the information extraction task was on named entity recognition, normalization of disorders, and normalization of acronyms/abbreviations. In 2014, we extended the challenge with a focus on disorder attribute identification and normalization from clinical text. In 2015 and 2016, we supplemented the tasks by aiming to release nurses' time from documentation to patient communication by considering first clinical speech recognition to capture the verbal shift-change handover and then information extraction to pre-fill a handover form from the speech recognized text by automatically identifying relevant text-snippets for each slot of the form.

To continue this evolution from a widely studied corpus type (written in English) towards a larger variety of corpora by considering spoken English in the handover tasks, we introduced a multilingual challenge in 2015, which considered information extraction from French clinical texts. This challenge was grown in the subsequent years [12,13]. In last year's lab [14] we began the evolution of the multilingual element task towards the inclusion of other European languages, such as Hungarian and Italian. In this year's task we continue this evolution.

Our goal in the coming years is to offer an information extraction task using comparable corpora in several languages in order to challenge participants with the issue of language adaptation and to encourage the development of systems that are able to address a multilingual setting or can easily be tuned to specialize to specific languages.

2.2 Information Retrieval and Personalization

In 2013 and 2014 the focus of the information retrieval task was on evaluating the effectiveness of search engines to support people when searching for information about known conditions, for example, to answer queries like "thrombocytopenia treatment corticosteroids lengt", with multilingual queries added in the 2014 challenge [2,4,3]. This task aimed to model the scenario of a patient being discharged from hospital and wanting to seek more information about diagnosed conditions or prescribed treatments.

In 2015 the information retrieval task changed to focus on studying the effectiveness of search engines to support individuals' queries issued for self-diagnosis purposes, and again offered a multilingual queries challenge [15]. In addition, we began adding personalization elements to the challenge on an incremental basis by assessing the readability of information and taking this into account in the evaluation framework.

This individualized information retrieval approach was continued in the 2016 and 2017 labs [21,16] and we also introduced gradual shifts from an ad-hoc search paradigm (that of a single query and a single document ranking) to a session based search paradigm. Along these lines we also revised how relevance is measured for evaluation purposes, taking into account instead whole-of-session usefulness. In 2018 [7] we continued this evolution, and introduced query intent elements.

Our next goals are as follows: (1) to further progress the evaluation methodology for session based and query intent search paradigms that we laid the foun-

dations for in the previous years, and (2) to introduce spoken query elements and supporting evaluation methodology.

2.3 Technology Assisted Reviews

The *Technology Assisted Reviews* (TARs) task, organized for the first time in 2017 and continued in 2018 [8,9], was a high-recall IR task in English that aimed at evaluating search algorithms that seek to identify all studies relevant for conducting a systematic review in empirical medicine. The task had a focus on *Diagnostic Test Accuracy* (DTA) reviews. The typical process of searching for scientific publications to conduct a systematic review consists of three stages:

1. specifying a number of inclusion criteria that characterize the articles relevant to the review and constructing a complex Boolean Query to express them,
2. screening the abstracts and titles that result from the Boolean query, and
3. screening the full documents that passed the Abstract and Title Screening.

The 2017 task focused on the second stage of the process, that is, Abstract and Title Screening. Building on this the 2018 task focused on the first stage (*subtask 1*) and second stage (*subtask 2*) of the process, that is, Boolean Search and Abstract and Title Screening. The task built two benchmark collections and implemented a number of evaluation metrics to automatically assess the quality of methods on these collection, all of which have been made available at `https://github.com/CLEF-TAR`.

Directions to take to further build the task in the coming years include the following: (1) developing metrics to evaluate systems on the ranking and thresholding tasks, (2) increasing the labelled data offered with the challenge, and (3) providing an infrastructure to support running of participants' algorithms in house, thus allowing for use of full text articles and live, iterative active learning technique development.

3 CLEF eHealth 2019 Tasks

Continuing the CLEF eHealth growth path from 2013–2018, in 2019 CLEF eHealth offers three tasks. Specifically, Task 1 on Multilingual Information Extraction, Task 2 on TARs in Empirical Medicine, and Task 3 on Consumer Health Search.

3.1 Task 1. Multilingual Information Extraction

This task builds upon the previous CLEF eHealth IE tasks. This year's task continues to explore the automatic assignment of ICD-10 codes to health-related documents with the focus on the German language and on *Non-Technical Summaries* (NTSs) of animal experiments. Specifically, in 2019, participants are challenged with the semantic indexing of NTSs using codes from the German version of the *International Classification of Diseases* (ICD-10). The NTSs are

short summaries which are currently publicly available in the AnimalTestInfo database[12], as part of the approval procedure for animal experiments in Germany [1]. The database currently contains more than 8,000 NTSs, which have been manually indexed by domain experts, and that was used to generate a training dataset. The task can be treated as a named entity recognition and normalization task, but also as a text classification task. Only fully automated means are allowed, that is, human-in-the-loop approaches are not permitted.

3.2 Task 2. Technology Assisted Reviews in Empirical Medicine

This task builds on the TAR task first introduced in 2017. The task is a ranking and classification task (similar to the 2017 and 2018 version), and includes two subtasks: (1) No Boolean Query and (2) Title and Abstract Screening. For the former users are provided with a set of topics and parts of the systematic review protocol. The goal of the participants is to rank PubMed abstracts and titles and provide a threshold on the ranking. For the latter users are provided with a set of topics, the original Boolean query used by the researchers that conducted the systematic review, and the results of that query. The goal of the participants is to rank PubMed abstracts and titles and provide a threshold on the ranking.

3.3 Task 3. Consumer Health Search

This task builds on the CLEF eHealth information retrieval tasks that have ran since the onset of CLEF eHealth. The main components of the *Consumer Health Search* (CHS) task are the document collection, the set of topics and the system evaluation. This year's challenge uses the new document collection introduced in last year's challenge, consisting of over 5 million Web pages. It is a compilation of Web pages of selected domains acquired from the CommonCrawl[13]. User stories for query (and query variant) generation are created using the discharge summaries and forum posts we used in previous years of the task. For the first time, queries are also offered as spoken queries, with automatic speech-to-text transcripts provided. The challenge is structured into 5 subtasks, specifically: ad-hoc search, personalization search, query variations, multilingual search and search intent.

4 CLEF eHealth Contributions

In its seven years of existence, the CLEF eHealth series has offered a recurring contribution to the creation and dissemination of text analytics resources, methods, test collections, and evaluation benchmarks in order to ease and support patients, their next-of-kins, clinical staff, and health scientists in understanding, accessing, and authoring eHealth information in a multilingual setting. In

[12] `https://www.animaltestinfo.de/` (last accessed on 18 October 2018)
[13] `http://commoncrawl.org/` (last accessed on 19 October 2018)

2012–2017 alone it has attracted over 700 teams to register their interest in its 15 tasks, leading to 130 task submissions, 180 papers, and their 1,300 citations for the 741 included authors from 33 countries across the world [18].

The annual workshops and evaluation labs offered by CLEF eHealth have matured and established their presence over the years. In total, 70 unique teams registered their interest and 28 teams took part in the 2018 tasks (14 in Task 1, 7 in Task 2 and 7 in Task 3). In comparison, in 2017, 2016, 2015, 2014, and 2013, the number of team registrations was 67, 116, 100, 220, and 175, respectively and the number of participating teams was 32, 20, 20, 24, and 53 [20,11,5,10,6,19].

Given the significance of the tasks, all problem specifications, test collections, and text analytics resources associated with the lab have been made available to the wider research community through our CLEF eHealth website[14].

5 Conclusion

In this paper, we have provided an overview of the CLEF eHealth evaluation lab series and presented the 2019 lab tasks. The CLEF eHealth workshop series was established in 2012 as a scientific workshop with an aim of establishing an evaluation lab [17]. This ambition was realized in the CLEF eHealth evaluation lab, which has ran since 2013. This annual lab offers shared tasks in the eHealth space each year in the domain of medical information retrieval, management and extraction [20,11,5,10,6,19].

The CLEF eHealth 2019 lab offers three shared tasks: Task 1 on multilingual information extraction to extend the 2018 task on French, Hungarian, and Italian corpora to German; Task 2 on technologically assisted reviews in empirical medicine building on the 2018 task in English; and Task 3 on patient-centered IR in mono- and multilingual settings that builds on the 2013–18 IR tasks. Test collections generated by each of the three CLEF eHealth 2019 tasks offer a specific task definition, implemented in a dataset distributed together with an implementation of relevant evaluation metrics to allow for direct comparability of the results reported by systems evaluated on the collections.

Acknowledgements

We gratefully acknowledge the people involved in the CLEF eHealth labs as participants or organizers. We also acknowledge the many organizations that have supported CLEF eHealth labs since 2012. The CLEF eHealth 2019 evaluation lab is supported in part by (in alphabetical order) the CLEF Initiative, and Data61/CSIRO.

[14] `https://sites.google.com/site/clefehealth/datasets` (last accessed on 18 October 2018)

References

1. Bert, B., Dörendahl, A., Leich, N., Vietze, J., Steinfath, M., Chmielewska, J., Hensel, A., Grune, B., Schönfelder, G.: Rethinking 3r strategies: Digging deeper into animaltestinfo promotes transparency in in vivo biomedical research. PLOS Biology **15**(12), 1–20 (12 2017). https://doi.org/10.1371/journal.pbio.2003217, `https://doi.org/10.1371/journal.pbio.2003217`
2. Goeuriot, L., Jones, G.J., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salantera, S., Suominen, H., Zuccon, G.: ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. CLEF 2013 Online Working Notes **8138** (2013)
3. Goeuriot, L., Jones, G.J., Kelly, L., Leveling, J., Lupu, M., Palotti, J., Zuccon, G.: An Analysis of Evaluation Campaigns in ad-hoc Medical Information Retrieval: CLEF eHealth 2013 and 2014. Springer Information Retrieval Journal (2018)
4. Goeuriot, L., Kelly, L., Lee, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Gareth J.F. Jones, H.M.: ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In: CLEF 2014 Evaluation Labs and Workshop: Online Working Notes. Sheffield, UK (2014)
5. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the clef ehealth evaluation lab 2015. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Springer Berlin Heidelberg (2015)
6. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: Clef 2017 ehealth evaluation lab overview. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 291–303. Springer Berlin Heidelberg (2017)
7. Jimmy, ., Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L.: 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. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2017 technologically assisted reviews in empirical medicine overview. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2017)
9. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2018 technologically assisted reviews in empirical medicine overview. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2018)
10. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2016. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 255–266. Springer Berlin Heidelberg (2016)
11. Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W., Martinez, D., Zuccon, G., Palotti, J.: Overview of the ShARe/CLEF eHealth evaluation lab 2014. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 172–191. Springer Berlin Heidelberg (2014)
12. Névéol, A., Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Kelly, L., Goeuriot, L., Rey, G., Robert, A., Tannier, X., Zweigenbaum, P.: Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, `http://ceur-ws.org/Vol-1609/` (2016)

13. Névéol, A., Anderson, R.N., Cohen, K.B., Grouin, C., Lavergne, T., Rey, G., Robert, A., Zweigenbaum, P.: CLEF eHealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french. In: CLEF 2017 Online Working Notes. CEUR-WS (2017)
14. Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikn, L., Ramadier, L., Rey, G., Zweigenbaum, P.: CLEF eHealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. In: CLEF 2018 Online Working Notes. CEUR-WS (2018)
15. Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanburyn, A., Jones, G.J., Lupu, M., Pecina, P.: CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information about Medical Symptoms. In: CLEF 2015 Online Working Notes. CEUR-WS (2015)
16. Palotti, J., Zuccon, G., Jimmy, Pecina, P., Lupu, M., Goeuriot, L., Kelly, L., Hanbury, A.: 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)
17. Suominen, H.: In: Forner, P., Karlgren, J., Womser-Hacker, C., Ferro, N. (eds.) CLEF 2012 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1178/ (2012)
18. 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 Research Protocols **7**(7), e10961 (2018). https://doi.org/10.2196/10961
19. Suominen, H., Kelly, L., Goeuriot, L., Névéol, A., Ramadier, L., Robert, A., Kanoulas, E., Spijker, R., Azzopardi, L., Li, D., Jimmy, Palotti, J., Zuccon, G.: Overview of the clef ehealth evaluation lab 2018. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 286–301. Springer Berlin Heidelberg (2018)
20. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L., Jones, G.J., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., Zuccon, G.: Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 212–231. Springer Berlin Heidelberg (2013)
21. Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L., Lupu, M., Pecina, P., Mueller, H., Budaher, J., Deacon, A.: 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)